The SAGE Handbook of Quantitative Methodology for the Social Sciences

David Kaplan Editor

The SAGE Handbook of Quantitative Methodology for the Social Sciences

To Allison, Rebekah, and Hannah

The SAGE Handbook of Quantitative Methodology for the Social Sciences

David Kaplan Editor
University of Delaware

SAGE Publications
International Educational and Professional Publisher
Thousand Oaks ▪ London ▪ New Delhi

For information:

Sage Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

Sage Publications Ltd.
1 Oliver's Yard
55 City Road
London, EC1Y 1SP
United Kingdom

Sage Publications India Pvt. Ltd.
B-42 Panchsheel Enclave
Post Box 4109
New Delhi 110–017 India

Printed in the United States of America on acid-free paper.

Library of Congress Cataloging-in-Publication Data

The Sage handbook of quantitative methodology for the social sciences / David Kaplan, editor.
p. cm.
Includes bibliographical references and index.
ISBN 0-7619-2359-4 (pbk.)
1. Social sciences—Research—Methodology. 2. Social sciences—Mathematical models. I. Kaplan, David, 1955- II. Sage Publications, Inc.
H62.S275 2004
001.4′2—dc22

2003028071

04 05 06 07 08 09 10 9 8 7 6 5 4 3 2 1

Acquiring Editor:	Lisa Cuevas Shaw
Editorial Assistant:	Benjamin Penner
Project Editor:	Claudia A. Hoffman
Copy Editor:	Gillian Dickens
Typesetter:	C&M Digitals (P) Ltd.
Indexer:	Molly Hall
Cover Designer:	Michelle Kenny

CONTENTS

PREFACE

This handbook was conceived as a way of introducing applied statisticians, empirical researchers, and graduate students to the broad array of state-of-the-art quantitative methodologies in the social sciences. Quantitative methodology is a highly specialized field, and with any highly specialized field, working through idiosyncratic language can be challenging—especially when concepts are conveyed in the language of mathematics and statistics. With that challenge in mind, the contributing authors of this handbook were asked to write about their areas of expertise in a way that would convey to the reader the utility of their respective methodologies. Mathematical language was not to be avoided per se, but as much descriptive meat was to be added to the mathematical bones as possible. Relevance to real-world problems in the social sciences was to be an essential ingredient of each chapter. The goal was for a researcher working in the area of, say, multilevel modeling to be able to read the chapter on, say, dual scaling and understand the basic ideas and the critical arguments for the utility of the method. In my view, the authors of these chapters rose to the requirements admirably. I hope you agree, and I now invite you to dip into the broad and deep pool of quantitative social science methodology.

This handbook is organized around six topical sections. The ordering of the sections is not accidental. Rather, it represents a view of the progression of quantitative methodology, beginning with the scaling of qualitative experience, through the properties of tests and measurements; advancing to the application of statistical methods applied to measures and scales; and closing with broad philosophical themes that transcend many of quantitative methodologies represented here.

Section I concerns the topic of scaling—the quantitative representation of qualitative experiences. Shizuhiko Nishisato opens this section with dual scaling. He begins by arguing that the main goal of data analysis is to extract as much information as possible from the linear and nonlinear relations among variables. Dual scaling (also referred to as optimal scaling) is a method to accomplish this goal by assigning optimal spaced weights to variables. Nishisato provides an interesting example of whether Likert category weights are appropriate for scaling two attitudinal items. He then moves to examples of dual scaling applied to incidence and dominance data, providing many interesting examples along the way. The next chapter is a discussion of multidimensional scaling and unfolding of symmetric and asymmetric proximity relationships by Willem Heiser and Frank Busing. These authors show how the methods of multidimensional scaling and unfolding provide a unified approach to the study of entire relational systems. Their chapter focuses on methods primarily for proximity relationships as separate from so-called dominance or order relationships commonly found in multivariate statistical methods. This section closes with a discussion of principal components analysis with nonlinear optimal scaling of nominal and ordinal data by Jacqueline Meulman, Anita Van der Kooij, and Willem Heiser. These authors consider the ubiquitous problem of scales having arbitrary measurement units, such as an ill-defined zero point or unequal/unknown distances among category points. Meulman and her colleagues show how categorical principal components analysis can be used to develop optimal quantitative values for qualitative scales.

As one proceeds from the scaling of qualitative experiences, the question arises as to the statistical and psychometric properties of measuring instruments. Section II addresses advances in this area. At the most fundamental level are issues of reliability and validity. Thus, this section opens with the chapter by Bruno Zumbo and André Rupp, who situate the concepts of reliability and validity in their historical context but also provide an overview of modern ideas in reliability and validity theory. Rather than

cataloging every new method under the general rubric of reliability and validity, Zumbo and Rupp provide a unifying view of reliability and validity through the lens of statistical modeling. Moving to more advanced ideas in the analysis of item response data, Ratna Nandakumar and Terry Ackerman write about the problem of test modeling. Their chapter provides a comprehensive overview of modeling test data, specifically within the item response theory framework. An important contribution of the Nandakumar and Ackerman chapter is the presentation of an algorithm for choosing an appropriate model for test data along with an illustration of their algorithm using simulated data. Louis Roussos and William Stout offer a discussion of practical issues and new ideas in differential item functioning. They note that with federal legislation such as the No Child Left Behind Act, the issue of test equity is of paramount importance, and methods of assessing differential item functioning are key to documenting the equity of tests. Finally, Hua-Hua Chang continues with advances in computerized adaptive testing (CAT). Given the well-documented achievements and advantages of CAT over paper-and-pencil test administrations, Chang focuses his attention on issues and problems with CAT—particularly issues of test compatibility and security.

With respect to the organization of this handbook, Sections I and II are fundamental to statistical modeling. Scaling qualitative experiences along with knowledge of the properties of measurement instruments are necessary first steps toward the interpretation of statistical models applied to data derived from the employment of those instruments. The next three sections are composed of chapters that detail advances in modern statistical methodology.

Section III concerns statistical models for categorical outcomes. David Rindskopf provides an overview of recent, as well as recycled, trends in the analysis of categorical variables. Rindskopf considers a method as recycled if it was developed long ago but resurrected in a context that is more general than the original idea. Rindskopf also offers some methods that he argues are candidates for recycling. This is followed by an overview of ordinal regression models by Valen Johnson and James Albert. A particularly interesting example used in the Johnson and Albert chapter concerns the modeling of ratings given to student essays—a problem of great significance to large-scale testing companies. Jay Magidson and Jeroen Vermunt continue with a discussion of latent class analysis, in which categorical (dichotomous) measurements are related to a categorical latent variable. Magidson and Vermunt offer formal treatment of the latent class factor model and a detailed discussion of latent class regression models. They show how the latent class cluster model, as applied to continuous variables, can be an improvement over common approaches to cluster analysis. Moving from models of categorical data for cross-sectional studies, John Willett and Judith Singer take up the analysis of ordinal outcomes in longitudinal settings—specifically, the analysis of discrete-time survival data. A particularly important part of Willett and Singer's chapter is the discussion of how researchers can be led astray when using methods other than discrete-time survival analysis to model the event occurrence.

Arguably, one of the most important recent developments in quantitative methodology for the social sciences has been the advent of models to handle nested data. Such data typically derive from the study of social organizations, such as schools. However, the analysis of individual change, as well as meta-analytic studies, can also yield nested data. Models for the analysis of nested data are the subject of Section IV. At the most basic level is the analysis of individual growth and change. Donald Hedeker begins by offering a didactic introduction to the analysis of growth and change from the multilevel modeling perspective, illustrating general ideas with data from a longitudinal study of the response to tricyclic antidepressants for psychiatric patients suffering nonendogenous and endogenous forms of depression. Moving to the application of multilevel modeling to organizational studies, Russell Rumberger and Gregory Palardy provide a comprehensive overview of multilevel modeling applied to the study of school effects. Their chapter takes the reader through a number of decisions that have to be made regarding research questions and data quality, at each point relaying these concerns back to basic substantive issues. Extensions of multilevel modeling to complex designs in program evaluation are taken up in the chapter by Michael Seltzer. Seltzer's chapter is particularly timely given the increased attention to the evaluation of social interventions in experimental and quasi-experimental field settings. Finally, the methodology of meta-analysis is discussed in the chapter by Spyros Konstantopoulos and Larry Hedges. The authors point out that the term *meta-analysis* is often used to connote the entire range of methods for research synthesis but that their chapter will focus on the statistical methods of meta-analysis. Although the authors provide a very general description of meta-analysis, this chapter fits nicely in the section on multilevel modeling insofar as multilevel models provide a convenient framework for estimating across-study variation in study-level effect sizes.

The bulk of the chapters in Sections III and IV concern the analysis of manifest outcomes. In Section V, attention turns specifically to the analysis of unobserved (i.e., latent) variables. The chapter that opens Section V is a discussion of unrestricted exploratory factor analysis by Rick Hoyle and Jamieson Duval. In addition to providing a review of commonly used methodologies for determining the number of factors, Hoyle and Duval show how two commonly used procedures yield incorrect conclusions regarding the number of factors. Following Hoyle and Duval is Gregory Hancock's overview of latent variable models for quasi-experimental, experimental, and nonexperimental designs. Hancock focuses specifically on the utility of structured means analysis and multiple-indicator, multiple-cause (MIMIC) analysis to remove problems of measurement error from hypothesis testing in designed studies. Moving from latent variable models for cross-sectional data, we turn to the method of dynamic factor analysis discussed in the chapter by John Nesselroade and Peter Molenaar. Their chapter concentrates on examining the history of factor-analytic approaches to time-series data and presents new developments aimed at improving applications of dynamic factor analysis to social and behavioral science research. The section closes with Bengt Muthén's chapter on growth mixture modeling—nicely tying in a number of methodologies discussed in Sections IV and V—including multilevel modeling, growth curve modeling, latent class analysis, and discrete-time survival modeling.

In considering the content of this handbook, I viewed it as important to provide the reader with a discussion of some of the major philosophical issues that underlie the use of quantitative methodology. Thus, Section VI covers a number of different foundational topics that are more or less applicable to all of the methodologies covered in this handbook. This section opens with a chapter by Richard Neapolitan and Scott Morris that focuses on probabilistic modeling with Bayesian networks. In their chapter, Neapolitan and Morris first provide a philosophical context, comparing the frequentist approach of von Mises to the subjective probability/Bayesian approach most closely associated with Lindley. From there, they move to Bayesian network models—also referred to as direct acyclic graph (DAG) models. The Neapolitan and Morris chapter is followed by an engaging critique of the "null hypothesis ritual" by Gerd Gigerenzer, Stefan Krauss, and Oliver Vitouch. In this chapter, Gigerenzer et al. make a compelling case for reconsidering the ritual aspects of null hypothesis testing and instead considering null hypothesis testing as a tool among many for empirical research. My contribution to the handbook overviews advances on the problem of defining and testing exogeneity. I examine the problem of exogeneity from within the econometric perspective, highlighting problems with existing ad hoc definitions of exogeneity commonly found in applied statistics textbooks, and point to statistical criteria that distinguish between three types of statistical exogeneity. This is followed by Stanley Mulaik's discussion of objectivity in science and structural equation modeling. Locating the problem of objectivity in the work of Emanuel Kant, Mulaik's chapter provides a sweeping examination of how various metaphors from the theory of object perception underlie the practice of structural equation modeling and how recent developments in the cognitive sciences provide an expansion of Kant's ideas. Finally, the handbook closes with a detailed discussion on causal inference by Peter Spirtes, Richard Scheines, Clark Glymour, Thomas Richardson, and Chris Meek. These authors closely examine such questions as the difference between a causal model and a statistical model, the theoretical limits on causal inference, and the reliability of certain methods of causal inference commonly used in the social sciences.

ACKNOWLEDGMENTS

This handbook is the result of a luncheon conversation I had with my friend C. Deborah Laughton at the 2000 annual meeting of the American Educational Research Association. Deborah's encouragement throughout all phases of this project has been essential in helping me maintain my sense of perspective and humor. It is not too much of an exaggeration to say this book would not have been possible without Deborah's support and friendship.

Next, I would like to thank the section editors: Rick Hoyle, Ratna Nandakumar, Stanley Mulaik, Shizuhiko Nishisato, David Rindskopf, and Michael Seltzer. I relied on these distinguished scholars to help me make sense of those chapters that were outside my area of expertise and to help me maintain the right balance of thoroughness and readability that I strived for in this handbook.

I would like to also acknowledge the staff at Sage Publications: Gillian Dickens, Claudia Hoffman, Alison Mudditt, Benjamin Penner, and Lisa Cuevas Shaw, whose editorial professionalism is evident throughout this handbook. Finally, I want to thank Laura Dougherty, Doug Archbald, Chris Clark, Andrew Walpole, and "Moose," who perhaps don't realize how much our musical interludes have kept me sane throughout this project.

Section I

SCALING

Chapter 1

DUAL SCALING

SHIZUHIKO NISHISATO

1.1. WHY DUAL SCALING?

Introductory and intermediate courses in statistics are almost exclusively based on the following assumptions: (a) the data are continuous, (b) they are a random sample from a population, and (c) the population distribution is normal. In the social sciences, it is very rare that our data satisfy these assumptions. Even if we manage to use a random sampling scheme, the data may not be continuous but qualitative, and the assumption of the normal distribution then becomes irrelevant. What can we do with our data, then? Dual scaling will offer an answer to this question as a reasonable alternative.

More important, however, the traditional statistical analysis is mostly what we call *linear analysis*, which is a natural fate of using continuous variables, for which such traditional statistical procedures as analysis of variance, regression analysis, principal component analysis, and factor analysis were developed. In traditional principal component analysis, for example, we can look into such a linear phenomenon as "blood pressure increases as one gets older" while failing to capture a nonlinear phenomenon such as "migraines occur more frequently when blood pressure is very low or very high." When we look at possible forms of relations between two variables, we realize that most relations are nonlinear and that it is not advantageous to restrict our attention only to the linear relation. Dual scaling captures linear and nonlinear relations among variables, without modeling the forms of relations for analysis.

Dual scaling is also referred to as "optimal scaling" (Bock, 1960) because all forms of relations among variables are captured through optimally spacing categories of variables. The main purpose of data analysis lies in delineating relations among variables, linear or nonlinear, or, more generally, in extracting as much information in data as possible. We will find that dual scaling is an optimal method to extract a maximal amount of information from multivariate categorical data. We will see later that dual scaling can be applied effectively to many kinds of psychological data such as observation data, teacher evaluation forms, attitude/aptitude data, clinical data, and all types of questionnaire data. This chapter contains a minimal package of information about all aspects of dual scaling.

1.2. HISTORICAL BACKGROUND

1.2.1. Mathematical Foundations in Early Days

Two major contributions to the area from the past are (a) algebraic *eigenvalue theory*, pioneered

AUTHOR'S NOTE: This work was supported by the Natural Sciences and Engineering Research Council of Canada. The paper was written while the author was a Visiting Professor at the School of Business Administration, Kwansei Gakuin University, Nishinomiya, Japan.

by mathematicians (e.g., Euler, Cauchy, Jacobi, Cayley, and Sylvester) in the 18th century, and (b) the theory of *singular value decomposition* (SVD) by Beltrami (1873), Jordan (1874), and Schmidt (1907).

The eigenvalue decomposition (EVD) was for orthogonal decomposition of a square matrix, put into practice as principal component analysis (Hotelling, 1933; Pearson, 1901). SVD was for the joint orthogonal decomposition of row structure and column structure of any rectangular matrix and reappeared much later in metric multidimensional scaling as the *Eckart-Young decomposition* (Eckart & Young, 1936). Both EVD and SVD are based on the idea of principal hyperspace, that is, space described in terms of principal axes.

1.2.2. Pioneers in the 20th Century

With these precursors, Richardson and Kuder (1933) presented the idea of what Horst (1935) called the *method of reciprocal averages* (MRA) for the analysis of multiple-choice data. Hirschfeld (1935) provided a formulation for weighting rows and columns of a two-way table in such a way that the regression of rows on columns and that of columns on rows could be simultaneously linear, which Lingoes (1964) later called *simultaneous linear regressions*. Fisher (1940) considered discriminant analysis of data in a contingency table, in which he, too, suggested the algorithm of MRA. Most important contributions in the early days were by Guttman (1941) for his detailed formulation for the scaling of multiple-choice data and Maung (1941) for elaborating Fisher's scoring method for contingency tables. Guttman (1946) further extended his approach of internal consistency to rank-order and paired-comparison data. Thus, solid foundations were laid by 1946.

1.2.3. Period of Rediscoveries and Further Developments

We can list Mosier (1946), Fisher (1948), Johnson (1950), Hayashi (1950, 1952), Bartlett (1951), Williams (1952), Bock (1956, 1960), Lancaster (1958), Lord (1958), Torgerson (1958), and many other contributors. Among others, there were four major groups of researchers: the Hayashi school in Japan since 1950, the Benzécri school in France since the early 1960s, the Leiden group in the Netherlands since the late 1960s, and the Toronto group in Canada since the late 1960s.

Because of its special appeal to researchers in various countries and different disciplines, the method has acquired many aliases, mostly through rediscoveries of essentially the same technique—among others, the method of reciprocal averages (Horst, 1935; Richardson & Kuder, 1933), simultaneous linear regressions (Hirschfeld, 1935; Lingoes, 1964), appropriate scoring and additive scoring (Fisher, 1948), principal component analysis of categorical data (Torgerson, 1958), optimal scaling (Bock, 1960), correspondence analysis (Benzécri, 1969; Escofier-Cordier, 1969), biplot (Gabriel, 1971), canonical analysis of categorical data (de Leeuw, 1973), reciprocal averaging (Hill, 1973), basic structure content scaling (Jackson & Helmes, 1979), dual scaling (Nishisato, 1980), homogeneity analysis (Gifi, 1980), centroid scaling (Noma, 1982), multivariate descriptive statistical analysis (Lebart, Morineau, & Warwick, 1984), nonlinear multivariate analysis (Gifi, 1990), and nonlinear biplot (Gower & Hand, 1996). Because all of these are based on singular value decomposition of categorical data, they are either mathematically identical or not much different from one another.

1.2.4. Dual Scaling

The name *dual scaling* (DS) was coined by Nishisato (1980) as a result of the discussion at the symposium on optimal scaling during the 1976 annual meeting of the Psychometric Society in Murray Hill, New Jersey (see Nishisato & Nishisato, 1994a). With the general endorsement among the participants, he adopted it in the title of his 1980 book. Franke (1985) states that he "uses Nishisato's term for its generality and lack of ambiguity" (p. 63).

Under the name *dual scaling*, Nishisato has extended its applicability to a wider variety of categorical data, including both incidence data and dominance data. This aspect of DS is reflected in Meulman's (1998) statement that "dual scaling is a comprehensive framework for multidimensional analysis of categorical data" (p. 289). For those interested in the history of quantification theory, see de Leeuw (1973), Benzécri (1982), Nishisato (1980), Greenacre (1984), Gifi (1990), Greenacre and Blasius (1994), and van Meter, Schiltz, Cibois, and Mounier (1994).

1.3. An Intuitive Introduction to Dual Scaling

1.3.1. Is Likert Scoring Appropriate?

Suppose subjects were asked two multiple-choice questions.[1]

Q1: What do you think of taking sleeping pills? (1) *strongly disagree*, (2) *disagree*, (3) *indifferent*, (4) *agree*, (5) *strongly agree*

Q2: Do you sleep well every night? (1) *never*, (2) *rarely*, (3) *sometimes*, (4) *often*, (5) *always*

The data are in Table 1.1. Likert scores are often used for ordered sets of categories (Likert, 1932). Suppose we assign $-2, -1, 0, 1, 2$ to the five ordered categories of each set in the above example. Our question here is if these Likert scores are appropriate. There is a simple way to examine it.

First, we calculate the mean of each category, using Likert scores. For example, the mean of category *never* is $[15 \times (-2) + 5 \times (-1) + 6 \times 0 + 0 \times 1 + 1 \times 2]/27 = -1.2$. Likewise, we calculate the means of row categories and those of column categories, which are summarized in Table 1.2. We now plot those averages against the original scores $(-2, -1, 0, 1, 2)$, as seen in Figure 1.1. The two lines are relatively close to a straight line, which indicates that the original scores are "pretty good." Suppose we use, instead of those subjective category weights, the weights derived by DS and calculate the weighted category means and plot these against the DS weights. We then obtain Figure 1.2.

Notice that the two lines are now merged into a single straight line. This is "mathematically optimal," as seen later. We will also see shortly that the slope of the line in Figure 1.2 is equal to the maximal "nontrivial" singular value for this data set.

But how do we arrive at the DS weights? It is simple: Once we obtain the mean category scores as in Figure 1.1, replace the original scores (e.g., -2, -1, etc.) with the corresponding mean scores, and then calculate the new mean category scores in the same way as before and plot the new category scores against the first mean scores, replace the old mean scores with the new mean scores, and calculate new mean category scores and plot them. This is a convergent process (Nishisato, 1980, pp. 60–62, 65–68). Horst (1935) called the above process the *method of reciprocal averages* (MRA), used by Richardson and Kuder (1933), also suggested by Fisher (1940), and fully illustrated by Mosier (1946). MRA is one of the algorithms for DS.

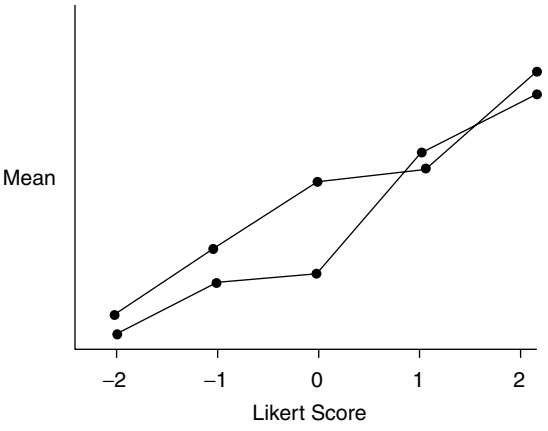
Table 1.1 Sleeping and Sleeping Pills

	Never	*Rarely*	*Sometimes*	*Often*	*Always*	*Sum*	*Score*
Strongly against	15	8	3	2	0	28	−2
Against	5	17	4	0	2	28	−1
Neutral	6	13	4	3	2	28	0
For	0	7	7	5	9	28	1
Strongly for	1	2	6	3	16	28	2
Sum	27	47	24	13	29	140	
Score	−2	−1	0	1	2		

Table 1.2 Likert Scores and Weighted Means

Score	*Mean*	*Score*	*Mean*
−2	−1.2	−2	−1.3
−1	−0.5	−1	−0.8
0	0.4	0	−0.6
1	0.5	1	0.6
2	1.3	2	1.1

Figure 1.1 Likert Scores



1.3.2. The Method of Reciprocal Averages (MRA)

Let us illustrate the process of MRA.[2] Suppose three teachers (White, Green, and Brown) were rated on their teaching performance by students (see Table 1.3).

1. With permission from Nishisato (1980).

2. With permission from Nishisato and Nishisato (1994a).

Figure 1.2 Dual-Scaling Optimal Weights

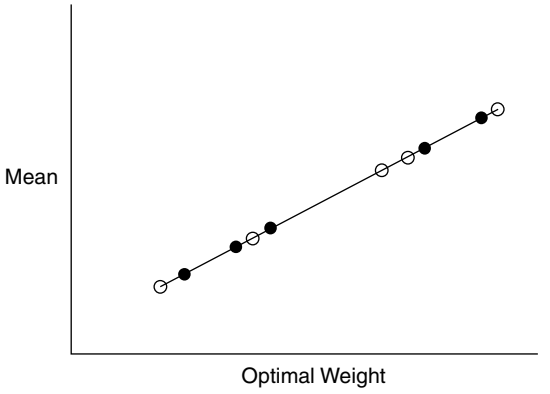


Table 1.3 Evaluating Teachers

Teacher	*Good*	*Ave*	*Poor*	*Total*
White	1	3	6	10
Green	3	5	2	10
Brown	6	3	0	9
Total	10	11	8	29

The MRA is carried out in the following way:

Step 1: The MRA starts with assigning arbitrary weights to columns (or rows, if preferred). Although such values are arbitrary, one must avoid identical weights for all columns (or rows), including zero. It is always a good strategy to use "reasonable" values. As an example, consider the following:

$$\begin{aligned} x_1(\text{good}) &= 1, \\ x_2(\text{average}) &= 0, \\ x_3(\text{poor}) &= -1. \end{aligned} \tag{1}$$

Step 2: Calculate the weighted averages of the rows:

$$\begin{aligned} y_1(\text{White}) &= \frac{1 \times x_1 + 3 \times x_2 + 6 \times x_3}{10} \\ &= \frac{1 \times 1 + 3 \times 0 + 6 \times (-1)}{10} = -0.5, \end{aligned} \tag{2}$$

$$y_2(\text{Green}) = \frac{3 \times 1 + 5 \times 0 + 2 \times (-1)}{10} = 0.1000, \tag{3}$$

$$y_3(\text{Brown}) = \frac{6 \times 1 + 3 \times 0 + 0 \times (-1)}{9} = 0.6667. \tag{4}$$

Step 3: Calculate the mean responses weighted by y_1, y_2, y_3:

$$\begin{aligned} M &= \frac{10y_1 + 10y_2 + 9y_3}{29} \\ &= \frac{10 \times (-0.5) + 10 \times 0.1 + 9 \times 0.6667}{29} \\ &= 0.0690. \end{aligned} \tag{5}$$

Step 4: Subtract M from each of y_1, y_2, y_3, and adjusted values should be indicated again by y_1, y_2, y_3, respectively:

$$y_1 = -0.5000 - 0.0690 = -0.5690, \tag{6}$$

$$y_2 = 0.1000 - 0.0690 = 0.0310, \tag{7}$$

$$y_3 = 0.6667 - 0.0690 = 0.5977. \tag{8}$$

Step 5: Divide y_1, y_2, y_3 by the largest absolute value of y_1, y_2, y_3, say, g_y. At this stage, $g_y = 0.5977$. Adjusted values should again be indicated by y_1, y_2, y_3:

$$\begin{aligned} y_1 &= \frac{-0.5690}{0.5977} = 0.9519, \\ y_2 &= \frac{0.0310}{0.5977} = 0.0519, \\ y_3 &= \frac{0.5977}{0.5977} = 1.0000. \end{aligned} \tag{9}$$

Step 6: Using these new values as weights, calculate the averages of the columns:

$$\begin{aligned} x_1 &= \frac{1y_1 + 3y_2 + 6y_3}{10} \\ &= \frac{1 \times (-0.9519) + 3 \times 0.0519 + 6 \times 1.0}{10} \\ &= 0.5204, \end{aligned} \tag{10}$$

$$\begin{aligned} x_2 &= \frac{3 \times (-0.9519) + 5 \times 0.0519 + 3 \times 1.0000}{11} \\ &= 0.0367, \end{aligned} \tag{11}$$

$$\begin{aligned} x_3 &= \frac{6 \times (-0.9519) + 2 \times 0.0519 + 0 \times 1.0000}{8} \\ &= -0.7010. \end{aligned} \tag{12}$$

Step 7: Calculate the mean responses weighted by x_1, x_2, x_3:

$$\begin{aligned} N &= \frac{10 \times 0.5204 + 11 \times 0.0367 + 8 \times (-0.7010)}{29} \\ &= 0. \end{aligned} \tag{13}$$

Step 8: Subtract N from each of x_1, x_2, x_3.

Table 1.4 Iterative Results

	Iter2 y	*Iter2 x*	*Iter3 y*	*Iter3 x*	*Iter4 y*	*Iter4 x*	*Iter5 y*	*Iter5 x*
1	−0.9954	0.7321	−0.9993	0.7321	−0.9996	0.7311	−0.9996	0.7311
2	0.0954	0.0617	0.0993	0.0625	0.0996	0.0625	0.0996	0.0625
3	1.0000	−1.0000	1.0000	−1.0000	1.0000	−1.0000	1.0000	−1.0000
g	0.5124	0.7227	0.5086	0.7246	0.5083	0.7248	0.5083	0.7248

Step 9: Divide each element of x_1, x_2, x_3 by the largest absolute value of the three numbers, say, g_x. Because −0.7010 has the largest absolute value, $g_x = 0.7010$. Adjusted values are indicated again by x_1, x_2, x_3:

$$x_1 = \frac{0.5204}{0.7010} = 0.7424,$$

$$x_2 = \frac{0.0367}{0.7010} = 0.0524,$$

$$x_3 = \frac{-0.7010}{0.7010} = -1.0000. \qquad (14)$$

Reciprocate the above averaging processes (Steps 2 through 9) until all the six values are stabilized. Iteration 5 provides the identical set of numbers as Iteration 4 (see Table 1.4). Therefore, the process has converged to the optimal solution in four iterations. Notice that the largest absolute values at each iteration, g_y and g_x, also converge to two constants, 0.5083 and 0.7248. Nishisato (1988) showed that the eigenvalue, ρ^2, is equal to the product, $g_y g_x = 0.5083 \times 0.7248 = 0.3648$, and the singular value, ρ, is the geometric mean,

$$\rho = \text{singular value} = \sqrt{g_y g_x}$$
$$= \sqrt{0.5083 \times 0.7248} = 0.6070. \qquad (15)$$

If we start with the cross-product symmetric table, instead of the raw data (the present example), the process will converge to one constant of g, which is the *eigenvalue*, and its positive square root is the *singular value* (Nishisato, 1980). See Nishisato (1994, p. 89) for why the final value of g is the eigenvalue.

Step 10: In the DUAL3 for windows (Nishisato & Nishisato, 1994b), the unit of weights is chosen in such a way that the sum of squares of weighted responses is equal to the number of responses. In this case, the constant multipliers for adjusting the unit of y (say, c_r) and x (c_c) are given by

$$c_r = \sqrt{\frac{29}{10y_2^1 + 10y_2^2 + 9y_3^2}} = 1.2325,$$

$$c_c = \sqrt{\frac{29}{10x_2^1 + 11x_2^2 + 8x_3^2}} = 1.4718. \qquad (16)$$

Table 1.5 Two Types of Optimal Weights

	Normed y	*Normed x*	*Projected y*	*Projected x*
1	−1.2320	1.0760	−0.7478	0.6531
2	0.1228	0.0920	0.0745	0.0559
3	1.2325	−1.4718	0.7481	−0.8933

The final weights are obtained by multiplying y_1, y_2, y_3 by c_r and x_1, x_2, x_3 by c_c. These weights are called *normed weights*. The normed weights, multiplied by the singular value—that is, ρy_i and ρx_j—are called *projected weights*, which reflect the relative importance of categories. The distinction between these two types of weights will be discussed later. In the meantime, let us remember that normed weights and projected weights are what Greenacre (1984) calls *standard coordinates* and *principal coordinates*, respectively, and that projected weights are the important ones because they reflect relative importance of the particular solution (component, dimension). The final results are in Table 1.5. These weights thus obtained are scaled in such a way that (a) the sum of responses weighted by y is zero, and the sum of responses weighted by x is zero; (b) the sum of squares of responses weighted by y is the total number of responses, and the same for x. Once the first solution is obtained, calculate the residual frequencies, and apply the MRA to the residual table to obtain the second solution. This process will be discussed later.

1.4. Two Types of Categorical Data

Nishisato (1993) classified categorical data into two distinct groups, *incidence data* (e.g., contingency tables, multiple-choice data, sorting data) and *dominance data* (e.g., rank-order, paired-comparison data).

1.4.1. Incidence Data

Elements of data are 1 (presence), 0 (absence), or frequencies, as we see in contingency tables,

multiple-choice data, and sorting data. DS of incidence data is characterized by (a) the use of the "chi-square metric" (Greenacre, 1984; Lebart et al., 1984; Nishisato & Clavel, 2003), (b) a lower rank approximation to input data, (c) "a trivial solution" (Gifi, 1990; Greenacre, 1984; Guttman, 1941; Nishisato, 1980, 1994), and (d) more than one dimension needed to describe the data (Nishisato, 2002, 2003). This last point is true even when all variables are perfectly correlated to one another. *Correspondence analysis* and *multiple correspondence analysis* were originally developed in France specifically for incidence data for the contingency table and multiple-choice data, respectively.

1.4.2. Dominance Data

Elements of data are greater than, equal to, or smaller than, as we see in rank-order data and paired-comparison data. Because the information is typically given in the form of inequality relations, without any specific amount of the discrepancy between the two attributes or stimuli indicated, it is not possible to approximate the value of the data directly as is done with the incidence data. Instead, the objective here is to derive new measurements for objects in such a way that the ranking of the measurements best approximates the corresponding ranking of the original dominance data. DS of dominance data is characterized by (a) the use of the Euclidean metric (Nishisato, 2002), (b) a lower rank approximation to the *ranks* of the data (Nishisato, 1994, 1996), (c) no trivial solution (Greenacre & Torres-Lacomba, 1999; Guttman, 1946; Nishisato, 1978; van de Velden, 2000), and (d) one dimension to describe the data when all variables are perfectly correlated to one another (Nishisato, 1994, 1996).

1.4.3. Scope of Dual Scaling

DS is applicable not only to the incidence data but also to the dominance data. The DUAL3 for Windows (Nishisato & Nishisato, 1994b), a computer program package for DS, handles both types of categorical data. Recently, Greenacre and Torres-Lacomba (1999) and van de Velden (2000) reformulated correspondence analysis for dominance data, which were not much different from Nishisato's (1978) earlier study. After all, they are all based on singular-value decomposition.

1.5. Scaling of Incidence Data

1.5.1. Contingency Tables

Contingency tables are often used to summarize data. For example, a small survey on the popularity of five movies, collected from three age groups, can be summarized into a 5×3 table of the number of people in each cell. Similarly, we often see a large number of tabulation tables on voting behavior, typically on two categorical variables (e.g., age and education). These are contingency tables.

1.5.1.1. Some Basics

Consider an n-by-m contingency table with typical element f_{ij}. DS first eliminates from this table the frequencies expected when rows and columns are statistically independent, that is, $f_{i.}f_{.j}/f_t$, where f_t is the total frequency in the table. This is called a trivial solution. Then, the residual table, consisting of typical elements for row i and column j, say,

$$f_{ij} - \frac{f_{i.}f_{.j}}{f_t} = f_{ij} - h_{ij}, \tag{17}$$

is decomposed into independent components, called solutions. Let $\min(n, m)$ be the smaller value of n and m. Then the n-by-m residual table can be exhaustively explained by at most $[\min(n, m) - 1]$ solutions. In other words, the total number of nontrivial solutions, that is, proper solutions $T(\text{sol})$, is given by

$$T(\text{sol}) = \min(n, m) - 1. \tag{18}$$

The variance of solution k is called the eigenvalue, ρ_k^2, which is a measure of information conveyed by solution k. The total information contained in the residual matrix, $T(\text{inf})$, is the sum of the $[\min(n, m) - 1]$ eigenvalues, which is equal to

$$T(\text{inf}) = \sum_{k=1}^{p} \rho_k^2 = \frac{\chi^2}{f_t}, \text{ where}$$
$$\chi^2 = \sum_{i}^{n}\sum_{j}^{m} \frac{(f_{ij} - h_{ij})^2}{h_{ij}}, \tag{19}$$

and h_{ij} is the frequency expected when the ith row and the jth column are statistically independent. The percentage of the total information explained by solution k is indicated by δ_k and is given by

$$\delta_k = \frac{100\rho_k^2}{T(\text{inf})}. \tag{20}$$

1.5.1.2. Example: Biting Habits of Laboratory Animals

The biting habits of four laboratory animals were investigated. The following data were obtained from Sheskin's (1997) book.[3] Because this is a small example, let us list the main output from the program DUAL3 (Nishisato & Nishisato, 1994b) (Table 1.7).

Because this data set is a 4×3 table, $T(\text{sol}) = 2$, and the analysis shows that δ_1 and δ_2 are 94.2% and 5.8%, respectively. The order-0 approximation is the trivial solution. The trivial solution is removed from the data, and the residual table is analyzed into components. The order-1 approximation is what one can predict from the trivial solution and Solution 1:

$$f^*_{ij(1)} = \frac{f_{i.} f_{.j}}{f_t}[1 + \rho_1 y_{i1} x_{j1}]. \tag{21}$$

Because the value of δ_1 is 94.2% (the contribution of Solution 1), this approximation to the input data is very good, and the residual table does not contain much more information to be analyzed. In the current example, the order-2 approximation perfectly reproduces the input data:

$$f^*_{ij(2)} = \frac{f_{i.} f_{.j}}{f_t}[1 + \rho_1 y_{i1} x_{j1} + \rho_2 y_{i2} x_{j2}]. \tag{22}$$

See also the residual table (Table 1.7), which shows no more information left to be analyzed. Notice that it is not clear what relations between the animals and biting habits are from the input table, but see the graph based on DS: The two-dimensional graph (Figure 1.3) shows, among other things, that (a) guinea pigs are flagrant biters, (b) mice are between flagrant biters and mild biters, (c) mild biters and nonbiters are relatively closely located, (d) gerbils are nonbiters, and (e) hamsters are between mild biters and nonbiters. The graph is much easier to understand than the original table.

1.5.2. Multiple-Choice Data

Multiple-choice data are ubiquitous in psychological research, particularly in personality, social, and clinical research. We should question, however, how arbitrarily such data are typically analyzed. When response options are ordered (e.g., never, sometimes, often, always), researchers often use the integer scores 1, 2, 3, and 4 for these ordered categories and analyze the data. This practice of using the so-called Likert scores is by no means effective in retrieving information in data. We will see this problem very shortly. In contrast, dual scaling can analyze such multiple-choice data in a very effective way in terms of information retrieval. We will see an example of dual-scaling analysis shortly.

Table 1.6 Sheskin's Data on Biting Habits of Laboratory Animals

Animals	*Not a Biter*	*Mild Biter*	*Flagrant Biter*
Mice	20	16	24
Gerbils	30	10	10
Hamsters	50	30	10
Guinea pigs	19	11	50

1.5.2.1. Some Basics

Consider n multiple-choice items, with item j having m_j options. Consider further that each of N subjects is asked to choose one option per item. Let m be the total number of options of n items. For DS, multiple-choice data are expressed in the form of (1,0) response patterns (see the example in 1.5.2.2) and also have a trivial solution. The aforementioned statistics of multiple-choice data are as follows:

$$T(\text{sol}) = m - n \text{ or } N - 1, \text{ whichever is smaller.} \tag{23}$$

$$T(\text{inf}) = \sum_{k=1}^{m-n} \rho_k^2 = \frac{\sum_{j=1}^{n} m_j}{n} - 1 = \overline{m} - 1. \tag{24}$$

The definition of δ_k is the same as the contingency table, but in practice we will modify it as we discuss later. Option weights are determined, as Lord (1958) proved, to yield scores with a maximal value of the generalized Kuder-Richardson internal consistency reliability, or Cronbach's α (Cronbach, 1951), which can be inferred from the following relations (Nishisato, 1980):

$$\alpha = 1 - \frac{1 - \rho^2}{(n-1)\rho^2} = \frac{n}{n-1}\left(\frac{\sum r_{jt}^2 - 1}{\sum r_{jt}^2}\right) \text{ since}$$

$$\rho^2 = \frac{\sum_j^n r_{jt}^2}{n}, \tag{25}$$

where r_{jt}^2 is the square of correlation between item j and the total score. It is known (Nishisato, 1980, 1994) that the average information in multiple-choice data, that is—$T(\text{inf})/T(\text{sol})$—is $1/n$ and that α becomes negative when ρ^2 is smaller than the average information. Therefore, Nishisato (1980, 1994) suggests

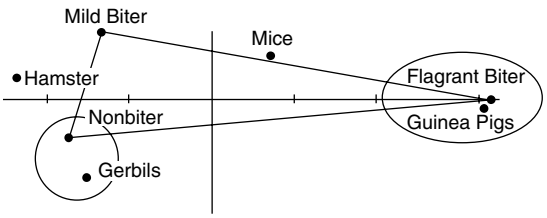
3. Reprinted with permission from Sheskin (1997).

Table 1.7 Approximation to Input

ORDER 0 APPROXIMATION			*RESIDUAL MATRIX*			
25.5	14.4	20.1	−6.5	1.6	3.9	
21.3	12.0	16.8	8.8	−2.0	−6.8	
38.3	21.5	30.2	11.8	8.5	−20.2	
34.0	19.1	26.9	−15.0	−8.1	23.1	
ORDER 1 APPROXIMATION			*RESIDUAL MATRIX*			*SOLUTION 1*
22.7	13.1	24.2	−2.7	2.9	−0.2	Eigenvalue = 0.20
26.1	14.2	9.7	3.9	−4.2	0.3	Singular value = 0.45
52.1	27.8	10.2	−2.1	2.2	−0.2	Delta = 94.2%
18.1	12.0	49.9	0.9	−1.0	0.1	CumDelta = 94.2
ORDER 2 APPROXIMATION			*RESIDUAL MATRIX*			*SOLUTION 2*
20.0	16.0	24.0	0.0	0.0	0.0	Eigenvalue = 0.01
30.0	10.0	10.0	0.0	0.0	0.0	Singular value = 0.11
50.0	30.0	10.0	0.0	0.0	0.0	Delta = 5.8%
19.0	11.0	50.0	0.0	0.0	0.0	CumDelta = 100.0

	PROJECTED WEIGHTS			*PROJECTED WEIGHTS*	
	Sol-1	*Sol-2*		*Sol-1*	*Sol-2*
Mice	0.14	0.12	Not a biter	−0.34	−0.10
Gerbils	−0.30	−0.21	Mild biter	−0.27	0.19
Hamsters	−0.47	0.06	Flagrant biter	0.63	−0.01
Guinea pigs	0.61	−0.03			

Figure 1.3 Biting Habits of Four Animals



stopping the extraction of solutions as soon as ρ^2 becomes smaller than $1/n$. Accordingly, we redefine the statistic δ_k as the percentage of ρ_k^2 over the sum of ρ_j^2 greater than $1/n$.

1.5.2.2. Example: Blood Pressure, Migraines, and Age

As mentioned earlier, Torgerson (1958) called DS "principal component analysis of categorical data." Because principal component analysis (PCA) is a method to find a linear combination of continuous variables (PCA) and that of categorical variables (DS), it would be interesting to look at differences between them. The following example is adopted from Nishisato (2000):

1. How would you rate your blood pressure? (Low, Medium, High): coded 1, 2, 3
2. Do you get migraines? (Rarely, Sometimes, Often): 1, 2, 3 (as above)
3. What is your age group? (20–34, 35–49, 50–65): 1, 2, 3
4. How would you rate your daily level of anxiety? (Low, Medium, High): 1, 2, 3
5. How would you rate your weight? (Light, Medium, Heavy): 1, 2, 3
6. What about your height? (Short, Medium, Tall): 1, 2, 3

Suppose we use the traditional Likert scores for PCA—that is, 1, 2, 3 as scores for the three categories of each question. DS uses response patterns of 1s and 0s. See the two data sets from 15 subjects in Table 1.8 and the product-moment correlation matrix for PCA in Table 1.9. Examine the correlation between blood pressure (BP) and age (Age) ($r = 0.66$) and that between BP and migraines (Mig) ($r = -0.06$) using the data in the contingency table format (Table 1.10).

Notice a linear relation between BP and Age and a nonlinear relation between BP and Mig. It seems that the nonlinear relation between BP and Mig is much clearer than the linear relation between BP and Age:

Table 1.8 Likert Scores for PCA and Response Patterns for DS

	PCA						*DS*					
Subject	*Bpr Q1*	*Mig Q2*	*Age Q3*	*Anx Q4*	*Wgt Q5*	*Hgt Q6*	*Bpr 123*	*Mig 123*	*Age 123*	*Anx 123*	*Wgt 123*	*Hgt 123*
1	1	3	3	3	1	1	100	001	001	001	100	100
2	1	3	1	3	2	3	100	001	100	001	010	001
3	3	3	3	3	1	3	001	001	001	001	100	001
4	3	3	3	3	1	1	001	001	001	001	100	100
5	2	1	2	2	3	2	010	100	010	010	001	010
6	2	1	2	3	3	1	010	100	010	001	001	100
7	2	2	2	1	1	3	010	010	010	100	100	001
8	1	3	1	3	1	3	100	001	100	001	100	001
9	2	2	2	1	1	2	010	010	010	100	100	010
10	1	3	2	2	1	3	100	001	010	010	100	001
11	2	1	1	3	2	2	010	100	100	001	010	010
12	2	2	3	3	2	2	010	010	001	001	010	010
13	3	3	3	3	3	1	001	001	001	001	001	100
14	1	3	1	2	1	1	100	001	100	010	100	100
15	3	3	3	3	1	2	001	001	001	001	100	010

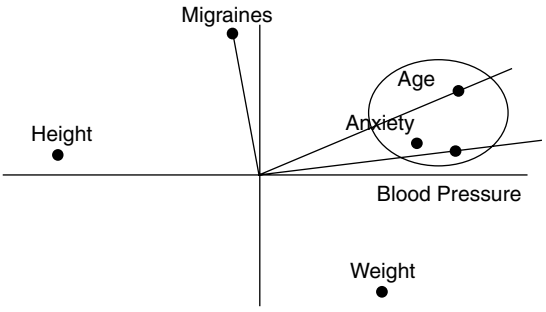
Table 1.9 Product-Moment Correlation Based on Likert Scores

	BP	*Mig*	*Age*	*Anx*	*Wgt*	*Hgt*
Blood pressure (BP)	1.00					
Migraine (Mig)	−.06	1.00				
Age (Age)	.66	.23	1.00			
Anxiety (Anx)	.18	.21	.22	1.00		
Weight (Wgt)	.17	−.58	−.02	.26	1.00	
Height (Hgt)	−.21	.10	−.30	−.23	−.31	1.00

Table 1.10 Relation of Blood Pressures to Age and Migraines

	Age			*Migraine*		
	20–34	*35–49*	*50–65*	*Rarely*	*Sometimes*	*Often*
High BP	0	0	4	0	0	4
Mid BP	1	4	1	3	3	0
Low BP	3	1	1	0	0	5

Figure 1.4 Two Solutions From Principal Component Analysis



"If you have frequent migraines, your blood pressure is either high or low." The first two principal components of Likert scores are plotted in Figure 1.4. Notice that it captures only linear relations. The data for DS are expressed in terms of chosen response patterns, and the units of analysis are response options, not items as in the case of PCA. PCA is a method to determine the most informative weighted combinations of items, whereas DS looks for the most informative weighted combinations of categories of items. This means that DS yields an inter-item correlation matrix for each solution, rather than one for the entire data set as in PCA.

The current data yield four solutions associated with positive values of reliability coefficient α (see Table 1.11).

The adjusted delta is the one redefined in terms of solutions associated with positive values of reliability α. CumDelta and CumAdjDelta are cumulative values of delta and adjusted delta, respectively. For the limited space, we will look at only the first two solutions and their projected option weights (see Table 1.12). Notice that the weights for options of BP and Mig for Solution 1 are weighted in such a way that the nonlinear relation is captured. Study the weights to convince yourself. Using these weights, inter-item

Table 1.11 Four Solutions

	Solution 1	*Solution 2*	*Solution 3*	*Solution 4*
Eigenvalue	0.54	0.37	0.36	0.31
Singular value	0.74	0.61	0.59	0.55
Delta	27	19	17	15
CumDelta	27	46	63	79
Adjusted delta	34	24	22	20
CumAdjDelta	34	58	80	100

Table 1.12 Projected Option Weight of Two Solutions

	Solution 1	*Solution 2*
Blood Pressure		
Low	−0.71	0.82
Medium	1.17	−0.19
High	−0.86	−0.74
Anxiety		
Low	1.55	1.21
Medium	0.12	0.31
High	−0.35	−0.33
Migraine		
Rarely	1.04	−1.08
Sometimes	1.31	0.70
Often	−0.78	0.12
Weight		
Light	−0.27	0.46
Medium	0.32	0.01
Heavy	0.50	−1.40
Age		
20–34	0.37	0.56
35–49	1.03	0.22
50–65	−0.61	−0.56
Height		
Short	−0.56	−0.63
Medium	0.83	−0.35
Tall	−0.27	0.98

correlation matrices are obtained for the two DS solutions (see Table 1.13).

BP and Mig are now correlated at 0.99 in Solution 1. This was attained by assigning similar weights to high BP, low BP, and frequent migraines, which are very different from the weights given to medium BP, rare migraines, and occasional migraines. The same correlation for Solution 2 is 0.06. Characteristics of the first two DS solutions can be obtained by putting options of similar weights together (see Table 1.14). "Nonlinear combinations" of response categories are involved in each solution. In DS, linear correlation is maximized by transforming categories linearly or nonlinearly, depending on the data, whereas PCA filters out all nonlinear relations in the process of analysis, which is why it is called linear analysis. The first two DS solutions are plotted in Figure 1.5. Unlike PCA solutions, three categories of a single variable are not forced to be on a single line but usually form a triangle, the area of which is monotonically related to the contribution of the variable to these dimensions. PCA can never reveal a strong relation between BP and Mig, but this relation is the most dominant one in DS. In DS, high and low BP are associated with frequent migraines, but the second dimension identifies a different association between low and high BP—the former with young, skinny, and tall subjects and the latter with old, heavy, and short subjects.

1.5.3. Sorting Data

Sorting data are not as popular as contingency tables and multiple-choice data, but in some areas, such as cognitive psychology, we often see references to sorting data. So, in this section, we will learn how sorting data are collected and optimally analyzed by dual scaling.

1.5.3.1. Some Basics

Sorting data are collected in the following way. Consider the first object to be a member of the first pile and assign 1 to it; go down the list, and each time you find an object similar to the first object, assign 1 to it. When you finish identifying all the objects with 1, go to the next object that has not been chosen so far and give it 2; go down the list and identify all the objects that are similar to the object with number 2. In this way, you classify all objects on the list into piles. Takane (1980) demonstrated that DS can be used to analyze sorting data by transposing the data or exchanging the roles of subjects and item options in multiple-choice data with objects and subject piles in sorting data, respectively. With this understanding, $T(\text{sol})$ and $T(\text{inf})$ are the same as those of multiple-choice data.

1.5.3.2. Example: Sorting 19 Countries Into Similar Groups

The data in Table 1.15 were collected from Nishisato's class in 1990. The last two columns of the table indicate the optimal (projected) weights of the countries on the first two solutions. Note that prior to DS analysis, the data are first transformed to (1, 0) response patterns, as was the case of multiple-choice data. One of the outcomes is the inter-subject correlation matrix, just like the inter-item correlation matrix in multiple-choice data. Table 1.16 shows the

Table 1.13 Correlation Matrices From Two DS Solutions

	Solution 1						*Solution 2*					
	BP	*Mig*	*Age*	*Anx*	*Wgt*	*Hgt*	*BP*	*Mig*	*Age*	*Anx*	*Wgt*	*Hgt*
BP	1.0						1.0					
Mig	.99	1.0					.06	1.0				
Age	.60	.58	1.0				.59	−.31	1.0			
Anx	.47	.52	.67	1.0			.07	.35	.35	1.0		
Wgt	.43	.39	.08	−.33	1.0		.28	.62	−.01	.19	1.0	
Hgt	.56	.57	.13	.19	.20	1.0	.31	.29	.32	.17	.38	1.0

Table 1.14 Characteristics of Two DS Solutions

Solution 1		*Solution 2*	
One End	*The Other End*	*One End*	*The Other End*
Low BP	Medium BP	High BP	Low BP
High BP	Rare migraine	Rare migraine	Occasional migraine
Frequent migraine	Middle age	Old	Young
Old age group	Low anxiety	Heavy	Tall
High anxiety	Medium height	Short	
Short			

Figure 1.5 First Two Dual-Scaling Solutions

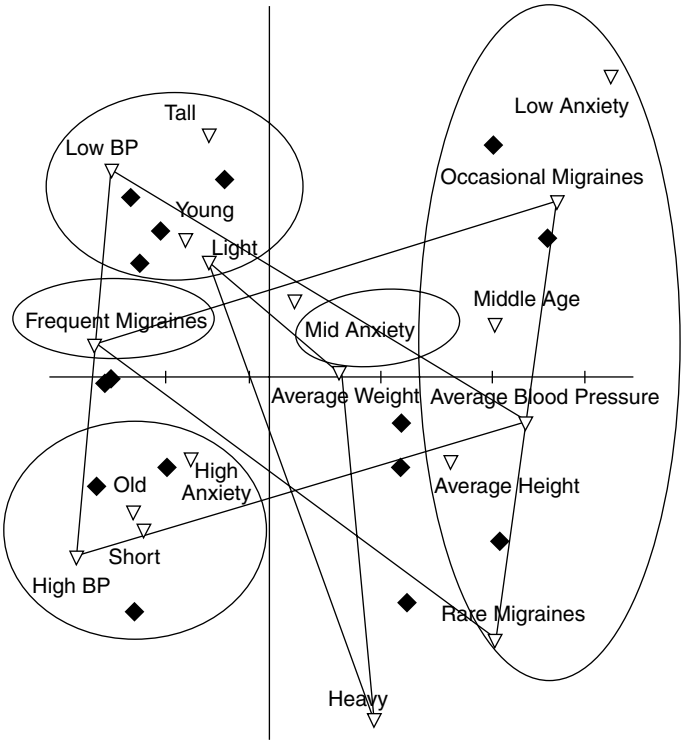


Table 1.15 Sorting of 19 Countries by Five Subjects

Country	*S1*	*S2*	*S3*	*S4*	*S5*	*Solution 1*	*Solution 2*
Britain	1	1	1	1	1	−0.50	−0.69
Canada	5	2	2	2	1	1.06	−0.81
China	2	3	3	3	2	1.53	0.52
Denmark	1	1	1	1	3	−0.73	−0.71
Ethiopia	3	5	5	4	4	−1.00	2.15
Finland	1	4	1	1	3	−0.81	−0.71
France	1	1	1	1	5	−0.73	−0.71
Germany	1	4	1	5	8	−0.50	−0.60
India	4	3	4	3	6	1.02	0.81
Italy	1	4	5	5	7	−0.93	−0.17
Japan	2	3	6	2	8	1.21	−0.01
New Zealand	4	1	6	1	1	0.24	−0.31
Nigeria	3	5	4	4	4	−0.76	2.34
Norway	1	4	1	1	3	−0.81	−0.71
Singapore	4	3	6	3	8	1.12	0.24
Spain	1	5	5	1	7	−0.92	0.34
Switzerland	1	4	1	5	5	−0.85	−0.71
Thailand	4	3	6	3	6	1.20	0.46
United States	5	2	2	2	8	1.17	−0.73

Table 1.16 Inter-Subject Correlation for Two DS Solutions

	Solution 1					*Solution 2*				
Subject 1	1.00					1.00				
Subject 1	0.90	1.00				0.63	1.00			
Subject 3	0.93	0.82	1.00			0.60	0.90	1.00		
Subject 4	0.88	0.99	0.81	1.00		0.98	0.67	0.63	1.00	
Subject 5	0.77	0.87	0.75	0.85	1.00	0.90	0.87	0.82	0.90	1.00

inter-subject correlation matrices associated with the two solutions. In both solutions, the correlation between subjects is relatively high. Figure 1.6 shows only the configuration of 18 of the 19 countries (France is missing because it occupies the same point as Denmark) captured by the first two solutions. The graph clearly shows geographical similarities of the countries.

One commonly observed characteristic of sorting data is that there are often too many dominant solutions to interpret. It must be a reflection of the freedom that the subjects can enjoy in terms of the number of piles and the sizes of piles that are completely in the hands of the subjects. The δ values of the first eight solutions are 19%, 18%, 16%, 11%, 9%, 7%, 6%, and 5%, an unusually gradual drop in percentage from solution to solution. This poses in practice a problem of how many solutions to extract and interpret.

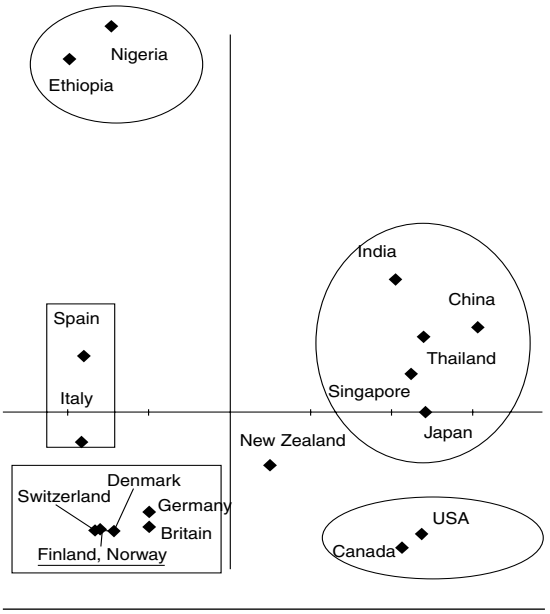
1.6. Scaling of Dominance Data

We will discuss only rank-order data and paired-comparison data. As for DS of successive categories data, see Nishisato (1980, 1986,1994), Nishisato and Sheu (1980), and Odondi (1997).

1.6.1. Rank-Order Data

Ranking is a very popular task in psychological research. For instance, we ask people to rank a number of candidates for a committee and choose the winner in terms of the average ranks of the candidates. Although this popular method for processing ranking data looks reasonable, it is far from even being good and is rather misleading. Why? We will see why such averaged ranks should not be used to evaluate candidates or

Figure 1.6 Sorting of 19 Countries



voters, which becomes obvious once we analyze the same ranking data with dual scaling.

1.6.1.1. Some Basics

Suppose that each of N subjects ranks all of n objects, according to the order of preference, with 1 being the first choice and n being the last choice. Assuming that the number of subjects is greater than that of the objects, the total number of solutions and the total information from the data are given by the following:

$$T(\text{sol}) = n - 1 \text{ and } T(\text{inf}) = \frac{n+1}{3(n-1)}. \qquad (26)$$

When dominance data are subjected to DS, the original rank-order data are first converted to a dominance table. Let us indicate by R_{ij} the rank given to object j by subject i. Then, assuming that each subject ranks n objects, the corresponding dominance number, e_{ij}, is given by the formula

$$e_{ij} = n + 1 - 2R_{ij}, \qquad (27)$$

where e_{ij} indicates the number of times subject i ranked object j before other objects minus the number of times the subject ranked it after other objects. So it indicates relative popularity of each object within each subject. The sum of dominance numbers for each subject is always zero, and the dominance number is bounded between $-(n-1)$ and $(n-1)$. Because dominance numbers are *ipsative* (i.e., each row sum is a constant), we must modify the process of MRA by redefining each row marginal to be $n(n-1)$ and that of column $N(n-1)$. The total number of responses in the dominance table is $Nn(n-1)$. These numbers are based on the fact that each element in the dominance table is the result of $(n-1)$ comparisons between each object and the remaining $(n-1)$ objects (Nishisato, 1978). Using these redefined marginals, we may use MRA for analysis.

The ipsative property of dominance numbers has another implication for quantification: There is no centering constraint on weights for subjects. Thus, the weights for subjects can be all positive or negative. This aspect of quantification of dominance data is very different from that of incidence data, in which both weights for subjects and those for stimuli are centered within each set.

1.6.1.2. Example: Ranking of Municipal Services

Table 1.17 contains ranking of 10 municipal services by 31 students, collected from Nishisato's class in 1982, together with the dominance table. If there were no individual differences, the reasonable scale values or satisfaction values of the 10 government services would be given by the average dominance numbers of the services over subjects. However, in DS, we assume that individual differences are worthwhile variates. The scale values of the services are calculated as averages differentially weighted by subjects' weights. Its main task is to determine appropriate weights for subjects, appropriate in the sense that the variance of the weighted means be a maximum. Individual differences are responsible for multidimensional data structure. $T(\text{sol})$ is 9, and the δ values are in Table 1.18. Considering a relatively sharp drop from Solution 2 to Solution 3, one may decide to look at two solutions, as is done here.

For dominance data, there exists a strict rule for plotting (Nishisato, 1996), namely, plot-normed weights of subjects and projected weights of objects. Then, in the total space, we obtain a configuration such that each subject ranks the closest object first, second closest second, and so on for all subjects and objects—that is, a solution to the Coombs problem of multidimensional unfolding (Coombs, 1964).

Figure 1.7 (p. 18) shows a plot of the first two solutions. A large number of subjects are furthest from postal service, which indicates that postal service is the least satisfactory. This is partly due to the fact that

Table 1.17 Ranking of 10 Government Services in Toronto and Dominance Table

	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
1	1	7	9	10	2	6	3	8	5	4	9	−3	−7	−9	7	−1	5	−5	1	3
2	6	10	9	5	3	1	7	2	4	8	−1	−9	−7	1	5	9	−3	7	3	−5
3	9	8	4	3	5	6	10	2	1	7	−7	−5	3	5	1	−1	−9	7	9	−3
4	2	10	5	6	3	1	4	8	7	9	7	−9	1	−1	5	9	3	−5	−3	−7
5	2	10	6	7	4	1	5	3	9	8	7	−9	−1	−3	3	9	1	5	−7	−5
6	1	3	5	6	7	8	2	4	10	9	9	5	1	−1	−3	−5	7	3	−9	−7
7	7	10	1	6	5	3	8	4	2	9	−3	−9	9	−1	1	5	−5	3	7	−7
8	2	10	6	7	4	1	5	3	9	8	7	−9	−1	−3	3	9	1	5	−7	−5
9	2	10	5	8	4	1	6	3	7	9	7	−9	1	−5	3	9	−1	5	−3	−7
10	2	10	5	9	8	7	4	1	3	6	7	−9	1	−7	−5	−3	3	9	5	−1
11	9	10	7	6	5	1	4	2	3	8	−7	−9	−3	−1	1	9	3	7	5	−5
12	6	10	7	4	2	1	3	9	8	5	−1	−9	−3	3	7	9	5	−7	−5	1
13	1	10	3	9	6	4	5	2	7	8	9	−9	5	−7	−1	3	1	7	−3	−5
14	8	6	5	3	10	7	9	2	1	4	−5	−1	1	5	−9	−3	−7	7	9	3
15	8	10	9	6	4	1	3	2	5	7	−5	−9	−7	−1	3	9	5	7	1	−3
16	3	5	10	4	6	9	8	2	1	7	5	1	−9	3	−1	−7	−5	7	9	−3
17	1	10	8	9	3	5	2	6	7	4	9	−9	−5	−7	5	1	7	−1	−3	3
18	5	4	9	3	10	8	7	2	1	6	1	3	−7	5	−9	−5	−3	7	9	−1
19	2	10	6	7	8	1	5	4	3	9	7	−9	−1	−3	−5	9	1	3	5	−7
20	1	4	2	10	9	7	6	3	5	8	9	3	7	−9	−7	−3	−1	5	1	−5
21	2	10	5	7	3	1	4	6	8	9	7	−9	1	−3	5	9	3	−1	−5	−7
22	6	3	9	4	10	8	7	2	1	5	−1	5	−7	3	−9	−5	−3	7	9	1
23	6	9	10	4	8	7	5	2	1	3	−1	−7	−9	3	−5	−3	1	7	9	5
24	5	2	1	9	10	4	8	6	3	7	1	7	9	−7	−9	3	−5	−1	5	−3
25	2	10	6	7	9	1	3	4	5	8	7	−9	−1	−3	−7	9	5	3	1	−5
26	7	10	9	5	2	6	3	1	4	8	−3	−9	−7	1	7	−1	5	9	3	−5
27	8	7	10	3	5	9	4	2	1	6	−5	−3	−9	5	1	−7	3	7	9	−1
28	3	8	6	7	5	10	9	2	4	1	5	−5	−1	−3	1	−9	−7	7	3	9
29	2	10	7	9	4	1	5	3	6	8	7	−9	−3	−7	3	9	1	5	−1	−5
30	2	10	9	1	4	7	5	3	6	8	7	−9	−7	9	3	−3	1	5	−1	−5
31	4	10	9	7	5	1	3	2	6	8	3	−9	−7	−3	1	9	5	7	−1	−5

Table 1.18 Nine Solutions and Their Contributions

	Solution								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Delta	37.9	22.4	13.4	10.6	4.9	4.2	2.7	2.2	1.9
CumDelta	37.9	60.2	73.6	84.2	89.0	93.2	95.9	98.1	100.0

the data were collected shortly after a major postal strike. There are groups who prefer theaters first and restaurants second, or vice versa, suggesting that those who go to theaters must go to good restaurants near the theaters. The most dominant group considers public libraries most satisfactory. One important message of this graphical analysis is that it is very difficult, if not impossible, to interpret the configuration of only services. When we plot subjects and see they are all scattered in the space, the configuration of the services suddenly becomes meaningful because they provide us with how they view those services in terms of satisfaction.

One can calculate the distance from each subject (normed) to each service (projected) in the two-dimensional graph and see if indeed the ranking of distances between each subject and each of the 10 services is close to the ranking in the input data. The ranking thus derived from the first two solutions is called rank-2 approximation to the input ranking. The DUAL3 (Nishisato & Nishisato, 1994b) provides these distances and approximated ranks. The distances between each of the first five subjects and the 10 services and the rank-2 and rank-8 approximations to input ranks are in Tables 1.19 and 1.20. The rank-9 approximation perfectly reproduces the

Table 1.19 Rank 2: Distances and Ranks of Distances

	Service									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Distances										
Subject 1	0.20	2.05	0.66	1.32	0.26	0.13	0.29	1.29	1.81	1.29
Subject 2	2.05	5.42	3.65	2.79	2.43	1.81	2.33	1.14	2.09	3.64
Subject 3	3.03	3.48	3.40	1.76	3.08	3.35	2.94	1.08	0.94	2.49
Subject 4	1.33	4.82	2.48	3.56	1.57	0.95	1.63	2.97	4.10	3.60
Subject 5	1.31	5.26	2.81	3.40	1.65	0.87	1.66	2.29	3.55	3.72
Ranks of distances										
Subject 1	2	10	5	8	3	1	4	7	9	6
Subject 2	3	10	9	7	6	2	5	1	4	8
Subject 3	6	10	9	3	7	8	5	2	1	4
Subject 4	2	10	5	7	3	1	4	6	9	8
Subject 5	2	10	6	7	3	1	4	5	8	9

Table 1.20 Rank 8: Distances and Ranks of Distances

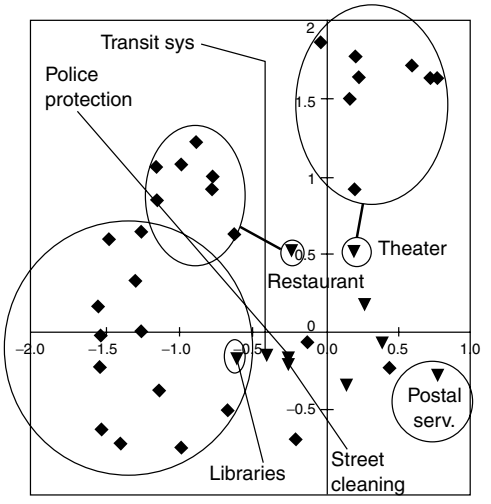
	Service									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Distances										
Subject 1	13.90	16.75	17.42	17.76	14.20	16.15	14.63	16.91	15.67	15.02
Subject 2	5.79	8.16	6.69	4.99	4.40	4.03	5.49	4.36	4.49	6.78
Subject 3	11.29	11.02	9.04	8.48	9.36	9.98	11.58	8.07	7.73	10.18
Subject 4	4.99	8.49	6.52	6.75	5.24	4.32	6.05	7.30	7.44	7.71
Subject 5	2.70	6.79	4.09	4.59	3.52	2.66	3.36	3.43	5.45	5.42
Ranks of distances										
Subject 1	1	7	9	10	2	6	3	8	5	4
Subject 2	7	10	8	5	3	1	6	2	4	9
Subject 3	9	8	4	3	5	6	10	2	1	7
Subject 4	2	10	5	6	3	1	4	7	8	9
Subject 5	2	10	6	7	5	1	3	4	9	8

Table 1.21 Average Squared Rank Discrepancies

	Rank k										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Solution 1*	*Solution 2*
Subject 1	8.8	7.8	9.0	4.6	4.2	1.4	1.6	0.0	0.0	0.65	−0.51
Subject 2	6.2	2.8	1.4	0.2	0.4	0.4	0.2	0.4	0.0	1.15	1.08
Subject 3	19.6	8.0	8.0	1.2	1.2	0.0	0.0	0.0	0.0	−0.16	1.51
Subject 4	1.4	1.0	1.2	1.6	1.6	1.6	0.6	0.2	0.0	1.39	−0.73
Subject 5	1.2	0.8	1.4	1.4	1.4	1.0	0.8	0.6	0.0	1.54	−0.21

input ranks. It is useful to look at average squared rank discrepancies between these approximated ranks and the original ranks (see Table 1.21). Notice that the rank-9 approximation reproduced the input ranks, thus showing no discrepancies. Table 1.21 also lists normed weights for those five subjects, which should be all equal to 1.00 if no individual differences were involved.

Figure 1.7 Ten Government Services



1.6.2. Paired-Comparison Data

The method of paired comparison (see Bock & Jones, 1968) has been one of the pillars in the history of psychological scaling. For a unidimensional preference scale to be constructed from paired-comparison data, we must avoid intransitive judgments (e.g., A is preferred to B, B to C, and C to A), and we must consider individual differences as random fluctuations of judgments. But in real data, we see many intransitive judgments and substantial individual differences. For us to analyze such paired-comparison data, therefore, we must consider a multidimensional scale and treat individual differences as legitimate variates for analysis. This mode of more realistic analysis than the traditional method of paired comparisons is what dual scaling offers. There is no need to worry about unidimensionality, for dual scaling yields as many dimensions as data dictate. We will see how paired-comparison data can be effectively analyzed by dual scaling.

1.6.2.1. Some Basics

For n objects, create all $n(n-1)/2$ possible pairs, present each pair to N subjects, and ask which object in the pair they like better. Collected in this way, such paired-comparison data have mathematically the same structure as the N-by-n rank-order data: $T(\text{sol})$ and $T(\text{inf})$ are identical to those of rank-order data. The only difference is that in rank order, one must arrange all objects in a single order, whereas in paired comparisons, one can anticipate so-called intransitive choices (e.g., A is preferred to B, B is preferred to C, and C is preferred to A). For subject i and pair (X_j, X_k), Nishisato (1978) defined a response variable as follows:

$$ {}_i f_{jk} = \begin{cases} 1 & \text{if } X_j > X_k \\ 0 & \text{if } X_j = X_k \\ -1 & \text{if } X_j < X_k \end{cases} . \tag{28} $$

The subjects-by-objects dominance table can be obtained by transforming ${}_i f_{jk}$ to e_{ij} by the following formula:

$$ e_{ij} = \sum_{\substack{k=1 \\ k \neq j}}^{n} {}_i f_{jk}. \tag{29} $$

Recall that the dominance numbers were easily obtained for rank-order data by a simpler formula than this. The meaning is the same; that is, e_{ij} is the number of times subject i preferred X_j to X_k minus the number of times subject i preferred other objects to X_j.

1.6.2.2. Wiggins's Christmas Party Plans

As a course assignment, Ian Wiggins, now a successful consultant in Toronto, collected paired-comparison data[4] from 14 researchers at a research institute on his eight Christmas party plans:

1. A potluck at someone's home in the evening
2. A potluck in the group room
3. A pub/restaurant crawl after work
4. A reasonably priced lunch in an area restaurant
5. Keep to one's self
6. An evening banquet at a restaurant
7. A potluck at someone's home after work
8. A ritzy lunch at a good restaurant (tablecloths)

Table 1.22 contains data in the form of subjects (14) by pairs (28 pairs), with elements being 1 if the subject prefers the first plan to the second one and 2 if the second plan is preferred to the first ("2" will be later changed to "−1" for analysis). Dominance numbers are in Table 1.23. As is the case with rank-order data, each element of the 14×8 dominance table is based on seven comparisons. Or, more generally, for the $N \times n$ dominance table, each element is based on $(n-1)$ comparisons. Therefore, the marginal frequency of responses for each row is $n(n-1)$ and that of each column is $N(n-1)$.

4. Data used with permission from Ian Wiggins.

Table 1.22 Wiggins's Christmas Party Plans Data

j	*1111111*	*222222*	*33333*	*4444*	*555*	*66*	*7*
k	*2345678*	*345678*	*45678*	*5678*	*678*	*78*	*8*
1	1121121	222222	21121	1121	121	21	2
2	2221212	121212	21112	1112	222	12	2
3	1111121	111121	11121	1121	222	21	1
4	2121112	111112	21222	1112	222	22	2
5	2221212	221222	21212	1111	222	12	2
6	1111111	221222	21222	1111	222	22	1
7	1111121	121121	21121	1121	222	22	1
8	1111121	121221	21221	1221	221	21	1
9	1221121	221122	11121	1121	222	22	1
10	1211222	221222	11111	1222	222	11	2
11	1211111	222222	11111	1111	222	22	2
12	2222122	121111	21111	1111	111	22	1
13	1211212	222222	11111	1212	222	11	2
14	2222121	211111	11111	2121	121	21	1

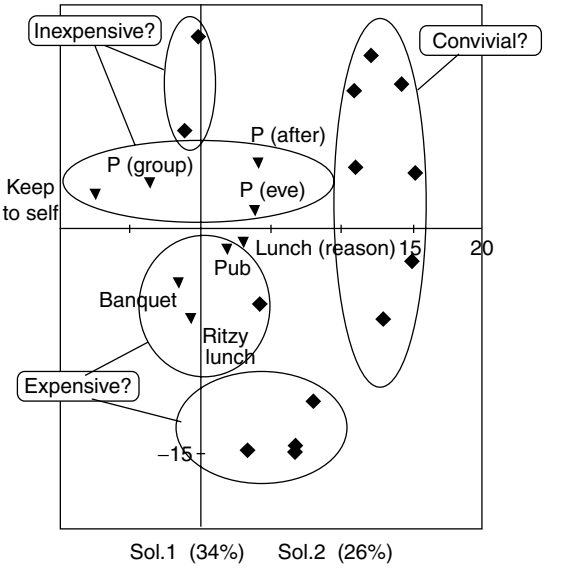
Table 1.23 Dominance Table

j	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	3	−7	1	5	−1	−3	5	−3
2	−3	1	−1	5	−7	1	−5	7
3	5	3	1	−1	−7	−3	7	−5
4	1	5	−5	3	−7	−3	−1	7
5	−3	−3	1	7	−7	3	−3	5
6	7	−5	−3	5	−7	−1	3	1
7	5	1	−1	3	−7	−5	7	−3
8	5	−1	−3	1	−5	3	7	−7
9	1	−3	5	3	−7	−5	7	−1
10	−1	−5	7	−3	−7	5	1	3
11	5	−7	7	3	−5	−3	−1	1
12	−5	5	3	7	1	−7	−1	−3
13	1	−7	7	−1	−5	5	−3	3
14	−3	5	7	−1	1	−5	3	−7

Table 1.24 Contributions of Seven Solutions to Total Information

	Solution						
	1	*2*	*3*	*4*	*5*	*6*	*7*
Delta	34	26	16	13	7	3	1
CumDelta	34	60	76	89	96	99	100

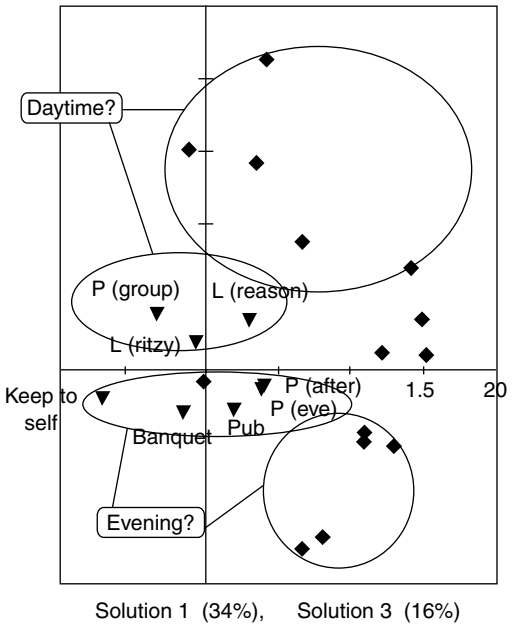
Figure 1.8 Solutions 1 and 2



From the dominance table, it is clear that Plan 5 is not very popular because the corresponding elements from 14 subjects are mostly negative. If we calculate the mean dominance numbers of the eight columns, they may provide good unidimensional estimates of preference values of the party plans, provided that individual differences are negligible. In DS, we weight subjects differentially in such a way that the variance of the eight weighted averages be a maximum. For the present data set, $T(\text{sol})$ is 7, and the corresponding δ values are in Table 1.24. Although weights are not listed here, Solution 4 is dominated only by one variable, that is, "pub/restaurant crawl." In contrast, the first three solutions present a variety of preference patterns. Therefore, let us look at the first three solutions. Figures 1.8 and 1.9 show the following: Dimension 1 divides party plans into the convivial side and the "Keep to one's self" side, Dimension 2 separates plans into expensive and nonexpensive, and Dimension 3 divides party plans into daytime parties and evening parties. Note that the weights of subjects on Solution 1 are mostly positive, but that those on Solutions 2 and 3 are much more evenly distributed than those on Solution 1. This is a reflection of the property of dominance data that the weights for subjects are not centered, due to the row-ipsative nature of dominance data, and are free to vary.

That subjects are scattered in the three-dimensional space means that different subjects prefer different party plans. As noted earlier, each subject in total space ranks the closest plan first. The graphs offer an interesting way to look at individual differences in judgment: DS can accommodate any patterns or combinations of different aspects of the party, such as daytime-inexpensive, daytime-expensive, evening-inexpensive, and evening-expensive.

Figure 1.9 Solutions 1 and 3



1.7. Forced Classification for Multiple-Choice Data

We have seen dual scaling of multiple-choice data, and it was noted that dual scaling maximizes the average of all possible inter-item correlation coefficients. There are occasions, however, when we are not interested in all the items but only one item. For instance, if we collect children's background medical and psychological information in addition to whether or not they have allergy problems, we would be interested in finding which of the medical and psychological variables may be related to the allergy problems. In this case, we are no longer interested in scaling data to maximize the average inter-variable correlation, but our interest now lies in the scaling method that maximizes the correlation between the allergy variable and the other variables. This task is carried out by the procedure called *forced classification*.

Nishisato (1984) proposed a simple procedure to carry out the above task, which is nothing but discriminant analysis with categorical data. It is based on two principles: principle of internal consistency (PIC) and principle of equivalent partitioning (PEP). Let us denote the data of n multiple-choice questions from N subjects as

$$\mathbf{F} = [\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_j, \ldots, \mathbf{F}_n], \tag{30}$$

where $\mathbf{F}_j$ is an N-by-m_j matrix, in which the row i consists of subject i's response to item j, with 1 being the choice and 0s the nonchoices out of m_j options. Each subject chooses only one option per item. Suppose that we repeat $\mathbf{F}_j$ k times in the data matrix. As k increases, the response patterns in $\mathbf{F}_j$ become more dominant in the data set, and eventually we will see that the response patterns in the repeated $\mathbf{F}_j$ determine the first solution (PIC). Instead of repeating $\mathbf{F}_j$ k times, it is known that the same dual-scaling results can be obtained from analysis of the following matrix (PEP):

$$[\mathbf{F}_1, \mathbf{F}_2, \ldots, k\mathbf{F}_j, \ldots, \mathbf{F}_n]. \tag{31}$$

This matrix is obtained from the original matrix by replacing each 1 in $\mathbf{F}_j$ with a k. Thus, the computation involved here is ordinary DS with an altered data matrix by multiplying the chosen submatrix by a large enough scalar k. Possible applications of this procedure are, for instance, the following:

1. to identify personality traits that are closely related to the school dropout,
2. to find out if academic performance is influenced by some environmental factors (school buildings, computers, etc.),
3. to see if the high blood pressure is related to the regions where people live,
4. to collect questions related to anxiety for the construction of an anxiety scale
5. to eliminate age effects, if any, from consumer data on purchase patterns of cosmetics after finding significant age effects.

Due to the limited space for this chapter, a numerical example of forced classification is not given here. Please refer to Nishisato and Gaul (1990) for its applications to marketing research and to Nishisato and Baba (1999) for the latest development.

1.8. Mathematics of Dual Scaling

1.8.1. Structure of Data

Given a two-way table of data with typical element f_{ij}, singular-value decomposition can be described as bilinear decomposition:

$$f_{ij} = \frac{f_{i.}f_{.j}}{f_{..}}[1 + \rho_1 y_{i1}x_{j1} + \rho_2 y_{i2}x_{j2} + \cdots + \rho_k y_{ik}x_{jk}], \tag{32}$$

where ρ_k is the kth largest singular value, y_{ik} is the ith element of singular vector y_k for the rows, and x_{jk} is the jth element of singular vector x_k for the columns of the table. These singular vectors can be viewed as weight vectors for the rows and the columns. The first term inside the bracket—that is, the element 1—is called a trivial solution associated with the case in which the rows and the columns are statistically independent. Another well-known expression of the singular-value decomposition is what Benzécri et al. (1973) call *transition formulas* and Nishisato (1980) refers to as *dual relations:*

$$y_{ik} = \frac{1}{\rho_k} \frac{\Sigma f_{ij} x_{jk}}{f_{i.}}; \; x_{jk} = \frac{1}{\rho_k} \frac{\Sigma f_{ij} y_{ik}}{f_{.j}}. \qquad (33)$$

These weights, y_{ik}, x_{jk}, are called *normed weights* (Nishisato, 1980) or *standard coordinates* (Greenacre, 1984). If we multiply the formulas by ρ_k, the resultant weights are called *projected weights* (Nishisato, 1980) or *principal coordinates* (Greenacre, 1984). The projected weights are

$$\rho_k y_{ik} = \sum_{j=1}^{n} \frac{f_{ij} x_{jk}}{f_{i.}}, \; \rho_k x_{jk} = \sum_{i=1}^{m} \frac{f_{ij} y_{ik}}{f_{.j}}. \qquad (34)$$

The above sets of formulas hold for any data matrix (f_{ij}).

To arrive at these formulas, one can define the task in many ways, which is probably one of the reasons why so many researchers have discovered the method independently and coined their own names. For example, one may state the problem in any of the following ways:

- Determine x_{jk} and y_{ik} in such a way that the data weighted by x_{jk} and the data weighted by y_{ik} attain the maximal product-moment correlation.
- Determine x_{jk} to make the between-row sum of squares, relative to the total sum of squares, be a maximum; determine y_{ik} so as to make the between-column sum of squares to the total sum of squares be a maximum.
- Determine those two sets of weights to make the regression of the rows on the columns and the regression of the columns on the rows be simultaneously linear.
- Determine those two sets of weights in such a way to make the sum of the squared differences between f_{ij} and $\frac{f_{i.} f_{.j}}{f_{..}} x \rho y_{ik} x_{jk}$ be a minimum.

All of these lead to the identical solution set (ρ_k, y_{ik}, x_{jk}). For detailed mathematical derivations, see Benzécri (1973), Nishisato (1980, 1994), Greenacre (1984), and Gifi (1990).

1.8.2. Row Space and Column Space Are Different

We are interested in the relations between rows and columns of a two-way table, for example, relations between subjects and chosen objects. Unfortunately, the space for row variables and the space for column variables are different, the discrepancy of which is related to the cosine of the singular values. In other words, when singular values are relatively large, the discrepancy between the row space and the column space is comparatively small. When we want to put both row and column variables in the same space, we must plot normed weights of rows (or columns) and projected weights of columns (or rows). Then, both sets of weights span the same space. We often talk about symmetric scaling to indicate that both projected row and projected column weights are plotted, in which case care must be exercised in judging their distances because of the discrepancy of the two spaces. Or, rather, symmetric scaling may be justified only when singular values are close to 1. Nonsymmetric scaling of one set of weights to be projected to the other set is the mathematically correct one, but we must often deal with a rather nasty problem of a large difference between the spread of normed weights and that of projected weights, the latter being often too much smaller than the former, making comparisons between them difficult. See Nishisato and Clavel (2003) for a discussion on the discrepant spaces and the calculation of distances between points in two different spaces.

1.8.3. Chi-Square Metric and Data Types

One of the difficult problems in quantifying incidence data lies in its use of the chi-square metric, which is necessitated by the sheer characteristics of the data. When Point A has one observation and Point B nine observations, the midpoint between them is 9 units away from A and one unit away from B. This is an example of a chi-square metric, which is a reciprocal function of the number of observations. In the above example, the distance between A and the midpoint times 1 (observation) is equal to the distance between the midpoint and B times 9. Thus, the point with more observations has a stronger pull than the point with fewer observations.

In contrast, each cell in the dominance table is represented by a constant number of observations (i.e., $n - 1$). Therefore, the chi-square metric is reduced to the Euclidean metric, where the midpoint between A and B is located halfway between A and B. It should be remembered, however, that the way in which DS

handles dominance data is to treat dominance numbers as cardinal numbers, rather than ordinal. At the present moment, we have not developed an ordinal way of handling dominance numbers. This is one problem for future research. Another point of caution is that both *chi-square metric* and *Euclidean metric* are defined for the Euclidean space.

1.9. Linear Analysis and Dual Scaling

In the principal coordinate system, each continuous variable is expressed as a straight line (axis), whereas categories of each variable in DS no longer lie on a straight line. In consequence, when data are in multidimensional space, the contribution or information of each variable in PCA is expressed by the length of its vector, which increases as the dimensionality increases, whereas the contribution of each variable in DS increases as the dimensionality increases in a distinctively different way from PCA. The DS contribution of each variable to the given space is not expressed by the length of any vector but by the area or volume formed by connecting the points of those categories of the variable.

If an item has three categories, the information of the variable in the given dimension is the area of a triangle obtained by connecting the three category points in the space. The area of the triangle monotonically increases as the dimensionality of the space for the data increases. If a variable has four categories, the information of the variable in three-dimensional or higher dimensional space is given by the volume of the form created by connecting four-category points. If the variable has n categories, the information of the variable in $n-1$ or higher dimensional space is given by the volume of the form created by connecting n points.

Thus, by stretching our imagination to the continuous variable, where the number of categories is considered very large but finite, we can conclude that the information of the variable in the given space must be expressed by the volume of a shape and not by the length of a vector. This conjecture can be reinforced by the fact that many key statistics associated with dual scaling are related to the number of categories of variables. Some of the examples are given below.

The total number of dimensions required to accommodate a variable with m_j categories is

$$N_j = m_j - 1. \tag{35}$$

The total number of dimensions needed for n variables is

$$N_T = \sum_{j=1}^{n}(m_j - 1) = \sum_{j=1}^{n} m_j - n = m - n. \tag{36}$$

The total amount of information in the data—that is, the sum of the squared singular values, excluding 1—is given by

$$\sum_{k=1}^{K} \rho_j^2 = \frac{\sum_{j=1}^{n} m_j}{n} - 1 = \overline{m} - 1. \tag{37}$$

Therefore, as the number of categories of each variable increases, so does the total information in the data set. The information of variable j with m_j categories is given by

$$\sum_{k=1}^{m-n} r_{jt(k)}^2 = m_j - 1. \tag{38}$$

These are all related to the number of categories of each variable. Thus, we can imagine what will happen as m_j increases to infinity or, in practice, to the number of observations (subjects) N. An inevitable conclusion, then, seems to be that the total information in the data set is much more than the sum of the lengths of vectors of the variables in multidimensional space: It is the sum of the volumes of hyperspheres associated with categories of individual variables.

The above conclusion (Nishisato, 2002) suggests how little information popular linear analyses such as PCA and factor analysis capture. Traditionally, the total information is defined by the sum of the eigenvalues associated with a linear model. But we have just observed that it seems inappropriate unless we are totally confined within the context of a linear model. In a more general context, in which we consider both linear and nonlinear relations among variables, DS offers the sum of the eigenvalues as a reasonable statistic of the total information in the data. As the brain wave analyzer filters a particular wave such as alpha, most statistical procedures—particularly PCA, factor analysis, other correlational methods, and multidimensional scaling—play the role of a linear filter and filter out most of the information from the data, that is, a nonlinear portion of the data. In this context, dual scaling should be reevaluated and highlighted as a means for analyzing both linear and nonlinear information in the data, particularly in the behavioral sciences, where it seems that nonlinear relations are more abundant than linear relations.

REFERENCES

Bartlett, M. S. (1951). The goodness of fit of a single hypothetical discriminant function in the case of several groups. *Annals of Eugenics, 16*, 199–214.

Beltrami, E. (1873). Sulle funzioni bilineari (On the bilinear functions). *Giornale di Mathematiche, 11*, 98–106.

Benzécri, J. P. (1969). Statistical analysis as a tool to make patterns emerge from data. In S. Watanabe (Ed.), *Methodologies of pattern recognition* (pp. 35–74). New York: Academic Press.

Benzécri, J. P., et al. (1973). *L'analyse des données: II. L'analyse des correspondances* (The analysis of data: Vol. 2. The analysis of correspondences). Paris: Dunod.

Benzécri, J. P. (1982). *Histoire et préhistoire de l'analyse des données* (History and prehistory of the analysis of data). Paris: Dunod.

Bock, R. D. (1956). The selection of judges for preference testing. *Psychometrika, 21*, 349–366.

Bock, R. D. (1960). Methods and applications of optimal scaling (University of North Carolina Psychometric Laboratory Research Memorandum, No. 25). Chapel Hill: University of North Carolina.

Bock, R. D., & Jones, L. V. (1968). *Measurement and prediction of judgment and choice*. San Francisco: Holden-Day.

Coombs, C. H. (1964). *A theory of data*. New York: John Wiley.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

de Leeuw, J. (1973). *Canonical analysis of categorical data*. Leiden, The Netherlands: DSWO Press.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, 16*, 211–218.

Escofier-Cordier, B. (1969). L'analyse factorielle des correspondances (Factor analysis of correspondences). Paris: Bureau Univesitaire de Recherche operationelle, Cahiers, Série Recherche, Université de Paris.

Fisher, R. A. (1940). The precision of discriminant functions. *Annals of Eugenics, 10*, 422–429.

Fisher, R. A. (1948). *Statistical methods for research workers*. London: Oliver and Boyd.

Franke, G. R. (1985). Evaluating measures through data quantification: Applying dual scaling to an advertising copytest. *Journal of Business Research, 13*, 61–69.

Gabriel, K. R. (1971). The biplot graphical display of matrices with applications to principal component analysis. *Biometrics, 58*, 453–467.

Gifi, A. (1980). *Nonlinear multivariate analysis*. Unpublished manuscript.

Gifi, A. (1990). *Nonlinear multivariate analysis*. New York: John Wiley.

Gower, J. C., & Hand, D. J. (1996). *Biplots*. London: Chapman & Hall.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Greenacre, M. J., & Blasius, J. (Eds.). (1994). *Correspondence analysis in the social sciences*. London: Academic Press.

Greenacre, M. J., & Torres-Lacomba, A. (1999). *A note on the dual scaling of dominance data and its relationship to correspondence analysis* (Working Paper Ref. 430). Barcelona, Spain: Departament d'Economia i Empresa, Universidat Pompeu Fabra.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In the Committee on Social Adjustment (Ed.), *The prediction of personal adjustment* (pp. 319–348). New York: Social Science Research Council.

Guttman, L. (1946). An approach for quantifying paired comparisons and rank order. *Annals of Mathematical Statistics, 17*, 144–163.

Hayashi, C. (1950). On the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics, 2*, 35–47.

Hayashi, C. (1952). On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics, 3*, 69–98.

Hill, M. O. (1973). Reciprocal averaging: An eigenvector method of ordination. *Journal of Ecology, 61*, 237–249.

Hirschfeld, H. O. (1935). A connection between correlation and contingency. *Cambridge Philosophical Society Proceedings, 31*, 520–524.

Horst, P. (1935). Measuring complex attitudes. *Journal of Social Psychology, 6*, 369–374.

Hotelling, H. (1933). Analysis of complex of statistical variables into principal components. *Journal of Educational Psychology, 24*, 417–441, 498–520.

Jackson, D. N., & Helmes, E. (1979). Basic structure content scaling. *Applied Psychological Measurement, 3*, 313–325.

Johnson, P. O. (1950). The quantification of qualitative data in discriminant analysis. *Journal of the American Statistical Association, 45*, 65–76.

Jordan, C. (1874). Mémoire sur les formes binlinieres (Note on bilinear forms). *Journal de Mathématiques Pures et Appliquées, deuxiéme Série, 19*, 35–54.

Lancaster, H. O. (1958). The structure of bivariate distribution. *Annals of Mathematical Statistics, 29*, 719–736.

Lebart, L., Morineau, A., & Warwick, K. M. (1984). *Multivariate descriptive statistical analysis*. New York: John Wiley.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychlogy, 140*, 44–53.

Lingoes, J. C. (1964). Simultaneous linear regression: An IBM 7090 program for analyzing metric/nonmetric or linear/nonlinear data. *Behavioral Science, 9*, 87–88.

Lord, F. M. (1958). Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika, 23*, 291–296.

Maung, K. (1941). Measurement of association in contingency tables with special reference to the pigmentation of hair and eye colours of Scottish children. *Annals of Eugenics, 11*, 189–223.

Meulman, J. J. (1998). Review of W. J. Krzanowski and F. H. C. Marriott, "Multivariate analysis: Part I. Distributions, ordinations, and inference." *Journal of Classification, 15*, 297–298.

Mosier, C. I. (1946). Machine methods in scaling by reciprocal averages. In *Proceedings, research forum* (pp. 35–39). Endicath, NY: International Business Corporation.

Nishisato, S. (1978). Optimal scaling of paired comparison and rank order data: An alternative to Guttman's formulation. *Psychometrika, 43*, 263–271.

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press.

Nishisato, S. (1984). Forced classification: A simple application of quantification technique. *Psychometrika, 49*, 25–36.

Nishisato, S. (1986). Multidimensional analysis of successive categories. In J. de Leeuw, W. Heiser, J. Meulman, & F. Critchley (Eds.), *Multidimensional data analysis*. Leiden, The Netherlands: DSWO Press.

Nishisato, S. (1988, June). *Effects of coding on dual scaling*. Paper presented at the annual meeting of the Psychometric Society, University of California, Los Angeles.

Nishisato, S. (1993). On quantifying different types of categorical data. *Psychometrika, 58*, 617–629.

Nishisato, S. (1994). *Elements of dual scaling*. Hillsdale, NJ: Lawrence Erlbaum.

Nishisato, S. (1996). Gleaning in the field of dual scaling. *Psychometrika, 61*, 559–599.

Nishisato, S. (2000). Data analysis and information: Beyond the current practice of data analysis. In R. Decker & W. Gaul (Eds.), *Classification and information processing at the turn of the millennium* (pp. 40–51). Heidelberg: Springer-Verlag.

Nishisato, S. (2002). Differences in data structure between continuous and categorical variables as viewed from dual scaling perspectives, and a suggestion for a unified mode of analysis. *Japanese Journal of Sensory Evaluation, 6*, 89–94 (in Japanese).

Nishisato, S. (2003). Geometric perspectives of dual scaling for assessment of information in data. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 453–463). Tokyo: Springer-Verlag.

Nishisato, S., & Baba, Y. (1999). On contingency, projection and forced classification of dual scaling. *Behaviormetrika, 26*, 207–219.

Nishisato, S., & Clavel, J. G. (2003). A note on between-set distances in dual scaling and correspondence analysis. *Behaviormetrika, 30*(1), 87–98.

Nishisato, S., & Gaul, W. (1990). An approach to marketing data analysis: The forced classification procedure of dual scaling. *Journal of Marketing Research, 27*, 354–360.

Nishisato, S., & Nishisato, I. (1994a). *Dual scaling in a nutshell*. Toronto: MicroStats.

Nishisato, S., & Nishisato, I. (1994b). *The DUAL3 for Windows*. Toronto: MicroStats.

Nishisato, S., & Sheu, W. J. (1980). Piecewise method of reciprocal averages for dual scaling of multiple-choice data. *Psychometrika, 45*, 467–478.

Noma, E. (1982). The simultaneous scaling of cited and citing articles in a common space. *Scientometrics, 4*, 205–231.

Odondi, M. J. (1997). *Multidimensional analysis of successive categories (rating) data by dual scaling*. Unpublished doctoral dissertation, University of Toronto.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazines and Journal of Science, Series 6, 2*, 559–572.

Richardson, M., & Kuder, G. F. (1933). Making a rating scale that measures. *Personnel Journal, 12*, 36–40.

Schmidt, E. (1907). Zür Theorie der linearen und nichtlinearen Integralgleichungen. Erster Teil. Entwickelung willkürlicher Functionen nach Systemaen vorgeschriebener (On the theory of linear and nonlinear integral equations. Part one. Development of arbitrary functions according to prescribed systems). *Mathematische Annalen, 63*, 433–476.

Sheskin, D. J. (1997). *Handbook of parametric and nonparametric procedures*. Boca Raton, FL: CRC Press.

Takane, Y. (1980). Analysis of categorizing behavior. *Behaviormetrika, 8*, 75–86.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: John Wiley.

van de Velden, M. (2000). Dual scaling and correspondence analysis of rank order data. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis* (pp. 87–99). Dordrecht, The Netherlands: Kluwer Academic.

van Meter, K. M., Schiltz, M., Cibois, P., & Mounier, L. (1994). Correspondence analysis: A history and French sociological perspectives. In M. Greenacre & J. Blasius (Eds.), *Correspondence analysis in the social sciences* (pp. 128–137). London: Academic Press.

Williams, E. J. (1952). Use of scores for the analysis of association in contingency tables. *Biometrika, 39*, 274–289.

Chapter 2

Multidimensional Scaling and Unfolding of Symmetric and Asymmetric Proximity Relations

Willem J. Heiser

Frank M. T. A. Busing

2.1. Introduction: Relations and Relational Systems

The behavioral and social sciences have produced an extensive methodology to study relations. In psychology, studying relations often implies studying the overall strength of a relation between variables. For example, one may ask how strongly—if at all—aggression depends on frustration. It is common to express the strength of such relations in a correlation coefficient, an F-ratio, or a chi-square value and to test their significance. However, there are also occasions in which interest is not so much in the overall strength of a single relation but in the *details of a complete relational system.* In one of the applications that will be discussed later, for example, knowledge diffusion in a social network of 12 psychological journals is studied by analyzing the citing behavior among all pairs of them. In another application that will serve to illustrate how to study relational differences, the starting point is a set of ratings provided by six samples of judges, who had to assess the relative friendliness or hostility between nations at the time of World War II. In each case, the goal is to model *all pairwise relations* to discover or confirm which factors are operative in the system, among many potential ones.

Because a relational system may involve relations between variables or between persons, between stimuli or between responses, between processes or between concepts, and even between combinations of all of these, there is room for a variety of relational systems. Therefore, it is useful to delineate the position of multidimensional scaling (MDS) and unfolding, which are a group of analysis techniques for relational data, compared to some cognate methodologies for analyzing relational systems.

AUTHORS' NOTE: The authors thank Natale Leroux for help with the data entry of the Breakfast data.

2.1.1. Proximity and Dominance Relations

To start off, MDS and unfolding provide models for proximity relations. When Roger Shepard introduced his revolutionary nonmetric multidimensional scaling technique (Shepard, 1962), he was interested in modeling the substitution errors between stimuli during identification learning as a way to describe and explain stimulus generalization processes by their psychological similarity (Green & Anderson, 1955; Rothkopf, 1957). But he also noted that for other classes of entities, the notion of similarity or substitutability seems inappropriate while we are still interested in relations of *closeness* or *remoteness*—for instance, when we study communication between persons in groups (Frank, 1996) or when we run experiments using word-association or free-recall tasks (Henley, 1969). He then coined the generic term *proximity* for all relations of this kind.

The second pioneer of multidimensional scaling and unfolding methods, Clyde Coombs, in his *Theory of Data* (1964), introduced the notion that all psychological observations can be interpreted as a relational system and distinguished proximity relations from *dominance* or *order* relations. *Proximity* refers to psychological nearness, which is symmetric (if orange is near red, then red must also be near orange), whereas *dominance* refers to a hierarchy or an ordering among the objects involved, which implies lack of symmetry (if Roger dominates Clyde, then Clyde cannot dominate Roger at the same time). Typical techniques for studying dominance relations are based on paired-comparison models if the relation is defined on one set of objects, for example, stimuli (cf. Critchlow & Fligner, 1991). Item response (IRT) models are used if the relation is defined on two sets of objects, where the prime example is persons and items (cf. Fischer & Molenaar, 1995; Van der Linden & Hambleton, 1997). In the sequel, we will only meet dominance relations in passing, when we discuss asymmetric data. They are mentioned here by way of contrast and because the unfolding technique has been associated with the analysis of dominance data by Carroll (1980), DeSarbo and Rao (1984), and DeSarbo and Carroll (1985). However, in their terminology, the term *dominance* refers to order relations among proximities, not among persons and items.

2.1.2. Unipolar and Bipolar Relations

To further demarcate our subject, we introduce a new distinction. A proximity measure may be either *bipolar* or *unipolar.* Bipolarity refers to the fact that a measure can have three designated markers on its scale: a maximal value associated with one pole, a neutral value, and a minimal value associated with a second pole. A prototypical example of a bipolar proximity measure is the correlation coefficient: As is well known, it measures the association between two variables, ranging from $+1.0$, indicating perfect (linear) association, through 0.0, indicating lack of (linear) association, to -1.0, indicating perfect negative (linear) association. At both poles, the variables are completely substitutable, whereas at the neutral point, positions on one variable cannot be predicted from positions on the other. Some more examples of bipolar proximity measures are the covariance, Kendall's τ, and any other coefficient measuring linear or monotonic association between variables (cf. Coxon, 1982, chap. 2).

Unipolarity is characterized by nonnegativity, that is, by the absence of negative association, and hence by the absence of the negative pole. A unipolar proximity measure, which can be keyed either as similarity or as dissimilarity, has only two designated markers on its scale, one of which is associated with a pole, whereas the other one is a neutral point. In the case of a similarity measure, there is some maximal value indicating equality or substitutability and a minimal value of zero indicating complete lack of similarity. Conversely, in the case of a dissimilarity measure, there is some maximal value indicating complete lack of similarity and a minimal value of zero indicating equality or substitutability. Thus, the single pole of a unipolar proximity scale is associated with maximal similarity or minimal dissimilarity, and the other side of the scale plays the role of the zero point in a bipolar proximity measure: lack of likeness, lack of resemblance, and lack of affinity or association. Unipolar proximity relations formed the primary context in which MDS and unfolding were developed, whereas bipolar proximity relations form the realm of factor analysis and structural equations modeling. Thus, difference in polarity leads to different use of geometry.

2.1.3. Empirical Relations Become Geometric Relations

Coombs (1964) has made a strong case for the notion that any system of empirical relations can be modeled as a system of geometric relations. What is the connection between polarity and the type of geometric model used for representing the system of relations? If proximity is measured on a bipolar scale, as is the case with correlations between variables, it is natural to require

a geometric model in which the three designated markers also have a definite representation. In the common factor analysis model (Mulaik, 1972; Yates, 1987), this requirement is indeed satisfied: The observed variables and the unobserved factors are represented as vectors, and their inter-correlations are represented as angles between these vectors. Two variables with a correlation of $+1.0$ will have an angle of zero degrees, two variables with a correlation of 0.0 an angle of 90 degrees, and two variables with a correlation of -1.0 an angle of 180 degrees. A multidimensional scaling model of the same variables would represent them by a set of points instead of vectors and would represent their inter-correlations by some decreasing function of the inter-point distances—for instance, by some optimal, monotonically decreasing transformation, which is characteristic for nonmetric MDS (Kruskal, 1964). Whereas bipolarity leads to a geometric model in which each element has an antipode (a unique opposite vector), unipolarity leads to a geometric model in which such a notion does not exist.

Although nonmetric multidimensional scaling of inter-correlations has been propagated and used with some success (Levy & Guttman, 1975; Paddock & Nowicki, 1986; Rounds, Davison, & Dawis, 1979; Schlesinger & Guttman, 1969), on the grounds that it tends to give low-dimensional representations that are easier to understand than traditional factor analysis results, it appears that there are also drawbacks. MDS provides no identification of factors or any other data-generating mechanism from which clusters of variables can be formed, whereas clustering the variables is often the ultimate motivation for psychologists to analyze their inter-correlations. In addition, only one of the three markers on the correlation scale is preserved. Pairs of variables with a correlation of 1 will obtain zero distance (they will coincide), but pairs of variables with zero correlation and with perfectly negative correlation are not easily distinguished or recognized in the representation.

2.1.4. Recent Applications of Multidimensional Scaling

Applications of multidimensional scaling in psychology are numerous. Some recent examples in cognitive psychology include work on category learning and cognitive skills (Griffith & Kalish, 2002; Lee & Navarro, 2002; Nosofsky & Palmeri, 1997), brain diagnostics, neural activity and evoked responses (Beckmann & Gattaz, 2002; Laskaris & Ioannides, 2002; Samson, Zatorre, & Ramsay, 2002; Welchew, Honey, Sharma, Robbins, & Bullmore, 2002), the nonvisual senses (Barry, Blamey, & Martin, 2002; Berglund, Hassmen, & Preis, 2002; Clark, Yang, Tsui, Ng, & Clark, 2002; Francis & Nusbaum, 2002; Kappesser & Williams, 2002; Sulmont, Issanchou, & Koster, 2002), and body images and body comparison processes (Fisher, Dunn, & Thompson, 2002; Viken, Treat, Nosofsky, McFall, & Palmeri, 2002).

But there are also many applications in less "hard" areas, such as social cognition and emotion recognition (Alvarado & Jameson, 2002; Green & Manzi, 2002; Pollick, Paterson, Bruderlin, & Sanford, 2001), clinical assessment via cognitive tasks (Sumiyoshi et al., 2001; Treat, McFall, Viken, & Kruschke, 2001; Treat et al., 2002), vocational and leisure interest questionnaires and personality assessment (du Toit & de Bruin, 2002; Hansen & Scullard, 2002; Pukrop et al., 2002; Shivy & Koehly, 2002), measurement of quality of life (Kemmler et al., 2002; Mackie, Jessen, & Jarvis, 2002; Takkinen & Ruoppila, 2001), cross-cultural psychology (Smith, Cowie, Olafsson, & Liefooghe, 2002; Struch, Schwartz, & Van der Kloot, 2002), communication behavior and social influence (Porter & Alison, 2001; Taylor, 2002), and crime behavior and coping with crime (Kocsis, Cooksey, & Irwin, 2002; Lundrigan & Canter, 2001; Magley, 2002). Applications of multidimensional unfolding lag seriously behind, undoubtedly due to the many technical problems that formed a serious obstacle to successful data analysis until recently.

2.1.5. Organization of This Chapter

The rest of this chapter is organized as follows. The next section introduces MDS and unfolding on an equal footing by considering one square proximity table that may be asymmetric. General strategies to deal with the asymmetry are discussed and applied to the same example mentioned earlier, concerning mutual citation frequencies among psychological journals. Next, we extend the discussion to strategies for studying relational differences to accommodate designs in which proximity data are collected under several different conditions. A number of these strategies are demonstrated in two further applications: an MDS study concerning national attitudes at the onset of Word War II and an unfolding study concerning preferences for food items. The chapter concludes with a discussion of some recent methodological developments.

2.2. Analyzing One Proximity Relation

The situation in which we have one proximity relation between the elements of one set of objects constitutes the classic multidimensional scaling setup, which is well documented in the literature (e.g., Everitt & Rabe-Hesketh, 1997; Kruskal & Wish, 1978). After a summary of the general MDS setup, we pay special attention to the analysis of asymmetric data and conclude this section with a related discussion of unfolding, analyzing one proximity relation between the elements of two sets of objects.

2.2.1. General MDS Setup

In brief, the objective of MDS is to find a configuration of n points $\{x_i,\ i = 1, \ldots, n\}$, where x_i has coordinates $\{x_{iu}, u = 1, \ldots, p\}$ specifying its location in a p-dimensional spatial model. Typically, the configuration is two-dimensional ($p = 2$), but this choice can only be justified, of course, by a reasonably good fit. The quality of the fit is assessed by determining distances $d(x_i, x_j)$ between all pairs of points (most common is the ordinary Euclidean distance). These inter-point distances should reflect the inter-object proximities: If two objects are relatively similar in the data, their corresponding points in the model must be close together, but if two objects are relatively dissimilar, their corresponding points must be far apart. Goodness of fit of the configuration is measured (quite indirectly) by the quality of the fit of the nonlinear regression equation

$$\varphi[\delta(a_i, a_j)] = d(x_i, x_j) + \varepsilon_{ij}. \qquad (1)$$

Here, $\delta(a_i, a_j)$ denotes the given dissimilarity value for objects a_i and a_j; $\varphi[.]$ denotes a transformation that maps the dissimilarity values into a set of transformed values $\hat{d}(a_i, a_j)$, called *pseudo-distances* or *d-hats;* and ε_{ij} are the residuals. In general, $\varphi[.]$ will be some selected type of function, reflecting the kind of information in $\delta(a_i, a_j)$ that we want to take into account in the analysis. For example, $\varphi[.]$ could be a linear function with positive slope and either with or without an intercept, or it could be a step function that assigns new, identically ordered values to the given $\delta(a_i, a_j)$, so that only the rank-order information is preserved. When the relation between the objects is given in terms of a similarity function $\rho(a_i, a_j)$, the transformation $\varphi[.]$ is required to be linear with negative slope, or monotonically decreasing.

2.2.1.1. Measures of Fit and Distributional Assumptions

In any case, according to (1), the pseudo-distances $\hat{d}(a_i, a_j) = \hat{\varphi}[\delta(a_i, a_j)]$ follow from the *optimal transformation,* that is, the transformation that optimizes the fit for given x_i and x_j, and so they are approximately equal to the $d(x_i, x_j)$ in some definite sense. Measuring quality of fit of an MDS solution by a least squares criterion was an idea introduced by Kruskal (1964), who actually used the root mean square error, which he called *Stress* (the reverse of fit). Mainly for computational convenience, Takane, Young, and de Leeuw (1977) squared the distances in (1), calling the resulting badness-of-fit function *S-Stress.* Noting that proximities are always nonnegative, Ramsay (1977, 1978, 1980, 1982) introduced an alternative regression equation based on the idea that the distances are not disturbed by additive random factors ε_{ij}, but by multiplicative, positive random factors υ_{ij}, which are asymmetrically distributed around 1, in such a way that $\log \upsilon_{ij}$ is normally distributed around zero. In a similar vein, Takane (1981, 1982), Takane and Carroll (1981), Takane and Sergent (1983), and Sergent and Takane (1987) suggested and tested several multidimensional scaling models based on a variety of distributional assumptions for specific data collection processes.

Here, we stay in the least squares framework, which provides maximum likelihood estimates under the assumption of normal errors. Reiterating the strong points, the least squares method is flexible, weights can be used to adjust for nonstandard error structures, and it is known to perform well under a broad range of circumstances. In fact, Storms (1995) has shown in a Monte Carlo study that violations of the assumed error distribution have virtually no effect on the estimated parameters. Spence and Lewandowsky (1989) and Heiser (1988a) studied robust methods for MDS but reached the conclusion that under moderate levels of error, standard MDS methods are not particularly vulnerable to outliers, especially when robust initial configurations are used.

2.2.1.2. Probabilistic Models

It is also possible to make distributional assumptions on the model side of the equation—that is, on x_i and x_j in (1)—from which a different class of methods has arisen (Ennis, Palen, & Mullen, 1988; MacKay, 1989, 1995, 2001; Mullen & Ennis, 1991; Zinnes & Griggs, 1974; Zinnes & MacKay, 1983; Zinnes & Wolff, 1977). These probabilistic models provide a

rather different mechanism of random variation, with the counterintuitive property that the expected value of a dissimilarity judgment over replications can be far off the model value of the corresponding distance (even to the extent that there is no monotonic relationship between expected dissimilarity and distance). For this reason, Monte Carlo studies using random perturbations of the point locations x_i to study the behavior of standard MDS methods, such as Young (1970), Sherman (1972), Spence (1972), and Spence and Domoney (1974), have questionable validity. The same remark applies to the studies by Girard and Cliff (1976), MacCallum and Cornelius (1977), and MacCallum (1979), who used a data generation mechanism with biases in the small and large distances.

2.2.1.3. The Problem of Asymmetry

A key assumption in the classic multidimensional scaling setup is *symmetry* of the proximity relation—that is, $\delta(a_i, a_j) = \delta(a_j, a_i)$—in accordance with the symmetry of the distance function used in the geometric model. Yet, quite often, relational data in their raw form are not symmetric. For instance, in stimulus identification experiments, confusion errors are counted, and it is not unusual to observe rather big asymmetries between the count of responding with a_j when stimulus a_i is presented and the count of responding with a_i after presentation of a_j (cf. Heiser, 1988b). These effects may be due to stimulus familiarity or a response bias or to similar processes in other contexts. They can be removed prior to analysis or explicitly incorporated in a model. For comprehensive reviews of the treatment of asymmetry, the reader is referred to Everitt and Rabe-Hesketh (1997, chap. 6) and Zielman and Heiser (1996).

Here, attention will be restricted to two strategies to analyze asymmetric relational data. The first splits the relation into two parts and finds two representations of a single set of objects; the second considers the row and column elements of the data matrix as two different kinds of entities and finds a single representation of two sets of objects. Taking frequencies as our leading case of data collection, we denote the raw observations by f_{ij}, with $i = 1, \ldots, n$ and $j = 1, \ldots, n$, and we assume $f_{ij} > 0$ for all i, j.

2.2.2. Making Two Representations of a Single Set of Objects

As a preliminary to any consideration of genuine asymmetry, it is often useful to remove the main effects from the data, which reflect the tendency of some objects having consistently higher frequencies than others, because they are more prominent, more popular, or otherwise more bulky. A simple correction for such main effects is to equalize all self-similarities by the standardization

$$s_{ij} = \frac{f_{ij}}{\sqrt{f_{ii} f_{jj}}}, \tag{2}$$

which ensures that $s_{ii} = 1$ for all i. The rationale of using this standardization is that, if the simple model $f_{ij} = \alpha_i \alpha_j \theta_{ij}$ holds, with α_i some object-specific main effect parameter and θ_{ij} an interaction parameter with equal diagonal elements ($\theta_{ii} = 1$), then these assumptions would give $s_{ij} = \theta_{ij}$ in (2). Note that this standardization does not affect the asymmetry in f_{ij} except for scale; that is, the odds across the diagonal remain the same: $s_{ij}/s_{ji} = f_{ij}/f_{ji}$.

2.2.2.1. Multiplicative Decomposition

Now consider the multiplicative decomposition of s_{ij} into a symmetric factor and an antisymmetric factor,

$$s_{ij} = r(a_i, a_j)\, t(a_i, a_j), \tag{3}$$

where the two constituting factors are defined as

$$r(a_i, a_j) = \sqrt{s_{ij} s_{ji}}, \tag{4a}$$

$$t(a_i, a_j) = \sqrt{\frac{s_{ij}}{s_{ji}}}. \tag{4b}$$

In these definitions, the objects are again identified explicitly by a_i and a_j, for reasons that will become apparent shortly. It is easily verified by substitution of $r(a_i, a_j)$ and $t(a_i, a_j)$ that equation (3) is always true, so that the decomposition can always be made without any further conditions. It is also clear from (4a) that the first factor is symmetric, $r(a_i, a_j) = r(a_j, a_i)$, and that it equals the geometric mean of the elements above and below the diagonal of the matrix $\mathbf{S} = \{s_{ij}\}$, whereas (4b) shows that the second factor is antisymmetric: Two corresponding elements across the diagonal have a perfectly inverse relationship, $t(a_i, a_j) = 1/t(a_j, a_i)$.

2.2.2.1.1. Shepard's universal law of generalization. Combining (4a) with (2), we obtain the symmetric similarity measure

$$r(a_i, a_j) = \sqrt{\frac{f_{ij} f_{ji}}{f_{ii} f_{jj}}}, \tag{5}$$

an expression first developed by Shepard (1957) for stimulus and response generalization processes. In this paper, he also gave the rationale for linking similarity to distance by the rule

$$r(a_i, a_j) = e^{-d(x_i, x_j)}. \tag{6}$$

If (6) is correct, then it follows that a nonmetric MDS of $r(a_i, a_j)$, based on (1), should yield the transformation $\varphi[.] = -\log$. Evidence in more than 10 studies (Shepard, 1987; also see Nosofsky, 1992), involving both human and animal subjects and both visual and auditory stimuli, has confirmed this hypothesis, and hence the exponential decay function (6) has been named the *universal law of generalization.*

2.2.2.1.2. Luce's choice model. Combining (4b) with (2), we find

$$t(a_i, a_j) = \sqrt{\frac{f_{ij}}{f_{ji}}}, \tag{7}$$

which can be interpreted as the root odds of responding with a_j if a_i is presented against the reverse; $t(a_i, a_j)$ is a natural measure of the dominance relation between a_i and a_j. The simplest model for a dominance relation is the Bradley-Terry-Luce (BTL) model, a theory of choice developed by Bradley and Terry (1952), which was extended and given an axiomatic basis by Luce (1959). It states that the probability of a_i dominating a_j depends only on the two nonnegative parameters associated with each object, α_i and α_j, and not on any other parameter:

$$p_{ij} = \frac{\alpha_i}{\alpha_i + \alpha_j}. \tag{8}$$

From (8), it follows that $p_{ij} + p_{ji} = 1$ and that the root odds defined in (7) under this model are $\sqrt{\alpha_i / \alpha_j}$, simply the root of the ratio of the two parameters. Summarizing the development so far, we can decompose any asymmetric set of similarities $\{s_{ij}\}$ into a symmetric component $\{r(a_i, a_j)\}$, on which we can do some form of multidimensional scaling, and an antisymmetric component $\{t(a_i, a_j)\}$, on which we can fit the BTL model, or some similar model, for paired-comparison data.

2.2.2.2. Additive Decomposition

Up to this point, all operations have been multiplications and divisions. However, it is often desirable when working with frequencies to use a log scale, as is done in log-linear analysis (Wickens, 1989). An additive version of the basic decomposition (3) is obtained by taking the logarithm of both sides of the equation, yielding

$$\mu_{ij} = \rho(a_i, a_j) + \tau(a_i, a_j), \tag{9}$$

where $\mu_{ij} = \log s_{ij}$, $\rho(a_i, a_j) = \log r(a_i, a_j)$, and $\tau(a_i, a_j) = \log t(a_i, a_j)$. The equivalents of (4a) and (4b) are

$$\rho(a_i, a_j) = \frac{1}{2}[\mu_{ij} + \mu_{ji}], \tag{10a}$$

$$\tau(a_i, a_j) = \frac{1}{2}[\mu_{ij} - \mu_{ji}]. \tag{10b}$$

In general, any matrix can be additively decomposed as in (9), that is, into the sum of a symmetric component (10a) and a skew-symmetric component (10b). Instead of a geometric mean (4a), we now have an arithmetic mean (10a), and instead of the antisymmetry property $t(a_i, a_j) = 1/t(a_i, a_j)$, we now have the skew-symmetry property $\tau(a_i, a_j) = -\tau(a_i, a_j)$. Additive decomposition of asymmetric matrices is well known through the work of Gower (1977), although the idea seems to be much older: Halmos (1958, p. 136) refers to it as the *Cartesian decomposition.* As pointed out by Gower, the components $\rho(a_i, a_j)$ and $\tau(a_i, a_j)$ are uncorrelated, so that we can analyze them separately by least squares.

2.2.2.3. Application: Citation Frequencies Among Psychological Journals

To illustrate this approach to asymmetry, we now reanalyze some data collected by Weeks and Bentler (1982) on citation patterns among 12 psychological journals. The raw frequencies are reproduced in Table 2.1, together with the list of journals used. An entry in Table 2.1 indicates the number of times that a paper in the row journal cites some paper in the column journal. It is clear that the *Journal of Personality and Social Psychology* (*JPSP*) generates by far the largest number of citations, including many self-citations, whereas the *American Journal of Psychology* (*AJP*) and *Multivariate Behavioral Research* (*MBR*) have a rather low number of citations (primarily due to the smaller number of articles per year), with *AJP* citing the *Journal of Experimental Psychology* (*JEP*) more frequently than itself and *MBR* citing *Psychometrika* (*PKA*) more frequently than itself. To avoid problems with zero frequencies, we added 0.5 to all entries of the table. Then s_{ij} was calculated according to (2); the symmetric similarities $\rho(a_i, a_j)$ according to (10a), in which the minimal value was added to make all

Table 2.1 Journal Citation Data

	AJP	*JABN*	*JPSP*	*JAPP*	*JCPP*	*JEDP*	*JCCP*	*JEP*	*PKA*	*PB*	*PR*	*MBR*
AJP	31	10	10	1	36	4	1	119	2	14	36	0
JABN	7	235	55	0	13	4	65	25	3	50	31	0
JPSP	16	54	969	28	15	21	89	62	16	149	141	16
JAPP	3	2	30	310	0	8	5	7	6	71	14	0
JCPP	4	0	2	0	386	0	2	13	1	22	35	1
JEDP	1	7	61	10	2	100	6	5	4	18	9	2
JCCP	0	105	55	7	3	10	331	3	19	89	22	8
JEP	9	20	16	0	32	6	1	120	2	18	46	0
PKA	2	0	0	0	0	6	0	6	152	31	7	10
PB	23	46	124	117	138	7	86	84	62	186	90	7
PR	9	2	21	6	3	0	0	51	30	32	104	2
MBR	0	7	14	4	0	0	24	3	95	46	2	56

SOURCE: Weeks and Bentler (1982).

NOTE: Rows represent journals giving citations; columns represent journals receiving citations. Data collected in 1979. Journals and their abbreviations: *AJP = American Journal of Psychology; JABN = Journal of Abnormal Psychology; JPSP = Journal of Personality and Social Psychology; JAPP = Journal of Applied Psychology; JCPP = Journal of Comparative and Physiological Psychology; JEDP = Journal of Educational Psychology* (numbers 1–3 only); *JCCP = Journal of Consulting and Clinical Psychology; JEP = Journal of Experimental Psychology (General); PKA = Psychometrika; PB = Psychological Bulletin; PR = Psychological Review; MBR = Multivariate Behavioral Research.*

Table 2.2 Journal Citation Data: Decomposition in Symmetric and Skew-Symmetric Parts

	AJP	*JABN*	*JPSP*	*JAPP*	*JCPP*	*JEDP*	*JCCP*	*JEP*	*PKA*	*PB*	*PR*	*MBR*
AJP	0	4.37	4.06	2.88	4.49	3.57	1.87	6.04	3.32	5.22	5.52	2.21
JABN	0.17	0	4.48	1.16	1.89	3.37	5.43	4.65	1.68	5.18	3.77	2.56
JPSP	−0.23	0.01	0	3.72	2.06	4.49	4.56	4.28	1.75	5.51	4.89	3.93
JAPP	−0.42	−0.80	−0.03	0	0.10	3.72	2.73	2.04	1.85	5.68	3.72	2.16
JCPP	1.05	1.65	0.91	0	0	1.47	1.85	4.31	1.01	5.07	3.75	1.51
JEDP	0.55	−0.26	−0.53	−0.11	−0.80	0	3.55	3.73	3.51	4.19	2.79	2.43
JCCP	0.55	−0.24	0.24	−0.16	−0.17	−0.24	0	2.18	2.37	5.61	2.63	4.40
JEP	1.27	0.11	0.67	1.35	−0.44	−0.08	0.42	0	3.13	5.31	5.82	2.51
PKA	0	0.97	1.75	1.28	0.55	−0.18	1.83	−0.48	0	5.31	4.52	5.57
PB	−0.24	0.04	0.09	−0.25	−0.91	0.45	0.02	−0.76	−0.34	0	5.70	4.94
PR	0.67	1.27	0.94	0.40	1.16	1.47	1.90	−0.05	−0.70	0.51	0	3.22
MBR	0	−1.35	0.06	−1.10	0.55	0.80	−0.53	−0.97	−1.10	−0.91	0	0
$\underline{\hat{\beta}_i}$	0.28	0.10	0.36	0.22	−0.31	0.28	0.30	−0.46	−0.66	0.12	−0.63	0.38

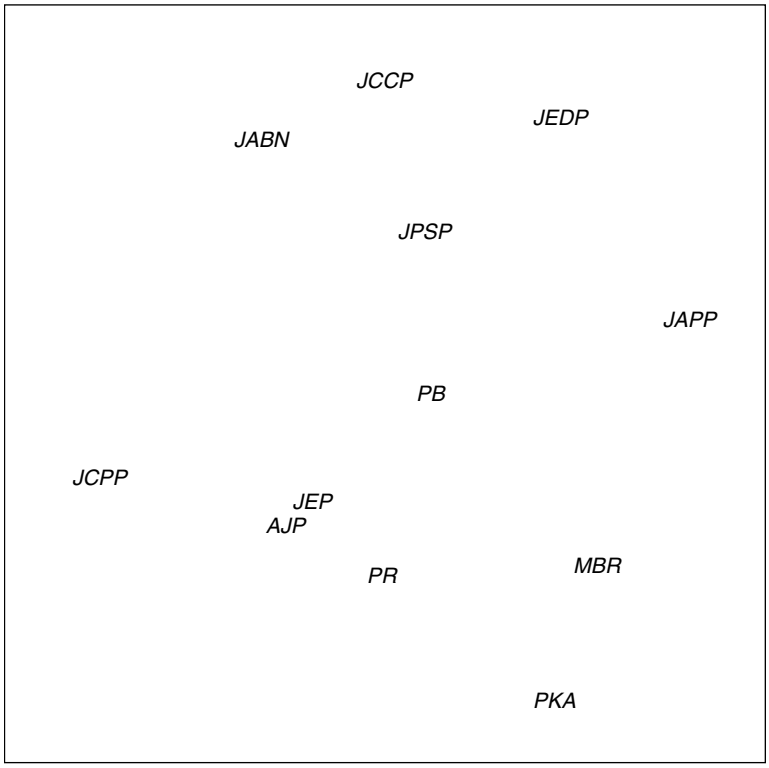
NOTE: Upper triangular part contains symmetric similarities; lower triangular part contains skew-symmetric dominances. Journals and their abbreviations: *AJP = American Journal of Psychology; JABN = Journal of Abnormal Psychology; JPSP = Journal of Personality and Social Psychology; JAPP = Journal of Applied Psychology; JCPP = Journal of Comparative and Physiological Psychology; JEDP = Journal of Educational Psychology; JCCP = Journal of Consulting and Clinical Psychology; JEP = Journal of Experimental Psychology (General); PKA = Psychometrika; PB = Psychological Bulletin; PR = Psychological Review; MBR = Multivariate Behavioral Research.*

quantities nonnegative; and the skew-symmetric dominance data $\tau(a_i, a_j)$ according to (10b). The results are given in Table 2.2 above the diagonal and below the diagonal, respectively.

2.2.2.3.1. MDS analysis of the symmetric part. The symmetric similarities in the upper-triangular section of Table 2.2 were then input to the MDS program PROXSCAL[1], with the ordinal transformation option chosen, and initialized with the classic Torgerson solution (Torgerson, 1958) on the quantities $\rho_{\max} - \rho(a_i, a_j)$, where $\rho_{\max}$ is the maximal similarity value. The two-dimensional solution is shown in Figure 2.1 (as we have $12\ (12-1)/2 = 66$ independent data values, we restrict attention here to $p = 2$, which requires $2\ (12-1) - 1 = 21$ free parameters to be estimated). The fit of the solution in terms of Kruskal's Stress-1 is 0.192, which is "fair" according to Kruskal's (1964) qualifications. In terms of the percentage of dispersion accounted for (%DAF)—which is defined as 100 times the sum of squared distances, divided by the sum of squared

1. PROXSCAL is distributed by SPSS, Inc., 233 S. Wacker Drive, 11th Floor, Chicago, IL 60606–6307 (www.spss.com), as part of the Categories package.

Figure 2.1 Two-Dimensional Ordinal MDS Solution for the Symmetric Part of the Journal Citation Data



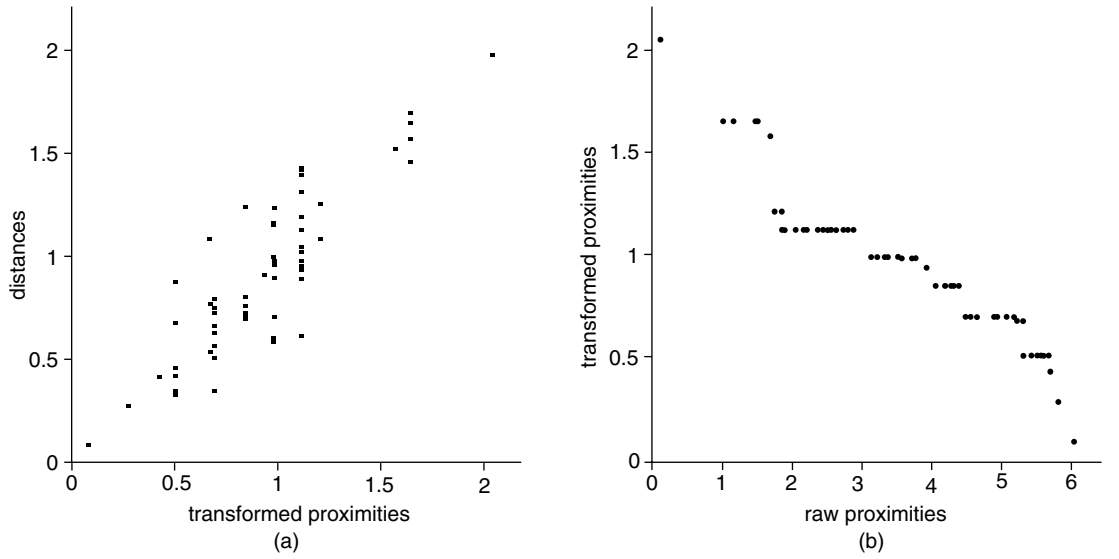
NOTE: Journals and their abbreviations: *AJP = American Journal of Psychology; JABN = Journal of Abnormal Psychology; JPSP = Journal of Personality and Social Psychology; JAPP = Journal of Applied Psychology; JCPP = Journal of Comparative and Physiological Psychology; JEDP = Journal of Educational Psychology; JCCP = Journal of Consulting and Clinical Psychology; JEP = Journal of Experimental Psychology (General); PKA = Psychometrika; PB = Psychological Bulletin; PR = Psychological Review; MBR = Multivariate Behavioral Research.*

pseudo-distances[2] (Heiser & Groenen, 1997), and which is comparable to percentage of variance accounted for, except that the mean is not taken out—the fit is 96.3%, which is quite satisfactory. To give a visual impression of the fit, we provide a regression plot in Figure 2.2a of the fitted distances against the transformed proximities, which are in turn plotted against the original similarities $\rho(a_i, a_j)$ in Figure 2.2b, in a so-called transformation plot. What Figure 2.2b shows is that the monotonically decreasing values of the transformed proximities (which preserve the order of the original proximities) are rather close to a linear transformation of $\rho(a_i, a_j) = \log r(a_i, a_j)$, with negative slope. The implication is that (6) is correct, a confirmation of Shepard's law. The location of the journals in Figure 2.1 is close to the result obtained by Weeks and Bentler (1982) with their specific model. It shows *Psychological Bulletin* (*PB*) in the center and, going counterclockwise, a clinical-social-educational cluster at the top, a physiological-cognitive cluster in the lower left corner, a quantitative-methodological cluster in the lower right corner, and finally the *Journal of Applied Psychology* (*JAPP*), which communicates least with the *Journal of Comparative and Physiological Psychology* (*JCPP*).

2.2.2.3.2. BTL analysis of the skew-symmetric part. As originally pointed out by Fienberg and Larntz (1976), the maximum likelihood estimates of the BTL parameters in their log form ($\beta_i = \log \alpha_i$) can be obtained by a standard log-linear analysis program (cf. Wickens, 1989, pp. 255–257). Simple least squares estimates of these β-parameters can

2. Dispersion accounted for is equal to 1 minus the quantity actually minimized in PROXSCAL.

Figure 2.2 Scatter Plots of Journal Citation Data



NOTE: Panel (a) shows the regression plot of fitted distances against transformed proximities, and panel (b) shows the transformation plot of transformed proximities against logged input frequencies.

be more easily obtained by just taking the column averages of a matrix that has the same lower triangular elements as Table 2.2, denoted by $\tau(a_i, a_j)$, and upper triangular elements defined as $\tau(a_j, a_i) = -\tau(a_i, a_j)$; such column averages are given in the last row of Table 2.2. The values of the estimated BTL scale values range from -0.66 to 0.38, which is a rather small range (they can be compared to z-values), indicating that the amount of asymmetry is modest. In fact, the relative amounts of symmetry and skew-symmetry in the table can be expressed quantitatively because the fact that $\rho(a_i, a_j)$ and $\tau(a_i, a_j)$ are uncorrelated implies that from (9), we can derive an additive decomposition of the sum of squares of the μ_{ij} values:

$$SSQ[\mu_{ij}] = SSQ[\rho(a_i, a_j)] + SSQ[\tau(a_i, a_j)]. \quad (11)$$

In the present example, we obtain $SSQ[\mu_{ij}] = 1030.22$, $SSQ[\rho(a_i, a_j)] = 988.88$, and $SSQ[\tau(a_i, a_j)] = 41.34$, from which the relative contributions of the symmetric and the skew-symmetric component are 96% and 4%, respectively. As the last line in Table 2.2 shows, *PKA, PR, JEP,* and *JCPP* are journals that tend to be cited, whereas *MBR, JPSP,* and the *Journal of Consulting and Clinical Psychology* (*JCCP*) tend to cite others more than others cite them.

2.2.3. Unfolding: Analyzing the Proximity Relation Between Two Sets of Objects

In the example of citation counts between journals, we might also consider the row elements as being different from the column elements because they have different roles: row journals are citing, whereas column journals are being cited. More generally, we might consider the proximity relation as being one between a set of row objects $\{a_i, i = 1, \ldots, n\}$ and a set of column objects $\{b_j, j = 1, \ldots, m\}$, to be represented as a set of row points $\{x_i, i = 1, \ldots, n\}$ and a set of column points $\{y_j, j = 1, \ldots, m\}$, respectively, with x_i having coordinates $\{x_{iu}\}$, as before, and y_j having coordinates $\{y_{ju}\}$.

2.2.3.1. General Definition of Unfolding

In the general unfolding situation, we do not necessarily have $n = m$, as is the case in the current citation example, and we might even have completely different types of objects in rows and columns. Most typically for unfolding, the set $\{a_i\}$ usually refers to persons, the set $\{b_j\}$ usually refers to attitude items or stimuli, and the proximity relation expresses the strength with which a particular person a_i endorses a particular item b_j, or the relative amount of time or money a_i would be willing to spend on b_j. In the spatial representation

sought, we determine the Euclidean distance between x_i and y_j by the formula

$$d(x_i, y_j) = \sqrt{\sum_u (x_{iu} - y_{ju})^2}. \qquad (12)$$

A related model for analyzing individual differences in rankings or ratings that is often subsumed under the unfolding concept (Carroll, 1972; Nishisato, 1994, 1996) is the so-called vector model, independently conceived by Tucker (1960) and Slater (1960). Because this chapter is restricted to distance models, whereas the Tucker-Slater model uses inner products between vectors to represent the data, the reader is referred to Heiser and de Leeuw (1981) for a detailed comparison between the two.

2.2.3.2. Unfolding a Square Table

Note that the Euclidean distance used in MDS is a constrained version of (12), for which it is required that the two sets of points coincide: We have $y_{ju} = x_{iu}$ for corresponding i and j. These constraints have two consequences, which make the distance used in unfolding fundamentally different from the distance used in MDS and which become particularly apparent in the analysis of square tables. First, although equation (12) is symmetric in the sense that $d(x_i, y_j) = d(y_j, x_i)$, this fact only implies that if we transpose the distance matrix and switch the two sets of objects at the same time, nothing has really changed. However, although we do have $d(x_i, x_j) = d(x_j, x_i)$ in the ordinary MDS model, we generally find $d(x_i, y_j) \neq d(x_j, y_i)$ in the unfolding model (as far as the range of subscripts permit). Thus, distances in unfolding are inherently asymmetric. Second, although in MDS we must have $d(x_i, x_i) = 0$, we generally have $d(x_i, y_i) \neq 0$ in unfolding. Thus, the unfolding model also allows modeling of the diagonal of a square table, unlike the MDS model. In the citation example, the diagonal represents the amount of self-citation, which can be a characteristic attribute of a journal and its readership. If we look at the raw data in Table 2.1, we already get the impression that the *Journal of Educational Psychology* (*JEDP*) has a relatively high amount of self-citations, whereas *JPSP*—which has a much higher absolute number of self-citations—is relatively often cited by or citing others.

2.2.3.3. Correcting the Data for Independent Main Effects

It is clear that in the unfolding case, too, it is a good idea to correct for the main effects. If the journals would cite each other completely in a random fashion, we would expect the joint frequencies to satisfy the usual formula for the expected frequencies (e_{ij}) under independence:

$$e_{ij} = N\left(\frac{f_{i+}}{N}\right)\left(\frac{f_{+j}}{N}\right) = \frac{f_{i+}f_{+j}}{N}, \qquad (13)$$

that is, the product of the estimated probability of citing and the estimated probability of being cited times the total number of citations N (here, the $+$ in the marginal totals replaces the index over which we have summed). As a measure of similarity to be used in the unfolding analysis, we define the odds of journal a_i citing journal b_j against what we expect under independence:

$$\rho(a_i, b_j) = \frac{f_{ij}}{e_{ij}} = \frac{Nf_{ij}}{f_{i+}f_{+j}}. \qquad (14)$$

These similarities are given in Table 2.3. Note that $\rho(a_i, b_j) = 1$ if journal a_i cites journal b_j as expected according to the size of the journals (like *JPSP* toward *Psychological Review* [*PR*]), $\rho(a_i, b_j) < 1$ if journal a_i does not cite journal b_j as much as expected according to the size (like *JCPP* and *JCCP* mutually), and $\rho(a_i, b_j) > 1$ if journal a_i does cite journal b_j relatively often (like *MBR* towards *PM*). The self-citations are also higher than expected. The odds that a paper in *JEDP* is citing another paper in *JEDP* (or being cited by it), rather than exchanging references with the other psychological journals, are 16 to 1. *MBR* and *PKA* are also quite self-directed, whereas *PB* and *JPSP* are most open. In the unfolding representation, $\rho(a_i, b_j) < 1$ will lead to a relatively large distance $d(x_i, y_j)$ and $\rho(a_i, b_j) > 1$ to a relatively small distance $d(x_i, y_j)$.

2.2.3.4. Unfolding the Citation Frequencies

Figure 2.3 gives the unfolding solution in two dimensions, with a single ordinal transformation across the whole table because all entries are comparable. In this figure, the open circles indicate the citing positions and the closed circles the cited, and corresponding points are connected with arrows. The quality of the solution, as measured by %DAF, is 94.3%, which is slightly lower than the MDS solution. The percentage of variance accounted for is 63.1%, which corresponds to a correlation between distances and pseudo-distances of 0.79. The optimal transformation (not shown) again confirms Shepard's law. The global position of the journals is similar to the MDS solution in Figure 2.2, except that the configuration is rotated counterclockwise almost 180 degrees, bringing *MBR* and *PKA* to the top of the plot. Perhaps the most striking feature

Table 2.3 Journal Citation Data: Odds of a Row Journal Citing a Column Journal Against the Expected Values Under Independence

	AJP	*JABN*	*JPSP*	*JAPP*	*JCPP*	*JEDP*	*JCCP*	*JEP*	*PKA*	*PB*	*PR*	*MBR*
AJP	6.81	0.47	0.17	0.05	1.32	0.56	0.04	5.51	0.12	0.45	1.55	0
JABN	0.83	6.01	0.51	0	0.26	0.30	1.33	0.63	0.10	0.86	0.72	0
JPSP	0.59	0.43	2.76	0.22	0.09	0.49	0.56	0.48	0.16	0.79	1.02	0.61
JAPP	0.38	0.05	0.30	8.57	0	0.64	0.11	0.19	0.20	1.31	0.35	0
JCPP	0.50	0	0.02	0	8.04	0	0.04	0.34	0.03	0.40	0.85	0.13
JEDP	0.26	0.39	1.22	0.56	0.09	16.31	0.27	0.27	0.28	0.67	0.45	0.53
JCCP	0	2.01	0.38	0.14	0.04	0.56	5.07	0.06	0.45	1.15	0.38	0.73
JEP	1.93	0.92	0.27	0	1.15	0.82	0.04	5.44	0.12	0.56	1.93	0
PKA	0.54	0	0	0	0	1.03	0	0.34	11.04	1.22	0.37	2.79
PB	1.38	0.59	0.57	1.52	1.38	0.26	0.89	1.06	0.99	1.61	1.05	0.43
PR	2.01	0.10	0.36	0.29	0.11	0	0	2.40	1.79	1.03	4.54	0.46
MBR	0	0.35	0.25	0.20	0	0	0.95	0.15	5.88	1.54	0.09	13.33

NOTE: Journals and their abbreviations: *AJP = American Journal of Psychology; JABN = Journal of Abnormal Psychology; JPSP = Journal of Personality and Social Psychology; JAPP = Journal of Applied Psychology; JCPP = Journal of Comparative and Physiological Psychology; JEDP = Journal of Educational Psychology; JCCP = Journal of Consulting and Clinical Psychology; JEP = Journal of Experimental Psychology (General); PKA = Psychometrika; PB = Psychological Bulletin; PR = Psychological Review; MBR = Multivariate Behavioral Research.*

of the current solution is that all open circles tend to be closer to the origin than their corresponding closed counterparts, causing all arrows to point outwards. The interpretation of this effect is that it reflects specialization: Almost all journals cite *PB* or *PR* regularly but then have the tendency to just cite within their own cluster. For example, the eccentric position of the closed circles of *JEDP, JCCP, JPSP,* and the *Journal of Abnormal Psychology* (*JABN*) indicates that they are not cited very much by anyone than themselves and some of their closest neighbors. In the cognitive cluster, there is more extensive cross-referencing but still primarily within their own cluster.

2.2.4. Some Concluding Remarks

In conclusion, the two strategies to asymmetry show many of the same characteristics of journal citing behavior, but there are also important differences. On one hand, the decomposition into a symmetric and a skew-symmetric part allows a more thorough analysis of the dominance relations between the journals in their role of senders and receivers, which is less evident in the unfolding solution. On the other hand, the unfolding analysis of the odds against independent citing gives a better understanding of the self-citation behavior of the journals, in connection with how easily they tend to reach each other.[3]

The present solution was obtained with the program PREFSCAL.[4] A caveat against the uncritical use of unfolding programs is in order here because one has to be aware of a phenomenon called *degeneration.* Unfolding programs, or unfolding options within MDS programs, calculate a solution to the nonlinear regression equation

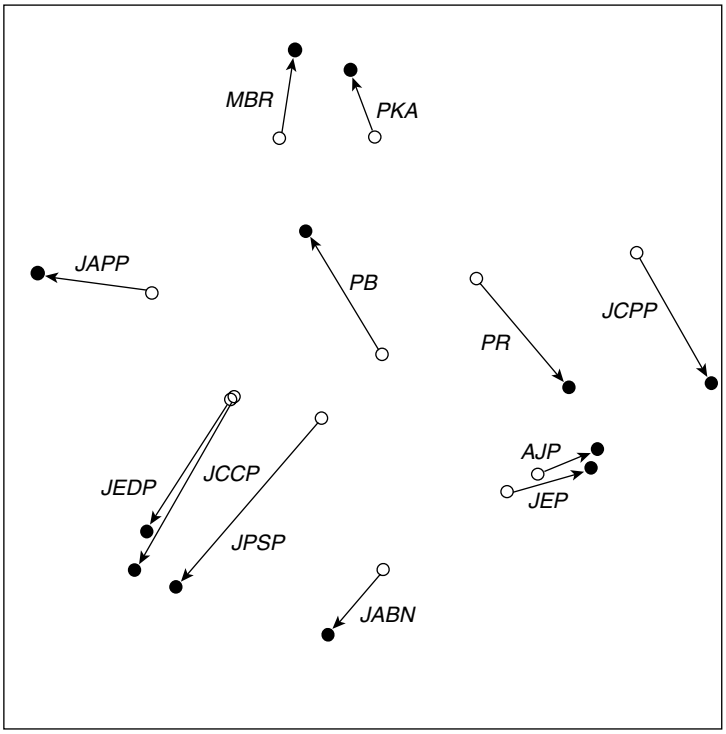
$$\varphi[\delta(a_i, b_j)] = d(x_i, y_j) + \varepsilon_{ij}, \tag{15}$$

which is equivalent to (1), except that the dissimilarities refer to two sets of objects, and the distances refer to two sets of points. No particular problems are to be expected if $\varphi[.]$ is specified as a linear function with a (positive) slope parameter but without an intercept or with an intercept but without a slope parameter. However, the simultaneous presence of both a slope parameter and an intercept will lead to problems because the distances in (15) can be made equal and the transformed dissimilarities as well, giving an uninformative solution with perfect fit (cf. Busing, Groenen, & Heiser, 2004; Heiser, 1989). Degenerate solutions often take the form of one set of objects clustering into one single point. If $\varphi[.]$ is more general (e.g., an ordinal transformation) or if there is a separate $\varphi_i[.]$ for each row in the data matrix (called "row conditionality" or "split-by-row" regression), degeneration occurs in most circumstances as well, provided the program is properly allowed to converge. The remedy against this phenomenon proposed by Kruskal and Carroll (1969)—the introduction of a normalization

3. Similarities and differences between senders and receivers can be made more apparent in the unfolding solution by connecting with an arrow all pairs of points that have an odds ratio greater than 1. Drawing such a graph would show, for example, that *PR* and *PB* are both good senders and good receivers and that *JCPP* is a good receiver but a poor sender.

4. PREFSCAL (beta version) can be obtained from the second author upon request (e-mail: busing@fsw.leidenuniv.nl).

Figure 2.3 Two-Dimensional Ordinal Unfolding Solution for the Journal Citation Data



NOTE: Arrows point from citing position to cited position of the same journal. Journals and their abbreviations: *AJP = American Journal of Psychology; JABN = Journal of Abnormal Psychology; JPSP = Journal of Personality and Social Psychology; JAPP = Journal of Applied Psychology; JCPP = Journal of Comparative and Physiological Psychology; JEDP = Journal of Educational Psychology; JCCP = Journal of Consulting and Clinical Psychology; JEP = Journal of Experimental Psychology (General); PKA = Psychometrika; PB = Psychological Bulletin; PR = Psychological Review; MBR = Multivariate Behavioral Research.*

factor based on the variance of the distances—appears to be not effective enough. Therefore, Busing et al. (2004) introduced a stronger normalization factor (actually, a penalty factor), which discourages solutions in which transformed dissimilarities (and hence distances) have small variation. This penalty approach does seem to work well, and it was used for the examples in this chapter.

2.3. How to Deal With Several Relations

Many research questions require the collection of several sets of proximity data. Relations between the same objects may be studied under several experimental conditions, using different individuals or subsamples, or at several points in time. For example, to study change in citation patterns, one could easily collect again the type of data analyzed in the previous section, covering a number of recent years. Using the generic term *source* to describe these multiple origins of the relational data, a natural question to ask is whether the sources vary systematically and, if so, in what way. Another interesting question, which is not often asked and which we will not discuss, is whether one relation can be predicted from a linear combination of several others, via a kind of multiple regression equation, and how to test in this situation the regression coefficients for significance. For this topic, the interested reader is referred to Krackhardt (1988).

2.3.1. General Strategies to Describe Relational Differences

The dissimilarities of source k ($k = 1, \ldots, K$) are denoted by $\delta_k(a_i, a_j)$. There are several strategies to study the differences between relations,

alternatively called *relational differences.* In principle, these strategies are valid for both MDS and unfolding alike, but because MDS is the more common type of analysis, we use $\delta_k(a_i, a_j)$ and $d(x_i, x_j)$ in the discussion below. Simply replacing these functions by $\delta_k(a_i, b_j)$ and $d(x_i, y_j)$, respectively, gives the unfolding version of the same strategy.

2.3.1.1. Individual Spaces

A strategy that keeps as closely as possible to the data, and one that would be especially suitable if very little is known about the differences to be expected, is to analyze all sources separately, that is, to fit the systems of equations

$$\varphi_k[\delta_k(a_i, a_j)] = d(x_{(k)i}, x_{(k)j}) + \varepsilon_{ijk} \qquad (16)$$

by repeated use of some standard MDS program (for $k = 1, \ldots, K$). In equation (16), $x_{(k)i}$ denotes the position of point i in the configuration of source k, and $\varphi_k[.]$ is the admissible transformation of source k. One could then compare the resulting individual spaces by visual inspection or by *generalized Procrustes analysis,* a technique that finds translations, rotations, and dilations (uniform rescaling) of the individual configurations to optimize their mutual match if they are superimposed; this technique has become especially popular in sensory research (Dijksterhuis & Gower, 1991).

2.3.1.2. Identity Model

If the sources are replications, or if we are only interested in what is common among them, we can fit just one geometric model to all of the sources simultaneously:

$$\varphi_k[\delta_k(a_i, a_j)] = d(x_i, x_j) + \varepsilon_{ijk}. \qquad (17)$$

This approach is almost equal to the even more simple strategy of averaging the individual dissimilarities and then scaling the average, but the difference is that in (17), there is a separate transformation ϕ_k for each source. These transformations allow us, for example, to quantify ordinal data at the source level while still summarizing them in one common configuration. If we fit the regression equation (17) by least squares and denote the optimal transformed proximities by $\hat{d}_k(a_i, a_j)$, then it can be shown that this approach amounts to fitting an MDS model to the average $(1/K) \sum_k \hat{d}_k(a_i, a_j)$.

2.3.1.3. Points-of-View (POV) Model

Suppose we have a way to group the sources into a limited number of, say, L classes, with $1 \leq L < K$, then we can average the proximities in each class. This idea goes back to Tucker and Messick (1963), and it has recently been further developed into an integrated method by Meulman and Verboon (1993). The Tucker and Messick process finds the classes in a first step by a principal components analysis of the $\delta_k(a_i, a_j)$, strung out into variables of length $n(n-1)/2$, followed by a rotation to simple structure, which yields component loadings μ_{kl} for source k and class l. Then the weighted average proximity is

$$\bar{\delta}_l(a_i, a_j) = {}^1\!/_C \sum\nolimits_k \mu_{kl} \varphi_k[\delta_k(a_i, a_j)], \qquad (18)$$

with C the sum of the weights across k, and on these quantities an MDS (or unfolding) model for each of the L classes (points of view) is fitted in a second step. Meulman and Verboon integrated these steps and showed that the POV model is a constrained version of the model to be discussed next.

2.3.1.4. Weighted Euclidean Model

In this model, differences among the sources in their relations among the objects are assumed to arise from a differential weighting of the coordinate axes. So there is one common space and not several, as in POV analysis. However, each source can have different weights associated with any dimension of this common space. If a weight is zero, the corresponding dimension does not affect the proximities of that source at all. Although the model had been considered earlier by others, it owes its fame (and its name, INDSCAL, for INDividual differences SCALing) to the influential paper by Carroll and Chang (1970), which provided a forceful justification and an ingenious computational method. Carroll and Chang emphasized that INDSCAL dimensions are *unique:* It does make a difference which set of dimensions are differentially weighted; that is, rotations are not permissible, even though each individual space is assumed to be Euclidean (and Euclidean distances by themselves do not change if the points are rotated). Thus, the weighted Euclidean model helps to discover dimensions that matter, in the sense that they cause relational differences among individuals (or other sources).

Carroll and Chang's (1970) computational method does not easily generalize to the nonmetric case, but Bloxom (1978) showed how to develop a least squares method based on fitting separate spaces, as in (16), with

ϕ_k linear transformations (making the method suitable for interval data) by posing the coordinate constraints

$$x_{(k)iu} = w_{ku} x_{iu}. \tag{19}$$

Thus, the weighted Euclidean model has the advantage that it can be simply interpreted through the coordinates $\{x_{iu}\}$ of the common space and the weights $\{w_{ku}\}$ for each source on each dimension. Bloxom's approach was generalized by de Leeuw and Heiser (1980) to the nonmetric case.

2.3.1.5. Generalized Euclidean Model

The previous model can be generalized by allowing differential rotation of axes for each source before weighting (Carroll & Chang, 1970). Thus, we may fit (16) with the additional constraints

$$z_{(k)iv} = \text{ROTA}_k(x_{iu}) \text{ for } k = 1, \ldots, K, \tag{20a}$$

$$x_{(k)iv} = w_{kv} z_{(k)iv}. \tag{20b}$$

The notation $\text{ROTA}_k(x_{iu})$ in (20a) indicates that the common coordinates $\{x_{iu}\}$ are expressed with respect to an idiosyncratic set of axes by rotation, which may be over different angles for each source k (hence the model has also been called the IDIOSCAL model). The result of the rotation are the source-specific coordinates $\{z_{(k)iv}\}$ with respect to the axes $v = 1, \ldots, p$, which still generate the same distances as the common space. Then the individual coordinates $x_{(k)iv}$ are obtained in (20b) by weighting the rotated common space. The dimensions of the common space are no longer unique in this model because any preliminary rotation of them would still lead to the same $\{z_{(k)iv}\}$ in (20a) if (20b) is to hold.

2.3.1.6. Reduced-Rank Model

The idea of the reduced-rank model (Bloxom, 1978) is that the individual spaces have dimensionality r_k that is less than the dimensionality p of the common space (hence the term *reduced rank*). For instance, the stimulus objects could be families varying in the number of boys and the number of girls. One group of subjects could view their proximity entirely in terms of the total number of children, whereas another group of subjects could view the proximity between families entirely in terms of sex bias, that is, the difference between the number of boys and the number of girls. The former group could be represented by projecting all families on a direction under 45 degrees of the common axes representing the number of boys and the number of girls, whereas the latter group could be represented by projecting all families on a direction perpendicular to the first. Thus, the common space has dimensionality 2, whereas the individual spaces are projected points after rotation and have dimensionality 1. In general, the process is described by

$$z_{(k)iv} = \text{PROJ}_k\,[\text{ROTA}_k(x_{iu})] \quad \text{for } k = 1, \ldots, K, \quad \text{with } v = 1, \ldots, r_k < p, \tag{21a}$$

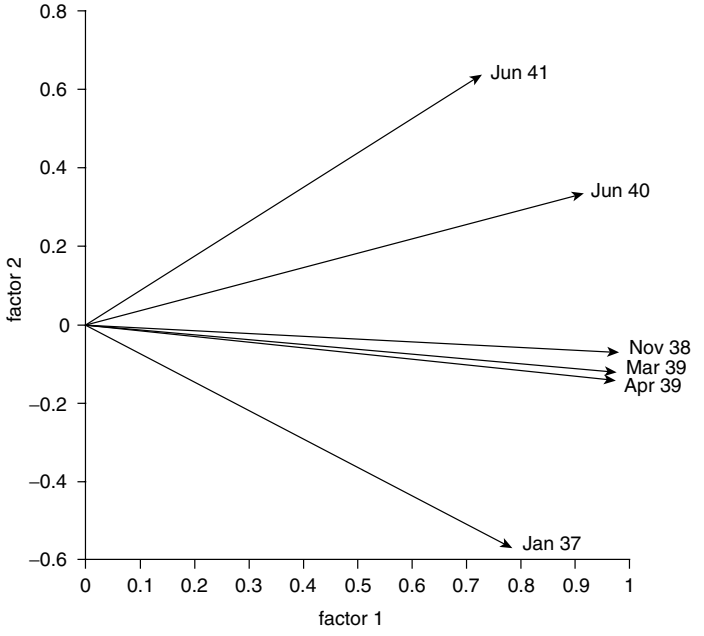
$$x_{(k)iv} = w_{kv} z_{(k)iv}. \tag{21b}$$

The weighted Euclidean model also allows for solutions in which the sources have lower dimensionality than the common space but only in terms of the original axes, not in terms of rotations of them.

2.3.2. Application: MDS of the Klingberg Great Powers Data

The first multidimensional scaling paper appearing in *Psychometrika* with an actual application was by Klingberg (1941). It described the measurement of the friendly or hostile relations among states through expert opinion, using a variety of data collection methods. Breaking new ground, Klingberg was not interested in the attitudes of the experts of international affairs toward certain states but instead attempted to elicit only their assessment of the attitudes of various states toward one another. He collected data in the period from January 1937 to June 1941, using six samples at six points in time. The January 1937 sample had size $N = 83$, and the experts were to give their opinion, for 88 pairs of states, of the chance that "war will exist between them within the next ten years" (only the 21 pairs of the seven Great Powers were reported). The November 1938 sample had size $N = 144$ and used the less complex task to order triads of states in terms of their relative friendliness or hostility. For the March 1939 sample, the "method of multidimensional rank order" was used, in which judges were asked to rank the seven Great Powers in order of the friendliness of 14 small states toward them (and of the other six Great Powers). For the later samples, the same data collection method was used. The first method immediately gives proximities in terms of estimated probabilities, whereas the other methods require some extra calculation (i.e., finding the average proportion of the judges who regarded any pair of states as more hostile than all the other pairs with which it was directly compared), but they are easier to carry out by the judges. Split-half methods were used to show that the reliability was high. Klingberg then demonstrated the power

Figure 2.4 Component Loadings of an Ordinal PCA on the Six Time Points of the Great Powers Data



of MDS to give an integrated view of the relations between the seven Great Powers at the onset of World War II by a three-dimensional analysis of the March 1939 sample.

2.3.2.1. Points-of-View Analysis of the Great Powers

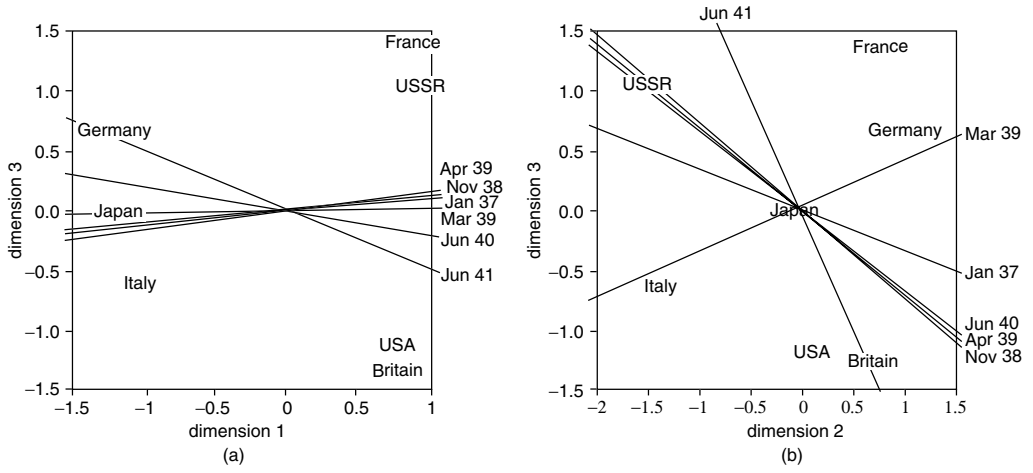
In a first attempt to trace the development of these state relations in time, we performed the initial step of a POV analysis with the program CATPCA,[5] which calculates principal components with optimal ordinal transformations of the variables (Meulman, Heiser, & SPSS, 1999). In this case, the variables are the six proximity matrices, strung out into arrays of $7(7-1)/2 = 21$ elements, read off from Chart A in Klingberg (1941). Figure 2.4 shows a plot of the component loadings of this ordinal PCA, which clearly exhibits a strong first factor (83.3% variance accounted for), with a weaker second factor (15.4%, together 98.7%). There is no evidence for two or more clusters of variables—for instance, before and after certain significant dates, such as the German occupation of Bohemia and Moravia on March 14, 1939, or the outbreak of the war with Great Britain in September 1939. Because clusters were lacking, the POV analysis was aborted. However, there does seem to be some evidence for a regular progression in time (along the second factor).

2.3.2.2. PROXSCAL Analyses of the Great Powers

The second analysis of these data was an ordinal PROXSCAL run under the weighted Euclidean model to see if the progression could be captured in a pattern of dimension weights. However, this analysis produced disappointing results, as there appeared very little variation in the weights, whatever reasonable dimensionality was chosen. By running separate two-dimensional analyses of the six tables, it became clear that most changes from year to year involve local contractions and expansions, not global ones. For instance, from January 1937 to November 1938, Germany and Italy became friendlier on one side of space, but the United States and France became less friendly on the other side of space. In March 1939, there was a complete polarization of Germany, Italy, and Japan against the other states, whereas from June 1940 to June 1941, the position of France changed

5. CATPCA is distributed by SPSS, Inc., 233 S. Wacker Drive, 11th Floor, Chicago, IL 60606–6307 (www.spss.com), as part of the Categories package.

Figure 2.5 Three-Dimensional Common Space of the Great Powers According to the Reduced-Rank Model



NOTE: Lines indicate the projected one-dimensional subspaces for each time point; common Dimensions 3 and 1 are shown in the left panel (a), and common Dimensions 3 and 2 are shown in the right panel (b).

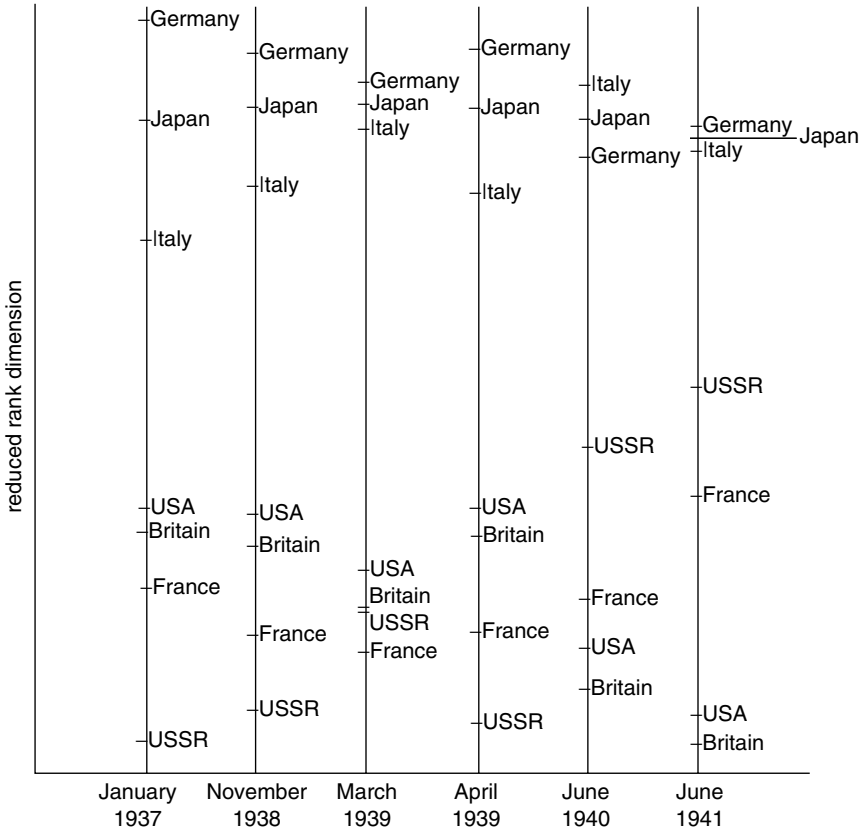
dramatically because it approached Germany and receded from Britain.

Next, the reduced-rank model was fitted in various dimensionalities with ordinal transformations for each source. As there are only $6 \times 21 = 126$ independent data values, the number of fitted parameters is an important consideration (we have $(n - 1)p - p(p - 1)/2$ free parameters for the common space and pr_k for each source). Although the models with $r_k = 2$ fitted well in four and three dimensions (Kruskal's Stress-1 of 0.069 and 0.092, respectively), the number of free parameters (df) was considered too large ($df = 66$ and $df = 51$, respectively). The models with $r_k = 1$ had Kruskal's Stress-1 of 0.129 in four dimensions ($df = 42$) and 0.135 in three dimensions ($df = 33$), so the last one was preferred. It has a %DAF of 98.1%, which is good. The common space is shown in Figure 2.5, in which the left panel displays Dimensions 3 and 1, and the right panel displays Dimensions 3 and 2. The first dimension corresponds closely to the first axis found by Klingberg (1941), who called it "dynamism" (national attitudes insistent on change). The direction that runs from northwest to southeast in the plot of Dimensions 2 and 3 was called "communism" (opposition to or fear of it) by Klingberg, and the direction perpendicular to it, contrasting Germany and France with the United States and Italy, was called "belligerency" (willingness and readiness to fight). Japan clearly has an ambivalent position in this plane. The lines in the plots represent the six individual sources, labeled with their dates; more precisely, the perpendicular projection of the points onto one of these lines gives an approximation to the corresponding individual spaces, which are given separately in Figure 2.6. Although dynamism was the most important factor throughout, at the earlier dates, communism was second in importance, but in March 1939, belligerency became decisive, causing a complete split into two blocks. After the fall of France in June 1940, its position had switched into the middle of the United States/Britain versus Germany/Japan axis by June 1941, finding itself still close to the USSR, which had moved toward Germany and Italy (this was just before the outbreak of the German-Russian war). In conclusion, it appears that the reduced-rank model picks up the rather abrupt local changes in interstate relationships among the Great Powers at the onset of World War II quite well.

2.3.3. Application: Unfolding of the Green and Rao (1972) Breakfast Data

A classic example of an unfolding data set is the one collected by Green and Rao (1972) in the context of a larger study involving dissimilarity judgments, stimulus construct ratings, and preferences of 42 respondents for 15 food items used at breakfast and snack time. What will be analyzed here are the rankings of the 15 food items according to six preference "scenarios," the first being for overall preference (1) and the remainder for the following menus and

Figure 2.6 One-Dimensional Representations of the Great Powers at Six Time Points According to the Reduced-Rank Model



serving occasions: (2) "When I am having a breakfast, consisting of juice, *bacon and eggs,* and beverage"; (3) "When I am having a breakfast, consisting of juice, *cold cereal,* and beverage"; (4) "When I am having a breakfast, consisting of juice, *pancakes, sausage,* and beverage"; (5) "Breakfast, with beverage only"; and (6) "At snack time, with beverage only."

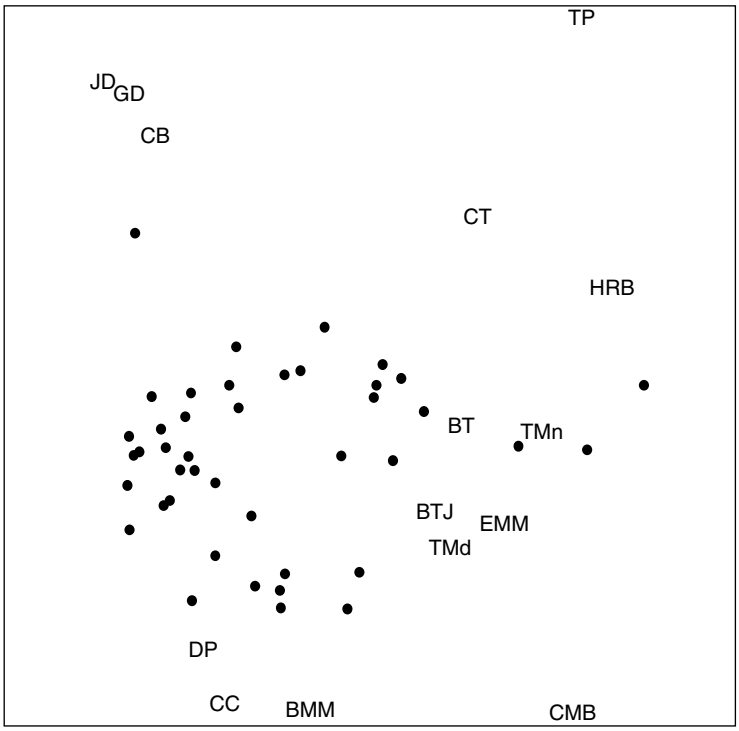
2.3.3.1. Degeneracy Problems With Previous Approaches

When Green and Rao (1972) tried an unfolding analysis on the overall preferences, they found disappointing results and concluded,

> The stimuli with the exception of toast pop-up appear to fall roughly on the circumference of a circle. Ideal-point concentrations are noted in the third and fourth quadrants, suggesting some polarization between groups of respondents who prefer sweet items and those who prefer non-sweet items. However, the poor goodness-of-fit values suggest that (a) either the program was unable to find an appropriate representation in low dimensionality or (b) the simple unfolding model is inadequate to account for these data. (p. 87)

As was noted in our earlier discussion of unfolding, difficulties that previous approaches had, due to insufficient awareness of the necessity to push a method (and the computer program that implements it) to its limits, can be overcome by a penalty approach that discourages solutions with small variation in its distances. Busing et al. (2004) demonstrated that the overall preference data alone could be unfolded rather well, without signs of degeneracy and with a reasonable fit (average product-moment correlation between distances and d-hats was 0.75, corresponding to a %VAF of 56%). Here, it will be demonstrated that the unfolding version of the weighted Euclidean model is well suited to fit the differences between the six scenarios as well.

Figure 2.7 Two-Dimensional Common Unfolding Space for the Breakfast Data According to the Weighted Euclidean Model



NOTE: The breakfast items (and plotting codes) are as follows: toast pop-up (TP), buttered toast (BT), English muffin and margarine (EMM), jelly donut (JD), cinnamon toast (CT), blueberry muffin and margarine (BMM), hard rolls and butter (HRB), toast and marmalade (TMd), buttered toast and jelly (BTJ), toast and margarine (TMn), cinnamon bun (CB), Danish pastry (DP), glazed donut (GD), coffee cake (CC), and corn muffin and butter (CMB).

2.3.3.2. PREFSCAL Analysis of the Breakfast Data

The PREFSCAL analysis was done completely nonmetrically, with separate ordinal transformations for each of the 6×42 rankings.[6] It required 27,278 iterations because the stopping criteria were set very strictly (no differences greater than 1.0E-8). Variance accounted for was 69%, which is higher than the fit of the overall preferences. Figures 2.7 and 2.8 show, respectively, the two-dimensional common space of stimuli and ideal points, as well as the dimension weights for the scenarios. In Figure 2.7 the respondent points spread reasonably well; in the stimulus configuration, we have horizontally the *toast factor:* toasted, crisp, or warm items to the right (several types of toast, English muffin) and soft and cooled down items to the left (donuts, coffee cake). The vertical dimension is the *yeast factor:* On top, we have breads and pastries made from dough that is yeast risen (donuts, hard rolls, toasted bread), and at the bottom, we have quick breads risen from eggs and baking powder (coffee cake, muffins) or made from puff pastry (Danish pastry). The former are what you eat at home, whereas the latter are what you get if you order a Continental breakfast in a fine hotel. In Figure 2.8, the six scenarios are spread out in an interesting way: The yeast factor is especially important to distinguish preferences conditionally upon the egg-type entrée and the other heavy breakfast entrees, whereas the very light breakfast and snack scenarios lead to individual differences along the toast factor. The overall preference is closest to the latter scenarios.

Figure 2.9 shows just how strong the effects of the scenarios influence the shape of the individual configurations. This result is an indication that a

6. Weights were used in the analysis to achieve relatively good fit for high preferences compared to low preferences because the former are considered to be more reliable than the latter. In particular, the weight for each cell was set equal to 1 over the dissimilarity value.

Figure 2.8 Source Weights for the Breakfast Data, Showing the Differences Between the Six Scenarios

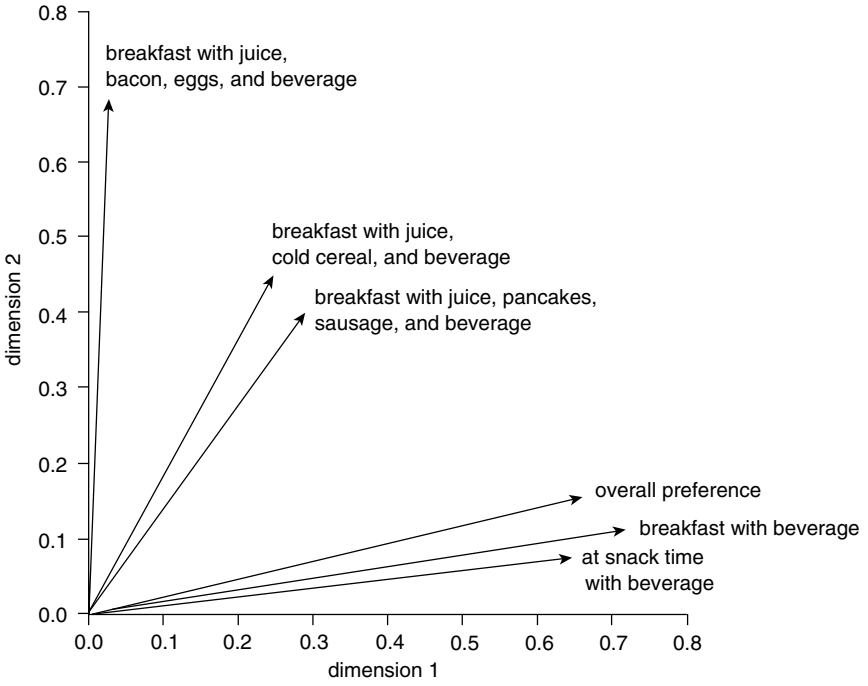
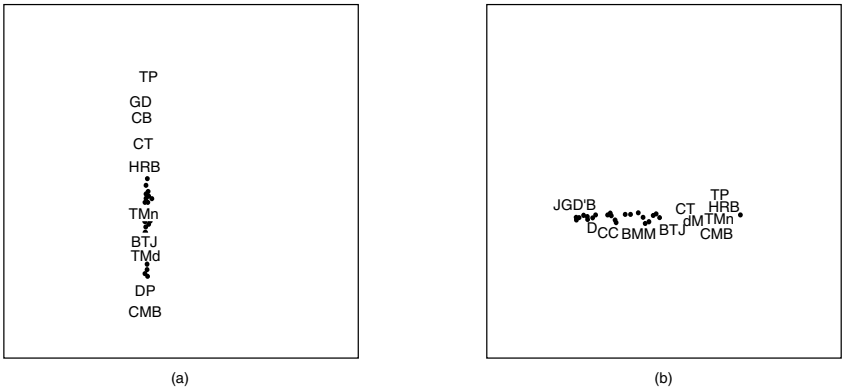


Figure 2.9 Two Extreme Individual Spaces for the Breakfast Data, Showing on the Left (a) Strong Emphasis on the Yeast Factor Under Scenario 2, and on the Right (b) Strong Emphasis on the Toast Factor Under Scenario 6



NOTE: The breakfast items (and plotting codes) are as follows: toast pop-up (TP), buttered toast (BT), English muffin and margarine (EMM), jelly donut (JI), cinnamon toast (CT), blueberry muffin and margarine (BMM), hard rolls and butter (HRB), toast and marmalade (TMd), buttered toast and jelly (BTJ), toast and margarine (TMn), cinnamon bun (CB), Danish pastry (DP), glazed donut (GD), coffee cake (CC), and corn muffin and butter (CMB).

three-dimensional model might be called for because separately, the sources are not so close to one-dimensionality, but such an analysis is not pursued here.

2.4. Discussion

The major tool that has been used in this chapter for presenting an overview of scaling methods has been a

nonlinear regression equation with proximities on the left-hand side and distances on the right-hand side. The scaling literature has contributed a lot to the general idea that a dependent or response variable—in this case, proximity—may have to be transformed before some model can be fitted to it. Optimal data transformations have now become much more common in other parts of statistics as well. At the model side, we have a two-way or three-way design. Pairs of stimuli or pairs of persons and stimuli form the basic independent variable, which may be extended by a replication factor (as was implicitly the case in our citation example, where the replications are the individual citations from any article in a given journal to any other article in the same or a different journal). It may also be crossed with another independent variable ("time" in the case of the Klingberg data and "scenario" in the case of the Green and Rao data). Common models for this situation are linear, bilinear, or multilinear. Characteristic for MDS and unfolding models is that they are truly nonlinear: If one moves one point x_i toward some other point y_j, the distance first decreases (and hence proximity between the corresponding objects is predicted to increase), but then if one moves x_i beyond y_j, distance increases (and hence proximity is predicted to decrease). The regression formulation allowed us to keep technical issues such as the estimation method and optimization in the background. Technical details should follow from general considerations in statistical theory.

In our discussion of the probabilistic approach to MDS, which assumes that the point locations arise from a stochastic process, it was mentioned that these models tend to violate the basic monotonic relationship between dissimilarity and distance. This objection does not apply to methods for fitting the weighted Euclidean (INDSCAL) model that take the point locations as fixed parameters but the individual dimension weights as stochastic. Winsberg and de Soete (1993) offered an approach in which the dimension weights come from a limited number of latent classes, whereas Clarkson and Gonzalez (2001) proposed a genuine random-effects model for the weights. These approaches have the advantage that the number of parameters is drastically reduced and does not increase with the number of sources (or subjects), without having to sacrifice the rotational invariance property. There are also new approaches in metric unfolding that work with latent classes for persons and various restrictions on the stimulus points (de Soete & Heiser, 1993; Wedel & DeSarbo, 1996).

Asymmetry always has been and still is an important issue. For a more technical discussion of distance models for contingency tables and their relation with the well-known RC (M)-association model, the reader is referred to de Rooij and Heiser (2004). Okada (1997) and de Rooij (2002) have recently proposed models that are also suitable for analyzing several asymmetric relations. In the unfolding of the journal citation data, it turned out that the arrows from corresponding senders to receivers all pointed in centrifugal directions. Special models have been developed for the case that they all point to the same direction (Zielman & Heiser, 1993) or to some common location (Adachi, 1999). With a little adjustment, the latter model might be relevant for the citation example.

A final area of development that deserves to be mentioned concerns the use of constraints. After Bloxom's (1978) seminal paper, which introduced a class of constraints that is completely covered by standard options in the PROXSCAL program (Meulman et al., 1999), there have been two major new directions for constrained MDS and unfolding. The first one is the incorporation of cluster constraints in distance models, which allows a more parsimonious description of large data sets (see Heiser & Groenen, 1997, for MDS; de Soete & Heiser, 1993, for unfolding). The second one is the idea of approximating multivariate observations with a distance model, possibly with optimal scaling of the variables, instead of just projecting the data points down into some space of reduced dimensionality (Commandeur, Groenen, & Meulman, 1999; Meulman, 1992, 1996). Working with a distance model for low-dimensional representations of high-dimensional data ensures that the relations between the objects are better represented in terms of their mutual proximities compared to the results of other multivariate techniques, which work with projection. Both developments have opened up an entirely new and exciting area of application for MDS and unfolding techniques.

REFERENCES

Adachi, K. (1999). Constrained multidimensional unfolding of confusion matrices: Goal point and slide vector models. *Japanese Psychological Research, 41,* 152–162.

Alvarado, N., & Jameson, K. A. (2002). Varieties of anger: The relation between emotion terms and components of anger expressions. *Motivation and Emotion, 26,* 153–182.

Barry, J. G., Blamey, P. J., & Martin, L. F. A. (2002). A multidimensional scaling analysis of tone discrimination ability in Cantonese-speaking children using a cochlear implant. *Clinical Linguistics & Phonetics, 16,* 101–113.

Beckmann, H., & Gattaz, W. F. (2002). Multidimensional analysis of the concentrations of 17 substances in the CSF of schizophrenics and controls. *Journal of Neural Transmission, 109,* 931–938.

Berglund, B., Hassmen, P., & Preis, A. (2002). Annoyance and spectral contrast are cues for similarity and preference of sounds. *Journal of Sound and Vibration, 250,* 53–64.

Bloxom, B. (1978). Constrained multidimensional scaling in *N* spaces. *Psychometrika, 43,* 397–408.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika, 39,* 324–345.

Busing, F. M. T. A., Groenen, P. J. F., & Heiser, W. J. (2004). Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. *Psychometrika, 69,* in press.

Carroll, J. D. (1972). Individual differences and multidimensional scaling. In R. N. Shepard, A. K. Romney, & S. B. Nerlove (Eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences: Vol. 1. Theory* (pp. 105–155). New York: Seminar Press.

Carroll, J. D. (1980). Models and methods for multidimensional analysis of preferential choice (or other dominance) data. In E. D. Lantermann & H. Feger (Eds.), *Similarity and choice: Papers in honor of Clyde Coombs* (pp. 234–289). Bern: Hans Huber.

Carroll, J. D., & Chang, J. -J. (1970). Analysis of individual differences in multidimensional scaling via an *N*-way generalization of "Eckart-Young" decomposition. *Psychometrika, 35,* 283–319.

Clark, W. C., Yang, J. C., Tsui, S. L., Ng, K. F., & Clark, S. B. (2002). Unidimensional pain rating scales: A multidimensional affect and pain survey (MAPS) analysis of what they really measure. *Pain, 98,* 241–247.

Clarkson, D. B., & Gonzalez, R. (2001). Random effects diagonal metric multidimensional scaling models. *Psychometrika, 66,* 25–43.

Commandeur, J. J. F., Groenen, P. J. F., & Meulman, J. J. (1999). A distance-based variety of nonlinear multivariate data analysis, including weights for objects and variables. *Psychometrika, 64,* 169–186.

Coombs, C. H. (1964). *A theory of data.* New York: John Wiley.

Coxon, A. P. M. (1982). *The user's guide to multidimensional scaling.* London: Heinemann.

Critchlow, D. E., & Fligner, M. A. (1991). Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on GLIM. *Psychometrika, 56*, 517–533.

de Leeuw, J., & Heiser, W. J. (1980). Multidimensional scaling with restrictions on the configuration. In P. R. Krishnaiah (Ed.), *Multivariate analysis* (Vol. 5, pp. 501–522). Amsterdam: North-Holland.

de Rooij, M. (2002). Distance models for three-way tables and three-way association. *Journal of Classification, 19*, 161–178.

de Rooij, M., & Heiser, W. J. (2004). Graphical representations and odds ratios in a distance association model for the analysis of cross-classified data. *Psychometrika, 69,* in press.

DeSarbo, W. S., & Carroll, J. D. (1985). Three-way metric unfolding via alternating least squares. *Psychometrika, 50*, 275–300.

DeSarbo, W. S., & Rao, V. R. (1984). GENFOLD2: A set of models and algorithms for the GENeral unFOLDing analysis of preference/dominance data. *Journal of Classification, 1*, 147–186.

de Soete, G., & Heiser, W. J. (1993). A latent class unfolding model for analyzing single stimulus preference ratings. *Psychometrika, 58*, 545–565.

Dijksterhuis, G., & Gower, J. C. (1991). The interpretation of generalized Procrustes analysis and allied methods. *Food Quality and Preference, 3*, 67–87.

du Toit, R., & de Bruin, G. P. (2002). The structural validity of Holland's R-I-A-S-E-C model of vocational personality types for young Black South African men and women. *Journal of Career Assessment, 10*, 62–77.

Ennis, D. M., Palen, J., & Mullen, K. (1988). A multidimensional stochastic theory of similarity. *Journal of Mathematical Psychology, 32*, 449–465.

Everitt, B. S., & Rabe-Hesketh, S. (1997). *The analysis of proximity data*. London: Arnold.

Fienberg, S. E., & Larntz, K. (1976). Log linear representation for paired and multiple comparison models. *Biometrika, 63*, 245–254.

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch Models: Foundations, recent developments, and applications*. New York: Springer-Verlag.

Fisher, E., Dunn, M., & Thompson, J. K. (2002). Social comparison and body image: An investigation of body comparison processes using multidimensional scaling. *Journal of Social and Clinical Psychology, 21*, 566–579.

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 349–366.

Frank, K. A. (1996). Mapping interactions within and between cohesive subgroups. *Social Networks, 18,* 93–119.

Girard, R. A., & Cliff, N. (1976). A Monte Carlo evaluation of interactive multidimensional scaling. *Psychometrika, 41,* 43–64.

Gower, J. C. (1977). The analysis of asymmetry and orthogonality. In J. R. Barra, F. Brodeau, G. Romier, & B. van Cutsem (Eds.), *Recent developments in statistics* (pp. 109–123). Amsterdam: North-Holland.

Green, B. F., & Anderson, L. K. (1955). The tactual identification of shapes for coding switch handles. *Journal of Applied Psychology, 39*, 219–226.

Green, P. E., & Rao, V. R. (1972). *Applied multidimensional scaling.* New York: Holt, Rinehart & Winston.

Green, R. J., & Manzi, R. (2002). A comparison of methodologies for uncovering the structure of racial stereotype subgrouping. *Social Behavior and Personality, 30,* 709–727.

Griffith, T. L., & Kalish, M. L. (2002). A multidimensional scaling approach to mental multiplication. *Memory & Cognition, 30,* 97–106.

Halmos, P. R. (1958). *Finite-dimensional vector spaces.* New York: Van Nostrand Reinhold.

Hansen, J. I. C., & Scullard, M. G. (2002). Psychometric evidence for the Leisure Interest Questionnaire and analyses

of the structure of leisure interests. *Journal of Counseling Psychology, 49,* 331–341.

Heiser, W. J. (1988a). Multidimensional scaling with least absolute residuals. In H. -H. Bock (Ed.), *Classification and related methods of data analysis* (pp. 455–462). Amsterdam: North-Holland.

Heiser, W. J. (1988b). Selecting a stimulus set with prescribed structure from empirical confusion frequencies. *British Journal of Mathematical and Statistical Psychology, 41,* 37–51.

Heiser, W. J. (1989). Order invariant unfolding analysis under smoothness restrictions. In G. de Soete, H. Feger, & K. C. Klauer (Eds.), *New developments in psychological choice modeling* (pp. 3–31). Amsterdam: North-Holland.

Heiser, W. J., & de Leeuw, J. (1981). Multidimensional mapping of preference data. *Mathématiques et Sciences Humaines, 19*, 39–96.

Heiser, W. J., & Groenen, P. J. F. (1997). Cluster differences scaling with a within-clusters loss component and a fuzzy successive approximation strategy to avoid local minima. *Psychometrika, 62,* 63–83.

Henley, N. M. (1969). A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior, 8*, 176–184.

Kappesser, J., & Williams, A. C. D. (2002). Pain and negative emotions in the face: Judgements by health care professionals. *Pain, 99*, 197–206.

Kemmler, G., Holzner, B., Kopp, M., Dunser, M., Greil, R., Hahn, E., et al. (2002). Multidimensional scaling as a tool for analyzing quality of life data. *Quality of Life Research, 11,* 223–233.

Klingberg, F. L. (1941). Studies in measurement of the relations among sovereign states. *Psychometrika, 6,* 335–352.

Kocsis, R. N., Cooksey, R. W., & Irwin, H. J. (2002). Psychological profiling of sexual murders: An empirical model. *International Journal of Offender Therapy and Comparative Criminology, 46,* 532–554.

Krackhardt, D. (1988). Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Social Networks, 10,* 359–381.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis. *Psychometrika, 29,* 1–28.

Kruskal, J. B., & Carroll, J. D. (1969). Geometrical models and badness-of-fit functions. In P. R. Krishnaiah (Ed.), *Multivariate analysis* (Vol. 2, pp. 639–671). New York: Academic Press.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling.* Beverly Hills, CA: Sage.

Laskaris, N. A., & Ioannides, A. A. (2002). Semantic geodesic maps: A unifying geometrical approach for studying the structure and dynamics of single trial evoked responses. *Clinical Neurophysiology, 113,* 1209–1226.

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review, 9,* 43–58.

Levy, S., & Guttman, L. (1975). On the multivariate structure of wellbeing. *Social Indicators Research, 2*, 361–388.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York: John Wiley.

Lundrigan, S., & Canter, D. (2001). A multivariate analysis of serial murderers' disposal site location choice. *Journal of Environmental Psychology, 21,* 423–432.

MacCallum, R. C. (1979). Recovery of structure in incomplete data by ALSCAL. *Psychometrika, 44,* 69–74.

MacCallum, R. C., & Cornelius, E. T., III. (1977). A Monte Carlo investigation of recovery of structure by ALSCAL. *Psychometrika, 42,* 401–428.

MacKay, D. B. (1989). Probabilistic multidimensional scaling: An anisotropic model for distance judgments. *Journal of Mathematical Psychology, 33,* 187–205.

MacKay, D. B. (1995). Probabilistic multidimensional unfolding: An anisotropic model for preference ratio judgments. *Journal of Mathematical Psychology, 39,* 99–111.

MacKay, D. B. (2001). Probabilistic multidimensional scaling using a city-block metric. *Journal of Mathematical Psychology, 45,* 249–264.

Mackie, P. C., Jessen, E. C., & Jarvis, S. N. (2002). Creating a measure of impact of childhood disability: Statistical methodology. *Public Health, 116,* 95–101.

Magley, V. J. (2002). Coping with sexual harassment: Reconceptualizing women's resistance. *Journal of Personality and Social Psychology, 83,* 930–946.

Meulman, J. J. (1992). The integration of multidimensional scaling and multivariate analysis with optimal transformations. *Psychometrika, 57,* 539–565.

Meulman, J. J. (1996). Fitting a distance model to homogeneous subsets of variables: Points of view analysis of categorical data. *Journal of Classification, 13,* 249–266.

Meulman, J. J., Heiser, W. J., & SPSS. (1999). *Categories.* Chicago: SPSS.

Meulman, J. J., & Verboon, P. (1993). Points of view analysis revisited: Fitting multidimensional structures to optimal distance components with cluster restrictions on the variables. *Psychometrika, 58,* 7–35.

Mulaik, S. A. (1972). *The foundations of factor analysis.* New York: McGraw-Hill.

Mullen, K., & Ennis, D. M. (1991). A simple multivariate probabilistic model for preferential and triadic choices. *Psychometrika, 56,* 69–75.

Nishisato, S. (1994). *Elements of dual scaling: An introduction to practical data analysis.* Hillsdale, NJ: Lawrence Erlbaum.

Nishisato, S. (1996). Gleaning in the field of dual scaling. *Psychometrika, 61,* 559–599.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology, 43,* 25–53.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104,* 266–300.

Okada, A. (1997). Asymmetric multidimensional scaling of two-mode three-way proximities. *Journal of Classification, 14,* 195–224.

Paddock, J. R., & Nowicki, S., Jr. (1986). The circumplexity of Leary's Interpersonal Circle: A multidimensional scaling perspective. *Journal of Personality Assessment, 50*, 279–289.

Pollick, F. E., Paterson, H. M., Bruderlin, A., & Sanford, A. J. (2001). Perceiving affect from arm movement. *Cognition, 82,* B51–B61.

Porter, L. E., & Alison, L. J. (2001). A partially ordered scale of influence in violent group behavior: An example from gang rape. *Small Group Research, 32,* 475–497.

Pukrop, R., Steinmeyer, E. M., Woschnik, M., Czernik, A., Matthies, H., Sass, H., et al. (2002). Personality, attenuated traits, and personality disorders. *Nervenarzt, 73,* 247–254.

Ramsay, J. O. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika, 42,* 241–266.

Ramsay, J. O. (1978). Confidence regions for multidimensional scaling analysis. *Psychometrika, 43,* 145–160.

Ramsay, J. O. (1980). Joint analysis of direct ratings, pairwise preferences, and dissimilarities. *Psychometrika, 45,* 149–165.

Ramsay, J. O. (1982). Some statistical approaches to multidimensional scaling data. *Journal of the Royal Statistical Society, Series A (General), 145,* 285–312.

Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology, 53*, 94–101.

Rounds, J. B., Jr., Davison, M. L., & Dawis, R. V. (1979). The fit between Strong-Campbell Interest Inventory general occupation themes and Holland's hexagonal model. *Journal of Vocational Behavior, 15*, 303–315.

Samson, S., Zatorre, R. J., & Ramsay, J. O. (2002). Deficits of musical timbre perception after unilateral temporal-lobe lesion revealed with multidimensional scaling. *Brain, 125,* 511–523.

Schlesinger, I. M., & Guttman, L. (1969). Smallest space analysis of intelligence and achievement tests. *Psychological Bulletin, 71*, 95–100.

Sergent, J., & Takane, Y. (1987). Structures in two-choice reaction time data. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 300–315.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika, 22*, 325–345.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function: I. *Psychometrika, 27*, 125–140.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237,* 1317–1323.

Sherman, C. R. (1972). Nonmetric multidimensional scaling: A Monte Carlo study of the basic parameters. *Psychometrika, 37,* 323–355.

Shivy, V. A., & Koehly, L. M. (2002). Client perceptions of and preferences for university-based career services. *Journal of Vocational Behavior, 60,* 40–60.

Slater, P. (1960). The analysis of personal preferences. *British Journal of Statistical Psychology, 13*, 119–135.

Smith, P. K., Cowie, H., Olafsson, R. F., & Liefooghe, A. P. D. (2002). Definitions of bullying: A comparison of terms used, and age and gender differences, in a fourteen-country international comparison. *Child Development, 73,* 1119–1133.

Spence, I. A. (1972). Monte Carlo evaluation of three nonmetric multidimensional scaling algorithms. *Psychometrika, 37,* 461–486.

Spence, I. A., & Domoney, D. W. (1974). Single subject incomplete designs for nonmetric multidimensional scaling. *Psychometrika, 39,* 469–489.

Spence, I. A., & Lewandowsky, S. (1989). Robust multidimensional scaling. *Psychometrika, 54,* 501–513.

Storms, G. (1995). On the robustness of maximum likelihood scaling for violations of the error model. *Psychometrika, 60,* 247–258.

Struch, N., Schwartz, S. H., & Van der Kloot, W. A. (2002). Meanings of basic values for women and men: A cross-cultural analysis. *Personality and Social Psychology Bulletin, 28,* 16–28.

Sulmont, C., Issanchou, S., & Koster, E. P. (2002). Selection of odorants for memory tests on the basis of familiarity, perceived complexity, pleasantness, similarity and identification. *Chemical Senses, 27,* 307–317.

Sumiyoshi, C., Matsui, M., Sumiyoshi, T., Yamashita, I., Sumiyoshi, S., & Kurachi, M. (2001). Semantic structure in schizophrenia as assessed by the category fluency test: Effect of verbal intelligence and age of onset. *Psychiatry Research, 105*(3), 187–199.

Takane, Y. (1981). Multidimensional successive categories scaling: A maximum likelihood method. *Psychometrika, 46,* 9–28.

Takane, Y. (1982). The method of triadic combinations: A new treatment and its applications. *Behaviormetrika, 11,* 37–48.

Takane, Y., & Carroll, J. D. (1981). Nonmetric maximum likelihood multidimensional scaling from directional rankings of similarities. *Psychometrika, 46,* 389–405.

Takane, Y., & Sergent, J. (1983). Multidimensional scaling models for reaction times and same-different judgments. *Psychometrika, 48,* 393–423.

Takane, Y., Young, F. W., & de Leeuw, J. (1977). Nonmetric individual differences in multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika, 42,* 7–67.

Takkinen, S., & Ruoppila, I. (2001). Meaning in life as an important component of functioning in old age. *International Journal of Aging & Human Development, 53,* 211–231.

Taylor, P. J. (2002). A cylindrical model of communication behavior in crisis negotiations. *Human Communication Research, 28,* 7–48.

Torgerson, W. S. (1958). *Theory and methods of scaling.* New York: John Wiley.

Treat, T. A., McFall, R. M., Viken, R. J., & Kruschke, J. K. (2001). Using cognitive science methods to assess the role of social information processing in sexually coercive behavior. *Psychological Assessment, 13,* 549–565.

Treat, T. A., McFall, R. M., Viken, R. J., Nosofsky, R. M., MacKay, D. B., & Kruschke, J. K. (2002). Assessing clinically relevant perceptual organization with multidimensional scaling techniques. *Psychological Assessment, 14,* 239–252.

Tucker, L. R. (1960). Intra-individual and inter-individual multidimensionality. In H. Gulliksen & S. Messick (Eds.), *Psychological scaling: Theory and applications* (pp. 155–167). New York: John Wiley.

Tucker, L. R., & Messick, S. (1963). An individual differences model for multidimensional scaling. *Psychometrika, 28,* 333–367.

Van der Linden, W., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer.

Viken, R. J., Treat, T. A., Nosofsky, R. M., McFall, R. M., & Palmeri, T. J. (2002). Modeling individual differences in perceptual and attentional processes related to bulimic symptoms. *Journal of Abnormal Psychology, 111,* 598–609.

Wedel, M., & DeSarbo, W. S. (1996). An exponential-family multidimensional scaling mixture methodology. *Journal of Business & Economic Statistics, 14,* 447–459.

Weeks, D. G., & Bentler, P. M. (1982). Restricted multidimensional scaling models for asymmetric matrices. *Psychometrika, 47,* 201–208.

Welchew, D. E., Honey, G. D., Sharma, T., Robbins, T. W., & Bullmore, E. T. (2002). Multidimensional scaling of integrated neurocognitive function and schizophrenia as a disconnexion disorder. *NeuroImage, 17,* 1227–1239.

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Winsberg, S., & de Soete, G. (1993). A latent class approach to fitting the weighted Euclidean model, CLASCAL. *Psychometrika, 58,* 315–330.

Yates, A. (1987). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis.* Albany: State University of New York Press.

Young, F. W. (1970). Nonmetric multidimensional scaling: Recovery of metric information. *Psychometrika, 35,* 455–473.

Zielman, B., & Heiser, W. J. (1993). Analysis of asymmetry by a slide-vector. *Psychometrika, 58*, 101–114.

Zielman, B., & Heiser, W. J. (1996). Models for asymmetric proximities. *British Journal of Mathematical and Statistical Psychology, 49,* 127–146.

Zinnes, J. L., & Griggs, R. A. (1974). Probabilistic multidimensional unfolding analysis. *Psychometrika, 48,* 27–48.

Zinnes, J. L., & MacKay, D. B. (1983). Probabilistic multidimensional scaling: Complete and incomplete data. *Psychometrika, 48,* 27–48.

Zinnes, J. L., & Wolff, R. P. (1977). Single and multidimensional same-different judgments. *Journal of Mathematical Psychology, 16,* 30–50.

Chapter 3

Principal Components Analysis With Nonlinear Optimal Scaling Transformations for Ordinal and Nominal Data

Jacqueline J. Meulman

Anita J. Van der Kooij

Willem J. Heiser

3.1. Introduction

This chapter focuses on the analysis of ordinal and nominal multivariate data, using a special variety of principal components analysis that includes nonlinear optimal scaling transformation of the variables. Since the early 1930s, classical statistical methods have been adapted in various ways to suit the particular characteristics of social and behavioral science research. Research in these areas often results in data that are nonnumerical, with measurements recorded on scales having an uncertain unit of measurement. Data would typically consist of qualitative or categorical variables that describe the persons in a limited number of categories. The zero point of these scales is uncertain, the relationships among the different categories is often unknown, and although frequently it can be assumed that the categories are ordered, their mutual distances might still be unknown. The uncertainty in the unit of measurement is not just a matter of measurement error because its variability may have a systematic component.

For example, in the data set that will be used throughout this chapter as an illustration, concerning feelings of national identity and involving 25,000 respondents in 23 different countries all over the world (International Social Survey Programme [ISSP], 1995), there are variables indicating how close the respondents feel toward their neighborhood, town, and country, measured on a 5-point scale with labels ranging from *not close at all* to *very close.* This response format is typical for a lot of behavioral research and definitely is not numerical (even though the categories are ordered and can be coded numerically).

3.1.1. Optimal Scaling Transformations

An important development in multidimensional data analysis has been the optimal assignment of quantitative values to qualitative scales. This form of optimal

quantification (optimal scaling, optimal scoring) is a very general approach to treat multivariate (categorical) data. Taking the linear regression model as a leading case, we may wish to predict a response variable from a number of predictor variables. This objective is achieved by finding a particular linear combination of the predictor variables that correlates maximally with the response variable. Incorporating optimal scaling amounts to further maximization of this correlation, not only over the regression weights but also over admissible nonlinear functions of the predictor variables. For instance, in the National Identity Study data, we may try to find nonlinear scale values of the response categories of the closeness variables that improve the multiple-correlation coefficient for predicting willingness to move because it may be that some response categories equally predict high willingness, whereas other categories strongly differentiate between small steps in low willingness. These nonlinear functions are called transformations, optimal scalings, scorings, or quantifications. In this chapter, we will use both the terms *nonlinear optimal scaling transformations* and *optimal quantifications.* The optimal scaling process turns qualitative variables into quantitative ones. Optimality is a relative notion, however, because it is always obtained with respect to the particular data set that is analyzed.

The nonlinear optimal scaling transformations of ordered categorical or continuous (ordinal) data can be handled by means of *monotonic* transformations, which maintain the order in the original data. Categorical (nominal) data in which the categories are not ordered will be given an optimal quantification (scoring). Nonmonotonic functions can also be used for continuous (numeric) and ordinal variables when nonlinear relationships among the variables are assumed. In these cases, it is often useful to collapse the data in a limited number of categories (sometimes called *binning*) and find an optimal quantification for the categories (see Section 3.6.2). However, if we do not want to lose the fine gradings, we can also fit a monotonic or nonmonotonic spline. A spline is a function that consists of piecewise polynomials of a low degree that are joined at particular points, called *knots.* Of course, special software is required to simultaneously transform and analyze the data.

3.1.2. Software for Nonlinear Principal Components: CATPCA

A state-of-the-art computer program, called CATPCA, that incorporates all the features that will be described in this chapter is available from SPSS Categories 10.0 onwards (Meulman, Heiser, & SPSS, 1999). In CATPCA, there is a large emphasis on graphical display of the results, and this is done in joint plots of objects[1] and variables, also called *biplots* (Gower & Hand, 1996). In addition to fitting points for individual objects, additional points may be fitted to identify groups among them, and graphical display can be in a *triplot,* with variables, objects, and groups of objects. Special attention will be given to particular properties that make the technique suited for data mining. Very large data sets can be analyzed when the variables are categorical at the outset or by binning.

Because CATPCA incorporates differential weighting of variables, it can be used as a "forced classification" method (Nishisato, 1984), comparable to "supervised learning" in machine learning terminology. Objects and/or variables can be designated to be supplementary; that is, they can be omitted from the actual analysis but fitted into the solution afterwards. When a prespecified configuration of points is given, the technique may be used for property fitting (external unfolding), that is, fitting external information on objects, groups, and/or variables into the solution (see Section 3.6.1). The information contained in the biplots and triplots can be used to draw special graphs that identify particular groups in the data that stand out on selected variables.

Summarizing, CATPCA can be used to analyze complicated multivariate data, consisting of nominal, ordinal, and numerical variables. A straightforward spatial representation is fitted to the data, and different groups of objects can be distinguished in the solution without having to aggregate the categorical data beforehand. We will discuss the various aspects of the analysis approach, giving attention to its data-analytical, graphical, and computational aspects.

3.1.3. Some Historic Remarks on Related Techniques

Historically, the idea of optimal scaling originated from different sources. On one hand, we find the history of the class of techniques that is nowadays usually called (*multiple*) *correspondence analysis,* a literal translation of Benzécri's *L'analyse des correspondances (multiples)* (Benzécri, 1973, 1992). This history can be traced in the work of Fisher (1948),

1. In the CATPCA terminology, the units of analysis are called objects; depending on the application, these can be persons, groups, countries, or other entities on which the variables are defined.

Guttman (1941), Burt (1950), and Hayashi (1952), among others, and in the rediscoveries since the 1970s (among others, see Benzécri, 1992; de Leeuw, 1973; Greenacre, 1984; Lebart, Morineau, & Warwick, 1984; Saporta, 1975; Tenenhaus & Young, 1985). The class of techniques is also known under the names *dual scaling* (Nishisato, 1980, 1994) and *homogeneity analysis* (Gifi, 1981/1990). In the course of its development, the technique has been given many different interpretations. In the original formulation of Guttman (1941), the technique was described as a principal components analysis of qualitative (nominal) variables. There is also an interpretation as a form of generalized canonical correlation analysis (Lebart & Tabard, 1973; Masson, 1974; Saporta, 1975), based on earlier work by Horst (1961a, 1961b), Carroll (1968), and Kettenring (1971).

Another major impetus to optimal scaling was given by work in the area of nonmetric multidimensional scaling (MDS), pioneered by Shepard (1962a, 1962b), Kruskal (1964), and Guttman (1968). In MDS, a set of proximities between objects is approximated by a set of distances in a low-dimensional space, usually Euclidean. Optimal scaling of the proximities was originally performed by monotonic regression; later on, spline transformations were incorporated (Ramsay, 1982). Since the so-called nonmetric breakthrough in MDS in the early 1960s, optimal scaling has subsequently been integrated in multivariate analysis techniques that hitherto were only suited for the analysis of numerical data. Some early contributions include Kruskal (1965), Shepard (1966), and Roskam (1968). In the 1970s and 1980s, psychometric contributions to the area became numerous. Selected highlights from the extensive psychometric literature on the subject include de Leeuw (1973); Kruskal and Shepard (1974); Young, de Leeuw, and Takane (1976); Young, Takane, and de Leeuw (1978); Nishisato (1980); Heiser (1981); Young (1981); Winsberg and Ramsay (1983); Van der Burg and de Leeuw (1983); Van der Burg, de Leeuw, and Verdegaal (1988); and Ramsay (1988). Attempts at systematization resulted in the ALSOS system by Young et al. (1976), Young et al. (1978), and Young (1981) and the system developed by the Leiden "Albert Gifi" group. Albert Gifi's (1990) book, *Nonlinear Multivariate Analysis,* provides a comprehensive system, combining optimal scaling with multivariate analysis, including statistical developments such as the bootstrap. Since the mid-1980s, the principles of optimal scaling have gradually appeared in the mainstream statistical literature (Breiman & Friedman, 1985; Buja, 1990; Gilula & Haberman, 1988; Hastie et al., 1994; Ramsay, 1988). The Gifi system is discussed among traditional statistical techniques in Krzanowski and Marriott (1994).

3.2. Graphical Representation

The way we will treat principal components analysis (PCA) is more like a multidimensional scaling (MDS) technique than a technique from the classic multivariate analysis (MVA) domain. The central concept in classical multivariate analysis is the covariance or correlation *among variables.* Consequently, the modeling of the covariance or correlation matrix is the main objective of the analysis; therefore, the persons on which the variables are defined are usually regarded merely as a replication factor. Thus, the role of the persons is confined to acting as intermediaries in obtaining covariance or correlation measures that describe the relationships among the variables. In the multidimensional scaling domain, techniques have been developed for the analysis of a (not necessarily) symmetric square table, with entries representing the degree of dissimilarity *among any kind of objects,* which may be persons. The objective, then, is to map the objects in some low-dimensional space, in which the distances resemble the initial dissimilarities as closely as possible. To make distinctions between MDS and classical MVA more explicit than they would be from a unifying point of view, consider factor analysis, one of the major data-analytic contributions to statistics originating from the behavioral sciences. Unfortunately, from a visualization point of view, the representation of persons became very complicated in the process. The factor-analytic model aggregates observations on persons into an observed covariance matrix for the variables, and the model involved for representing this covariance matrix is focused on the fitting of a matrix incorporating the common covariances among the variables and another (diagonal) matrix that displays the unique variance of each variable. By formulating the data-analytic task through this particular decomposition, the factor scores that would order the persons with respect to the underlying latent variables are undetermined: Although various approaches exist to have the persons reappear, their scores cannot be determined in a unique manner.

In contrast, principal components analysis can be discussed by focusing on the joint representation of persons and variables in a joint low-dimensional space. The variables in the analysis are usually represented as

vectors (arrows) in this low-dimensional space. Each variable is associated with a set of component loadings, one for each dimension, and these loadings, which are correlations between the variables and the principal components, give coordinates for the variables to represent them as vectors in the principal component space. The squared length of such a vector corresponds to the percentage of variance accounted for and thus equals the sum of squares of the component loadings across the dimensions. If we sum the squared component loadings in each dimension over the variables, we obtain the eigenvalues. In the CATPCA approach discussed in the sequel of this chapter, a variable can also be viewed as a set of category points. When a variable is visualized as a vector, these category points are located on a line, where the direction is given by the component loadings. There is, however, an alternative to representing the category points on a straight line, which is by displaying them as points in the middle, the *centroid,* of the cloud of associated object points in the low-dimensional representation space. These two ways of representing a variable will be called the *vector* and the *centroid model,* respectively.

3.2.1. The Vector Model

A very first description of the vector model can be found in Tucker (1960); Kruskal (1978) used the term *bilinear model,* and Gabriel (1971) invented the name *biplot.* A comprehensive book on biplots is by Gower and Hand (1996). The prefix *bi-* in *bilinear* and *biplot* refers to two sets of entities, the objects and the variables (and not to two dimensions, as is sometimes erroneously assumed). In PCA, the observed values on the M variables are approximated by the inner product of the P-dimensional component scores and component loadings for the variables, with P much smaller than M. Usually, the classic reference to lower rank approximation is Eckart and Young (1936), but it might be worthwhile to note that this reference is challenged by Stewart (1993), who remarks that the contribution of Schmidt (1907) was much earlier, which is also noted by Gifi (1990). Because the fit is defined on an inner product, one has to make a coherent choice of normalization.[2] Usually, the component scores are normalized to have means of zero and variances equal to 1; the coherent normalization implies that the component loadings are correlations between the variables and the P dimensions of the space fitted to the objects. Component loadings give coordinates for a variable vector in the space, and the angles between the vectors then approximate the correlations between the variables. The inner product of the matrix of component scores and a variable vector approximates a column of the data matrix, and the length of the variable vector in the space equals the correlation between the variable and its approximation.

In the classical PCA biplot, persons are represented as points, and variables are represented as vectors in the same low-dimensional space. In contrast, in the analysis of preference data, in which Tucker's (1960) vector model originated, the persons are represented as vectors and the items are represented as points (for an extended treatment of the vector model in the context of preference analysis, see Carroll, 1968, 1972; Heiser & de Leeuw, 1981). Because we include nonlinear optimal scaling transformations for the variables in principal components analysis, the vector/bilinear model represents not the original categorical variable but the transformed variable, which is given optimal (non)monotonic quantifications for its categories.

3.2.2. The Centroid Model

Unlike the vector model that is based on projection, the centroid model is most easily viewed in terms of distances between object points and category points. In the centroid model, each category obtains coordinates that represent the category in the same space as the objects. The centroid model originates from multiple-correspondence analysis (MCA), where a nominal variable is represented as a set of category points, which are in the centroids of the associated objects. The categories of a particular variable partition the cloud of object points into subclouds. When these subclouds overlap considerably, we say that the corresponding variable is a relatively bad discriminator. On the other hand, well-separated subclouds are associated with a good discriminator. When we have chosen the centroid model for two or more variables, and when the solution has a decent fit, the category points that are associated with the same objects will be close together, whereas categories of the same variable will be far apart (each representing a subcloud of object points through its centroid). The weighted mean squared distance of the category points toward the origin gives a measure similar to variance accounted for and has been called the *discrimination measure* (Gifi, 1990).

2. Because the inner product between two vectors $\mathbf{a}$ and $\mathbf{b}$ is defined as $\mathbf{a}'\mathbf{b}$, it remains unchanged if we transform $\mathbf{a}$ into $\tilde{\mathbf{a}} = \mathbf{T}\mathbf{a}$ and $\tilde{\mathbf{b}}$ into $\mathbf{b} = \mathbf{S}\mathbf{b}$, with $\mathbf{S} = (\mathbf{T}')^{-1}$, because $\mathbf{a}'\mathbf{b} = \mathbf{a}'\mathbf{T}'\mathbf{S}\mathbf{b} = \tilde{\mathbf{a}}'\tilde{\mathbf{b}}$. Choosing principal axes and a coherent normalization settles the choice of $\mathbf{T}$ and $\mathbf{S}$ (also see Section 3.2.4).

A special feature of the CATPCA approach is the possibility to fit the vector (bilinear) model and the centroid (distance) model for different variables (or even for the same variable) in a single analysis, a feature not available in other software programs that perform nonlinear principal components analysis.

3.2.3. Clustering and Forced Classification

The CATPCA method accommodates differential weights for separate variables. In this way, the centroid model can be used for *forced classification* (a term coined by Nishisato, 1984), which can also be called *supervised learning.* Forced classification is obtained by applying a (very) large weight for the particular variable that we have selected for the classification. Applying this large weight in combination with the centroid model will cause the object points that belong together to cluster into subclouds in the low-dimensional space. The larger the weight that is given, the tighter the clustering will be. This feature is especially attractive when the number of objects is very large and when they can be identified as members of a particular subgroup, such as citizens of different countries (as in the example given below) or members of a particular social group. In these cases, we would not be so much interested in the individual results but in the results for the groups. Because we are dealing with categorical data, it would not make sense to average the data beforehand. The use of a weighted classification variable takes care of this averaging during the analysis, and the size of the weight controls the subsequent clustering of the object points around their centroid.

In this way, we make certain that the classification variable plays a significant role in the first few dimensions of the principal components analysis solution. This property is extremely useful when we would use PCA as a first step in a discriminant analysis to diminish the number of predictors. Such a particular strategy is often used when the number of predictors exceeds the number of objects in the data matrix, as is the case, among others, in genometrics (the analysis of microarray gene expression data), proteometrics, and chemometrics but also in Q-sort data, with judges acting as variables, and with a classification variable available for the objects. In the same manner, CATPCA can be used as a prestep in a multiple regression analysis when the number of predictors exceeds the number of objects. In the latter case, the response variable is included in the analysis, with a much larger weight than the other variables and with the application of the vector model.

3.2.4. Different Normalizations

Different normalization options are possible for the display of objects and variables in the low-dimensional Euclidean space. The most commonly used normalization option in principal components analysis is to display the objects in an orthonormal cloud of object points, in which the dimensions themselves have equal variance. Then, the representation of the variables accounts for the differential fit in subsequent dimensions, with the first dimension accounting for most of the variance and subsequent dimensions displaying the variance accounted for (VAF) in a decreasing order. When the object scores are normalized, however, one loses a straightforward distance interpretation with respect to the objects. To attain the latter, one should normalize the component loadings and leave the object scores free (but keeping the inner product fixed). Therefore, an alternative option is provided that should be used if we wish CATPCA to perform a principal coordinates analysis as described in Gower (1966), which is equivalent to the classical MDS method usually attributed to Torgerson (1958). In principal coordinates analysis, the emphasis is on the representation of the objects, and the cloud of object points displays the differential fit in subsequent dimensions (the cloud is not orthonormal but shows a definite shape). The interpretation of nonlinear PCA in terms of distances between objects is given, among others, in Heiser and Meulman (1983) and Meulman (1986, 1992). Whether the object points or the (category points of the) variables are normalized depends algebraically on the allocation of the eigenvalues in the use of the singular-value decomposition to represent both sets of entities in the low-dimensional space. Therefore, in CATPCA, the impact of the eigenvalues (symbolizing the fit) could also be distributed symmetrically over objects and variables (enhancing the joint display, especially when the overall fit is not very large) or handled in a completely customized way to optimize the quality of the joint representation.

3.2.5. Different Biplots and a Triplot

For the display of the results, a variety of biplots is available in CATPCA. A biplot can display the objects (as points) and the variables (as vectors),

the objects and groups among them (represented by centroids), or the variables with groups of objects (represented by centroids). Combining these three options reveals relationships between objects, groups of objects, and variables, and we call this display a *triplot.* The ultimate summary of the analysis combines the information in the biplots and triplots in one-dimensional displays. These are obtained by taking centroids of the objects, according to a particular (classification) variable, and projecting these centroids on the vectors representing variables of particular interest in the analysis. In this way, the graph identifies particular groups in the data that stand out on the selected variables. The use of the projected centroids representation is demonstrated in Section 3.4.6.

3.3. MVA With Different Nonlinear Optimal Scaling Transformations

In the nonlinear transformation process in CATPCA, an appropriate quantification level has to be chosen for each of the variables. The most restricted transformation level is called *numerical;* it applies a linear transformation to the original integer scale values, so that the resulting variables will be standardized. The numerical scaling level fits category points on a straight line through the origin, with equal distances between the points. Instead of a linear transformation, we have the choice between different nonlinear transformations, and these can either be monotonic with the original order of the categories or nonmonotonic.

3.3.1. Nominal Transformation and Multiple Nominal Quantifications

When the only fact we will take into account is that a particular subset of the objects is in the same category (whereas others are in different ones), we call the transformation *nominal* (or *nonmonotonic*); the quantifications only maintain the class membership, and the original categories are quantified to give an optimal ordering. The nonlinear transformation can be carried out either by a least squares identity regression (which amounts to averaging over objects in the same category) or by fitting a nonmonotonic regression spline. Geometrically, the nominal scaling level fits category points in an optimal order on a straight line through the origin. The direction of this straight line is given by the corresponding component loadings.

What has been labeled the centroid model above (a categorical variable represented by a set of points located in the centroid of the objects that are in the associated categories) is also called a *multiple* nominal quantification. The quantification is called multiple because there is a separate quantification for each dimension (the average of the coordinates of the objects in the first dimension, the second dimension, etc.) and nominal because there is no prespecified order relationship between the original category numbers and the order in any of the dimensions. An example of the difference between a nominal and a multiple nominal quantification will be given later on. We choose a nominal transformation when we wish the category points to be represented on a vector and a multiple quantification when we wish them to be in the centroids of the associated objects.

3.3.2. Monotonic and Nonmonotonic Splines

Within the domain of either monotonic or nonmonotonic transformations, two approaches are available: optimal least squares transformations or optimal spline transformations. As indicated above, the class of monotonic transformations has its origin in the nonmetric multidimensional scaling literature (Kruskal, 1964; Shepard, 1962a, 1962b), in which original dissimilarities were transformed into pseudo-distances to be optimally approximated by distances between object points in low-dimensional space. Free monotonic transformations have been implemented since then to generalize multivariate analysis techniques as well (e.g., see Gifi, 1990; Kruskal, 1965; Kruskal & Shepard, 1974;Young et al., 1978). We call these transformations free monotonic because the number of parameters that is used is free. Because this freedom could lead to overfitting of the MVA model over the transformation of the variables, a more restricted class of transformations was introduced into the psychometric literature. The most important ones form the class of regression splines, and these were introduced in multiple regression analysis and principal components analysis in Winsberg and Ramsay (1980, 1983; for a nice overview, see Ramsay, 1988). For splines, the number of parameters is determined by the degree of the spline that is chosen and the number of interior knots. Because splines use fewer parameters, they usually will be smoother and more robust, albeit at the cost of less goodness of fit with respect to the overall loss function that is minimized.

3.3.3. Goodness of Fit: Component Loadings, Variance Accounted For, Eigenvalues, and Cronbach's α

Principal components analysis studies the interdependence of the variables. Nonlinear transformations maximize the average interdependence, and this optimality property can be expressed in various forms. When variables obtain an ordinal (monotonic spline) transformation or a nominal (nonmonotonic spline) transformation, the technique maximizes the sum of the P largest eigenvalues of the correlation matrix between the transformed variables (where P indicates the number of dimensions that are chosen in the solution). The sum of the eigenvalues, the overall goodness-of-fit index, is equal to the total variance accounted for (in the transformed variables). The variance accounted for in each dimension for each variable separately is equal to the squared component loading, and the component loading itself is the correlation between the transformed variable and a principal component (given by the object scores) in a particular dimension.

There is a very important relationship between the eigenvalue (the total sum of squared component loadings in each dimension) and probably the most frequently used coefficient for measuring internal consistence in applied psychometrics: Cronbach's α (e.g., see Heiser & Meulman, 1994; Lord, 1958; Nishisato, 1980). The relationship between α and the total variance accounted for, as expressed in the eigenvalue λ, is

$$\alpha = M(\lambda - 1)/(M - 1)\lambda, \qquad (1)$$

where M denotes the number of variables in the analysis. Because λ corresponds to the largest eigenvalue of the correlation matrix, and because CATPCA maximizes the largest eigenvalue of the correlation matrix over transformations of the variables, it follows that CATPCA maximizes Cronbach's α. This interpretation is straightforward when the CATPCA solution is one-dimensional. Generalized use of this coefficient in more-dimensional CATPCA is described in Section 3.4.2.

3.4. CATPCA in Action, Part 1

Throughout this chapter, the principles behind categorical principal components analysis (CATPCA), or principal components analysis with nonlinear optimal scaling transformations, will be illustrated by using a large-scale multivariate data set from the ISSP (1995) that can be considered exemplary for data collected in the social and behavioral sciences. The ISSP is a continuous annual cross-national data collection project that has been running since 1985. It brings together preexisting social science projects and coordinates research goals, thereby adding a cross-national perspective to the individual national studies. Since 1985, the ISSP grew from 6 to 30 participating countries in 1998. The ISSP Internet pages give access to detailed information about the ISSP data service provided by the Zentral Archiv, Cologne. The homepage of the ISSP-Secretariat provides information on ISSP history, membership, publications, and the ISSP listserver.

The original data concern feelings of national identity from about 28,500 respondents in 23 different countries all over the world. Because the number of respondents in the sample in each of the participating countries is not proportional to the population size, a random sample from the original data was taken so that all countries have the same weight in the analysis, with all being represented by 500 respondents. This selection makes the total number of individuals in our examples equal to 11,500.

For the first application, we have selected three groups of variables from the National Identity Study. The first group of five variables indicates how close the respondents feel toward their neighborhood (CL-1), their town (CL-2), their county (CL-3), their country (CL-4), and their continent (CL-5). (The data were recoded so that a score of 1 indicates *not close at all* and a score of 5 indicates *very close.*) The next five variables indicate whether the respondents are willing to move from their neighborhood to improve their work or living conditions, either to another neighborhood (MO-1), another city (MO-2), another county (MO-3), another country (MO-4), or another continent (MO-5), with the score 1 indicating *very unwilling* and the score of 5 indicating *very willing.* The third set of variables concerns statements about immigrants, asking the respondents on a scale from 1 to 5 whether they *strongly disagree* (1) or *strongly agree* (5) with the following statements: "Foreigners should not be allowed to buy land [in this country]" (I-Land), "Immigrants increase crime rates" (I-Crime), "Immigrants are generally good for the economy" (I-Econ), "Immigrants take jobs away from people who were born [in this country]" (I-Jobs), and " Immigrants make [this] country more open to new ideas and cultures" (I-Ideas). Also, respondents were asked to scale themselves with respect to the statement, "The number of immigrants to [my country] nowadays should be *reduced a lot* (1) . . . *increased a lot* (5)." More than 50% of the respondents

have one or more missing values on these 16 variables; therefore, a missing data treatment strategy other than deleting all cases with missing data is required, and it was decided to use the straightforward CATPCA option of imputing the modal category for each of the variables. (See Section 3.6.3 on the treatment of missing data for more sophisticated approaches available in the optimal scaling framework.)

3.4.1. VAF and Cronbach's α

The results of a two-dimensional solution with monotonic spline transformations will be presented that explains 41% of the variance of the scores of the 11,500 respondents on the 16 variables. The percentage of variance accounted for (PVAF) in the first dimension (26.7%) is almost twice the PVAF in the second dimension (14.4%). The VAF in the first dimension equals .267 $\times$ 16 (number of variables) = 4.275, and in the second dimension, .144 $\times$16 = 2.305. As explained above, the VAF is closely related to Cronbach's α.

As illustrated in Heiser and Meulman (1994), the relationship between α and the VAF (eigenvalue) is not linear but monotonically increasing, and it is severely nonlinear when M, the number of variables, grows. For $M = 16$, as in our example, the VAF in the first dimension corresponds to a value of $\alpha = .817$, and the VAF in the second dimension corresponds to a value of $\alpha = .604$. If we take the total variance accounted for (6.580) as the value of λ in equation (1), $\alpha = .905$ (the maximum is 1). This use of equation (1) clearly gives a much more general interpretation of α than was originally intended but provides an indication of the global fit of the CATPCA solution. The VAF per dimension is equal to the sum of squares of the component loadings and equal to the associated eigenvalue of the correlation matrix between the optimally transformed variables. Note that the value of α for a particular dimension becomes negative when the associated eigenvalue is less than 1.0. The largest eigenvalue of the correlation matrix between the original variables is 4.084, so the increase in VAF is $1 - 4.084/4.275 = 4.5\%$, which is not a dramatic overall increase. For most of the individual variables, however, the transformation is clearly nonlinear, as shown in Figure 3.1.

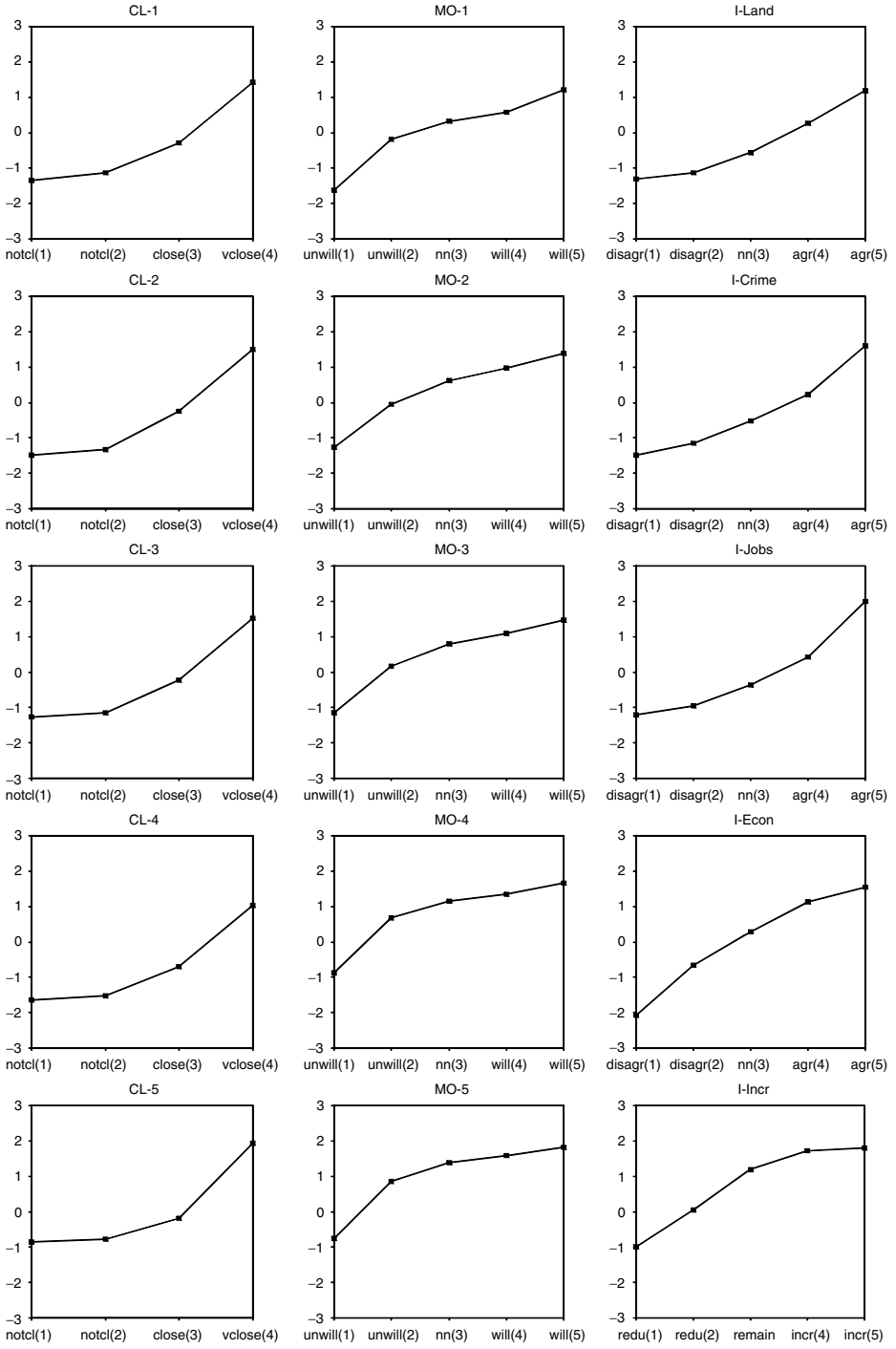
3.4.2. Nonlinear Transformations

In Figure 3.1, the transformations for CL-1 unto CL-5, MO-1 unto MO-5, and I-Land unto I-Incr are displayed in its columns; the optimal quantifications are given on the vertical axes versus the original values on the horizontal axes. The nonlinear transformations for CL-1 unto CL-5 show convexity, indicating that there is less distinction between the *not close at all* = ncl(1) and *not close* = ncl(2) categories, which are contrasted to the *very close* = ncl(4) category; the *close* = ncl(3) category is almost always near to the mean of 0. The MO-1 unto MO-5 quantifications show the opposite pattern: The nonlinear transformations approximate a concave function, grouping the *willing, very willing* categories, which are contrasted to the *very unwilling* category. The *unwilling* category has quantifications close to the mean, except for MO-4 and MO-5, which show the most concave functions. When we then inspect the quantifications for I-Land, I-Crime, and I-Jobs (the statements in which a high score expresses a negative attitude toward immigrants), we see that the transformations are convex again, contrasting the flat part for the (*strongly*) *disagree* categories at the lower end from the steep part toward the *strongly agree* category at the upper end. So these transformations resemble those for the CL variables. Looking at the quantifications for I-Econ and I-Incr, which express a positive attitude toward immigrants, we see that their quantifications give concave functions, just as for the MO variables: *strongly disagree* (at the lower end) is contrasted with *agree* and *strongly agree* (at the upper end) for I-Econ, and *reduced a lot* is contrasted with *increase* and *increase a lot* at the upper end for I-Incr ("the number of immigrants should be . . ."). The overall conclusion is that the steep parts of each of the transformations express negative feelings toward immigrants because they occur at the upper end for the negatively stated attitudes and at the lower end for the positively stated attitudes. Simultaneously, this pattern is reflected in the transformations for the CL variables, with the steep part indicating that one feels very close to one's living environment, and the MO variables, with the steep part indicating that one is very unwilling to move.

3.4.3. Representing Variables as Vectors

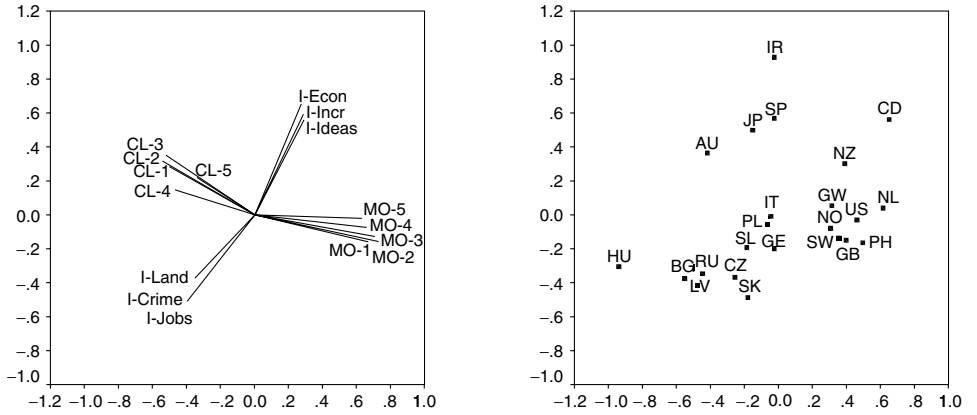
The optimal quantification process turns a qualitative, nominal (or ordinal) variable into a quantitative, numerical variable. The resulting nonlinearly transformed variable can be represented as a vector in the space that is determined for the objects. The coordinates for such a vector are given by the associated component loadings that give the correlation between the transformed variable and the dimensions of the

Figure 3.1 Spline Transformation of CL Variables (First Column), MO Variables (Second Column), and IM Variables From the CATPCA of the 1995 ISSP National Identity Study



object space. The graph of the component loadings is given in Figure 3.2 (left-hand panel), which shows vectors going in four different directions from the origin (the point 0, 0). Going clockwise, the first group of vectors points in the north-northeast direction, containing I-Econ, I-Incr, and I-Idea; the second

Figure 3.2 Loadings for MO, CL, and IM Variables (Left-Hand Panel) and Category Points for Country (Right-Hand Panel) From the CATPCA Analysis

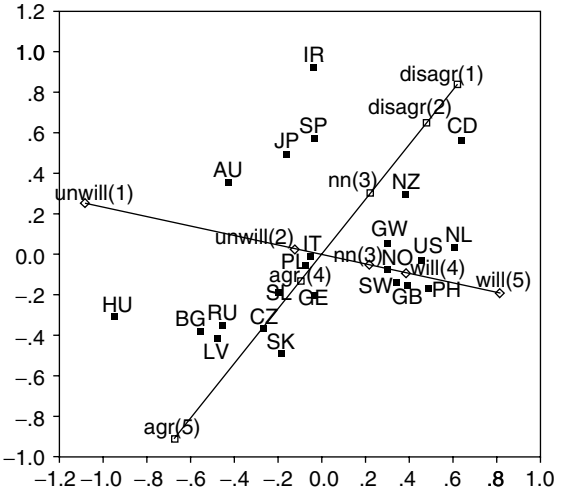


NOTE: Countries are identified as follows: IT = Italy; PL = Poland; SL = Slovenia; GE = East Germany; HU = Hungary; BG = Bulgaria; LV = Latvia; RU = Russia; CZ = Czech Republic; SK = Slovak Republic; AU = Austria; JP = Japan; SP = Spain; IR = Ireland; GW = West Germany; NO = Norway; SW = Sweden; GB = Great Britain; PH = Philippines; US = United States; NL = Netherlands; NZ = New Zealand; CD = Canada.

group points to the east-southeast, comprising the MO variables. The I-Land, I-Crime, and I-Jobs variables point in the south-southwest direction, and the CL variables, finally, point toward the west-northwest. From the transformation plots described above, we know that these directions indicate positive attitudes toward immigrants, willingness to move, very negative attitudes toward immigrants, and feeling very close to one's environment, respectively. It should be noted that each of these four groups of vectors has starting points representing the opposite meaning extending at the opposite side of the origin. So very close to the I-Econ, I-Incr, and I-Idea vectors, we should also envision the starting points of I-Land, I-Crime, and I-Jobs representing positive attitudes, as in the flat parts of the corresponding transformation plots. The reverse, therefore, is also true: The lower, very negative sides of the vectors for I-Econ, I-Incr, and I-Idea are very close to the plotted very negative sides of the vectors for I-Land, I-Crime, and I-Jobs. This whole story can be repeated for the MO and CL vectors that extend either to the right or to the left from the origin (also, see Figure 3.3).

The *very unwilling to move* lower endpoints are close to the *very close* upper endpoints, whereas the *not close* lower endpoints are near the *willing to move* upper endpoints. Now that we have interpreted the extremes of the optimally scaled categories depicted in the transformation plots, we can also interpret the full range of quantifications with respect to their original category labels. Before this will be done in Section 3.4.5, however, we will first inspect a

Figure 3.3 Joint Category Points for Country, MO-1, and I-Crime



NOTE: Countries are identified as follows: IT = Italy; PL = Poland; SL = Slovenia; GE = East Germany; HU = Hungary; BG = Bulgaria; LV = Latvia; RU = Russia; CZ = Czech Republic; SK = Slovak Republic; AU = Austria; JP = Japan; SP = Spain; IR = Ireland; GW = West Germany; NO = Norway; SW = Sweden; GB = Great Britain; PH = Philippines; US = United States; NL = Netherlands; NZ = New Zealand; CD = Canada.

different type of variables that can be introduced into the analysis described thus far.

3.4.4. Supplementary Variables

In the analysis of the CL, MO, and IM variables, we added a supplementary variable labeled *country*. This variable indicates from which of the 23 different countries the respondent originates. A supplementary variable has no influence on the actual analysis, but its quantifications are computed afterwards to establish its relationship with the solution obtained. In the case of the National Identity Study data, the number of respondents is too large to inspect the object scores on an individual level. Having the Country variable as a supplementary variable, however, gives the opportunity to display clusters of respondents from the same country by a single point. When the respondents from a particular country are very heterogeneous, their individual points will be scattered all over the two-dimensional space, and their associated country point, computed as the centroid of the appropriate individual points, will be located close to the origin of the configuration. To obtain these centroids for the 23 different countries in the National Identity Study, we have to specify that the country variable should obtain multiple nominal quantifications. The result is shown in the right-hand panel of Figure 3.2. In this graph, we see various clusters of points in three different directions, starting from the origin, which itself is close to Italy (IT) and Poland (PL) (and Slovenia [SL] and East Germany [GE]). First, a cluster of points contains Hungary (HU), Bulgaria (BG), Latvia (LV), Russia (RU), the Czech Republic (CZ), and the Slovak Republic (SK) in the lower left corner. Going in the upper left direction, we see Austria (AU), Japan (JP), Spain (SP), and Ireland (IR). Finally, going from the origin straight to the right, we have West Germany (GW), Norway (NO), Sweden (SW), Great Britain (GB), the Philippines (PH), the United States (US), and the Netherlands (NL). New Zealand (NZ) and Canada (CD) are almost on a straight line from the origin toward the upper right corner of the graph. Having these coordinates for the 23 countries, we can construct a biplot of the country points and the vectors for the CL, MO, and IM variables.

3.4.5. A Biplot of Centroids and Vectors

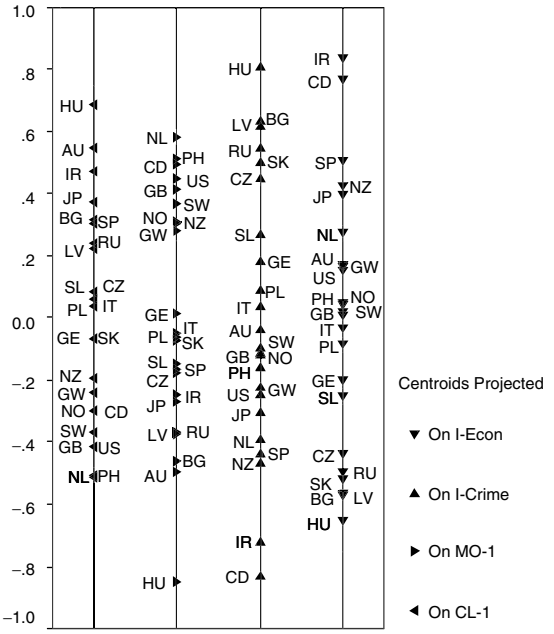
As described above, the CATPCA methodology allows a variety of different biplots. Because the number of objects in the National Identity Study is too large to inspect the relationship between the objects and the variables on the individual level, we represent the individual points by the centroids that are obtained by the supplementary country variable. There are two different ways for joint representation of country points and the vectors for the variables. The most straightforward one is a graph with the centroids from the right-hand panel of Figure 3.2 superimposed on the component loadings depicted in the left-hand panel. Elements of this plot (not shown) can be highlighted by the joint representation of the centroids and category points for selected variables. For illustration in our case, MO-1 and I-Crime were chosen, and the resulting plot is given in Figure 3.3. Here we notice the three most important clusters: Cluster 1 contains HU, BG, RU, LV, SK, and CZ; Cluster 2 contains AU, JP, SP, and IR; and Cluster 3 contains GW, NO, SW, GB, PH, US, and NL, located between the vectors given for MO-1 and I-Crime. In contrast to the component plot in Figure 3.2, a variable is now represented by the full set of category points on a straight line through the origin. For I-Crime, the category points "disagr(1) = *strongly disagree*" and "disagr(2) *disagree*" are both located at the side of the vector that points toward the north, whereas "agr(5) = *strongly agree*" is located at the opposite end, pointing to the south. The category "agr(4) = *agree*" is located close to the origin (compare the quantification close to zero in the transformation plot). The vector for the MO-1 variable contrasts "unwill(1) = *very unwilling*" on the left with "will(4) = *willing*" and "will(5) = *very willing*" on the right; here, the category "unwill(2) = *unwilling*" is close to the origin.

From the location of the country points with respect to the vectors for the variables, we can derive the relative positions by projection; for example, Ireland (IR) and Canada (CD) score high on the *disagree* end of the "Immigrants increase crime" vector. With respect to the cluster structure described above, the first cluster (with Russia [RU] in the center) agrees with the statement that immigrants increase the crime rate, and it is unwilling to move. Cluster 2, containing Japan, is also unwilling to move but (strongly) disagrees with the I-Crime statement. The third cluster, containing the United States, mildly disagrees but is willing to move (from its neighborhood).

3.4.6. Projected Centroids

The relative joint position of countries on statements is most clearly represented in the "projected centroids"

Figure 3.4 Projected Centroids for Country on Selected Variables (From Right to Left)



NOTE: Countries are identified as follows: IT = Italy; PL = Poland; SL = Slovenia; GE = East Germany; HU = Hungary; BG = Bulgaria; LV = Latvia; RU = Russia; CZ = Czech Republic; SK = Slovak Republic; AU = Austria; JP = Japan; SP = Spain; IR = Ireland; GW = West Germany; NO = Norway; SW = Sweden; GB = Great Britain; PH = Philippines; US = United States; NL = Netherlands; NZ = New Zealand; CD = Canada.

plot, shown in Figure 3.4. Here the 23 countries have been projected on the vectors for the statements as in the biplot, but now these projections themselves are shown on straight parallel lines representing the statements. From left to right, the following statements were used: CL-1, MO-1, I-Crime, and I-Econ. As we know from Figure 3.2 (left-hand panel), CL-1 and MO-1 are each other's opposite, and so this is also seen in Figure 3.4, with HU, AU, IR, JP, and BG scoring high on CL-1 (and low on MO-1) and NL, PH, CD, US, and the other countries from Cluster 3 scoring high on MO-1 (and low on CL-1). The two other variables represented show contrasts between Cluster 1 (scoring high on I-Crime and low on I-Econ) and CD, IR, NZ, SP, and NL (scoring low on I-Crime and high on I-Econ).

We should remind ourselves, however, that the data analyzed are from 1995 and that points of view will most probably have changed since then for at least some of the countries in this study.

3.5. Technical Background of Nonlinear Principal Components Analysis

3.5.1. Indicator Matrices

The nonlinear transformation approach deals with categorical variables in the following way. A categorical variable $\mathbf{h}_m$ defines a binary indicator matrix $\mathbf{G}_m$ with N rows and C_m columns, where C_m denotes the number of categories. Elements h_{im} then define elements $g_{ic(m)}$ as follows:

$$g_{ic(m)} = \begin{cases} 1 \text{ if } h_{im} = c_m \\ 0 \text{ if } h_{im} \neq c_m \end{cases}, \qquad (2)$$

where $c_m = 1, \ldots, C_m$ is the running index indicating a category number in the mth variable. If category quantifications are denoted by the vector $\mathbf{y}_m$ (with C_m elements), then a transformed variable $\mathbf{q}_m$ can be written as $\mathbf{G}_m\mathbf{y}_m$. For instance, in a standard linear model, with M predictor variables in $\mathbf{X}$ and b_m denoting the regression weight for the mth variable, the linear combination of the predictors that correlates maximally with the response $\mathbf{z}$ can be written as $\hat{\mathbf{z}} = \sum_{m=1}^{M} b_m\mathbf{x}_m$. Incorporating the nonlinear scaling of the predictor variables $\varphi_m(\mathbf{x}_m)$, for $m = 1, \ldots, M$ with $\varphi_m(\mathbf{x}_m)$ an admissible nonlinear function of $\mathbf{x}_m$, the optimal linear combination is now written as $\hat{\mathbf{z}} = \sum_m^M b_m\varphi_m(\mathbf{x}_m) = \sum_m^M b_m\mathbf{G}_m\mathbf{y}_m$. By mapping a categorical variable into an indicator matrix, invariance is ensured under the one-to-one nonlinear transformation of the original variable. The idea to replace a categorical variable by an indicator matrix can already be found in Guttman (1941). The term *indicator matrix* was coined by de Leeuw (1968); other names used are *attribute* or *trait matrix* (Lingoes, 1968), *response-pattern table* (Nishisato, 1980), *incidence matrix,* or *dummy variables* (in experimental design).

3.5.2. The Joint Objective Function

In this section, we will describe the objective function that jointly fits the vector model and the centroid model. We suppose that there are M_V variables fitted according to the vector model and M_B variables fitted according to the centroid model; thus, we have $M_V + M_B = M$. We start by defining the following terminology. The $N \times M$ matrix $\mathbf{Q}$ contains the scores for the N objects on M variables. The nature of the individual variables $\mathbf{q}_m$ will be discussed shortly. The $N \times P$ matrix $\mathbf{X}$ contains the coordinates

for the N objects in a P-dimensional representation space, and the matrix $\mathbf{A}$ (of size $M_V \times P$) gives the coordinates in the same space for the endpoints of the vectors that are fitted to the variables in the bilinear (vector) model. Thus, $\mathbf{a}_m$ contains the coordinates for the representation of the mth variable. Consequently, the part of the objective function that minimizes the value of the objective function with respect to the bilinear/vector model can be written as follows:

$$\overline{L}_V(\mathbf{Q}; \mathbf{X}; \mathbf{A}) = M_V^{-1} \sum_{m \in K_V} \|\mathbf{q}_m - \mathbf{X}\mathbf{a}_m\|^2, \quad (3)$$

where K_V denotes the index set that contains the indices of the variables that are fitted with the vector model, and $\|\cdot\|^2$ means taking the sum of squares of the elements. Assuming the data in $\mathbf{q}_m$ to have C_m different values, we can also write

$$\overline{L}_V(\mathbf{y}_V; \mathbf{X}; \mathbf{A}) = M_V^{-1} \sum_{m \in K_V} \|\mathbf{G}_m\mathbf{y}_m - \mathbf{X}\mathbf{a}_m\|^2, \quad (4)$$

where $\mathbf{G}_m$ is an indicator matrix that classifies each of the objects in one and only one category. The optimal category quantifications that will be obtained are contained in the C_m vector $\mathbf{y}_m$, where C_m denotes the number of categories for the mth variable. The vector $\mathbf{y}_V$ collects the quantifications for the M_V different variables and has length $\sum_{m \in K_V} C_m$.

The projection of the object points $\mathbf{X}$ onto the vector $\mathbf{a}_m$ gives the approximation of the nonlinearly scaled (optimally quantified) variable $\mathbf{q}_m = \mathbf{G}_m\mathbf{y}_m$ in P-dimensional Euclidean space. Minimization of the loss function $\overline{L}_V$ for the bilinear/vector model can be shown to be equivalent to the minimization of

$$L_V(\mathbf{y}_V; \mathbf{A}; \mathbf{X}) = M_V^{-1} \sum_{m \in K_V} \|\mathbf{G}_m\mathbf{y}_m\mathbf{a}_m' - \mathbf{X}\|^2 \quad (5)$$

(see Gifi, 1990). Here a P-dimensional matrix $\mathbf{X}$ is being approximated by the inner product $\mathbf{G}_m\mathbf{y}_m\mathbf{a}_m'$, which gives the coordinates of the categories of the mth variable located on a straight line through the origin in the joint P-dimensional space. The major advantage of this reformulation of the objective function is its capacity of capturing the centroid model in the same framework. The latter can simply be written as

$$L_B(\mathbf{Y}_B; \mathbf{X}) = M_B^{-1} \sum_{m \in K_B} \|\mathbf{G}_m\mathbf{Y}_m - \mathbf{X}\|^2, \quad (6)$$

where K_B denotes the index set of the variables for which a centroid model is chosen. The $C_m \times P$ matrix $\mathbf{Y}_m$ contains the coordinates of the categories in the P-dimensional space, and $\mathbf{Y}_B$ collects the quantities for the M_B variables stacked upon each other. The objective function for the centroid model implies that to obtain perfect fit, an object point in $\mathbf{X}$ should coincide with its associated category point in one of the rows of $\mathbf{Y}_m$.

At this point, we can write the joint objective function for CATPCA as a weighted linear combination of the separate losses:

$$L(\mathbf{Y}; \mathbf{A}; \mathbf{X}) = (M_V + M_B)^{-1}[M_V L_V(\mathbf{y}_V; \mathbf{A}; \mathbf{X}) + M_B L_B(\mathbf{Y}_B; \mathbf{X})], \quad (7)$$

where the first part is minimized for variables indexed by m for which a vector representation is chosen, and the second part is minimized for the representation of categorical variables. The optimal $\hat{\mathbf{X}}$ is found as

$$\hat{\mathbf{X}} = M^{-1}\left[\sum_{m \in K_V} \mathbf{G}_m\mathbf{y}_m\mathbf{a}_m' + \sum_{m \in K_B} \mathbf{G}_m\mathbf{Y}_m\right],$$

after which the object scores are orthonormalized as $\hat{\mathbf{X}}'\hat{\mathbf{X}} = N\mathbf{I}$ (thus, they are uncorrelated).

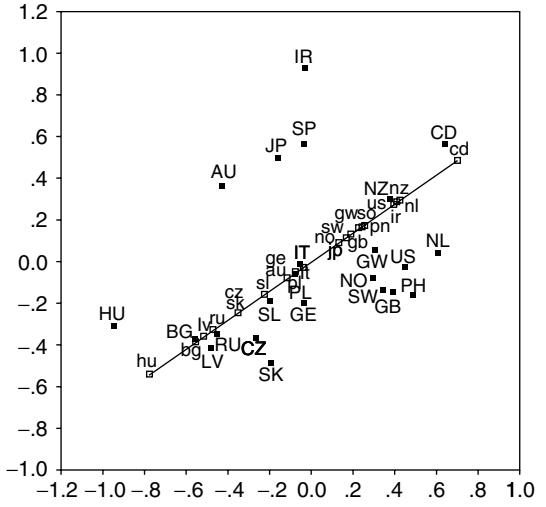
3.5.3. Quantifications and Geometry

In this section, we will describe the iterative process that turns multiple quantifications $\mathbf{Y}_k$ into vector coordinates $\mathbf{y}_m\mathbf{a}_m'$, possibly incorporating ordinal and numerical information from the original variables. Recall that in Figure 3.3, a joint representation was given for centroids (for the categories of the country variable) and for vector coordinates (for the categories of the MO-1 and I-Crime variables). The very same representation can also be given for one and the same variable. This idea is illustrated by including a copy of the supplementary Country variable in the analysis as well and giving this supplementary copy not multiple nominal quantifications but a nominal transformation that positions category points on a vector. The result is illustrated in Figure 3.5, in which the uppercase labels are for the centroids from the previous analysis, and the lowercase labels are for the additional vector coordinates. We see that in the cloud of the country points, the dominant direction is from northeast to southwest, from CD to HU and through Clusters 1 and 3. Computationally, the transition from centroids into vector coordinates involves the following steps.

3.5.3.1. From Centroids to Unordered Vector Coordinates

For each variable, we start with fitting a centroid model according to (6), which gives the minimum over

Figure 3.5 Centroids (Multiple Nominal Quantification) and Vector Coordinates (Nominal Transformation) for Country in CATPCA



NOTE: Countries are identified as follows: IT = Italy; PL = Poland; SL = Slovenia; GE = East Germany; HU = Hungary; BG = Bulgaria; LV = Latvia; RU = Russia; CZ = Czech Republic; SK = Slovak Republic; AU = Austria; JP = Japan; SP = Spain; IR = Ireland; GW = West Germany; NO = Norway; SW = Sweden; GB = Great Britain; PH = Philippines; US = United States; NL = Netherlands; NZ = New Zealand; CD = Canada.

$\mathbf{Y}_m$ as $\mathbf{Y}_m = \mathbf{D}_m^{-1}\mathbf{G}_m'\mathbf{X}$, where $\mathbf{D}_m = \mathbf{G}_m'\mathbf{G}_m$ contains the marginal frequencies of the categories of the mth variable. Next, for the vector model, the centroids $\mathbf{Y}_m$ are projected on a best-fitting line, denoted by $\mathbf{a}_m$, a vector through the origin. The least squares fit that is the minimum of

$$\begin{aligned}&\|\mathbf{G}_m\mathbf{Y}_m - \mathbf{G}_m\mathbf{y}_m\mathbf{a}_m'\|^2\\ &\quad = \operatorname{tr}(\mathbf{Y}_m - \mathbf{y}_m\mathbf{a}_m')'\mathbf{D}_m(\mathbf{Y}_m - \mathbf{y}_m\mathbf{a}_m')\end{aligned} \tag{8}$$

over both $\mathbf{y}_m$ and $\mathbf{a}_m$ determines the category quantifications $\mathbf{y}_m$ and (the orientation of) the vector $\mathbf{a}_m$. The coordinates $\mathbf{y}_m\mathbf{a}_m'$, the outer product of the category quantifications $\mathbf{y}_m$ and the vector $\mathbf{a}_m$, represent the category points on this line, which represents the mth variable in the joint space of objects and variables. The $\mathbf{a}_m$ are also called the component loadings, and they give the correlations between the variables and the dimensions of the principal components space. Setting the partial derivatives in (8) with respect to the component loadings $\mathbf{a}_m$ to zero gives the optimal $\hat{\mathbf{a}}_m$ as

$$\hat{\mathbf{a}}_m = \frac{\mathbf{Y}_m'\mathbf{D}_m\mathbf{y}_m}{(\mathbf{y}_m'\mathbf{D}_m\mathbf{y}_m)}. \tag{9}$$

Next, setting the partial derivatives in (8) with respect to $\mathbf{y}_m$ to zero shows that the optimal unnormalized $\tilde{\mathbf{y}}_m$ is found as

$$\tilde{\mathbf{y}}_m = \frac{\mathbf{Y}_m\mathbf{a}_m}{\mathbf{a}_m'\mathbf{a}_m}. \tag{10}$$

To satisfy the normalization conventions $\mathbf{q}_m'\mathbf{q}_m = N$, the standardized variable $\mathbf{q}_m$ should contain quantifications $\hat{\mathbf{y}}_m$ that are rescaled:

$$\hat{\mathbf{y}}_m = N^{1/2}\tilde{\mathbf{y}}_m(\tilde{\mathbf{y}}_m'\mathbf{D}_m\tilde{\mathbf{y}}_m)^{-1/2}. \tag{11}$$

Note that the length of the vector $\mathbf{a}_m$ has to be diminished to the same extent as the size of the quantifications $\hat{\mathbf{y}}_m$ is increased to keep $\mathbf{y}_m\mathbf{a}_m'$ the same. Equation (10) symbolizes the projection of the centroids $\mathbf{Y}_m$ on the vector $\mathbf{a}_m$ and defines the category coordinates for a nominal transformation. It is very unlikely that the category quantifications in $\mathbf{y}_m$ will be proportional to, or even only in the same order as the original integer scale values $1, \ldots, C_m$. In many cases, however, we would like to maintain the original numeric and/or rank-order information in the transformation, which can be dealt with as follows.

3.5.3.2. From Nominal to Ordinal and Numerical Transformations

If the ordinal, rank-order information should be maintained, an ordinal, monotonic transformation is chosen for variable m, and the quantifications $\mathbf{y}_m$ have to be constrained to be monotonic with the order of the original categories. As described above, this requirement can be satisfied by the use of one of two major classes of monotonic transformations. The first, also historically, is the class of least squares monotonic transformations, obtained by a monotonic regression of the values in $\hat{\mathbf{y}}_m$ upon the original scale values $1, \ldots, C_m$, taking the marginals on the diagonal of D_m into account. The second class is defined by monotonic regression splines. As indicated in Section 3.3.2, transformations by regression splines use fewer parameters than transformations obtained by monotonic regression. For monotonic regression, the number of parameters to be fitted is $C_m - 2$; for regression splines, the number of parameters is determined by the degree of

the spline that is chosen and the number of interior knots. If the number of categories is small, monotonic regression and regression splines will basically give the same result. When the number of categories is large, it is usually advised to use regression splines because monotonic regression may result in overfitting: The variance accounted for will increase, but so will the instability. (Note: There is a trade-off between the number of categories and the number of objects in those categories. If the number of objects is large, and all categories are sufficiently filled, monotonic regression will usually not result in overfitting.)

When it is decided to give the mth variable a numerical transformation, the implication is that the distances between the category points $\mathbf{y}_m \boldsymbol{a}_m'$ have to be equal, and the category quantifications $\mathbf{y}_m$ will be proportional to the original category numbers. This can be done by linear regression of the $\hat{\mathbf{y}}_m$ on the original scale values and will result in a standardized version of the set of the integer scale values $1, \ldots, C_m$, $\mathbf{G}_m \mathbf{y}_m = \alpha_m \mathbf{h}_m + \beta_m$, where the multiplicative constant and the intercept are fitted taking into account the marginal frequencies. If the distances between the categories have to be stretched very much to obtain unit variance, the VAF (expressed in the squared length of the vector a_m) will be very small. It is important to realize that this also applies to ordinary PCA with continuous variables (which can be considered as a CATPCA with N categories, where N is the number of objects, as usual).

3.6. Some Additional Options of the Program CATPCA

3.6.1. External Fitting of Variables

The CATPCA program not only provides an option for the analysis of supplementary variables, as we saw in Section 3.4.4, but for supplementary objects as well. As was true for supplementary variables, supplementary objects are not active in the analysis but enter into the representation afterwards. Another interesting application for the supplementary variables option is the following. CATPCA offers the possibility of reading a fixed configuration of object points, and thus the CATPCA method may be used for so-called property fitting or external unfolding (Carroll & Chang, 1967; Meulman, Heiser, & Carroll, 1986). In this way, external information on objects (contained in so-called external variables) is fitted into the fixed representational space by the use of the vector model (or the centroid model). The option accommodates the same variety of transformation levels as a standard CATPCA analysis (with nominal, ordinal, and numerical treatment of the variables, including the use of splines).

3.6.2. Making Continuous Variables Discrete—Binning

Although the CATPCA algorithm is tuned to the analysis of categorical variables, continuous variables can be introduced into the analysis as well, and this is after they have been made discrete using one of a variety of options provided. This process is comparable to fitting a histogram to a continuous distribution. The grouping options described below can also be used to merge a large initial number of categories into less, which is especially warranted when the distribution of the objects over the original categories is very skew or when some of the categories have very few observations.

3.6.2.1. Grouping in a Specified Number of Categories for a Uniform or Normal Distribution

In Max (1960), optimal discretization points were computed to transform a continuous variable into a categorical one, in which the number of categories can vary from 2 to 36. These discretization points are optimal with respect to an assumed distribution, particularly a univariate standard normal distribution or a univariate uniform distribution. As an illustration, we use the age variable from the National Identity Study: Respondents varied in age from 14 to 98; the modal age category is 30. When this variable is made discrete with seven categories, assuming the population distribution is normal, the following ranges (with corresponding marginal frequencies in parentheses) are obtained: 14–17 (107), 18–30 (2,596), 31–40 (2,335), 41–49 (2,002), 50–59 (1,794), 60–72 (1,916), and 73–98 (699). If, on the other hand, a uniform distribution would be assumed, the following categories and marginal frequencies result: 14–25 (1,653), 26–33 (1,691), 34–39 (1,444), 40–46 (1,657), 47–55 (1,731), 56–65 (1,639), and 66–98 (1,634).

3.6.2.2. Grouping in Equal Intervals With Specified Size

When it is preferred to have a continuous variable replaced by a categorical variable in which the original values are grouped into intervals of equal size, this

is a feasible option as well. Of course, the choice of a specific range for the interval determines the number of categories (bins in a histogram). For the age variable, choosing intervals of 10 years gives the following: 14–23 (1,216), 24–33 (2,128), 34–43 (2,394), 44–53 (2,066), 54–63 (1,669), 64–73 (1,397), 74–83 (493), 84–93 (79), and 94–98 (7). With this option, the groupings for the higher age ranges have rather low marginal frequencies. Comparing this distribution with the two previous ones, we would prefer the uniform option.

3.6.2.3. Ranking

This particular form of preprocessing is appropriate for at least two different situations. In the first place, it should be noted again that the optimal scaling framework guarantees that any ordinal transformation of the original data, among which is replacing numeric values by ranks, will leave the analysis results the same when variables are treated ordinally. When there are no ties in the original variable, the number of categories in the new variable will be N, the number of objects. However, such an ordinal analysis might involve too many parameters to be fitted. When the number of categories approaches the number of objects, it is often a better choice to fit a monotonic spline of a low degree with a limited number of knots. Another use of ranking is to give the resulting rank-order variables a numerical transformation level. In the latter case, the principal components analysis amounts to the analysis of the Spearman rank correlations. If the ranking operation is applied to a variable that contains a unique identification for the objects in the analysis, then the resulting variable, defined as *supplementary,* can be used to identify individual objects in various plots (e.g., in the projected centroids). Of course, this labeling is only feasible and useful when the number of objects is not too large.

3.6.2.4. Multiplying

The distributional properties of a continuous variable that contains noninteger values can be maintained as closely as possible by the particular linear transformation that transforms the real-valued variable into a discrete variable containing integers. The result of this process is a variable that could be treated as numerical; when all the variables in the analysis are treated this way, we are back to classical principal components analysis. However, when one assumes monotonic (instead of linear) relationships between such a variable and other variables in the analysis, it is advised to fit a monotonic spline transformation. When relationships are completely nonlinear, nonmonotonic splines should be fitted to allow these relationships to be revealed in the analysis.

3.6.3. Missing Data

To handle incomplete data in the analysis, a sophisticated option is available that only takes into account the nonmissing data when the loss function is minimized. The indicator matrix for a variable with incomplete data will, in this case, contain rows with only zeros for an object having a missing observation. The loss function in Section 3.5.2 is extended by the use of (internally generated) object weights, collected in a diagonal matrix in which the diagonal elements indicate the number of nonmissing observations for each of the objects. Although this option is very attractive (missing data are completely ignored), it also has a number of drawbacks that need not be severe, however (see Meulman, 1982). Because objects have a different number of observations, the *weighted* mean of the object scores is now equal to 0, and because the mean itself is not 0, various optimality properties of nonlinear PCA are no longer valid. The maximum/minimum value of the component loadings is no longer equal to 1.0 and –1.0, and therefore a component loading can no longer be interpreted as a correlation. (We can still project a transformed variable in the space of the objects, however.) Also, the property that nonlinear PCA optimizes the sum of the P largest eigenvalues of the correlation matrix between the transformed variables is no longer true. (However, when this correlation matrix is computed, there are various choices available for imputing values for the missing data.) Indications on how many data elements can be missing without too much disturbance are given by Nishisato and Ahn (1994).

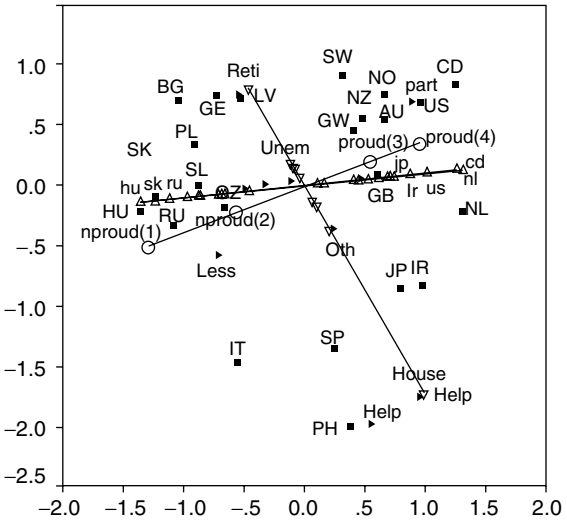
Alternatively, there are other straightforward strategies for treating the missing data in the primary analysis. The first is to exclude objects with missing values; the second provides a very straightforward imputation method, using the value of the modal category. Also, a separate, additional category can be fitted for all objects having a missing value on a particular variable. For all transformation levels, this extra category is positioned optimally with respect to the nonmissing categories. If other, more advanced, missing data strategies are called for (such as the imputation strategy of Van Buuren & Van Rijckevorsel, 1992), these would have to be part of a preprocessing process performed before the actual CATPCA analysis.

3.7. CATPCA IN ACTION, PART 2

Having the CATPCA methodology available gives various interesting possibilities compared to a standard correspondence analysis in which two variables are fitted according to the centroid model (Gifi, 1990). First, consider the nominal variables, country and employment status (Emp-Stat), from the National Identity Study. A standard correspondence analysis would display the category points for both variables in a joint low-dimensional space. An extended correspondence analysis may include the same two multiple nominal variables but with a third ordinal variable included as well. This idea will be illustrated by using the Country and Emp-Stat variables, which are now joined with the Democ variable (also from the National Identity Study). The Democ variable indicates, on a scale from 1 to 4, whether the respondent is *very proud* (4), *somewhat proud* (3), *not very proud* (2), or *not proud at all* (1) with respect to his or her country's democracy. The distribution of the original variable shows that the modal respondent is "somewhat proud" (n =4,140); the smallest category is "very proud" (n = 1,361), followed by the category "not proud at all" (n = 1,606), with "not very proud" the second largest category (n = 3,496). Where does this variable fit into the country × Emp-Stat space? The answer is given in Figure 3.6, the joint plot of the categories for Country, Emp-Stat, and Democ. Moreover, we added in this plot the vector representations for Country and Emp-Stat as well, obtained by including copies of these as supplementary variables to be fitted with the vector model.

The centroid representation for Country and Emp-Stat shows their relationship in terms of their category points. The vector representation for Emp-Stat shows that the two extreme categories on a one-dimensional scale would be "Retired" and "Unemployed" at the north-northwest endpoint and "Housewives" at the end pointing south-southeast. From the vector representation, it is easy to see that the category "House = house wives" scores relatively high in the Philippines (PH), Spain (SP), Ireland (IR), Japan (JP), Italy (IT), and the Netherlands (NL). The categories "Retired" and "Unemployed" score high in East Germany (GE), Bulgaria (BG), and Sweden (SW). The one-dimensional projection of the country category points shows that the major direction goes from west to east. The relationship between Country and Emp-Stat in an ordinary correspondence analysis changes when Democ is taken into account. The ordinal transformation of Democ (not shown) turned out to be close to linear but gives emphasis to the modal category "somewhat proud," which is quantified with a higher value than its numeric counterpart if the original variable containing the scores 1 to 4 had been standardized. The vector for Democ is orthogonal to the direction that connects the categories "Retired" and "Housewives" and is mostly related to the vector representation of Country, contrasting the "very proud of democracy" countries of Canada, the United States, and the Netherlands with the "not proud at all" countries of Italy, Russia, the Slovak Republic, and Hungary.

Figure 3.6 Use of CATPCA: Extended Correspondence Analysis of Country and Em-Stat With Democ (nproud(1) = *not proud at all* to proud(4) = *very proud*) as an Extra Ordinal Variable



NOTE: Countries are identified as follows: IT = Italy; PL = Poland; SL = Slovenia; GE = East Germany; HU = Hungary; BG = Bulgaria; LV = Latvia; RU = Russia; CZ = Czech Republic; SK = Slovak Republic; AU = Austria; JP = Japan; SP = Spain; IR = Ireland; GW = West Germany; NO = Norway; SW = Sweden; GB = Great Britain; PH = Philippines; US = United States; NL = Netherlands; NZ = New Zealand; CD = Canada.

3.8. DISCUSSION

3.8.1. Optimal Scaling and (Multiple) Correspondence Analysis

Although we stated earlier that it is beyond the scope of the present chapter to discuss the technique called multiple correspondence analysis (MCA), we need to mention explicitly the relationship between

principal components with nonlinear optimal scaling transformations and MCA. When the transformation level is chosen to render multiple nominal quantifications for all variables, the two techniques are completely equivalent. So the current CATPCA program, with all its special options for discretization, missing data, supplementary objects and variables, and a variety of plots, can be used to perform MCA as well. In terms of the loss function in Section 3.5.2, we have $M_V = 0$ and $M_B = M$, and we minimize (6) for all variables.

The classic case of simple correspondence analysis concerns two categorical variables, displayed in a cross table, with the categories of the first variable in the rows and the categories of the second variable in the columns. The cells of the table then contain the frequencies of the joint occurrence of category C_A from variable A and category C_B from variable B, and correspondence analysis displays the residuals from independence between the two variables (their interdependence). There are some details that should be taken into account with respect to normalizing the dimensions of the space, but a standard correspondence analysis and a CATPCA are basically equivalent when the two variables are given multiple nominal quantifications. The similarity is largest when the object scores in CATPCA are standardized so that the categories are the average of the object scores, and geometrically the category points will be in the centroid of the object points.

When we have *two* variables, a number of optimal scaling techniques are in fact equivalent. CATPCA with two *nominal* variables, combined with optimization in one dimension, is equivalent to simple regression with optimal scaling, and maximizes the Pearson correlation coefficient over all possible nominal quantifications (Hirschfeld, 1935). When the two variables have a nonlinear relationship, the regression is linearized because categories are allowed to be reordered (permuted), and distances between them are optimally scaled. The term *optimal scaling,* in this context, is due to Bock (1960); also, see Fisher (1940, 1948) for the maximal mutual discrimination principle, as well as the overview in de Leeuw (1990). Applying ordinal (spline) transformations maximizes the correlation coefficient under monotonicity restrictions. When one of the variables is treated as numeric, and the other is given a nominal transformation, the CATPCA technique would be equivalent to linear discriminant analysis but with one single predictor. Obviously, allowing an ordinal transformation instead of the numerical transformation level generalizes the latter technique, maximizing the between to total variation ratio under monotonic transformation of the predictor variable.

3.8.2. Special Applications

In the following subsections, we will briefly discuss some special types of applications of CATPCA. For a selection of concrete applications, sometimes using the precursor program PRINCALS, the user is referred to the following: Arsenault, Tremblay, Boulerice, and Saucier (2002); Beishuizen, Van Putten, and Van Mulken (1997); de Haas, Algera, Van Tuijl, and Meulman (2000); de Schipper, Tavecchio, Van IJzendoorn, and Linting (2003); Eurelings-Bontekoe, Duijsens, and Verschuur (1996); Hopman-Rock, Tak, and Staats (2001); Huyse et al. (2000); Theunissen et al. (2003); Vlek and Stallen (1981); Zeijl, te Poel, du Bois-Reymond, Ravesloot, and Meulman (2000); and Van der Ham, Meulman, Van Strien, and Van Engeland (1997).

3.8.2.1. Preferential Choice Data

In preferential choice data, respondents give a ranking of objects (sometimes called *stimuli*) according to some attribute, giving an explicit comparison. Consumers, for example, can be asked to rank a set of product brands, or psychologists may be asked to rank a number of psychology journals (see Gifi, 1990, pp. 183–187). Such rankings are usually collected in a data matrix, with the stimuli, options, or objects in the rows and the persons (judges) in the columns acting as the variables of the analysis. This situation was actually the very same one in which Tucker's (1960) vector model was applied in Carroll (1972) to preference data. In the latter mentioned application, the analysis was metric because no optimal scaling of the rankings was possible. Because rankings are ordinal by definition, optimal scaling by monotonic (spline) transformations appears most appropriate.

3.8.2.2. Q-Sort and Free-Sort Data

Another situation for which the persons act as variables is in the so-called analysis of Q-sort data. Here, a number of judges have to group N given objects in a predetermined number of piles (categories), in which the categories have a particular order and the frequencies have to follow a normal distribution as closely as possible. Again, this is a very natural situation for a CATPCA analysis with ordinal transformations. When the M judges are merely given a set of objects and have the liberty to group them

in as many categories as they like, without any given order of the categories, we use the term *free-sort* data. Nominal quantification options are called for in this case, either in the form of nominal (nonmonotonic spline) transformations, when the judges seem to group on one (unknown) dimension, or in the form of multiple nominal quantifications, when judges use more than one latent dimension and when different orderings of the categories are allowed for each dimension. (Nominal or nonmonotonic spline transformations will give the same reordering in each dimension.) Examples of multiple nominal quantifications in free-sort data can be found, among others, in Van der Kloot and Van Herk (1991) and Meulman (1996). In the latter paper, groupings were analyzed in the form of a free-sort of statements about the so-called rape myth.

3.8.2.3. The Analysis of Ratings Scales and Test Items

The application of CATPCA in one dimension is extremely useful because it explores the homogeneity between a set of variables that are assumed to measure the same property (latent characteristic). Optimal scaling minimizes the heterogeneity and maximizes the largest eigenvalue of the correlation matrix. For an extensive treatment of this particular application with its relationship to differential weighting of variables and classical psychometrics, see Heiser and Meulman (1994).

3.8.3. CATPCA and the Correlation Matrix Between the Transformed Variables

In ordinary PCA, the results in a two-dimensional solution are identical to those in the first two dimensions of a three-dimensional solution. This property is called *nestedness.* When quantifications have been chosen to be optimal in one dimension, the largest eigenvalue of the correlation matrix is maximized. When they are optimal for P dimensions, the sum of the first P eigenvalues is optimized. The latter does imply that the first eigenvalue, by itself, does not need to be as large as possible, and because this is true by definition for the one-dimensional solution, it implies that CATPCA solutions with different dimensionalities are not necessarily nested. Inspection of the eigenvalues of the transformed correlation matrix shows the distribution of the total sum of the eigenvalues (which is equal to M, the number of variables) over the optimized and nonoptimized dimensions. When the CATPCA includes variables with multiple nominal quantifications and a more-dimensional solution is obtained, the situation is somewhat more complicated. The first CATPCA dimension optimizes the largest eigenvalue between the transformed variables, including the first set of the multiple nominal quantifications, whereas the second dimension optimizes the largest eigenvalue of the same correlation matrix, but now including the second set of the multiple nominal quantifications. Therefore, if the primary objective is to maximize the homogeneity, either in one dimension for all variables together or in two dimensions, when the variables seem to form two groups (as in our example in Section 3.4.3), unordered variables should be given a nominal (or nonmonotonic spline) transformation.

3.8.4. Prospects

Because unordered or ordered categorical variables are so common in the behavioral sciences, the prospects for nonlinear principal components analysis seem to be good, especially in contexts where a relatively large number of responses have been collected and their mutual relationships have to be sorted out, as in survey research. Another clear application area for CATPCA is instrument development, where it can supplement the usual factor analysis and Cronbach's α calculations for item selection. Because CATPCA directly analyzes the data matrix and not the derived correlation matrix, there need not be the usual concern to have at least 15 times as many observations as the number of variables. In fact, CATPCA is eminently suited for analyses in which there are (many) more variables than objects.

Finally, we would like to mention that there is similar optimal scaling software in the SPSS Categories module for related multivariate analysis techniques. Among these are CATREG for (multiple) regression analysis with optimal scaling, CORRESPONDENCE for correspondence analysis, and OVERALS for nonlinear canonical correlation analysis (Meulman et al., 1999). Like CATPCA, these methods allow one to pursue classic objectives of multivariate analysis when the data do not satisfy the classic quantitative measurement requirements but are qualitative.

REFERENCES

Arsenault, L., Tremblay, R. E., Boulerice, B., & Saucier, J. F. (2002). Obstetrical complications and violent delinquency: Testing two developmental pathways. *Child Development, 73,* 496–508.

Beishuizen, M., Van Putten, C. M., & Van Mulken, F. (1997). Mental arithmetic and strategy use with indirect number problems up to hundred. *Learning and Instruction, 7,* 87–106.

Benzécri, J.-P. (1973). *L'analyse des Données, Tome II, L'analyse des Correspondances* (Data analysis: Vol. 2. Correspondence analysis). Paris: Dunod.

Benzécri, J.-P. (1992). *Correspondence analysis handbook.* New York: Marcel Dekker.

Bock, R. D. (1960). *Methods and applications of optimal scaling* (Report 25). Chapel Hill: L. L. Thurstone Lab, University of North Carolina.

Breiman, L., & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association, 80,* 580–598.

Buja, A. (1990). Remarks on functional canonical variates, alternating least squares methods and ACE. *Annals of Statistics, 18,* 1032–1069.

Burt, C. (1950). The factorial analysis of qualitative data. *British Journal of Psychology, 3,* 166–185.

Carroll, J. D. (1968). Generalization of canonical correlation analysis to three or more sets of variables. *Proceedings of the 76th Annual Convention of the American Psychological Association, 3,* 227–228.

Carroll, J. D. (1972). Individual differences and multidimensional scaling. In R. N. Shepard, A. K. Romney, & S. B. Nerlove (Eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences* (Vol. 1, pp. 105–155). New York: Seminar Press.

Carroll, J. D., & Chang, J. J. (1967, April). *Relating preferences data to multidimensional scaling solutions via a generalization of Coomb's unfolding model.* Paper presented at the annual meeting of the Psychometric Society, Madison, WI.

de Haas, M., Algera, J. A., Van Tuijl, H. F. J. M., & Meulman, J. J. (2000). Macro and micro goal setting: In search of coherence. *Applied Psychology, 49,* 579–595.

de Leeuw, J. (1968). *Canonical discriminant analysis of relational data* (Research Report RN-007–68). Leiden, The Netherlands: University of Leiden.

de Leeuw, J. (1973). *Canonical analysis of categorical data.* Unpublished doctoral dissertation, University of Leiden, Leiden, The Netherlands. (Reissued in 1986 by DSWO Press, Leiden, The Netherlands.)

de Leeuw, J. (1990). Multivariate analysis with optimal scaling. In S. Das Gupta & J. Sethuraman (Eds.), *Progress in multivariate analysis.* Calcutta: Indian Statistical Institute.

de Schipper, J. C., Tavecchio, L. W. C., Van IJzendoorn, M. H., & Linting, M. (2003). The relation of flexible child care to quality of center day care and children's socio-emotional functioning: A survey and observational study. *Infant Behavior & Development, 26,* 300–325.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, 1,* 211–218.

Eurelings-Bontekoe, E. H. M., Duijsens, I. J., & Verschuur, M. J. (1996). Prevalence of DSM-III-R and ICD-10 personality disorders among military conscripts suffering from homesickness. *Personality and Individual Differences, 21,* 431–440.

Fisher, R. A. (1940). The precision of discriminant functions. *Annals of Eugenics, 10,* 422–429.

Fisher, R. A. (1948). *Statistical methods for research workers* (10th ed.). Edinburgh, UK: Oliver & Boyd.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal components analysis. *Biometrika, 58,* 453–467.

Gifi, A. (1990). *Nonlinear multivariate analysis.* Chichester, UK: John Wiley. (Original work published 1981)

Gilula, Z., & Haberman, S. J. (1988). The analysis of multivariate contingency tables by restricted canonical and restricted association models. *Journal of the American Statistical Association, 83,* 760–771.

Gower, J. C. (1966). Some distance properties of latent roots and vector methods used in multivariate analysis. *Biometrika, 53,* 325–338.

Gower, J. C., & Hand, D. J. (1996). *Biplots.* London: Chapman & Hall.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis.* London: Academic Press.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst et al. (Eds.), *The prediction of personal adjustment* (pp. 319–348). New York: Social Science Research Council.

Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika, 33,* 469–506.

Hastie, T., Tibshirani, R., & Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association, 89,* 1255–1270.

Hayashi, C. (1952). On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics, 2,* 93–96.

Heiser, W. J. (1981). *Unfolding analysis of proximity data.* Unpublished doctoral dissertation, University of Leiden, Leiden, The Netherlands.

Heiser, W. J., & de Leeuw, J. (1981). Multidimensional mapping of preference data. *Mathématiques et Sciences Humaines, 19,* 39–96.

Heiser, W. J., & Meulman, J. J. (1983). Analyzing rectangular tables by joint and constrained multidimensional scaling. *Journal of Econometrics, 22,* 139–167.

Heiser, W. J., & Meulman, J. J. (1994). Homogeneity analysis: Exploring the distribution of variables and their nonlinear relationships. In M. Greenacre & J. Blasius (Eds.), *Correspondence analysis in the social sciences: Recent developments and applications* (pp. 179–209). New York: Academic Press.

Hirschfeld, H. O. (1935). A connection between correlation and contingency. *Proceedings of the Cambridge Philosophical Society, 31,* 520–524.

Hopman-Rock, M., Tak, E. C. P. M., & Staats, P. G. M. (2001). Development and validation of the Observation List for early signs of Dementia (OLD). *International Journal of Geriatric Psychiatry, 16,* 406–414.

Horst, P. (1961a). Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology, 17,* 331–347.

Horst, P. (1961b). Relations among *m* sets of variables. *Psychometrika, 26,* 129–149.

Huyse, F. J., Herzog, T., Lobo, A., Malt, U. F., Opmeer, B. C., Stein, B., et al. (2000). European consultation-liaison

psychiatric services: The ECLN Collaborative Study. *Acta Psychiatrica Scandinavica, 101,* 360–366.

International Social Survey Programme (ISSP). (1995). *National identity study.* Cologne, Germany: Zentralarchiv für Empirische Sozialforschung.

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika, 58,* 433–460.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29,* 1–28.

Kruskal, J. B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society, Series B, 27,* 251–263.

Kruskal, J. B. (1978). Factor analysis and principal components analysis: Bilinear methods. In W. H. Kruskal & J. M. Tanur (Eds.), *International encyclopedia of statistics* (pp. 307–330). New York: Free Press.

Kruskal, J. B., & Shepard, R. N. (1974). A nonmetric variety of linear factor analysis. *Psychometrika, 39,* 123–157.

Krzanowski, W. J., & Marriott, F. H. C. (1994). *Multivariate analysis: Part I. Distributions, ordination and inference.* London: Edward Arnold.

Lebart, L., Morineau, A., & Warwick, K. M. (1984). *Multivariate descriptive statistical analysis.* New York: John Wiley.

Lebart, L., & Tabard, N. (1973). *Recherche sur la Description Automatique des Données Socio-Economiques* (Research on the automatic description of socioeconomic data). Paris: CORDES-CREDOC.

Lingoes, J. P. (1968). The multivariate analysis of qualitative data. *Multivariate Behavioral Research, 3,* 61–94.

Lord, F. M. (1958). Some relation between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika, 23,* 291–296.

Masson, M. (1974). Analyse non-linéaire des données (Nonlinear data analysis). *Comptes Rendues de l'Academie des Sciences (Paris), 287,* 803–806.

Max, J. (1960). Quantizing for minimum distortion. *Proceedings IEEE (Information Theory), 6,* 7–12.

Meulman, J. J. (1982). *Homogeneity analysis of incomplete data.* Leiden, The Netherlands: DSWO Press.

Meulman, J. J. (1986). *A distance approach to nonlinear multivariate analysis.* Leiden, The Netherlands: DSWO Press.

Meulman, J. J. (1992). The integration of multidimensional scaling and multivariate analysis with optimal transformations of the variables. *Psychometrika, 57,* 539–565.

Meulman, J. J. (1996). Fitting a distance model to homogeneous subsets of variables: Points of view analysis of categorical data. *Journal of Classification, 13,* 249–266.

Meulman, J. J., Heiser, W. J., & Carroll, J. D. (1986). *PREFMAP-3 users' guide.* Murray Hill, NJ: AT&T Bell Laboratories.

Meulman, J. J., Heiser, W. J., & SPSS. (1999). *SPSS Categories 10.0.* Chicago: SPSS.

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications.* Toronto: University of Toronto Press.

Nishisato, S. (1984). Forced classification: A simple application of a quantification method. *Psychometrika, 49,* 25–36.

Nishisato, S. (1994). *Elements of dual scaling: An introduction to practical data analysis.* Hillsdale, NJ: Lawrence Erlbaum.

Nishisato, S., & Ahn, H. (1994). When not to analyse data: Decision making on missing responses in dual scaling. *Annals of Operations Research, 55,* 361–378.

Ramsay, J. O. (1982). Some statistical approaches to multidimensional scaling data. *Journal of the Royal Statistical Society, Series A, 145,* 285–312.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science, 3*(4), 425–461.

Roskam, E. E. C. I. (1968). *Metric analysis of ordinal data in psychology.* Voorschoten: VAM.

Saporta, G. (1975). *Liaisons entre Plusieurs Ensembles de Variables et Codage de Données Qualitatives* (Connections between several sets of variables and coding of qualitative data). Unpublished doctoral dissertation, Université Paris VI, Paris.

Schmidt, E. (1907). Zur Theorie der linearen und nichtlinearen Integralgleichungen (On the theory of the linear and nonlinear integral equations). *Mathematische Annalen, 63,* 433–476.

Shepard, R. N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function: I. *Psychometrika, 27,* 125–140.

Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function: II. *Psychometrika, 27,* 219–246.

Shepard, R. N. (1966). Metric structures in ordinal data. *Journal of Mathematical Psychology, 3,* 287–315.

Stewart, G. W. (1993). On the early history of the singular value decomposition. *SIAM Review, 35,* 551–566.

Tenenhaus, M., & Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data. *Psychometrika, 50,* 91–119.

Theunissen, N. C. M., Meulman, J. J., Den Ouden, A. L., Koopman, H. M., Verrips, G. H., Verloove-Vanhorick, S. P., et al. (2003). Changes can be studied when the measurement instrument is different at different time points. *Health Services and Outcomes Research Methodology, 4* (2).

Torgerson, W. S. (1958). *Theory and methods of scaling.* New York: John Wiley.

Tucker, L. R. (1960). Intra-individual and inter-individual multidimensionality. In H. Gulliksen & S. Messick (Eds.), *Psychological scaling: Theory and applications* (pp. 155–167). New York: John Wiley.

Van Buuren, S., & Van Rijckevorsel, L. A. (1992). Imputation of missing categorical data by maximizing internal consistency. *Psychometrika, 57,* 567–580.

Van der Burg, E., & de Leeuw, J. (1983). Non-linear canonical correlation. *British Journal of Mathematical and Statistical Psychology, 36,* 54–80.

Van der Burg, E., de Leeuw, J., & Verdegaal, R. (1988). Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling features. *Psychometrika, 53,* 177–197.

Van der Ham, T., Meulman, J. J., Van Strien, D. C., & Van Engeland, H. (1997). Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. *British Journal of Psychiatry, 170,* 363–368.

Van der Kloot, W. A., & Van Herk, H. (1991). Multidimensional scaling of sorting data: A comparison of three procedures. *Multivariate Behavioral Research, 26,* 563–581.

Vlek, C., & Stallen, P. J. (1981). Judging risks and benefits in the small and in the large. *Organizational Behavior and Human Performance, 28,* 235–271.

Winsberg, S., & Ramsay, J. O. (1980). Monotonic transformations to additivity using splines. *Biometrika, 67,* 669–674.

Winsberg, S., & Ramsay, J. O. (1983). Monotone spline transformations for dimension reduction. *Psychometrika, 48,* 575–595.

Young, F. W. (1981). Quantitative analysis of qualitative data. *Psychometrika, 46,* 357–387.

Young, F. W., de Leeuw, J., & Takane, Y. (1976). Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika, 41,* 505–528.

Young, F. W., Takane, Y., & de Leeuw, J. (1978). The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika, 43,* 279–281.

Zeijl, E., te Poel, Y., du Bois-Reymond, M., Ravesloot, J., & Meulman, J. J. (2000). The role of parents and peers in the leisure activities of young adolescents. *Journal of Leisure Research, 32,* 281–302.

Section II

Testing and Measurement

Chapter 4

Responsible Modeling of Measurement Data for Appropriate Inferences

Important Advances in Reliability and Validity Theory

Bruno D. Zumbo

André A. Rupp

4.1. Introduction

If the statistical, conceptual, and practical activities of measurement were a crop seeded by Spearman, Yule, Pearson, and others working the early fields of social and behavioral research, we could proudly say that those seedlings have resulted in a bountiful harvest. The annual yield of measurement research continues to grow, and the number of new journals and books devoted to and surveying the field and reporting advances has increased over the past decade. The goal of indexing data quality has a longstanding tradition in statistical modeling, and its ubiquity in psychometric modeling thus comes as no surprise, which is why research in reliability and validity theory continues to be of relevance today, as a quick glance at the reference list of this chapter reveals. Before we begin to describe the process of harvesting the statistical crops that have been sown, however, let us first take a look at the analyst's task in measurement itself.

Analysts of test data are typically faced solely with an array of numbers, which often consists of 0s and 1s when all items on a test are scored dichotomously. It is the objective of the analyst to use this array for a variety of meaningful inferences about the examinees and the measurement instrument itself, which should be appreciated as a daunting task. Statistical modeling has always been concerned with decomposing observational values into a component that is *deterministic* and a component that is *stochastic* so that relationships between manifest and unobserved variables can be explicitly stated and uncertainty about model parameters can be estimated and used to qualify the inferences that are possible under a given model. Psychometric models are, of course, descendants of this tradition (see Goldstein & Wood, 1989;

McDonald, 1982; Mellenbergh, 1994; Rupp, 2002) but are unique because they are located at the *intersection* of examinee and item spaces, *both* of which are typically of interest to measurement specialists. For example, classical test theory (CTT) (e.g., Lord & Novick, 1968) generically decomposes the observed score into a deterministic part (i.e., true score) and a stochastic part (i.e., error), generalizability theory (*g*-theory) (e.g., Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) further unpacks the stochastic part and redefines part of error as systematic components, and item response theory (IRT) (e.g., Lord, 1980; van der Linden & Hambleton, 1997) reformulates the two model components by inducing latent variables into the data structure. Structural equation models (SEM) (e.g., Muthén, 2002) and exploratory as well as confirmatory factor analysis models (EFA and CFA, respectively) (e.g., McDonald, 1999) decompose the covariance matrix of multivariate data into deterministic (i.e., reproduced covariance matrix) and stochastic (i.e., residual matrix) components, which is a model that can be equivalently written as a formulization involving latent variables.

Even though latent trait indicator values and observed composite scores are typically highly correlated, the injection of a latent continuum into the data matrix has given us the property of item and examinee parameter invariance for perfect model fit across populations and conditions, has allowed us to define conditional standard errors of measurement similar to *g*-theory (Brennan, 1998b), and has opened up the road for adaptive testing through the use of item and test information functions (e.g., Segall, 1996; van der Linden & Hambleton, 1997). Still, these advances have not come without a price. Improvements in the level of modeling and in quantifying measurement error have come at the expense of large sample sizes that are typically required for parameter estimation in both frequentist and Bayesian frameworks (see Rupp, Dey, & Zumbo, in press). For example, categorical data, particularly dichotomous data, require the use of estimation methods such as weighted least squares, which make, for example, reliability estimates based on small sample sizes suspect (Raykov, 1997a).

The focus in this chapter is on reliability and validity, two topics that have generated many papers and books, even if one were to focus on the past 25 years only. As it is nearly impossible to review all of the developments in a single book chapter, we aim to provide a broad overview of recent developments in reliability and validity theory and periodically provide more detail to demonstrate the vast array of measurement methodologies and approaches currently available to aid us in illuminating our understanding of social and behavioral phenomena. We will view these developments through a statistical modeling lens to highlight the consequences of choosing—perhaps even abusing—a particular modeling framework for inferential decisions.

We assume a basic exposure to measurement and test theory, but we will define basic key terms. For an accessible overview and advances in the statistical basis of reliability theory, the interested reader can consult Feldt and Brennan (1989), Knapp (2001), and Traub (1994), and for validity theory and practice, the reader can consult Messick (1995) and the papers in Zumbo (1998). Because our chapter presumes a working knowledge of modeling frameworks used in practical measurement problems, the reader might refer to Hambleton, Swaminathan, and Rogers (1991) or Lord (1980) as useful references for IRT, Kaplan (2000) or Byrne (1998) as useful references for structural equation modeling, and Comrey (1973), Everitt (1984), or McDonald (1999) as useful references for factor analysis (FA) methods.

Our discussion begins with an overview of frequently used key terms in the measurement literature to aid the understanding of our subsequent discussions, clarify some common misconceptions, and allow for more precise statements. We then present some important and practically relevant findings from the literature on reliability theory in roughly the past decade, with a strong focus on developments for reliability coefficients, standard errors of measurement, and other local quantifiers of measurement error. Finally, a section on validity theory illustrates how models for cognitively diagnostic assessment have forced measurement specialists to rethink their approaches to defining and measuring what constitutes valid inferences from test scores. But first, let us lay some groundwork with a brief discussion of terminology relevant for modeling data from measures.

4.2. Commonly Used and Misunderstood Terms in Measurement

Although the definitions presented in this section are fundamental, it is remarkable how often they are used inconsistently in the measurement literature. This is probably partly an artifact of inconsistent historical usage but can also be traced back to a discrepancy that typically exists between the everyday usage of these terms and their precise meaning in a mathematical modeling context.

First, there is the word *reliability* itself. In nonacademic contexts, *reliable* is commonly understood to mean "a consistent dependability of judgment, character, performance, or result" (see Braham, 1996, p. 1628). For applied measurement specialists, reliability is a desired property of tests, which should be dependable measurement instruments of the constructs that they are supposed to measure or dependable measurement instruments for performance evaluation (Klieme & Baumert, 2001). Even though these notions are intuitively appealing, they are relatively imprecise and need to be translated into properties that can be mathematically tested and estimated through sample quantities. Consequently, reliability in a nonmathematical sense is often understood to be so much more than reliability in a strictly mathematical sense because, under the latter lens, reliability is basically translated into the estimation of a coefficient based on variance components in a statistical model. Such a *reliability coefficient* assesses consistent *scores* but, per se, says little about the assessment instrument itself, related inferences, and social consequences because those aspects are embedded in the larger value-laden ethical and social context of test use (Messick, 1995).

As Zimmerman and Zumbo (2001) note, formally, test data are the realization of a stochastic event defined on a product space $\Omega = \Omega_I \times \Omega_J$, where the orthogonal components, Ω_I and Ω_J, are the probability spaces for items and examinees, respectively. The joint product space can be expanded to include other spaces as well, such as spaces induced by raters or occasions, a concept that was formalized in g-theory from an observed-score perspective and the facets approach to measurement from an IRT perspective. Hence, modeling of test data minimally requires sampling assumptions about items and examinees, as well as the specification of a stochastic process that is supposed to have generated the data (for readers interested in a measure-theoretic Hilbert-space approach to the analysis of test data, see Zimmerman & Zumbo, 2001). Therefore, two distinct directions of generalizability are typically of interest, which require an understanding of the reliability and validity properties of scores and inferences. First, it is of interest to make statements about the functioning of a particular assessment instrument for groups of examinees who share characteristics with those examinees who have already been scored with it. Second, it is of interest to make statements about the functioning of item sets that share characteristics with those items that are already included on a particular test form. For example, it is often of interest to show that the scores and resulting inferences for different examinee groups are comparably reliable and valid if the same instrument is administered to the different groups, a parallel version of the instrument is administered to the different groups, or selected subsets of items are administered to the different groups. This also specifically implies that researchers should report estimates of reliability coefficients and other parameters for their own data, rather than relying on published reports from other data, and that comparable validity needs to be continually assessed rather than being taken for granted based on a single assessment calibration. Let us take a look at some commonly used terms to describe the process of modeling assessment data.

It is useful to first distinguish between *test-level models* (e.g., CTT, g-theory models), in which modeling takes place at the observed total-score level, and *item-level models* (e.g., IRT for binary or rating scale item data and factor analysis models for continuous item data), in which modeling takes place at the item-score level along with the total-score level. For the latter models, the *primary* modeling unit is the *item,* which can be a written, aural, or graphical stimulus that entices examinees to produce behavioral responses. Yet the seemingly unambiguous notion of an item is rather fluid and context dependent. For example, items can be collected, either naturally through their placement alongside reference information on an assessment or statistically through definition, into item bundles or *testlets,* which can then be treated as a single item in subsequent mathematical analyses (note that potential response dependencies can be modeled explicitly as well; see Bradlow, Wainer, & Wang, 1999; Wang, Bradlow, & Wainer, 2002). Moreover, in other testing contexts with complex work products the definition of a single item can become extremely challenging if not impossible, and it might be preferable and necessary in the future to think of *measurement opportunities* more generally instead. For a recent description of the variety of items currently being used in measurement practice, see Zenisky and Sireci (2002).

Items can be assembled for different purposes such as personality trait assessment or knowledge assessment, and it is the latter scenario that typically leads to instruments that are commonly called *tests.* In addition, the term *scale* is also often used in the social science literature on personality assessment interchangeably with the term *questionnaire.* The terms *test, scale,* and *measure* are used interchangeably in this chapter, but it is acknowledged that *tests* are, in common language, used to imply some educational achievement or knowledge test with correct or incorrect responses.

A subject's response to an item then becomes a *behavioral observation* in an abstract conceptual sense that needs to be quantified with a *score,* which, in turn, becomes a *statistical observation.* Some typical forms of scores are the (weighted) linear composite or *total score* that arises from individual items being scored *dichotomously* or *polytomously.* Measurement specialists then resort to specific *modeling frameworks* to account for the fact that behavioral observations are imperfect representations of the *latent variable* whose relative absence or presence the assessment instrument is supposed to quantify and, as such, contain *measurement error.* Indeed, the choice of measurement model has fundamental implications for how measurement error is viewed, and these differences lead modelers to choose particular model-specific statistics to quantify this error. Error, then, albeit a universally present phenomenon of observed behavioral responses, is conceived and quantified differently in alternate micro-universes created by different modeling frameworks. Interestingly, the well-known psychometric statement $X = T + E$ is axiomatic for all models in such frameworks.

In any modeling framework, the observable or *manifest* scores created by the interaction of examinees with items on a measure are considered to be *indicators* or markers of unobservable or *latent* variables. In this chapter, we will use the term *latent variable* to refer to a random variable that is deliberately constructed or derived from the responses to a set of items and that constitutes the building block of a statistical model (e.g., θ scores in IRT or factor scores in FA). In other words, the scores are indicators of the latent variable, which is itself supposed to be an indicator of an underlying *latent trait* that is inherent in the examinees and supposedly tapped into by the items. However, these quantities and objects are not identical: The latent variable is a *psychometric* construction, whereas the latent trait is a *psychological* phenomenon. Put in a nutshell, a *construct* is defined with reference to a *nomological network* of other phenomena, empirical findings, and theories linking latent variables to abstract constructs (Embretson, 1983; Messick, 1995), whereas a latent variable is a mathematical construction. This often leads to confusion for applied specialists when psychometric dimensionalities of tests do not coincide with believed psychological dimensionalities, although this apparent discrepancy is perfectly expected if the precise distinction above is made.

Let us illustrate this distinction with an example. The Center for Epidemiological Studies–Depression (CES-D) is a 20-item scale introduced originally by Lenore S. Radloff to measure depressive symptoms in the general population. If we were studying the measurement properties of the CES-D via CFA models or IRT, the items would be considered indicators of a latent variable (which most researchers would call "depression"), but the latent variable is *not* depression itself as it is merely a mathematical construction. The latent variable is *related* to the construct of depression, however, which is defined as per the complex interrelations of ideas, definitions, and empirical findings in the clinical literature. Likewise, if one were empirically scoring the CES-D by summing the item responses, the resultant composite scale score is *not* depression itself either but again only *related* to that construct as an observable indicator of it. Even more precisely, the score is an indicator of the *severity* of depressive symptoms.

It should be noted, however, that the measurement literature is generally somewhat vague and inconsistent in its use of the term *latent variable.* The term has a number of different meanings in the measurement and statistics literature, each of which can lead to quite different variables. There are at least three common uses of relevance to this chapter. The first definition, which is the closest description of a (unobserved) latent variable in classical test theory, is that latent variables are real variables that could, in principle, be measured (e.g., proficiency or knowledge in a domain, such as mathematics, or level of depressive symptomatology). A second form of a latent variable is when observed scores arise by recording whether an underlying variable had values above or below fixed thresholds (e.g., a response to a Likert-type question). The former definition can be conceptualized within a framework of the latter definition, although it does not necessarily have to be so. The third definition, which is the most commonly used meaning in the social sciences, describes a latent variable as a constructed variable that comes prior to the items (or indicators) of which we measure. With item responses at hand and the use of a statistical model, one can predict a score on this latent variable for each person in the sample. This third meaning is most commonly used in factor analysis, latent variable modeling, and covariance structure models and is therefore the one used in this chapter. In terms of the psychometric approach to factor analysis, a latent variable is a reason for or summary of behavioral or cognitive manifestations. In the statistical framework, a latent variable is defined by local or conditional independence (statistical entity with no real theoretical purpose). Statistically, it is assumed that if two variables are correlated, they have something unobserved in common (i.e., the latent variable). Therefore,

uncorrelated errors (i.e., the residual correlation among the items over and above the factors) are a key defining feature of latent variable models.

Finally, it is useful to differentiate between observed-score and latent variable models. When an observed composite score is decomposed into two independent additive components, true score and error, without any further assumptions about the structure of the true score, researchers have termed this *CTT.* At the same time, different sets of assumptions about the error structure and true scores for repeated assessments and different sampling schemes for items and examinees have led to the definition of *parallel, essentially parallel, τ-equivalent, essentially τ-equivalent,* and *congeneric* test scores. Moreover, if no particular statistical model is assumed for the responses, models in CTT are typically referred to as *weak true-score models,* and if a statistical model is assumed (e.g., binomial, compound binomial), they are referred to as *strong true-score models.* If the relationship of the observed score to the true-score and error components is of a specific functional form that depends on at least one latent variable and can be formulated in a *generalized linear* (*latent variable*) *model framework,* we typically speak of latent variable models. Latent variables belong to the class of unobservable random variables, but they are a specific subset because their existence is *postulated,* and their *metric* is established through the specification of the model and the parameter estimation strategy. If response data are modeled at the item level, measurement specialists refer to these models as IRT models, which have become increasingly popular in the past two decades due to increasing computer power and their flexible mathematical formulation. It is interesting to note that there is no substantive theory in IRT but that, generally, the model *is* the theory, which, some argue, makes the rational link between the latent variable and the underlying construct it potentially indexes harder to establish as one can alternatively conceive of a latent variable as a mere data-processing filter that allows for ordered inferences about examinees and items (see Junker, 1999). In general, observed and latent variable frameworks benefit from one another and are compatible as, for example, methods of covariance structure analysis that are well suited to test assumptions about error structures associated with CTT.

At this point, it is important to take a small sidebar to highlight an essential difference between factor analysis (as it is commonly used) and IRT in item calibration. Although FA and IRT can be written as generalized linear latent variable models, the statistical estimation problem is compounded in IRT because the item responses are binary or ordered polytomous random variables, and the estimation strategy necessitates the estimation of the latent variable score for each individual in order to estimate the parameters of the item response function (i.e., calibrate the items). This is in stark contrast to most factor analysis models, wherein the latent variable is integrated out of the estimation equation by, in essence, marginalizing over the latent variable (i.e., focusing on reproducing the observed covariance matrix).

Once items have been calibrated, examinees have been scored, and quantifiers of measurement error have been computed, inferences are being made grounded in the mathematical model that was used. Ideally, those inferences ought to be accurate and result in *fair conferences* for the examinees and the assessment discipline. Investigations of the *degree* to which scores are consistent across administration conditions fall under the umbrella term of *reliability theory,* whereas investigations of the *degree* to which inferences made from test scores and the consequences of decisions based thereon are appropriate fall under the umbrella term of *validity theory.* Specifically, reliability is a question of *data quality,* whereas validity is a question of *inferential quality.* Of course, reliability and validity theory are interconnected research arenas, and quantities derived in the former bound or limit the inferences in the latter. This is seen explicitly in CTT statistics, for example, where it can be easily shown that a validity correlation coefficient is never greater than the square root of the test reliability coefficient. Moreover, to increase both reliability of scores and validity of inferences, a surge in models for *cognitively diagnostic assessment* has forced measurement specialists to refocus their attention on the *cognitive processes* that examinees are engaged in when responding to items. This has led to a renewed dissection of what forms of *evidence* support valid inferences and has brought the focus of investigations back to the examinees.

The title of this chapter was chosen to highlight that, when dealing with matters of reliability and validity, we are, in essence, dealing with matters of making inferences from test or scale scores. In other words, data on reliability and validity gathered in the process of measurement aid social and behavioral researchers in judging the *appropriateness* and *limitations* of their inferences from the test or scale scores. In the next section, we provide an overview of reliability theory and the statistical properties of test and scale scores. In the section that follows, we provide an overview of validity theory and then end the chapter with some pointers to future developments.

4.3. A Unified Look at Reliability and Error of Measurement as a Basis for Valid Inferences

Quantifying measurement error can take different forms, depending on the scoring framework that is used for modeling the data. Traditionally, CTT has been used predominantly by test developers as well as applied specialists. In CTT, reliability is quantified using *reliability coefficients,* and uncertainty in scores is quantified using unconditional and conditional *standard error of measurement.* In recent years, the ever-growing literature on latent variable models, particularly IRT models, might seem to suggest to some that CTT models are passé. This would be an inappropriate perception of testing reality, however, fueled more by academic research practice than by testing practice across a wide range of situations, and we will thus briefly address this controversy. For example, Brennan (1998a) writes, "Classical test theory is alive and well, and it will continue to survive, I think, for both conceptual and practical reasons" (p. 6).

Nevertheless, the growing interest in IRT by theoreticians and practitioners alike over the past 30 years has been nothing short of spectacular. This is evidenced in the number of sessions at measurement and testing conferences and the large proportion of publications in measurement and testing journals devoted to theoretical developments or applications of IRT. Although it is true that IRT is frequently being used in moderate- to large-scale testing programs and projects, CTT statistics continue to be widely used in the development and evaluation of tests and measures in many areas of the educational, social, and behavioral sciences that are concerned with tests and measures of limited volume of production and distribution. For example, an overwhelming majority of tests and measures reviewed in source books such as the *Mental Measurements Yearbook* series, produced by the Buros Institute of Mental Measurements, or the *Measures of Personality and Social Psychological Attitudes* book by Robinson, Shaver, and Wrightsman (1991) predominantly report CTT statistics. The primary reason for using CTT in small-volume testing programs and in research environments is the large sample sizes that are needed when one seeks to apply latent variable modeling approaches such as IRT and SEM (e.g., Bedeian, Day, & Kelloway, 1997; Bentler & Dudgeon, 1996; Junker, 1999). With observed-score measures being alive and well, it is thus worthwhile to investigate the recent developments that have taken place on these measures in the past decade. We will start appropriately with one of the oldest and most versatile indicators of score consistency, the reliability coefficient.

4.3.1. Recent Developments in the Theory of Reliability Coefficients

In the past 10 years, particularly due to the impact of increasing computer power, psychometric modeling has seen an explosion of sophisticated models that require the computer-intensive simultaneous estimation of numerous model parameters that has fueled a rethinking of the role of reliability coefficients. It is worth stating, though, that the dominating role of entities such as the information function in IRT has not changed modelers' desire for *conceptual reliability.* It has, however, changed the ways in which we look at the *mathematical formalization of reliability.*

As stated before, reliability is typically measured by a reliability coefficient, often denoted $\rho_{XX'}$, which in CTT or observed-score models is defined as the ratio of true-score variance to observed-score variance or the proportion of variation in the data that can be explained by differences among individuals or objects of measurement. Because the observed score is decomposed into two additive unobserved components, leading to ambiguities about the relative contribution of each unobserved component to total observed variance, the reliability coefficient cannot be computed directly. Instead, estimators have to be defined that provide reliability coefficient estimates based on test data from one or multiple measurement occasions. However, it is noteworthy that the definition of a reliability coefficient itself, in the context of multiple measurement occasions, poses subtle challenges to measurement specialists, who have been haunted for more than 40 years by complications that arise from *difference scores.* Some have called for a ban in difference scores because of their supposed low reliability, but today this ban has been lifted. It is recognized that although the frequently cited limitations of difference scores are real, these limitations mostly hold for restrictive situations and that there are many scenarios for which difference scores are most appropriate (Zumbo, 1999b).

A reliability coefficient is a particularly natural index in observed-score models, and the definition of a reliability coefficient in latent variable models such as IRT or SEM is much more artificial. For both latent variable models and observed-score models, the formulization of conditional measurement error and information is a natural pathway that connects different models. Yet the reliability coefficient is intricately related to the error of measurement. For example, variance ratios in random-effects models prevalent in *g*-theory or the asymptotic variance of the ability trait distribution in IRT models depend directly on quantities that measure the error in the associated models. Nevertheless, the reliability coefficient itself is sometimes preferred as an index of the amount of measurement uncertainty inherent in test

scores because it is unitless and is a single informative number that is practically easy to compute and included in most standard software packages (see Feldt & Brennan, 1989). Moreover, it is easily interpreted. Let us now turn to a few commonly encountered estimators of the population reliability coefficient.

4.3.2. Estimators of the Reliability Coefficient and Their Properties

> A fundamental fact concerning unreliability is that, in general, it cannot be estimated from only a single trial. Two or more trials are needed to prove the existence of variation in the score of a person on an item, and to estimate the extent of such variation if there is any. The experimental difficulties in obtaining independent trials have led to many attempts to estimate the reliability of a test from only a single trial by bringing in various hypotheses. Such hypotheses usually do not afford a real solution, since ordinarily they cannot be verified without the aid of at least two independent trials, which is precisely what they are intended to avoid. (Guttman, 1945, p. 256)

It is typically argued that reliability estimators fall into three distinct classes: (a) internal consistency coefficients, (b) alternative-forms reliability coefficients, and (c) test-retest coefficients. However, because reliability coefficients that involve multiple occasions for testing or rating can be estimated using intra-class coefficients, it seems more appropriate to distinguish only internal consistency coefficients and intra-class coefficients. Moreover, the intra-class coefficient in CTT is essentially a Spearman-Brown extrapolation of Cronbach's α (Feldt, 1990), which is itself the average of all split-half internal consistency correlation coefficients under appropriate model assumptions (Cronbach, 1951) and is, as such, preferred over a split-half coefficient computed for some arbitrary random split. Cronbach's α can be computed from data on a single administration of a test and does not require parallel forms, a test-retest scenario, or multiple judges for which an intra-class correlation coefficient can be used. For tests or items that are at least essentially τ equivalent with uncorrelated errors, α equals the correlation coefficient, and for congeneric tests, it is a lower bound (Lord & Novick, 1968; see Komaroff, 1997).

Coefficient α is among the most commonly reported statistics in all of social and behavioral sciences. What makes it so useful to researchers and test developers? First, it provides a conservative lower bound estimate of the theoretical reliability in the worst of situations (i.e., when essential τ equivalence does not hold). That is, the proportion of observed-score variance that is due to true individuals' differences is in truth at least the magnitude of coefficient α. Second, it provides this estimate without having to resort to repeated testing occasions and without necessitating parallel forms of a test. Third, it is easily computed and available on most statistical computer programs. The biggest limitation of coefficient α is that it results in an undifferentiated error of measurement. Generalizability theory, on the other hand, acknowledges that there are several sources for measurement error, which depend on the various factors modeled in the measurement experiment, and that one may want to model these various sources. Of course, it should be noted that in differentiating the error of measurement, one is actually also redefining the consistent or true-score part of the data.

It seems that Guttman's fears were not warranted and that we have overcome the problem of estimating reliability, a property of scores from repeated administrations, from scores from a single administration. Unfortunately, the situation may not be that simple if assumptions underlying the scoring model used are violated. In considering the assumptions of measurement models (and particularly uncorrelated errors), Rozeboom (1966) reminds us in his classic text on test theory that statistical assumptions are *empirical commitments:*

> However pleasant it may be to shuffle through the internal statistics of a compound test in search of a formula which gives the closest estimate of a test's reliability under conditions of uncorrelated errors, this is for practical applications like putting on a clean shirt to rassle a hog. (p. 415)

More than 35 years ago, Maxwell (1968) showed analytically that correlated errors lead to biased estimates of the correlation coefficient if an intra-class correlation coefficient is used as an estimator and argued that this bias is most likely to be an overestimate. It has been confirmed via simulation studies that Cronbach's α underestimates $\rho_{XX'}$ under violation of essential τ equivalence and that it overestimates $\rho_{XX'}$ if errors are correlated (Zimmerman, Zumbo, & LaLonde, 1993; see Raykov, 1998b, for composite tests and Zumbo, 1999a, for a simulation framework), but these effects can be partly attenuated if both assumptions are violated simultaneously (Komaroff, 1997). Nevertheless, it appears that α is relatively robust against moderate violations of these assumptions (see Bacon, Sauer, & Young, 1995; Feldt, 2002). Similar results have been found for g-theory designs with multiple time points. In such designs, underestimation was present for uncorrelated errors with increasing variances over time, overestimation was

present for correlated errors with equal variances over time, and both directions of estimation bias were present for correlated errors with unequal variances over time (Bost, 1995). It is important to note that correlated errors may arise for a variety of reasons. Given the advent of new item formats, one of the most common reasons for correlated errors is linked items. That is, historically, measurement specialists have advocated that items be disjoint statements that would not result in extra covariation in latent variable modeling due to item format. Items that are linked, however, may induce extra covariation among the items that appear as correlated errors (for an example, see Higgins, Zumbo, & Hay, 1999). We recommend that researchers faced with correlated errors arising from item format see Gessaroli and Folske (2002) for a useful, yet general, approach for estimating reliability.

In latent variable modeling, correlated errors are equivalent to introducing an additional latent variable (i.e., factor) that loads on the manifest variables (e.g., MacCallum, Wegener, Uchino, & Fabrigar, 1993; Raykov, 1998a). Today, FA methods, particularly CFA, continue to be useful tools to assess the degree of correlated errors (e.g., Reuterberg & Gustafsson, 1992) and have recently been used to construct adjusted αs that reduce and sometimes eliminate the inflation effect (Komaroff, 1997). Moreover, SEM allows for the estimation of a reliability coefficient for congeneric tests that is not a lower bound for the true reliability coefficient (unlike Cronbach's α) (Raykov, 1997a), along with a bootstrap estimation of its standard error that does not depend on normality assumptions (Raykov, 1998b). Unfortunately, large sample sizes are required for the stable estimation of model parameters, and not all estimation methods are recommendable (see Coenders, Saris, Batista-Foguet, & Andreenkova, 1999). Researchers need to be aware of the additional assumptions that are required for proper estimation in a covariance structure analysis (Bentler & Dudgeon, 1996). Among these are multivariate normality of the response data required for some estimation approaches, which is unlikely to hold for categorical data, and large sample sizes required for asymptotic theory, which are unlikely to exist for small-scale assessments.

Estimating reliability coefficients and assessing model assumptions has also been done for more than three decades using FA methods (e.g., Feldt, 2002; Fleishman & Benson, 1987; Jöreskog, 1970, 1971; Kaiser & Caffrey, 1965). It has been shown repeatedly that the assumption of uncorrelated errors, coupled with unidimensionality and the use of the simple total score in observed-score modeling, corresponds to an orthogonal factor model with a single dominant factor that has loadings for each item in the test. Under this model, the reliability coefficient is estimated as the sum of squared loadings (i.e., the communalities) divided by the sum of squared loadings plus error loadings (i.e., communalities plus unique variances).

Along with FA models, SEMs allow for flexible testing of multiple assumptions such as type of model (i.e., parallel, τ equivalent, congeneric), correlation of errors, invariance across time, and invariance across subgroups (e.g., Feldt, 2002; Fleishman & Benson, 1987; Raykov, 1997a, 1997b, 1998a, 1998b, 2000, 2001). In an SEM framework, the reliability coefficient can be estimated as an internal parameter or an external parameter of the model, and test or item weights can either be preset by the investigator or estimated as factor loadings simultaneously with all other model parameters. The general approach for testing assumptions about error structures using SEM requires at least four items or tests due to the identification requirements of the model so that all hypothesis tests, including the one about congenerity, can be performed (e.g., Raykov, 1997a). In addition to coefficient α, the omega coefficient with equal and unequal weights has been proposed; unequal weights are preferred by some authors because the coefficient never increases when items are dropped. Note, however, that reliability estimates are not necessarily recommended as sole yardsticks for test construction (Bacon et al., 1995). More recently, SEM has been advocated by some to model the type of correlation structure via integrated time-series models, but the practical utility of that approach remains limited at this point (Green & Hershberger, 2000). Finally, note that, just as attenuated correlation coefficients have been shown to be sensitive to the true-score distributions for examinees (Zimmerman & Williams, 1997), coefficient α is sensitive to the score distribution of examinees, which has led to the proposal of a robust generalization of α that is insensitive to tail fluctuations in this distribution (Wilcox, 1992).

So what is a practitioner to do when coefficient α needs to be estimated? It appears that for small sample sizes, sophisticated latent trait models would provide unreliable results, and the effort of estimating these is probably not worth it. If the sample size is large (e.g., at least 200 examinees for moderate tests as a guiding principle) and one has complex item formats, then latent trait models such as SEM may be useful to estimate reliability and related quantities. It is important to always be aware of the model assumptions that are lurking in the background when choosing a par-

ticular scoring model (Zumbo, 1994), however, and for larger sample sizes and high-stakes assessment scenarios, these should be investigated to obtain the most accurate estimate of reliability and measurement error. We recommend Gessaroli and Folske's (2002) approach.

4.3.3. Hypothesis Tests for Reliability Coefficients

The intra-class correlation coefficient, which can be used for test-retest, parallel forms, subtest, and inter-rater reliability, has found wide applications in social and behavioral research (Alsawalmeh & Feldt, 1992). Its distribution theory and the distribution theory for Cronbach's α have recently been developed in more detail (Feldt, 1990; van Zyl, Neudecker, & Nel, 2000). Hence, approximate tests have been developed for two independent intra-class reliability coefficients (Alsawalmeh & Feldt, 1992), two independent coefficient αs (Alsawalmeh & Feldt, 1999; Charter & Feldt, 1996), and two dependent coefficient αs (Alsawalmeh & Feldt, 2000). Similarly, tests for disattenuated correlation coefficients can be easily formulated in an SEM framework (Hancock, 1997).

Note, however, that not all distributional results are easily applicable across a wide range of situations. For example, the asymptotic distribution of the maximum likelihood (ML) estimator of α derived by van Zyl et al. (2000) requires no assumptions about the covariance structures of the items; yet, as an asymptotic result, it requires large sample sizes. Furthermore, the multivariate normal distribution of the item response data is unlikely to hold for dichotomously scored items.

Because the meaningful interpretation of hypothesis test results depends on the power of the test, it is essential to understand that the power of a test is not a function of the reliability coefficient but a relation of it (Williams, Zimmerman, & Zumbo, 1995; Zimmerman, Williams, & Zumbo, 1993a, 1993b). As these authors remind us, power is a function of the absolute value of observed variance, and its relative decomposition is irrelevant, even though it influences the magnitude of the reliability coefficient. However, formulas for computing the power and required sample size of a test for comparing coefficient αs for two populations can indeed depend on the direct magnitude of the respective sample values for the coefficient αs due to the sampling theory involved (Feldt & Ankenmann, 1998). In summary, the class of statistical tests for population reliability coefficients has been broadened, and even though the individual papers need to be referred to for the exact ways of conducting the tests, these tests are often not difficult.

4.3.4. Maximizing Reliability Coefficients and Composite Scores

It has long been acknowledged that Cronbach's α is not an indicator of test homogeneity or unidimensionality (e.g., Green, Lissitz, & Mulaik, 1977; Miller, 1995), and violations of the assumption of test homogeneity have been researched (e.g., Feldt & Qualls, 1996). If tests are measuring several related constructs, modelers in CTT deal with this by constructing composite test scores that receive appropriate weights using a table of specifications. Using a composite-score analysis instead of a total-score analysis may have a strong effect on the reliability estimate for the data, though. Formulas exist, most commonly for congeneric tests, which maximize reliability measures under different conditions (e.g., Armstrong, Jones, & Wang, 1998, for coefficient α; Goldstein & Marcoulides, 1991, and Sanders, Theunissen, & Baas, 1989, for generalizability coefficients; Knott & Bartholomew, 1993, for a normal factor model; Li, 1997, for a composite score; Li, Rosenthal, & Rubin, 1996, for cost considerations; Rozeboom, 1989, for using regression weights on a criterion variable; Segall, 1996, for linearly equated tests; Wang, 1998, for congeneric models).

Maximizing reliability is akin to determining the ideal sample size for a designed experiment under power considerations, and so, just as in traditional statistical design, practical consideration will eventually be the ultimate determining factor for test construction or the analysis method as some tests proposed to maximize reliability seem to have unrealistic characteristics (e.g., 700 multiple-choice items; see Li et al., 1996). In addition, most formulas for composite reliability coefficients require knowledge of the componential reliability coefficients. If reliability information is not available on the subcomponents that are supposed to be weighted, a multivariate covariance structure analysis approach may be called for, and formulas for weights that maximize reliability have been derived for some cases (Wang, 1998).

Coefficient α and intra-class correlation coefficients are not the only means of indexing measurement precision. In fact, they are only single numbers that capture the quality of the scores in a rather

superficial sense. To obtain more precise information about how measurement error actually affects the scores and hence decisions about examinees, we need to turn to score-level measures of precision.

4.3.5. Local Estimates of Precision in Scores

Scoring test data eventually brings about consequences for examinees. These consequences are mathematically dependent on accurately estimating the error associated with examinees' scores, which is most crucial for examinees with an observed score somewhere around the cut-score in criterion-referenced assessment or along the entire continuum for norm-referenced assessment. It has long been recognized that the score error is not constant along the continuum, even though in early work in CTT, unconditional raw-score SEM was reported and used. However, responsible data analysts and decision makers are aware that score error varies along the ability continuum, and more evidence from different estimation methods has been accumulated in the past decade to support this. Generally speaking, for observed-score models, curves depicting the conditional SEM will be somewhat inverse U-shaped, with smaller standard errors near the upper and lower tails of the true-score continuum and larger standard errors in the center of the true-score continuum. In contrast, the local precision curve for a test analyzed via IRT methods has the opposite, regular U shape. That is, there is less error in the center of the latent continuum near the point of maximum test information and more error for extreme values on the latent continuum. Thus, local measures of precision need to be considered in observed and latent variable models. Moreover, it is clear that a *conditional raw-score standard error of measurement* (*CRS-SEM*) should be used for fair decision making based on raw scores and that a *conditional scale-score standard error of measurement* (*CSS-SEM*) should be reported if raw scores are transformed via linear or nonlinear transformations to some other practically meaningful scale such as the percentiles, grade point equivalent, or stanine scales.

Although in the 1989 chapter by Feldt and Brennan, CRS-SEM only received a two-page treatment nested within a section on "special" issues in reliability and CSS-SEM was not discussed in much detail, during the past decade, researchers in the field of measurement have produced a series of papers that meticulously investigated different approaches to estimating local or conditional standard errors for scoring models on different scales and the behavior of these approaches in different calibration situations (e.g., Brennan, 1998b; Brennan & Lee, 1999; Feldt, 1996; Feldt & Qualls, 1996, 1998; Kolen, Hanson, & Brennan, 1992; Kolen, Zeng, & Hanson, 1996; Lee, 2000; Qualls-Payne, 1992; see also May & Nicewander, 1994). In general, most methods produce similar results that lead only to slight differences in confidence interval width if the conditional standard errors are used for their construction. As usual, CTT methods are comparatively easier to compute and do not rely as heavily on larger sample sizes for stable parameter estimation.

From earlier discussions, it should be clear that the explicit treatment of specific error structures in scoring models has been one of the most important contributions in the past decade. Within an observed-score context of conditional standard error, this has most notably resulted in a synthesis of conditional standard error estimation approaches for g-theory designs and estimations that include CTT scenarios as special cases (Brennan, 1998b). Within a latent trait framework, the dependency of responses for items presented with the same stimulus in testlets has driven researchers to develop a Bayesian estimation framework for dichotomous and polytomous items on the same test scored with IRT models (Bradlow et al., 1999; Wainer, Bradlow, & Du, 2000; Wainer & Thissen, 1996; Wainer & Wang, 2001; Wang et al., 2002; see also Sireci, Thissen, & Wainer, 1991, for reliability estimation as well as Lee & Frisbie, 1999, for a g-theory approach). These studies have found that incorporating testlet effects into an IRT model or g-theory model always improved estimation accuracy by incorporating within-testlet response pattern information into parameter estimates and is necessary if strong testlet effects are present to prevent biased ability estimates and thus incorrect decisions. This conclusion was further supported in a direct comparison of CRS-SEM estimates with models that accounted for testlet effects producing more accurate CRS-SEM under all conditions, even though g-theory estimation, as an alternative to IRT testlet models, worked well under mild testlet dependencies (Lee, 2000; see also Lee & Frisbie, 1999). Again, the message is that for larger sample sizes, it is particularly important to assess whether model assumptions are likely to hold, but for both smaller and larger sample sizes, conditional standard errors should be computed and used for decision making. It appears that the particular method for computing CRS-SEM or CSS-SEM does not matter much for most practical decisions and that the one that is simplest to implement should be chosen.

4.3.6. Relationships Between Error Estimates in Different Scoring Frameworks

We want to close the discussions about reliability and measurement error with a section on relationships between observed-score and IRT models as some concepts are often confused. As we have just seen, the notion of a local measure of precision, which is captured by the information function in IRT, also exists in CTT through conditional standard errors for raw and scale scores. Moreover, it is similarly possible to compute information functions in CTT (Feldt & Brennan, 1989; Mellenbergh, 1996) as well as unconditional standard errors and reliability coefficients in IRT (e.g., Samejima, 1994). In particular, the IRT equivalent to the unconditional standard error in CTT is the expectation of the asymptotic conditional standard error:

$$\text{SEM} = \sigma_\varepsilon = \int_{-\infty}^{\infty} [I(\theta)]^{-1/2} f(\theta) d\theta.$$

For practical estimation purposes, the information function in the above equation is replaced by the estimated test information function, and the ability distribution can be empirically estimated if conditional unbiasedness of $\hat{\theta}$ holds; otherwise, test information functions adjusted for bias should be used (Samejima, 1994). The reliability coefficient can now be predicted from a single administration of a test using the observed variation in θ and the estimated standard error as described above (in the formula, SEM indicates standard error):

$$\hat{\rho}_{\hat{\theta}_1,\hat{\theta}_2} = \frac{\hat{\text{Var}}(\hat{\theta}) - \hat{\text{SEM}}}{\hat{\text{Var}}(\hat{\theta})} = \frac{\hat{\text{Var}}(\theta)}{\hat{\text{Var}}(\hat{\theta})}.$$

The relationship between the multiple-occasion estimators of the reliability coefficient in CTT and IRT models has been investigated for some time, and some authors even go so far as to declare the reliability coefficient redundant (Samejima, 1994, p. 243). This statement seems a bit extreme because the appropriateness of an IRT estimate of reliability depends on the accuracy of the fitted model (see Meijer, Sijtsma, & Molenaar, 1995, p. 334, for this argument in a nonparametric context), and fitting a more complex IRT model may require more data than are available at a given moment. In addition, even though in IRT, standard errors are larger at the extreme ends of the scale (Lord, 1980), this is dependent on the choice of transformation from the true score to the latent trait scale, and dramatic differences between conditional standard errors can be observed for different choices of transformation (see Brennan, 1998b).

As another similarity between CTT and IRT models, recall that the reliability coefficient in CTT is the ratio of true-score variance to total observed variance or the ratio of signal to signal plus noise. Put differently, the signal-to-noise ratio equals the correlation coefficient divided by 1 minus the correlation coefficient. Therefore, a local reliability coefficient can be defined as a function of the item information function, which is itself proportional to the local signal-to-noise ratio (Nicewander, 1993).

Conditional standard errors for absolute decisions (and thus dependability coefficients) or relative decisions (and thus generalizability coefficients) can also be formulated in *g*-theory (Brennan, 1998b, 2001). In *g*-theory, the class of model specifications, albeit all generalized linear models (GLIMs), has been enlarged, but typically larger sample sizes are required for accurate estimation of variance components. In IRT, the class of GLIMs uses different link functions, but choices have to be made now between logit and probit models, the number of parameters in the model, and whether to choose a parametric or nonparametric formulation. In the latter case, reliability estimation is not even common practice, and even though a reliability coefficient that is related to a scalability coefficient can be estimated in Mokken's nonparametric alternatives to the Rasch model, their complementary uses remain unclear (Meijer et al., 1995; Meijer, Sijtsma, & Smid, 1990).

Finally, it needs to be highlighted that one of the advantages of reliability estimation in CTT is the relative simplicity of the model, whose only major alternatives consisted of different assumptions about its unobserved components. Claims that CTT is merely a special case of IRT (Nicewander, 1993) or FA seem to be overstatements and seem to ignore the difference between score-level and item-level modeling, as well as between a latent variable and a more general unobserved variable such as the true score in CTT. To the contrary to the overstatement, it can be argued that IRT is a first-order approximation to CTT. The overstatement also ignores the role that parameter estimation strategy has in defining a psychometric model. Put simply, the liberalization of CTT into *g*-theory—along with its reformulation and extension, in latent variable terms, in FA and SEM—and the advent of IRT have come at the price of stronger requirements on the data, which have affected reliability estimation. For larger sample sizes, we can definitely investigate more complex assessment scenarios through *g*-theory, as well as more complex dependency structures through FA and SEM, and achieve invariance properties for adaptive testing in IRT (see Rupp, 2003; Rupp & Zumbo, 2003, in press, on quantifying a lack of invariance

in IRT models), but for smaller sample sizes, these advances are often of limited benefit to the practitioner. In addition, no matter how sophisticated the model statement and estimation routines have become, a test of poor validity and thus poor conceptual reliability will always remain unaltered. This brings us to our final section.

4.4. Validity and the Practice of Validation

Validity theory aids us in the inference from the true score or latent variable score to the construct of interest. In fact, one of the current themes in validity theory is that construct validity is the totality of validity theory and that its discussion is comprehensive, integrative, and evidence based. In this sense, construct validity refers to the degree to which inferences can be made legitimately from the observed scores to the theoretical constructs about which these observations are supposed to contain information. In short, construct validity involves generalizing from our behavioral or social observations to the *concept* of our behavioral or social observations. The practice of validation aims to ascertain the extent to which an interpretation of a test is *conceptually* and *empirically* warranted and should be aimed at making explicit hidden ethical and social values that influence that process (Messick, 1995).

It is hard not to address validity issues when one is discussing errors of measurement. Yet the developments in validity theory have not been as dramatic over the past 15 years as have been the developments in reliability estimation and measurement model development. For a cursory overview, several papers are available that describe important current developments in validity theory (Hubley & Zumbo, 1996; Johnson & Plake, 1998; Kane, 2001). In brief, the recent history of validity theory is perhaps best captured by the following observations.

• As one can see in Zumbo's (1998) volume, there is a move to consider the *consequences* of inferences from test scores. That is, along with the elevation of construct validity to an overall validity framework for evaluating test interpretation and use came the consideration of the role of ethical and social consequences as validity evidence contributing to score meaning. This movement has been met with some resistance. In the end, Messick (1998) made the point most succinctly when he stated that one should not be simply concerned with the obvious and gross negative consequences of score interpretation, but rather one should consider the more subtle and systemic consequences of "normal" test use. The matter and role of consequences still remains controversial today and will regain momentum in the current climate of large-scale test results affecting educational financing and staffing in the United States and Canada.

• Although it was initially set aside in the move to elevate construct validity, criterion-based evidence is gaining momentum again in part due to the work of Sireci (1998).

• Of all the threats to valid inferences from test scores, test translation is growing in awareness due to the number of international efforts in testing and measurement (see Hambleton & Patsula, 1998).

• The use of cognitive models as an alternative to traditional test validation is gaining a great deal of momentum. One of the limitations of traditional quantitative test validation practices (e.g., factor-analytic methods, validity coefficients, and multitrait multimethod approaches) is that they are descriptive rather than explanatory. In other words, they are statistical and not psychological. Models for cognitively diagnostic assessment, particularly the work of Susan Embretson and Kikumi Tatsuoka, has expanded the evidential basis for test validation as well as the nomothetic span of the nomological network. The basic idea is that if one could understand why an individual responded a certain way to an item, then that would go a long way toward bridging the inferential gap between test scores and constructs.

Given that cognitive models present one of the most exciting new developments with implications for validity theory, the next section discusses them in more detail.

4.4.1. Cognitive Models for a Stronger Evidentiary Bases of Test Validation

It is informative to start this discussion by addressing the use of the term *cognitive psychology* in the literature on cognitive models. For many assessment situations, researchers use the word *cognition* to loosely refer to any process that is somehow grounded in our minds and therefore eventually our brains. Yet there is little doubt that measurement specialists are not interested in the biological or neuroscientific bases of cognitive processes for typical cognitively diagnostic assessments, so that we really often mean a "soft" form of cognitive psychology in a measurement context.

Some will undoubtedly argue that it is not the job of a psychometric data modeler to worry about what is done with the numerical estimates once they are handed down, but it is exactly this neglect of meaningful inferences at the expense of sophisticated estimation techniques that has often eradicated the psychology in "psycho"-metrics. The realization that it is time to put psychology back into the equation so that investigators who desire "reliable" tests are assured by modelers that their data do indeed provide evidence for dependable meaningful inferences. To appreciate the relevance and importance of cognitive models, one has to understand that we have not made many significant advances toward explicit validation of the inferences drawn from test scores through mathematical models. This holds true despite the injection of a latent continuum that allows modelers to extract information from test data more flexibly and accurately at the item level in IRT. For example, Junker (1999) suggests that

> despite the persistence of the "latent trait" terminology in their work, few psychometricians today believe that the latent continuous proficiency variable in an IRT model has any deep reality as a "trait"; but as a vehicle for effectively summarizing, ranking, and selecting based on performance in a domain, latent proficiency can be quite useful. (p. 10)

The main goal of modeling test data should always be to make valid inferences about the examinees, but the validity of these inferences cannot be mechanically increased by inducing latent constructs into the data structure.

Cognitive models seek to explicitly represent the cognitive processes that examinees engage in when responding to items through parameters in mathematical models, which typically consist of augmented IRT models, classification algorithms based on regular IRT models, or Bayesian inference networks that have IRT models as a central component. One approach to cognitively diagnostic assessment is the *rule-space methodology* that attempts to classify examinees into distinct attribute states based on observed item response data, an appropriate IRT model, and the attribute specification for the items (Tatsuoka, 1983, 1991, 1995, 1996; Tatsuoka & Tatsuoka, 1987). Despite a lack of consensus in the literature about what is meant exactly by an *attribute* and the sensitivity of the classification to the appropriateness of the chosen IRT model, this approach forces test developers to specify prerequisite cognitive characteristics of examinees—ideally before designing a test (Gierl, Leighton, & Hunka, 2000). Other approaches based on item attribute incidence or Q-matrices have been developed (e.g., DiBello, Stout, & Roussos, 1995), but their main weakness to date remains the vagueness and lack of guidance in attribute specification (e.g., Junker & Sijtsma, 2001). Developments in cognitive models have often taken place primarily in educational achievement and psychoeducational assessment contexts, though. An exception was Zumbo, Pope, Watson, and Hubley (1997) in personality assessment, in which they studied the relation of the abstractness and concreteness of items to the psychometric properties of a personality measure. Other advances are currently made in the development of simulation-based assessment software that emphasizes a deeper and richer understanding of the cognitive processes required for performing certain tasks in which data are analyzed through Bayesian networks (Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999).

More sophisticated models for cognitive assessment do not come without a price. One of the components of this price is again sample size because more complex IRT models, cognitive state models, or Bayesian inference networks typically require a larger number of parameters to be estimated. More important, however, the more useful models for cognitively diagnostic assessment are built on a solid understanding of the cognitive processes underlying the tasks that are being assessed. As an excellent example, consider the work by Embretson (1998), who used the cognitive process analysis of the Raven's Advanced Progressive Matrix test by Carpenter, Just, and Shell (1990) to model examinees' responses, extract diagnostic information, and generate similar items. Comprehensive models of cognitive abilities are still relatively rare, and even though advances have been made, it is necessary to note that their most important cornerstone, the analysis of cognitive processes, is still their weakest element.

The issue is less a lack of models for new kinds of test data but rather a lack of awareness in the applied world that these models exist along with a mismatch of assessment instruments and modeling practice. In other words, if test developers are interested in providing examinees and institutions with richer profiles of abilities and developmental progress, the nature of the assessment methods has to change to provide richer data sets from which relevant information can be more meaningfully extracted. What is meant by *more meaningful* will, of course, in the end depend on the use of the assessment data, but in general, authorities in the field are nowadays beginning to agree that we need more than simple test responses scored 0 and 1 to validate the inferences that are being made from the test data. As Embretson's (1998) work demonstrates,

the key to useful cognitive models is that they need to be explanatory and not just another set of descriptive models in cognitive terms rather than mathematical terms (Zumbo & MacMillan, 1999). Put differently, a change of terminology is insufficient to claim true advances in gathering more meaningful and weighty validity evidence.

A similar push for explanatory power has also taken place in the area of differential item functioning, where attitudinal, background, and cognitive variables are used to account for differential achievement profiles to investigate the inferential comparability of scores across populations (Klieme & Baumert, 2001; Watermann & Klieme, 2002). The developments that are currently taking place serve in part as a consciousness-raising device to help test developers and users to reflect more closely on how valid their inferences from test data really are and how these inferences can be improved. This continues the path toward a comprehensive and unified validation process of assessment instruments that has been eloquently laid out by Messick (1989, 1995).

4.4.2. Implications of Cognitive Models for Modeling Novel Dependency Structures

In traditional psychometric models, dependencies among item responses over and above what can be accounted for by the unobserved variables have been a dreaded feature of test data, and every effort has always been made to eliminate this dependency through test design or modeling efforts. This may be the wrong lens that is applied to the data, and it appears that cognitively diagnostic assessments—along with models for testlet structures and more complicated error dependencies—are the new figures that are slowly taking shape under a new perspective on items and responding to items. We have begun to shift our thinking back to the individual examinees because we are starting to realize that the goal of any assessment, be it strictly cognitively diagnostic or not, is to arrive at better inferences about examinees' abilities. Furthermore, item difficulty and discrimination are properties of the examinees that respond to the items because the items are windows into the minds of the examinees and are not qualities inherent in items independent of populations of examinees.

All of this is to say that the current push toward cognitively diagnostic assessment seems to be more than just an extension of currently existing models and statistical methodologies to richer domains. In fact, it is our chance to clean our windows into the minds of examinees and to refocus our lenses toward the examinees as the unit of investigation that matters most. From a mathematical perspective, this means looking for different types of information in data structures that may posit new challenges to the modelers. In particular, if cognitive processes are highly interrelated in complex neural networks at a biological-chemical level, then we can expect that item responses are probably also interrelated to a much higher degree than gives us comfort. Indeed, what we need is an extension of the models that are currently used in covariance structure analysis because the future seems to lie in accepting covariation and interrelationships rather than dreading them.

This can be seen not only in the models and scenarios discussed so far but also by looking at the variety of item types that can be found in new tests across multiple disciplines (Zenisky & Sireci, 2002). As these authors show, traditional test formats have been augmented with a whole new battery of items that require the test taker to engage in more sophisticated complex cognitive processes. We certainly have choices when scoring these item types as we really also have when dealing with the items that are used in cognitively diagnostic assessments. We could theoretically score them all 0–1 or on a simple graded scale and apply traditional models in CTT, g-theory, or IRT to the responses. We might find, however, that dependency structures in the data sets might compromise our simple analyses because the items are not isolated items anymore. Indeed, to use such items more successfully, it would make much more sense to focus on the interdependencies and go from there.

It should also be noted that the nature of dependencies that are deliberately build into more complex item types has crept up with traditional tests as well. For example, researchers have been busy investigating the data structure for CTT models in terms of the degree of test parallelism. As one dimension of complexity, researchers have defined parallel, essentially parallel, τ-equivalent, essentially τ-equivalent, and congeneric tests; as a second dimension, they consider uncorrelated and correlated errors; and as a third, they investigate sampling type (i.e., Type 1, Type 2, Type 12). With all these considerations at hand, psychometricians have been busy trying to find the best estimators of quantities such as the reliability coefficient or conditional standard errors for different data structures. Nevertheless, we are faced these days with data structures that do not adhere to any of the criteria above (e.g., spherical covariance matrices; see Barchard & Hakstian, 1997; Hakstian & Barchard, 2000), which compel us to search for better descriptions of the data structure at hand.

4.5. Conclusion

The emphasis in this chapter has been on measurement error, reliability, and validity through the lens of scoring data from tests within a particular scoring framework. We have highlighted on several occasions the distinctions between observed variable frameworks (i.e., CTT and *g*-theory) and latent variable frameworks (i.e., EFA, CFA, SEM, and IRT). We believe it is important to understand that the use of a particular scoring model always remains the choice of the data analyst and is not necessitated by the data. More often than not, the choice of a particular scoring model is the result of personal beliefs, training, and working conventions (Rupp & Zumbo, 2003). Yet it has severe consequences for how we define, quantify, and use measurement error and the decisions that we base thereon. Choosing a scoring model is an *empirical commitment* that demands the data analyst take responsibility for the consequences imparted on the examinees by this choice.

To underscore this responsibility one last time, consider for a moment a few issues that can arise with some popular scoring models. When working within a latent variable framework, it is certainly irresponsible to blindly fit IRT models to any kind of data—even if the models formally match the type of scores given (e.g., dichotomous, polytomous)—without ensuring that *sufficiently large* and *representative* calibration samples are available so that *stable* and *representative* parameter estimates can be obtained. If the parameters are not well estimated, decisions will be biased. In addition, if the intention is to use a one-shot calibration at one point in time with one set of examinees, it is logically inconsistent to justify the use of an IRT model because model parameters possess the feature of *invariance.* Invariance refers to the identity of item and examinee parameters from *repeated* calibration for *perfect* model fit and is not needed in this case. Hence, it should not be cited as the *primary* reason for using such a model.

Another example comes from the area of cognitively diagnostic assessment. Without any detailed data collected on examinees and any detailed attempts to develop realistic processing models, truly cognitively diagnostic assessment is not possible. In addition, an augmented IRT model for cognitive assessment needs to be *judiciously* chosen based on the cognitive theory underlying the test response processes and not simply because it is an interesting extension of basic IRT models (for excellent examples, see Embretson, 1998; Maris, 1995).

In the area of observed-score modeling, it is equally irresponsible to use the unconditional raw-score standard error when a large body of evidence has shown for years that CRS-SEM varies along the score continuum. Similarly, CSS-SEM needs to be computed separately if scores are transformed to scales such as stanine, percentile, or grade-equivalent scales as it also varies and is generally not equal to the CRS-SEM. Using inappropriate measures of error can lead to incorrect and unfair decisions for some, if not most, students. On a more subtle level, most observed-score scoring methods rely on assumptions about the score matrix such as parallelism, essential τ equivalence, or congenerity. In some cases, failing to adjust reliability coefficients or other measures of error to the right model can lead to biased statements about a test, overconfidence in test use, and unfair decisions about examinees. Moreover, factor-analytic procedures and software can nowadays easily be used to test for these assumptions and to produce appropriate error estimates for larger sample sizes.

It is the responsibility of mathematically trained psychometricians to inform those who are less versed in the theory about the consequences of their decisions to ensure that examinees are assessed fairly. Because models (which, in part, include the parameter estimation strategy) are empirical commitments, it is measurement specialists who need to take partial responsibility for the decisions that are being made with the models they provide to others. Everyone knows that a useful and essential tool such as an automobile, a chainsaw, or a statistical model can be very dangerous if put into the hands of people who do not have sufficient training and handling experience or lack the willingness to be responsible users.

All of this is not to say that decision-making disasters will immediately occur if the above things are not adhered to in the fullest. However, it can also be too tempting to take that exact fact to be less stringent and less careful about our practices, and we believe it is important that we all in the psychometric community work together to ensure fair and sound decision making. Technological advances have opened up doors for us to do more sophisticated and complex simulation work, analyze richer and more nested data structures than ever before, and synthesize findings across analyses. At the same time, it is important to remember that examinees are typically not interested in the particular scoring models used for obtaining their score but rather in a fair assessment, which simply translates to fair decisions based on their responses. The term *fair* is of course heavily value laden and can take on different shades of meaning for different examinees,

but nevertheless responsible data models consider the consequences of test score interpretation for which they provide the numerical ingredients.

Numerous questions about the reliability of tests have been asked in the past decade, and important advances have been made in the area of estimating conditional standard error for nonlinear scale transformations, estimating bias of reliability coefficient estimates such as coefficient α under simultaneous violations of assumptions, deriving algorithms for maximizing α, deriving tests for αs from different populations, and establishing relationships between CTT, *g*-theory, IRT, and SEM that show the interrelatedness of these procedures. In other words, we have been able to make convincing arguments for the unification of measurement models (see McDonald, 1999; Rupp, 2002; Zimmerman & Zumbo, 2001), and we have made convincing arguments for advantages of *g*-theory over CTT, IRT over *g*-theory, IRT over CTT, and SEM over *g*-theory, CTT, and IRT and so on. Important research in this area still needs to happen, and a wealth of unanswered research questions can be found in the concluding sections of the more than 100 articles that we could find in journals over the past 10 years.

We believe that this is fruitful work but that it is at least as important to reflect on our testing practice in the new millennium. Cognitively diagnostic assessments will play an important part, but we believe that they will neither replace traditional assessments entirely in the near future nor answer all of the problems encountered by psychometricians at the moment. But they are the psychometric discipline's way of pointing out that data modelers are ready to face new challenges posed by the need for richer information about examinees, concurrent new item types, redefinitions of the construct of an item itself, and a higher degree of interrelatedness of responses from a mathematical as well as from a soft cognition perspective. Reliability and validity will always be important in test development. Reliability indices are not irrelevant, as some proclaim, because they serve different purposes than conditional SEM and test information functions, and validity will always be the cornerstone of test development and use, particularly if we move to a more unified test development–data modeling–test use process. Measurement specialists are beginning to talk and reach out to each other more and more across disciplines and cultural boundaries. Content experts, psychometric data analysts, and cognitive psychologists may not always be at the same table yet, but at least they are more often pooling their expertise in the same metaphorical room, and that is certainly a good thing. We are far from a *practical* revolution in testing, but we seem to be at an exciting juncture for pausing and reflecting on what to focus on.

REFERENCES

Alsawalmeh, Y. M., & Feldt, L. S. (1992). Test of the hypothesis that the intraclass reliability coefficient is the same for two measurement procedures. *Applied Psychological Measurement, 16,* 195–205.

Alsawalmeh, Y. M., & Feldt, L. S. (1999). Testing the equality of two independent α coefficients adjusted by the Spearman-Brown formula. *Applied Psychological Measurement, 23,* 363–370.

Alsawalmeh, Y. M., & Feldt, L. S. (2000). A test of the equality of two related α coefficients adjusted by the Spearman-Brown formula. *Applied Psychological Measurement, 24,* 163–172.

Armstrong, R. D., Jones, D. H., & Wang, Z. (1998). Optimization of classical reliability in test construction. *Journal of Educational and Behavioral Statistics, 23,* 1–17.

Bacon, D. R., Sauer, P. L., & Young, M. (1995). Composite reliability in structural equations modeling. *Educational and Psychological Measurement, 55,* 394–406.

Barchard, K. A., & Hakstian, R. A. (1997). The robustness of confidence intervals for coefficient alpha under violation of the assumption of essential parallelism. *Multivariate Behavioral Research, 32,* 169–191.

Bedeian, A. G., Day, D. V., & Kelloway, E. K. (1997). Correcting for measurement error attenuation in structural equation models: Some important reminders. *Educational and Psychological Measurement, 57,* 785–799.

Bentler, P. M., & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology, 47,* 563–592.

Bost, J. E. (1995). The effects of correlated errors on generalizability and dependability coefficients. *Applied Psychological Measurement, 19,* 191–203.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153–168.

Braham, C. G. (Ed.-in-Chief). (1996). *Random House Webster's dictionary.* New York: Ballantine.

Brennan, R. L. (1998a). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice, 17,* 5–9, 30.

Brennan, R. L. (1998b). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement, 22,* 307–331.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Brennan, R. L., & Lee, W. (1999). Conditional scale-score standard errors of measurement under binomial and compound-binomial assumptions. *Educational and Psychological Measurement, 59,* 5–24.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming.* Mahwah, NJ: Lawrence Erlbaum.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of processing in the Raven's Progressive Matrices Test. *Psychological Review, 97,* 404–431.

Charter, R. A., & Feldt, L. S. (1996). Testing the equality of two alpha coefficients. *Perceptual and Motor Skills, 82,* 763–768.

Coenders, G., Saris, W. E., Batista-Foguet, J. M., & Andreenkova, A. (1999). Stability of three-wave simplex estimates of reliability. *Structural Equation Modeling, 6*(2), 135–157.

Comrey, A. L. (1973). *A first course in factor analysis.* New York: Academic Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability scores and profiles.* New York: John Wiley.

DiBello, L. V., Stout, W. F., & Roussos, L. (1995). Unified cognitive psychometric assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–389). Hillsdale, NJ: Lawrence Erlbaum.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380–396.

Everitt, B. S. (1984). *An introduction to latent variable models.* New York: Chapman & Hall.

Feldt, L. S. (1990). The sampling theory for the intraclass reliability coefficient. *Applied Measurement in Education, 3,* 361–367.

Feldt, L. S. (1996). Estimation of measurement error variance at specific score levels. *Journal of Educational Measurement, 33,* 141–156.

Feldt, L. S. (2002). Estimating the internal consistency reliability of tests composed of testlets varying in length. *Applied Measurement in Education, 15,* 33–48.

Feldt, L. S., & Ankenmann, R. D. (1998). Appropriate sample sizes for comparing alpha reliabilities. *Applied Psychological Measurement, 22,* 170–178.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan.

Feldt, L. S., & Qualls, A. L. (1996). Bias in coefficient alpha arising from heterogeneity of test content. *Applied Measurement in Education, 9,* 277–286.

Feldt, L. S., & Qualls, A. L. (1998). Approximating scale score standard error of measurement from the raw score standard error. *Applied Measurement in Education, 11,* 159–177.

Fleishman, J., & Benson, J. (1987). Using LISREL to evaluate measurement models and scale reliability. *Educational and Psychological Measurement, 47,* 925–939.

Gessaroli, M. E., & Folske, J. C. (2002). Generalizing the reliability of tests comprised of testlets. *International Journal of Testing, 2,* 277–296.

Gierl, M., Leighton, J. P., & Hunka, S. M. (2000). Exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice, 19,* 34–44.

Goldstein, H., & Wood, R. (1989). Five decades of item response modeling. *British Journal of Mathematical and Statistical Psychology, 42,* 139–167.

Goldstein, Z., & Marcoulides, G. A. (1991). Maximizing the coefficient of generalizability in decision studies. *Educational and Psychological Measurement, 51,* 79–88.

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling, 7,* 251–270.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37,* 827–838.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255–282.

Hakstian, A. R., & Barchard, K. A. (2000). Toward more robust inferential procedures for coefficient alpha under sampling of both subjects and conditions. *Multivariate Behavioral Research, 35,* 427–456.

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. In B. D. Zumbo (Ed.), *Validity theory and the methods used in validation: Perspectives from the social and behavioral sciences* (pp. 153–171). Amsterdam: Kluwer Academic.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hancock, G. R. (1997). Correlation/validity coefficients disattenuated for score reliability: A structural equation modeling approach. *Educational and Psychological Measurement, 57,* 598–606.

Higgins, N. C., Zumbo, B. D., & Hay, J. L. (1999). Construct validity of attributional style: Modeling context-dependent item sets in the Attributional Style Questionnaire. *Educational and Psychological Measurement, 59,* 804–820.

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *Journal of General Psychology, 123,* 207–215.

Johnson, J. L., & Plake, B. S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement, 58,* 736–753.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika, 57,* 239–251.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133.

Junker, B. W. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment.* Unpublished manuscript. Available: www.stat.cmu.edu/∼brian/nrc/cfa

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258–272.

Kaiser, H. F., & Caffrey, J. (1965). Alpha factor analysis. *Psychometrika, 30,* 1–14.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319–342.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions.* Thousand Oaks, CA: Sage.

Klieme, E., & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education, 16,* 385–402.

Knapp, T. R. (2001). *The reliability of measuring instrument*s. Vancouver, British Columbia: Edgeworth Laboratory for Quantitative Educational and Behavioral Science Series. Available: www.educ.ubc.ca/faculty/zumbo/series/knapp/index.htm

Knott, M., & Bartholomew, D. J. (1993). Constructing measures with maximum reliability. *Psychometrika, 58,* 331–338.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29,* 285–307.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33,* 129–140.

Komaroff, E. (1997). Effect of simultaneous violations of essential τ-equivalence and uncorrelated error on coefficient α. *Applied Psychological Measurement, 21,* 337–348.

Lee, G. (2000). A comparison of methods of estimating conditional standard errors of measurement for testlet-based scores using simulation techniques. *Journal of Educational Measurement, 36,* 91–112.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12,* 237–255.

Li, H. (1997). A unifying expression for the maximal reliability of a composite. *Psychometrika, 62,* 245–249.

Li, H., Rosenthal, R., & Rubin, D. B. (1996). Reliability of measurements in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods, 1,* 98–107.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

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

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.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60,* 523–547.

Maxwell, A. E. (1968). The effect of correlated errors on estimates of reliability coefficients. *Educational and Psychological Measurement, 28,* 803–811.

May, K., & Nicewander, W. A. (1994). Reliability and information functions for percentile ranks. *Journal of Educational Measurement, 31,* 313–325.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6,* 379–396.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum.

Meijer, R. R., Sijtsma, K., & Molenaar, I. W. (1995). Reliability estimation for single dichotomous items based on Mokken's IRT model. *Applied Psychological Measurement, 19,* 323–335.

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and Rasch approach to IRT. *Applied Psychological Measurement, 14,* 283–298.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115,* 300–307.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1,* 293–299.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Messick, S. (1998). Test validity: A matter of consequence. In B. D. Zumbo (Ed.), *Validity theory and the methods used in validation: Perspectives from the social and behavioral sciences* (pp. 35–44). Amsterdam: Kluwer Academic.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling, 2,* 255–273.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (1999). A cognitive task analysis with implications for designing simulation-based performance assessment. *Computers in Human Behavior, 15*(3–4), 335–374.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29,* 81–117.

Nicewander, W. A. (1993). Some relationships between the information function of IRT and the signal/noise ratio and reliability coefficient of classical test theory. *Psychometrika, 58,* 139–141.

Qualls-Payne, A. L. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement, 29,* 225–231.

Raykov, T. (1997a). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21,* 173–184.

Raykov, T. (1997b). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32,* 329–353.

Raykov, T. (1998a). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement, 22,* 375–385.

Raykov, T. (1998b). A method for obtaining standard errors and confidence intervals of composite reliability for congeneric items. *Applied Psychological Measurement, 22,* 369–374.

Raykov, T. (2000). A method for examining stability in reliability. *Multivariate Behavioral Research, 35,* 289–305.

Raykov, T. (2001). Bias of coefficient α for fixed congeneric measures with correlated errors. *Applied Psychological Measurement, 25,* 69–76.

Reuterberg, S., & Gustafsson, J. (1992). Confirmatory factor analysis and reliability: Testing measurement model assumptions. *Educational and Psychological Measurement, 52,* 795–811.

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (Eds.). (1991). *Measures of personality and social psychological attitudes.* San Diego: Academic Press.

Rozeboom, W. W. (1966). *Foundations of the theory of prediction.* Homewood, IL: Dorsey.

Rozeboom, W. W. (1989). The reliability of a linear composite of nonequivalent subtests. *Applied Psychological Measurement, 13,* 277–283.

Rupp, A. A. (2002). Feature selection for choosing and assembling measurement models: A building-block-based organization. *International Journal of Testing, 3–4,* 311–360.

Rupp, A. A. (2003). *Quantifying subpopulation differences for a lack of invariance using complex examinee profiles: An exploratory multi-group approach using functional data analysis.* Manuscript submitted for publication.

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (in press). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to item response modeling. *Structural Equation Modeling.*

Rupp, A. A., & Zumbo, B. D. (2003). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. *Alberta Journal of Educational Research, 49*, 264–276.

Rupp, A. A., & Zumbo, B. D. (in press). A note on how to quantify and report whether invariance holds for IRT models: When Pearson correlations are not enough. *Educational and Psychological Measurement.*

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18,* 229–244.

Sanders, P. F., Theunissen, T. J. J. M., & Baas, S. M. (1989). Minimizing the number of observations: A generalization of the Spearman-Brown formula. *Psychometrika, 54,* 587–598.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61,* 331–354.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Sireci, S. G. (1998). The construct of content validity. In B. D. Zumbo (Ed.), *Validity theory and the methods used in validation: Perspectives from the social and behavioral sciences* (pp. 83–117). Amsterdam: Kluwer Academic.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237–247.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item-response theory. *Journal of Educational Measurement, 20,* 345–354.

Tatsuoka, K. K. (1991). *Boolean algebra applied to determination of universal set of knowledge states* (Tech. Rep. No. RR-91–44-ONR). Princeton, NJ: Educational Testing Service.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–259). Hillsdale, NJ: Lawrence Erlbaum.

Tatsuoka, K. K. (1996). Use of generalized person-fit indices, Zetas for statistical pattern classification. *Applied Measurement in Education, 9,* 65–75.

Tatsuoka, K. K., & Tatsuoka, M. (1987). Bug distribution and statistical pattern classifications. *Psychometrika, 52,* 193–206.

Traub, R.E. (1994). *Reliability for the social sciences: Theory and applications.* Thousand Oaks, CA: Sage.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika, 65,* 271–280.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3pl model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–269). Boston: Kluwer Academic.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22–29.

Wainer, H., & Wang, X. (2001). *Using a new statistical model for testlets to score TOEFL* (Tech. Rep. No. TR-16). Princeton, NJ: Educational Testing Service.

Wang, T. (1998). Weights that maximize reliability under a congeneric model. *Applied Psychological Measurement, 22,* 179–187.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26,* 109–128.

Watermann, R., & Klieme, E. (2002). Reporting results of large-scale assessments in psychologically and educationally meaningful terms. *European Journal of Psychological Assessment, 18,* 190–203.

Wilcox, R. R. (1992). Robust generalizations of classical test reliability and Cronbach's alpha. *British Journal of Mathematical and Statistical Psychology, 45,* 239–254.

Williams, R. H., Zimmerman, D. W., & Zumbo, B. D. (1995). Impact of measurement error on statistical power: Review of an old paradox. *Journal of Experimental Education, 63,* 363–370.

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large scale assessment. *Applied Measurement in Education, 15,* 337–362.

Zimmerman, D. W., & Williams, R. H. (1997). Properties of the Spearman correction for attenuation for normal and realistic non-normal distributions. *Applied Psychological Measurement, 21,* 253–270.

Zimmerman, D. W., Williams, R. H., & Zumbo, B. D. (1993a). Reliability of measurement and power of significance tests based on differences. *Applied Psychological Measurement, 17,* 1–9.

Zimmerman, D. W., Williams, R. H., & Zumbo, B. D. (1993b). Reliability, power, functions, and relations: A reply to Humphreys. *Applied Psychological Measurement, 17,* 15–16.

Zimmerman, D. W., & Zumbo, B. D. (2001). The geometry of probability, statistics, and test theory. *International Journal of Testing, 1,* 283–303.

Zimmerman, D. W., Zumbo, B. D., & LaLonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement, 53,* 33–49.

Zumbo, B. D. (1994). The lurking assumptions in using generalizability theory to monitor an individual's progress. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. Boss (Eds.), *Modern theories of measurement: Problems & issues* (pp. 261–278). Ottawa, Ontario: University of Ottawa.

Zumbo, B. D. (Ed.). (1998). Validity theory and the methods used in validation: Perspectives from the social and behavioral sciences [Special issue]. *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement, 45* (1–3).

Zumbo, B. D. (1999a). *A glance at coefficient alpha with an eye towards robustness studies: Some mathematical notes and a simulation model* (Paper No. ESQBS-99–1). Prince George: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioural Science.

Zumbo, B. D. (1999b). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 269–304). Greenwich, CT: JAI.

Zumbo, B. D., & MacMillan, P. D. (1999). An overview and some observations on the psychometric models used in computer-adaptive language testing. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency* (pp. 216–228). Cambridge, UK: Cambridge University Press.

Zumbo, B. D., Pope, G. A., Watson, J. E., & Hubley, A. M. (1997). An empirical test of Roskam's conjecture about the interpretation of an ICC parameter in personality inventories. *Educational and Psychological Measurement, 57,* 963–969.

Chapter 5

Test Modeling

Ratna Nandakumar

Terry Ackerman

Discoveries with item response theory (IRT) principles, since the 1960s, have led to major breakthroughs in psychological and educational assessment. For example, using IRT principles, it is possible to determine the relative standing of an examinee on the latent continuum by administering any sample of items from a given domain of knowledge. This is possible through the principle of invariance in IRT, which means that item properties such as difficulty and discrimination can be determined irrespective of the ability level of the examinee. Hence, any set of items from a given domain can be used to estimate an examinee's position along the latent continuum. This is in sharp contrast to the traditional classical test theory (CTT), in which item statistics are a function of the specific group of examinees who took the item, and the examinee's performance is a function of the items on the test. That is, in CTT, the same item may have different *p*-values depending on the level of the examinees' ability taking the item. Similarly, in CTT, it is not possible to generalize the performance of an examinee beyond a given set of test items.

The advantages of IRT techniques are associated with strong models used to characterize examinee performance on a test, as opposed to the weak models of CTT that are tautologies and not testable. One can realize the potentials of IRT modeling and its consequences only if there is a close match between the model and data. Application of IRT techniques to data without ensuring the model-data fit can lead to unfair and unjustified ranking of examinees on the latent continuum of domain of interest.

The fundamental underlying assumptions of item response models are monotonicity, dimensionality, and local independence. Monotonicity implies that item performance is monotonically related to the ability. That is, a high-ability examinee has a greater probability of responding correctly to the item than a low-ability examinee. Because achievement test items inherently satisfy this assumption, it is implicitly assumed.[1] Local independence (LI) implies that item responses are conditionally independent. The conditional ability vector that ensures item independence is key to determining the dimensionality of data. For example, if local independence is achieved by conditioning on a unidimensional latent trait, then the response data are said to be unidimensional. If local independence is achieved by conditioning on a two-dimensional latent trait vector, then the response data are said to be two-dimensional. Hence, local independence and dimensionality assumptions are intertwined. One can only statistically test either of the assumptions assuming the other.

In addition to these basic foundational assumptions, a given model may have other assumptions. For

1. Normally, during the test construction process, if an item does not satisfy the assumption of monotonicity, it is deleted from the test.

example, among parametric models, there are models associated with different item types, such as dichotomous items (item is scored correct vs. incorrect) and polytomous items (arising from scoring essays and performance-type tasks). Each model has a set of assumptions associated with it. For a list of IRT models for different item formats and their development, refer to van der Linden and Hambleton (1997). To date, a great majority of tests are intended to be unidimensional ($d = 1$). That is, the purpose of the test is to assess an examinee's trait level based on his or her responses to unidimensional test items. Examinee test performance on a unidimensional test can be summarized with a single scale score. It is also well known that any unidimensional test is typically influenced by transient dimensions (abilities) common to just a few of the items. It is well documented (Hambleton & Swaminathan, 1985; Humphreys, 1985, 1986; Stout, 1987) that summarizing examinees' performance with a single scale score in the presence of transient abilities is harmless. However, when transient abilities are not insignificant, such as a paragraph comprehension test, or when a test is intentionally multidimensional, then a single scale score is not a meaningful format to summarize examinee performance. A multidimensional or other appropriate model is needed to summarize examinee performance. Hence, given test data, we need to empirically determine if unidimensional modeling and the resulting single-scale score summary is meaningful. If unidimensional modeling is not appropriate, ways to go about selecting an appropriate model are needed.

The focus of this chapter is to illustrate modeling of dichotomous data. Both unidimensional and multidimensional modeling are considered. In the following sections, assumptions of local independence and dimensionality are defined; several tools for assessing these assumptions will be described, and these tools will be illustrated with several realistic data sets. Based on these tools and indices, guidelines for determining an appropriate model for given data will be delineated.

5.1. Definition of Local Independence and Dimensionality

The purpose of a majority of standardized tests is to measure a single construct, ability, or dimension. Hence, a major question facing any test development, analysis, and interpretation is whether it is appropriate to summarize the performance of an examinee to test items using a single scaled score. That is, can the test be modeled using a monotone, locally independent, unidimensional model? The answer is simple. If the test items are tapping a single construct or one dominant dimension, and if the examinee subpopulation taking the test is homogeneous with respect to the construct being measured, then a single scaled score will summarize examinees' performance on the test. Although the answer is simple, ways of determining that the test indeed is measuring a dominant construct is not so simple. Assuming that the assumption of monotonicity is checked and satisfied during the test development process,[2] let us examine the definitions of local independence and dimensionality.

Let $\mathbf{U_n} = (U_1, U_2, \ldots, U_n)$ denote the item response pattern of a randomly sampled examinee on a test of length n. The random variable U_i takes a value of 1 if the item is correctly answered and 0 if the item is incorrectly answered. Let Θ denote the latent ability, possibly multidimensional, underlying item responses.

Definition 1. The test items $\mathbf{U_n}$ are said to be *locally independent* if

$$\text{Prob}(\mathbf{U_n} = \mathbf{u_n}|\boldsymbol{\Theta} = \boldsymbol{\theta}) = \prod_{i=1}^{n} \text{Prob}(U_i = u_i|\boldsymbol{\theta}) \quad (1)$$

for each response pattern $\mathbf{u_n} = (u_1, u_2, \ldots, u_n)$ and for all $\boldsymbol{\theta}$. That is, conditional on examinee ability, responses to different items are independent.

The dimensionality d of a test $\mathbf{U_n}$ is the minimal dimensionality required for $\boldsymbol{\Theta}$ to produce a model that is both monotone and locally independent (Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996). When $\boldsymbol{\Theta}$ consists of a single component, θ, then the test is said to be unidimensional. The definition of local independence provided above is referred to as the strong local independence (SLI) as it involves complete independence among items conditioned on examinee ability. On the other hand, weak local independence (WLI) involves conditional item pair covariance to be zero for all items pairs. That is, $\text{cov}(U_i, U_j|\theta) = 0$.

Definition 2. The test items $\mathbf{U_n}$ are said to be *weakly locally independent* if

$$\text{Prob}(U_i = u_i, U_j = u_j|\Theta = \theta) = \text{Prob}(U_i = u_i|\Theta = \theta)\text{Prob}(U_j = u_j|\Theta = \theta) \quad (2)$$

for all $n(n-1)/2$ items pairs and for all θ. WLI is also referred to as pairwise local independence (McDonald, 1994, 1997). Obviously, SLI implies WLI. It is

2. Monotonicity of items is established by high positive point-biserial correlation between the item score and the test score.

commonly accepted that, if the unidimensionality can be achieved through pairwise local independence, then unidimensionality is closely approximated through SLI (Stout, 2002).

From a factor-analytic point of view, it is not realistic to construct a strictly unidimensional test. In any test, it is not uncommon to find transient abilities common to one or more items (Humphreys, 1985; Tucker, Koopman, & Linn, 1969). In this sense, unidimensionality refers to the dominant ability measured by the test. The WLI, although very useful for empirical investigation of a dimensional structure underlying test data, does not capture the concept of dominant dimensions underlying data.

Stout (1987, 1990, 2002) theoretically conceptualized the separation of dominant dimensions from inessential or transient dimensions and referred to them as *essential dimensions,* meaning what the test is essentially measuring. Stout (1987) also developed a statistical test of essential unidimensionality. In his conceptual formulation and definition of essential dimensionality, Stout (1990) used the "infinite-length test" abstraction. That is, to understand the structure underlying test data resulting from administering a finite test to a finite group of examinees, Stout derived theoretical foundational results based on the abstraction of an infinite-length test U_∞ administered to a large group of examinees. Using this conceptual framework of infinite-length test, essential dimensionality is defined as follows.

Definition 3. A test U_∞ is essentially unidimensional with respect to the unidimensional latent random variable Θ if, for all θ,

$$\frac{\sum_{1\le i<j\le n} |\mathrm{Cov}(U_i, U_j|\Theta=\theta)|}{\binom{n}{2}} \to 0, \qquad (3)$$

as $n \to \infty$. The above definition implies that the average covariance, in the limit, approaches 0 as the test length increases to ∞. In other words, transient or nonessential traits common to one or more items may result in nonzero conditional covariance. However, the average covariance approaches 0. Essential dimensionality is a weaker form of strict dimensionality based on either SLI or WLI.

The definition of essential dimensionality has further led to theoretical results establishing the usefulness of number-correct score as a consistent estimator of unidimensional ability on the latent true-score scale (Stout, 1990) and to nonparametric estimation of item response functions (Douglas & Cohen, 2001).

5.2. Geometrical Representation of Multidimensional Structure

Although, in reality, dimensionality is determined by test items together with the examinee population taking the test, the geometrical description of items in the latent space provides an intuitive understanding of how item *direction* with respect to the test *direction* contributes to the dimensional structure underlying test data. In explaining the dimensional structure of test items geometrically, only two-dimensional test items are considered.

An item can be geometrically represented by a vector, which, if extended, passes through the origin of a coordinate system. The coordinate axes represent the two dimensions, θ_1 and θ_2, underlying test data. The origin of the coordinate system is the population multidimensional trait-level mean. The direction of the vector represents the θ_1, θ_{2-}, composite that has the maximum discrimination, which is appropriately defined for the model in use. The length of the vector is a measure of the magnitude of the item's discrimination, denoted by MDISC $= (a_1^2 + a_2^2)^{1/2}$, where a_1 and a_2 are the discriminating parameters associated with the two dimensions. The location of the base of the item vector corresponds to that level of multidimensional ability at which the probability of correct response to the item is 0.5. The item vector is orthogonal to the $p = .5$ equiprobability contour (Ackerman, 1996; Reckase, 1997). For example, in a two-dimensional space, items are located only in the first or third quadrants. This is because item discriminations can only take positive values. Easy items are located in the third quadrant and difficult items in the first quadrant. Figure 5.1 shows vector representation of items in a two-dimensional space. Item 1 is an easy item with low discrimination, whereas Item 2 and Item 3 are more difficult and high-discriminating items. The angle direction of the item measured from the θ_1-axis represents a composite of dimensions that the item is best measuring. For example, in a two-dimensional space, if the angle distance of an item from the θ_1-dimension is small, then the item is measuring mostly the θ_1-dimension (Item 3 in Figure 5.1). On the other hand, if the item vector is at 45 degrees, then the item's ability composite measures both dimensions equally (Item 1 in Figure 5.1).

Intuitively speaking, a test of items whose vectors cluster in a narrow sector (i.e., where all the items are measuring similar ability composites) is considered to be essentially unidimensional. If all test items lie on the coordinate axis (as opposed to a narrow sector), then the test would be considered strictly unidimensional.

Figure 5.1 Vector Representation of Two-Dimensional Items

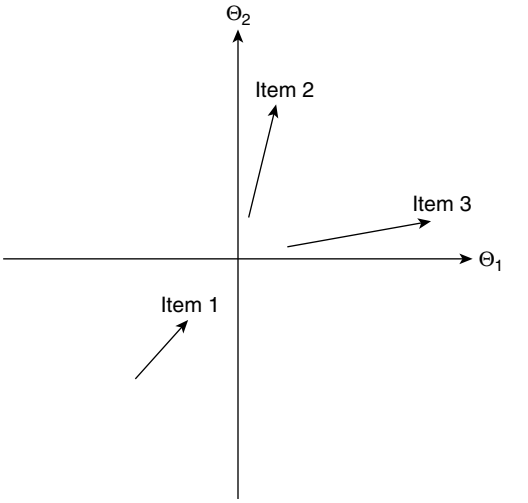
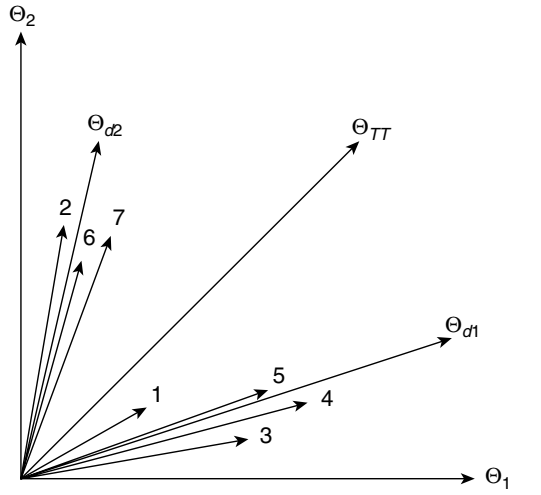


Figure 5.2 An Example of an Approximate Simple Structure Test



The way item vectors cluster together with respect to the coordinate axes, in a multidimensional space, determines the dimensional structure of the test. For the two-dimensional latent space, Figure 5.2 provides an example of a test with two clusters, whose direction of best measurement is represented by vectors Θ_{d1} and Θ_{d2}.[3] The direction of best-measurement vector Θ_{d1} is a weighted average of item discrimination vectors comprising its cluster. The same is true for Θ_{d2}. The direction of best measurement of the total test comprising the two clusters is represented by the vector Θ_{TT}.

A test is considered to have *simple structure* if all items in the test lie along the coordinate axes. In this case, although the dimensional clusters may be correlated, each is an independent item cluster. If, on the other hand, test items are spread along a narrow sector surrounding the coordinate axes, then each narrow sector of items is considered exhibiting an *approximate simple structure*. Figure 5.2 illustrates an example of an approximate simple structures test with two item clusters. Mathematically speaking, *approximate simple structure* can be defined as a k-dimensional latent coordinate axis existing within a d-dimensional latent space ($d \geq k$) such that items only lie within narrow sectors surrounding the coordinate axis. In such a case, there are k-dominant dimensions (Stout et al., 1996).

Zhang and Stout (1999a) have proved theoretical results for using conditional covariances as the basis for determining the dimensional structure underlying multidimensional data. The central theme of their results is that the dimensional structure of test data can be completely discovered using item pair conditional covariances (CCOV), conditional on the test vector represented by Θ_{TT}, provided there is an approximate simple structure underlying test data. The pattern of CCOV_{ij} is positive if items i and j measure similar ability composites, negative if items i and j measure different ability composites, and 0 if one of the items measures the same composite as Θ_{TT}. For example, in the case of a two-dimensional structure, as in Figure 5.2, the CCOV of an item pair is positive if the item vectors in the pair lie on the same side of the conditioning variable's direction of best measurement, Θ_{TT} (e.g., Items 3 and 4). The CCOV is negative if the item vectors lie on the opposite sides of Θ_{TT} (e.g., Items 1 and 2). The CCOV is zero if one of the items lies near the direction of best measurement, Θ_{TT}. This reasoning has been generalized to higher dimensions by Zhang and Stout through $d-1$ dimensional hyperplanes orthogonal to Θ_{TT} and by projecting each item onto this hyperplane.

The magnitude of CCOV indicates the degree of closeness of items' directions of best measurement to each other and their closeness to the conditional vector, Θ_{TT}. CCOV increases as the angle between item pair vectors decreases and as the angle either of the items makes with the Θ_{TT}-axis increases. The CCOV also relates to the degree of discrimination of the vectors. The CCOV increases in proportion to the items' discrimination vectors. Hence, CCOVs form

3. Coordinate axes are not necessarily orthogonal. For example, if $\text{cov}(\Theta_i, \Theta_j) > 0$, then the coordinate axes are not orthogonal.

the basis for establishing the dimensional structure underlying given data. Methods for assessing dimensional structure based on CCOVs are described and illustrated below.

5.3. Methods to Assess the Dimensional Structure Underlying Test Data

This section describes nonparametric methodologies for empirically determining the dimensional structure underlying test data based on CCOVs. It is assumed that one would use these procedures after the test is well developed and its reliability and validity have been established. As explained earlier, it is very important to assess the dimensional structure of the test to determine the test scoring and related issues such as equating and differential item functioning. If the unidimensional model is not appropriate, then recommendations will be made about finding an appropriate model.

Nonparametric tools DIMTEST and DETECT will be used to illustrate the steps involved in determining the correct model for data. We chose these methods because they are not dependent on any particular parametric model for scoring and describing data, and they are simple and easy to use. DIMTEST and DETECT are described below, followed by a flowchart to correctly determine the appropriate model for given data.

5.3.1. DIMTEST

DIMTEST (Stout, 1987; Nandakumar & Stout, 1993; Stout, Froelich, & Gao, 2001) is a nonparametric statistical procedure designed to test the hypothesis that the test data were generated from an LI, $d = 1$ model. The procedure for testing the null hypothesis consists of two steps. In Step 1, n test items are partitioned into two subtests, AT and PT. The AT subtest is of length $m(4 \leq m <$ half the test length), and the PT subtest is of length $n - m$. The AT subtest consists of items that are believed to be dimensionally homogeneous, and the PT subtest consists of the remaining items of the test. One way to select items for AT and PT subtests is through linear factor analysis of the tetrachoric correlation matrix (Froelich, 2000; Hattie, Krakowski, Rogers, & Swaminathan, 1996; Nandakumar & Stout, 1993). This is an automated procedure that uses part of the sample to select items for AT and PT subtests. Items loading on the same dimension are selected into the AT subtest. Expert opinion is another way to select items into these subtests (Seraphine, 2000). Because of the manner in which items are selected, when multidimensionality is present in test data, items in the AT subtest will be predominantly measuring the same unidimensional construct, whereas the remaining items in the PT subtest will be multidimensional in nature. If, on the other hand, the test is essentially unidimensional, then items in both the AT and PT subtests will be measuring the same valid unidimensional construct.

In Step 2, the DIMTEST statistic, T, is computed as follows. Examinees are grouped into subgroups based on their score on the PT subtest consisting of $n - m$ items. The kth subgroup consists of examinees whose total score on the PT subtest, denoted by X_{PT}, is k. In each subgroup k, two variance components, $\hat{\sigma}_k^2$ and $\hat{\sigma}_{U,k}^2$, are computed using items in the AT subtest:

$$\hat{\sigma}_k^2 = \frac{1}{J_k} \sum_{j=1}^{J_k} (Y_j^{(k)} - \bar{Y}^{(k)})^2$$

and

$$\hat{\sigma}_{U,k}^2 = \sum_{i=1}^{m} \hat{P}_i^{(k)} (1 - \hat{p}_i^{(k)}),$$

where

$$Y_j^{(k)} = \sum_{i=1}^{m} U_{ij}^{(k)}, \quad \bar{Y}^{(k)} = \frac{1}{J_k} \sum_{j=1}^{J_k} Y_j^{(k)},$$

$$\hat{p}_i^{(k)} = \frac{1}{J_k} \sum_{j=1}^{J_k} U_{ij}^{(k)},$$

and U_{ij}^k denotes the response of the jth examinee from subgroup k to the ith assessment item in AT, and J_k denotes the number of examinees in the subgroup k. After eliminating sparse subgroups containing too few examinees, let K denote the total number of subgroups used in the computation of the statistic T.

For each examinee subgroup k, compute

$$T_{L,k} = \hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2 = 2 \sum_{i<l \in AT} \widehat{\mathrm{Cov}}(U_i, U_l | X_{PT} = k),$$

where $\widehat{\mathrm{Cov}}(U_i, U_l | X_{PT} = k)$ is an estimate of the covariance between items U_i and U_l for examinees whose score on the PT subtest is k.

The statistic T_L is given by

$$T_L = \frac{\sum_{k=1}^{K} T_{L,k}}{\sqrt{\sum_{k=1}^{K} S_k^2}},$$

where S_k^2 is the appropriately computed asymptotic variance (Nandakumar & Stout, 1993; Stout et al., 2001) of the statistic $T_{L,k}$. For finite test lengths, the statistic T_L is known to exhibit positive bias (Stout, 1987). The positive bias in T_L is eliminated using a bootstrap technique as follows: For each item, an estimate of its unidimensional item response function (IRF) is computed using a kernel-smoothing procedure (Douglas, 1997; Ramsay, 1991). Using the estimated IRFs, examinee responses are generated for each of the items. Using the generated data and the original AT and PT subtest partition, another DIMTEST statistic is computed, denoted by T_G (see Froelich, 2000, for details). This process of the random generation of unidimensional data with kernel-smoothed estimates of items and the computation of T_G is repeated N times, and the average is denoted by $\bar{T}_G$. $\bar{T}_G$ denotes the inflation or bias in T_L that is due to the finite test length administered to a finite sample of examinees. The final bias-corrected DIMTEST statistic T is given by

$$T = \frac{T_L - \bar{T}_G}{\sqrt{(1 + 1/N)}}. \tag{4}$$

The statistic T follows the standard normal distribution as the number of items and the number of examinees tend to infinity. The null hypothesis of unidimensionality is rejected at level α if T is larger than the $100(1-\alpha)$th percentile of the standard normal distribution.

A number of studies have found the DIMTEST to be a reliable and consistent methodology for assessing unidimensionality. It is also extremely powerful compared to other methodologies in its power to detect multidimensionality (Hattie et al., 1996; Nandakumar, 1993, 1994; Nandakumar & Stout, 1993). The current version of DIMTEST, with recent revisions by Stout et al. (2001), is even more powerful than the former version and can be applied on test sizes as small as 15 items.

5.3.2. DETECT

DETECT (Kim, 1994; Zhang & Stout, 1999a, 1999b) is a statistical methodology for determining the multidimensional structure underlying test data. It partitions the test items into clusters in such a manner that items within clusters are dimensionally cohesive. The DETECT methodology uses the theory of conditional covariances to arrive at the partitioning of test items into clusters. As a result, items within a cluster have positive CCOVs with each other; and items from different clusters have negative CCOVs. The DETECT procedure also quantifies the degree of multidimensionality present in given test data. It is important to note that the number of dimensions and the degree of multidimensionality are two distinct pieces of information. For example, one could have a two-dimensional test in which the two item clusters are dimensionally far apart or close together. In the former case, the degree of multidimensionality is more than in the latter case. For example, in Figure 5.2, clusters represented by vectors Θ_{d1} and Θ_{d2} are the two dimensions underlying test data comprising all test items. The angle between these two vectors determines the degree of multidimensionality present in test data. If the angle between vectors Θ_{d1} and Θ_{d2} is small, the degree of multidimensionality present in test data is small, implying that the two clusters are dimensionally similar. If, on the other hand, the angle between the vectors is large, then two item clusters are dimensionally apart.

The theoretical computation of the DETECT index is briefly described here (for details, see Zhang & Stout, 1999b). Let n denote the number of dichotomous items of a test. Let $P = \{A_1, A_2, \ldots, A_k\}$ denote a partition of the n test items into k clusters. The theoretical DETECT index $D(P)$, which gives the degree of multidimensionality of the partition P, is defined as

$$D(P) = \frac{2}{n(n-1)} \times \sum_{1 \le i < j \le N} \delta_{ij} E[\mathrm{Cov}(X_i, X_j | \Theta_{TT} = \theta)], \tag{5}$$

where Θ_{TT} is the test composite, X_i and X_j are scores on items i and j, and

$$\delta_{ij} = \begin{cases} 1 & \text{if items } i \text{ and } j \text{ are in the same cluster of } P \\ -1 & \text{otherwise.} \end{cases} \tag{6}$$

The index $D(P)$ is a measure of the degree of multidimensionality present in the partition P. It is obvious that numerous ways exist to partition items of a test into clusters, and each partition produces a value of $D(P)$. Let P^* be a partition such that $D(P^*) = \max\{D(P) | P \text{ is a partition}\}$. Then P^* is treated as the optimal simple dimensionality structure of the test, and $D(P^*)$ is treated as the maximum amount of multidimensionality present in the test data. For example, for a purely unidimensional test, the optimal dimensionality structure of the test is that all the items will be partitioned into one single cluster, and $D(P^*)$ for the test will be close to 0. It has been shown by Zhang and Stout (1999b) that when there is a true simple structure underlying test data, $D(P)$ will be maximized only for the correct partition.

To determine if the partition P^*, which produced the maximum DETECT index $D(P)$, is indeed the correct simple structure of the test, we can use the following ratio:

$$R(P^*) = \frac{D(P^*)}{\tilde{D}(P^*)}, \tag{7}$$

where

$$\tilde{D}(P^*) = \frac{2}{n(n-1)} \times \sum_{1 \le i < j \le n} |E[\mathrm{Cov}(X_i, X_j | \Theta_{TT} = \theta)]|. \tag{8}$$

When there is an approximate simple structure underlying test data, then the ratio $R(P^*)$ is close to 1. The extent to which $R(P^*)$ differs from 1 is indicative of the degree to which the test structure deviates from the simple structure.

Because the true ability of an examinee is unobservable, $E[\mathrm{Cov}(X_i, X_j|\Theta_{TT} = \theta)]$ of equation (5) cannot be computed directly but must be estimated using observable data. There are two natural estimates of $E[\mathrm{Cov}(X_i, X_j|\Theta_{TT} = \theta)]$:

$$\widehat{\mathrm{Cov}}_{ij}(T) = \sum_{m=0}^{N} \frac{J_m}{J} \widehat{\mathrm{Cov}}(X_i, X_j | T = m), \tag{9}$$

where the conditional score $T = \sum_{l=1}^{N} X_l$ is the total score of all test items, J is the total number of examinees, and J_m is the number of examinees in subgroup m with the total score $T = m$. The other is the estimator based on the total score of remaining items given by

$$\widehat{\mathrm{Cov}}_{ij}(S) = \sum_{m=0}^{N-2} \frac{J_m}{J} \widehat{\mathrm{Cov}}(X_i, X_j | S = m), \tag{10}$$

where the score $S = \sum_{l=1, l \neq i, j}^{N} X_l$ is the total score of the remaining items, other than items i and j, and J_m is the number of examinees in subgroup m with the conditional score $S = m$.

When a test is unidimensional, $\widehat{\mathrm{Cov}}_{ij}(T)$ tends to be negative because items X_i and X_j are part of T. Therefore, $\widehat{\mathrm{Cov}}_{ij}(T)$ as an estimator of $E[\mathrm{Cov}(X_i, X_j | \Theta_T = \theta)]$ results in a negative bias (Junker, 1993; Zhang & Stout, 1999a). On the other hand, $\widehat{\mathrm{Cov}}_{ij}(S)$ tends to be positive and results in a positive bias (Holland & Rosenbaum, 1986; Rosenbaum, 1984; Zhang & Stout, 1999a).

Because $\widehat{\mathrm{Cov}}_{ij}(T)$ tends to have a negative bias and $\widehat{\mathrm{Cov}}_{ij}(S)$ tends to have a positive bias as estimators of $E[\mathrm{Cov}(X_i, X_j|\Theta_T = \theta)]$ in the unidimensional case, Zhang and Stout (1999b) proposed an average of these two estimates, resulting in the following index as an estimator of the theoretical DETECT index $D(P)$:

$$D_{ZS}(P) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le N} \delta_{ij} \widehat{\mathrm{Cov}}^*_{ij}, \tag{11}$$

where

$$\widehat{\mathrm{Cov}}^*_{ij} = \frac{1}{2} [\widehat{\mathrm{Cov}}_{ij}(S) + \widehat{\mathrm{Cov}}_{ij}(T)]. \tag{12}$$

An estimate of $R(P)$ can be similarly obtained. The DETECT software adopts a special technique, called the genetic algorithm, to divide items of a test into different dimensional clusters. The genetic algorithm iteratively mutates items to different dimensional clusters until the maximum degree of multidimensionality of the test $D_{\max}$, an estimate of $D(P^*)$, is obtained. The dimensional cluster pattern that produces $D_{\max}$ is treated as the final dimensionality structure of the test. The process is accelerated when the initial cluster solution for the genetic algorithm is obtained via cluster analysis developed by Roussos, Stout, and Marden (1993).

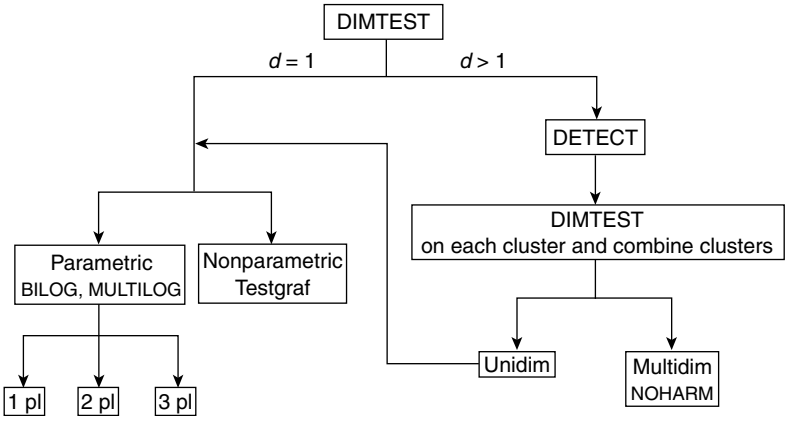
To interpret the results of DETECT in applications, Zhang and Stout (1999b) provided the following rule of thumb based on simulation studies. Divide the examinee sample into two parts: sample1 and sample2 (cross validation sample). Using sample1, find item partition, P_1^*, that maximizes the detect index for sample1, called $D_{\max}$. Using sample2, find P_2^*, that maximizes the detect index for sample2. Then using the item partition P_2^*, from the cross validation sample, compute the detect value for sample1, called D_{ref}. Generally is less than or equal to $D_{\max}$. A test is judged to be essentially unidimensional if D_{ref} is less than 0.1 or $\frac{D_{\max} - D_{ref}}{D_{ref}} > .5$.

5.4. Data Modeling

An algorithm is proposed below to model test data. As emphasized hitherto, the goal is to determine if unidimensional scoring is meaningful for given data. Although any appropriate methodology can be used to carry out the steps of the algorithm, DIMTEST and DETECT are recommended, as they are specifically developed for this purpose, easy to use, and nonparametric.

The flowchart in Figure 5.3 details the steps for test modeling, which are described in the algorithm following the flowchart. These steps are illustrated through the analyses of simulated data in the following section.

Figure 5.3 Flowchart Describing Steps for Test Modeling



5.4.1. An Algorithm for Test Modeling

Step 1. Use DIMTEST to determine if dimensionality, d, underlying test data is essentially 1.

Step 2. If $d = 1$, then fit a unidimensional model to data. Choose an appropriate unidimensional model. Exit.

Step 3. If $d > 1$, then investigate if test items can be decomposed into unidimensional clusters using DETECT.

Step 4. Test each cluster using DIMTEST to determine if $d = 1$.

Step 5. Combine clusters, if necessary, based on expert opinion and item content of the AT subtest of DIMTEST. Again test the hypothesis $d = 1$.

Step 6. If $d = 1$, go to Step 2. If $d > 1$ for any of the clusters, either delete them from the test or explore multidimensional modeling.

If unidimensional modeling is appropriate either on the whole test or on subtests (Step 2), one can fit either a parametric model or a nonparametric model. If a parametric model is desired, there are several models to choose from. Some of the commonly used models are the one-parameter logistic model (1PL), the two-parameter logistic model (2PL), or the three-parameter logistic model (3PL). Parameters of these models can be estimated using standard computer software such as BILOG (Mislevy & Bock, 1989), MULTILOG (Thissen, 1991), and RUMM (Sheridan, Andrich, & Luo, 1998). For more detailed information about fitting various parametric models, estimating parameters, and scoring, refer to Embretson and Reise (2000) and Thissen and Wainer (2001). An alternative is nonparametric modeling. Nonparametric estimation of item response functions can be carried out using the software TESTGRAF (Douglas & Cohen, 2001; Ramsay, 1993). If unidimensional modeling is not appropriate either for the whole test or after splitting into subtests (Step 6), multidimensional modeling of data is necessary. Currently, multidimensional models and estimation of their parameters are limited. One program that has shown a lot of promise in estimating multidimensional parameters is NOHARM (Fraser, 1986). For details about fitting multidimensional models, see Reckase (1997), McDonald (1997), and Ackerman, Neustel, and Humbo (2002).

5.4.2. Illustration of Test Modeling

Data modeling will be illustrated using simulated data. Unidimensional and two-dimensional data were simulated. All data sets had 30 items and 2,000 examinees, which are typical values usually encountered in applications. One unidimensional test and four two-dimensional tests were generated. Unidimensional data were generated using a unidimensional two-parameter logistic model (Hambleton & Swaminathan, 1985).

$$P_i(\theta_j) = \frac{1}{1 + \exp[-1.7[a_i(\theta_j - b_i)]]}, \qquad (13)$$

where $P_i(\theta_j)$ is the probability of a correct response to the dichotomous item i by an examinee with ability (θ_j), a_i is the discrimination parameter of the dichotomous item i, and b_i is the difficulty parameter of item i.

Figure 5.4 Item Vectors Representing the Simple Structure Test

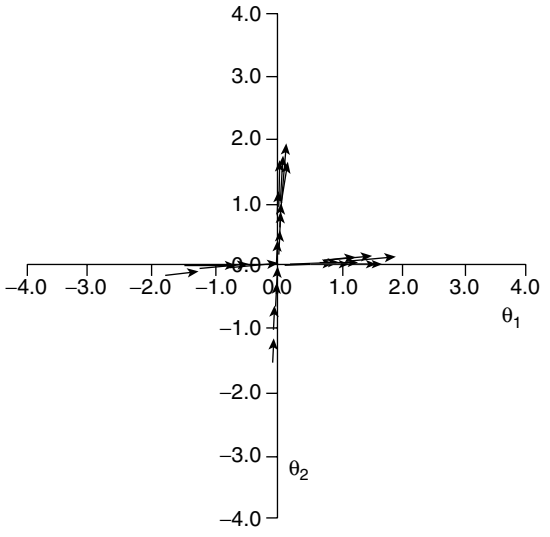
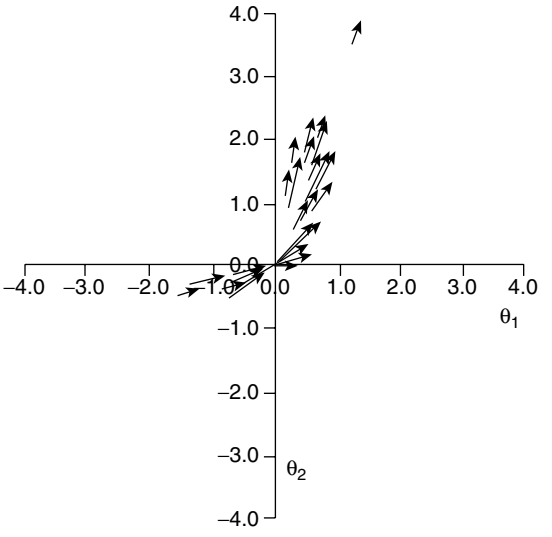


Figure 5.5 Item Vectors Representing the Complex Structure Test



Examinee abilities were randomly generated from the standard normal distribution with mean 0 and the standard deviation 1. Item parameters were randomly selected from a pool of parameter estimates from several nationally administered standardized achievement tests.

Two types of two-dimensional data were generated: simple structure and complex structure. Item parameters for the simple structure were such that items of each dimension were located within 15 degrees from the respective axes, as illustrated in Figure 5.4. Item parameters for the complex structure were selected from a two-dimensional calibration of an American College Test (ACT) mathematics test in which items span the entire two-dimensional space, as illustrated in Figure 5.5.

Two levels of correlation between dimensions (ρ_{θ_1,θ_2}) were considered: .5 and .7. This resulted in four two-dimensional tests: simple structure with $\rho = .5$, simple structure with $\rho = .7$, complex structure with $\rho = .5$, and complex structure with $\rho = .7$. For each two-dimensional test, the first half of the items (Items 1 to 15) measured predominantly the first dimension, and the second half measured predominantly the second dimension. Each examinee's abilities θ_1 and θ_2 were randomly generated from a bivariate normal distribution with an appropriate correlation coefficient between the abilities. Two-dimensional data were generated using the following two-dimensional, two-parameter compensatory model (Reckase, 1997; Reckase & McKinley, 1983):

$$P_i(\theta_{1j}, \theta_{2j}) = \frac{1}{1 + \exp[-1.7(a_{1i}\theta_{1j} + a_{2i}\theta_{2j} + b_i)]}, \tag{14}$$

where $P_i(\theta_{1j}, \theta_{2j})$ is the probability of a correct response to the dichotomous item i by an examinee j with ability $(\theta_{1j}, \theta_{2j})$, a_{1i} is the discrimination parameter of the dichotomous item i on dimension θ_1, a_{2i} is the discrimination parameter of the item i on dimension θ_2, and b_i is the difficulty parameter of item i. The simulated data sets are described in Table 5.1.

5.4.3. Results of Data Analyses

For each data set, the correct model was arrived at by following the steps described in the algorithm for test modeling, as illustrated in Figure 5.3. Results of the analyses are tabulated in Tables 5.2 and 5.3. These results will be summarized below in detail for each of the tests.

Uni.dat: DIMTEST results ($T = 0.85$ and $p = .20$) showed that it is essentially unidimensional. Hence, unidimensional modeling is appropriate for these data.

Table 5.1 Description of Simulated Data

Test	*# Items*	*# Examinees*	ρ^a	*Dimensionality*
uni.dat	30	2,000	—	$d = 1$
simplr5.dat	30	2,000	0.5	$d = 2$, simple structure
simplr7.dat	30	2,000	0.7	$d = 2$, simple structure
realr5.dat	30	2,000	0.5	$d = 2$, complex structure
realr7.dat	30	2,000	0.7	$d = 2$, complex structure

a. Denotes the correlation between latent abilities for two-dimensional tests.

Table 5.2 DIMTEST and DETECT Results

	DIMTEST		*DETECT*			
Test	*T*	*p*	D_{max}	*R*	*# Clusters*	*Item Clusters*
uni.dat	0.85	.20	—	—	—	—
simplr5.dat	9.69	.00	1.33	0.98	2	1–15, 16–30
simplr7.dat	6.0	.00	1.58	0.74	2	1–15, 16–30
realr5.dat	2.63	.00	0.16	0.29	3	(1, 4, 6, 7, 10, 11, 13, 14, 15, 27); (2, 5, 8, 9, 12, 19, 23, 29); (3, 16, 17, 18, 20, 21, 22, 24, 25, 26, 28, 30)
realr7.dat	0.86	.19	—	—	—	—

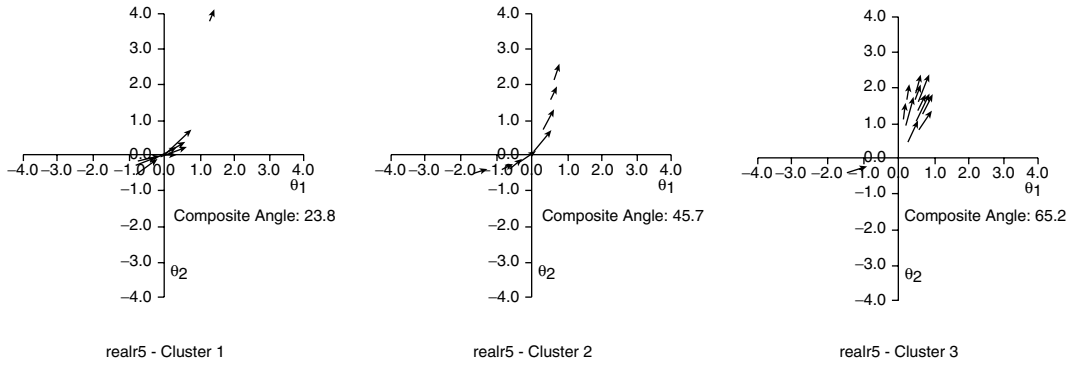
Table 5.3 Further Analyses of Two-Dimensional Data

		DIMTEST		*DETECT*	
Test	*Item Cluster*	*T*	*P*	D_{max}	*R*
simplr5.dat	1–15	–0.77	.78	—	—
	16–30	0.03	.49	—	—
simplr7.dat	1–15	–1.36	.91	—	—
	16–30	0.90	.18	—	—
realr5.dat	1, 4, 6, 7, 10, 11, 13, 14, 15, 27	–.76	.78	—	—
	2, 5, 8, 9, 12, 19, 23, 29	—	—	0.01	0.02
	3, 16, 17, 18, 20, 21, 22, 24, 25, 26, 28, 30	1.04	.15		
realr5.dat clusters 1 and 2	1, 2, 4 to 15, 19, 23, 27, 29	0.52	.30	—	—

simplr5.dat: DIMTEST results ($T = 9.69$ and $p = .00$) indicated the presence of more than one dominant dimension underlying these test data. DETECT analyses resulted in a two-cluster solution with a high value of $D_{\max}$ (1.33) and an R-value close to 1, indicating two dimensions with a simple structure solution. As expected, Items 1 to 15 formed one cluster, and the rest of the items formed the second cluster. Further analyses on these clusters, shown in Table 5.3, showed that each of these clusters is unidimensional ($T = -0.77$ and $p = .78$ for Items 1 to 15; $T = 0.03$ and $p = .49$ for Items 16 to 30). Hence, these subtests are amenable to unidimensional modeling.

simplr7.dat: These test data were also assessed as multidimensional ($T = 6.0$ and $p = .00$) by DIMTEST. DETECT analyses on these data resulted in a two-cluster solution. However, the $D_{\max}$ (0.58) and R-values (0.74) were not high, indicating that the simple structure solution is not as explicit as it was for simplr5.dat. This is due to high correlation between latent abilities. Nonetheless, it is noteworthy that DETECT was able to correctly classify items into clusters given the high degree of correlation between abilities. Further analysis on the clusters, shown in Table 5.3, showed that each cluster is unidimensional ($T = -1.36$ and $p = .91$ for Items 1 to 15; $T = 0.90$ and $p = .18$ for Items 16 to 30).

Figure 5.6 Item Vectors Representing the Three Clusters in the Test: realr5



realr5.dat: DIMTEST results ($T = 2.63$ and $p = .00$) indicated that the data violated the unidimensionality assumption. Subsequent DETECT analyses showed three clusters. Although the DETECT procedure split the test items into three clusters, the corresponding $D_{\max}$ (0.16) and R-values (0.29) were small, indicating that the degree of multidimensionality was not of a concern. In fact, the $D_{\max}$ value was within the range of what is expected for a unidimensional test. Here, the unidimensionality assumption is violated. However, there is not enough evidence of multidimensionality to warrant significant separate clusters.

To understand the nature of multidimensionality, each of the clusters was further analyzed for unidimensionality using DIMTEST. As the results suggest in Table 5.3, Clusters 1 and 3 were confirmed as unidimensional by DIMTEST ($T = -0.76$ and $p = .78$ for Cluster 1; $T = 1.04$ and $p = .15$ for Cluster 3). Because Cluster 2 contained too few items to apply DIMTEST, its dimensionality was estimated using DETECT. Note that the $D_{\max}$ value (0.01) associated with Cluster 3 was very small and resembles a value associated with unidimensional tests. Hence, one may treat this cluster as unidimensional.

DIMTEST also provided clues regarding the source of the multidimensionality. If the null hypothesis of $d = 1$ is rejected, it means that items in the subtest AT are contributing to multidimensionality. Upon observing the AT subtest of DIMTEST results of realr5.dat, it was found that there was an overlap of items between Cluster 3 and the AT subtest. Hence, it was conjectured that Cluster 3 was dimensionally distinct from Clusters 1 and 2. Hence, Clusters 1 and 2 were combined to confirm if the combined subtest is unidimensional. DIMTEST analysis confirmed unidimensionality of this subtest ($T = 0.52$ and $p = .30$). Hence, there are two unidimensional subtests of realr5.dat.

Figure 5.6 shows a graphical display of vector plots of items in the three clusters identified by DETECT. Contrasting Figures 5.5 and 5.6, it can be seen that the item vectors in Figure 5.5 (in which abilities have a correlation of 0.5) are split into three clusters by the DETECT procedure. The test composite vector of Cluster 1 is at 23.8 degrees from the θ_1-axis, the test composite vector of Cluster 2 is at 45.7 degrees from the θ_1-axis, and the test composite vector of Cluster 3 is at 65.2 degrees from the θ_1-axis. Both the DIMTEST and DETECT procedures are sensitive to the differences among these three clusters. As the detailed analyses revealed, Clusters 1 and 2 can be combined to form a unidimensional subtest, whereas Cluster 3 is an independent cluster dimensionally different from the other two clusters.

realr7.dat: DIMTEST analyses of this test revealed unidimensionality ($T = 0.86$ and $p = .19$). This is not surprising as the items span the entire two-dimensional space in which the two abilities are highly correlated. Hence, this group of items is best captured by a unidimensional vector encompassing all items in the space. Unidimensional scoring is the best way to summarize these data.

In summary, unidimensional modeling was appropriate for the following test data: uni.dat and realr7.dat. The former is an inherently unidimensional test, whereas the latter resembles a unidimensional test because of high correlation between abilities coupled with items spanning the entire two-dimensional space, as in Figure 5.5. For both of these tests, the DIMTEST results indicated unidimensionality. Two-dimensional

data sets—simplr5.dat and simplr7.dat, both simple-structure tests—were assessed as multidimensional based on DIMTEST analyses. DETECT results confirmed this fact by indicating a high degree of multidimensionality, as evidenced by large $D_{\max}$ and R-values. It is remarkable that DETECT, despite highly correlated abilities for simplr7.dat, correctly partitioned test items into clusters/subtests. The subtests of simplr5.dat and simplr7.dat were further assessed by DIMTEST as unidimensional. Hence, unidimensional modeling for each of these subtests is meaningful. Among all simulated test data, the dimensionality structure of realr5.dat turned out to be the most complex. For these test data, even though the DIMTEST analyses indicated the presence of multidimensionality, DETECT analyses indicated a very low degree of multidimensionalily. Further investigation and detection of the source of multidimensionality in realr5.dat led to the identification of two subtests, which were each unidimensional. Hence, all three two-dimensional tests could be split into subtests for unidimensional modeling or could be combined for two-dimensional modeling and scoring.

5.5. Summary and Conclusions

The aim of a test is to accurately capture the examinee's position on a continuum of latent trait(s) of interest. To accomplish this, one must use a model that best explains given data, which is an interaction between items and the examinee population taking the test. Most commonly used models to explain test data comprise monotone, local independent, and unidimensional assumptions. However, increasingly, tests are designed to measure more than one dominant trait. Hence, it has become ever more important to empirically investigate the suitability of the unidimensional modeling of test data. This chapter has provided a modeling algorithm using a series of procedures to investigate whether test data are amenable to monotone, local independent, and unidimensional modeling. The proposed algorithm for test modeling was illustrated using simulated test data. Although the algorithm described here provides a framework for test modeling, the process is more of an art than a science. Often, data in the real world may not strictly satisfy the criteria proposed here for test modeling. For example, results of DIMTEST and DETECT may lead to conclusions that test data do not adhere to unidimensional modeling. At the same time, test data may not warrant multidimensional modeling (e.g., realr5.dat). In such a situation, it is important to go beyond statistical analyses and consult content experts and test specifications to decide the most appropriate modeling of test data. Clearly, modeling test data involves many decisions and thus is more a craft than an exact science.

Another important aspect of test modeling is to consider implications of dimensionality considerations. There are well-established methodologies and a choice of software for fitting unidimensional models and estimating parameters of items and examinees. Hence, the selection of multidimensional models over unidimensional models needs careful examination. Other important factors to consider are the cost, improvement in accuracy and understanding of the results, and communication of results with the public.

References

Ackerman, T. (1996). Graphical representation of multidimensional item response theory. *Applied Psychological Measurement, 20,* 311–329.

Ackerman, T. A., Neustel, S., & Humbo, C. (2002, April). *Evaluating indices used to access the goodness-of-fit of the compensatory multidimensional item response model.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Douglas, J. A. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika, 62,* 7–28.

Douglas, J. A., & Cohen, A. (2001). Nonparametric ICC estimation to assess fit of parametric models. *Applied Psychological Measurement, 25,* 234–243.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Fraser, C. (1986). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, Australia: University of New England.

Froelich, A. G. (2000). *Assessing unidimensionality of test items and some asymptotics of parametric item response theory.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Amsterdam: Kluwer Nijhoff.

Hattie, J., Krakowski, K., Rogers, J., & Swaminathan, H. (1996). An assessment of Stout's index of essential dimensionality. *Applied Psychological Measurement, 20,* 1–14.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14,* 1523–1543.

Humphreys, L. G. (1985). General intelligence: An integration of factor, test, and simplex theory. In B. B. Wolman (Ed.), *Handbook of intelligence* (pp. 201–224). New York: John Wiley.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology, 71,* 327–333.

Junker, B. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21,* 1359–1378.

Kim, H. R. (1994). *New techniques for the dimensionality assessment of standardized test data.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 258–269). New York: Springer-Verlag.

Mislevy, R. J., & Bock, R. D. (1989). *BILOG: Item analysis and test scoring with binary logistic models.* Chicago: Scientific Software.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. *Applied Psychological Measurement, 1,* 29–38.

Nandakumar, R. (1994). Assessing latent trait unidimensionality of a set of items: Comparison of different approaches. *Journal of Educational Measurement, 31,* 1–18.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18,* 41–68.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611–630.

Ramsay, J. O. (1993). *TESTGRAF: A program for the graphical analysis of multiple choice test and questionnaire data: TESTGRAF user's guide.* Montreal, Quebec: Department of Psychology, McGill University.

Reckase, M. D. (1997). A liner logistic multidimensional model for dichotomous item response data. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer-Verlag.

Reckase, M. D., & McKinley, R. L. (1983, April). *The definition of difficulty and discrimination for multidimensional item response theory models.* Paper presented at the annual meeting of American Educational Research Association, Montreal, Quebec.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49,* 425–435.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1993, April). *Dimensional and structural analysis of standardized tests using DIMTEST with hierarchical cluster analysis.* Paper presented at the annual NCME meeting, Atlanta, GA.

Seraphine, A. E. (2000). The performance of the Stout T procedure when latent ability and item difficulty distributions differ. *Applied Psychological Measurement, 24,* 82–94.

Sheridan, B., Andrich, D., & Luo, G. (1998). *RUMM: Rasch unidimensional measurement models.* Duncraig, Australia: RUMM Laboratory.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589–617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55,* 293–325.

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67,* 485–518.

Stout, W., Froelich, A. G., & Gao, F. (2001). Using resampling to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 357–376). New York: Springer-Verlag.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20,* 331–354.

Thissen, D. (1991). *MULTILOG user's guide* – (Version 6). Chicago: Scientific Software.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring.* Mahwah, NJ: Lawrence Erlbaum.

Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika, 34,* 421–459.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

Yu, F., & Nandakumar, R. (2001). Poly-Detect for quantifying the degree of multidimensionality of item response data. *Journal of Educational Measurement, 38,* 99–120.

Zhang, J., & Stout, W. F. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64,* 129–152.

Zhang, J., & Stout, W. F. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 213–249.

Chapter 6

Differential Item Functioning Analysis

Detecting DIF Items and Testing DIF Hypotheses

Louis A. Roussos

William Stout

Standardized testing permeates our educational system from primary school to graduate school. In particular, high-stakes standardized testing has spread from postsecondary education down to secondary and primary schools. For example, a total of 19 states currently require students to pass a standardized high school exit exam to graduate. Moreover, the U.S. government has passed the No Child Left Behind Act of 2001, which mandates that all states have in place by 2005–2006 annual tests in reading and math in Grades 3 to 8 (U.S. Department of Education, 2002). Therefore, the educational measurement community must ensure that the highest standards are maintained for the development, administration, scoring, and usage of these tests. One important standard (some would say the *most* important) is that of *test equity.*

Test equity is the assurance of test validity in regard to particular subgroups of the test-taking population. In other words, all subgroups of the population should experience equally valid assessment for test takers of equal proficiency on the construct or constructs intended to be measured by the test. (For the rest of this chapter, we will restrict our discussion to tests intended to measure a single dominant construct—by far the most common situation.) The most typical subgroups of interest are based on ethnicity or gender; but groups based on other variables, such as instructional background or testwiseness, would clearly also be considered worthy of study.

Test equity is primarily achieved by ensuring that a test measures only *construct-relevant* differences between subgroups of the test-taking population (Messick, 1989). Procedures that help ensure test equity are implemented at every stage in the life of a test: development, administration, scoring, and usage. This chapter focuses on one particular procedure—differential item functioning (DIF) analysis. DIF analysis is most helpful when it occurs at the test development stage, but it is also (sometimes *only*) used at the test scoring stage.

DIF is said to occur in a test item when test takers of equal proficiency on the construct intended to be measured by a test, but from separate subgroups of the population, differ in their expected score on the item. (For ease of exposition, most of our examples and discussion will be for tests consisting of items that are dichotomously scored, although the techniques we describe apply just as easily to polytomously

scored items.) DIF items give an unfair advantage to one group over another. DIF clearly involves both substantive (e.g., the construct the test is intended to measure) and statistical (e.g., statistically significant differences) components. Unfortunately, throughout most of the history of DIF statistics, they have been used in almost total isolation from the substantive aspects. Until recently, the standard DIF procedure has been the application of a DIF statistic in an automatic one-item-at-a-time purely statistical analysis.

In this chapter, we first review DIF terminology and nomenclature, including a broad framework for organizing the various theoretical, statistical, and practical components of DIF analysis. Next we focus on the practical implementation of DIF analysis procedures. Here we first describe the traditional approach to DIF analysis, reviewing its successes as well as its limitations, and then we present a detailed description of a more sophisticated DIF analysis approach based on the latest advancements in DIF analysis research, which addresses the limitations of the traditional approach. Next, we briefly review recent research articles and papers that demonstrate how the latest advances in DIF analysis have resulted in new and significant progress in understanding the root causes of DIF and, thus, in increasing test equity for takers of standardized tests. Finally, we summarize the chapter and encourage an optimal (and practical) approach to DIF analysis that combines the advantages of the simpler traditional procedure with the advantages accrued from employing the more sophisticated procedure.

6.1. DIF TERMINOLOGY

For dichotomously scored items, DIF occurs in an item when test takers of equal proficiency on the construct the test is intended to measure, but from separate subgroups of the test-taking population, differ in their probability of a correct response on the item.

In current practice, standardized tests yield scores on a unidimensional scale. The construct that the test is intended to measure is the construct that corresponds to the substantive interpretation given to the test score. This construct is referred to as the *primary dimension* of the test. The term *dimension* here is used to refer to any substantive characteristic of an item that can affect the probability of a correct response on the item. It is generally accepted that all tests, in truth, measure multiple dimensions, but the primary dimension is the only one that all the items have in common. The remaining dimensions on the test are referred to as *secondary dimensions.* Each secondary dimension is measured by a (usually small) minority of the test items.

The item that is being tested for DIF is commonly referred to as the *studied item.* The items whose scores are used to match the test takers on the primary dimension of the test are called the *matching subtest* or the *matching criterion.* The subgroups of interest for DIF analyses are most commonly based on ethnicity or gender. The subgroups are typically studied in pairs, with one group labeled the *reference group* (e.g., Caucasians or males) and the other group labeled the *focal group* (e.g., various minority groups or females). The term *focal* refers to the particular group of interest for the DIF analysis, and *reference* refers to the group with whom the focal group (or groups) is to be compared. When multiple items are being studied as a set to test the statistical significance of the sum of their individual DIF estimates, the set of items is referred to as a studied item *bundle,* and the sum of their DIF estimates is referred to as an estimate of differential *bundle* functioning, or *DBF.*

It has long been recognized and accepted that the general cause of DIF is the presence of multidimensionality in items displaying DIF (e.g., see Ackerman, 1992); that is, such items measure at least one secondary dimension in addition to the primary dimension that the item is intended to measure. Although the presence of DIF automatically implies the presence of a secondary dimension, the presence of a secondary dimension does *not* automatically imply the presence of DIF. Some secondary dimensions cause DIF and some do not, depending on how the reference and focal groups differ in their proficiency on the secondary dimension. This is discussed more below, and the reader is referred to Ackerman (1992) and Roussos and Stout (1996) for more in-depth discussions.

When secondary dimensions do cause DIF, they are further categorized as either an *auxiliary* dimension or a *nuisance* dimension. An auxiliary dimension is a secondary dimension that is intended to be measured by the item (perhaps as mandated by test specifications and typically closely associated with the primary dimension), whereas a nuisance dimension is a secondary dimension that is not intended to be measured by the item (e.g., the context of a word problem in a situation where the context is not included in the test specifications). DIF that is caused by an auxiliary dimension is referred to as *benign* DIF, whereas DIF caused by a nuisance dimension is referred to as *adverse* DIF. DIF caused by an auxiliary dimension is considered benign because the item is *intended* to measure the auxiliary dimension; however, as pointed out by Linn (1993), "The burden should be on those

who want to retain an item with high DIF to provide a justification in terms of the intended purposes of the test" (p. 353). Moreover, benign DIF is not necessarily ignorable if auxiliary dimensions with large benign DIF can be replaced by equally valid auxiliary dimensions with less DIF or if the distribution of items across the various auxiliary dimensions can be modified to reduce the overall amount of benign DIF. DIF caused by a nuisance dimension is considered adverse because the difference in probability of a correct response on the item between different groups is due purely to group differences on an irrelevant construct.

Before we discuss specific implementation procedures for conducting DIF analyses, it is important to recognize that the explication of an implementation procedure is the last step of a three-step process in the development of a DIF analysis approach. The three steps are as follows:

1. Conceptualization of a DIF parameter
2. Formulation of a DIF statistic
3. Implementation of a DIF analysis procedure

In terms of item response theory (IRT), there exist infinitely many ways to specify a parameter that represents the amount of DIF in an item. It is important that this parameter be explicitly specified so that researchers can test whether a corresponding DIF statistic effectively estimates the parameter (see Roussos, Schnipke, & Pashley, 1999, for a striking example of how a faulty parameter can have unforeseen harmful consequences). Thus, the next step is to develop a statistic to estimate the DIF parameter. The statistic should be thoroughly investigated both theoretically (to ensure that its expected value approaches the value of the DIF parameter as the number of items and number of examinees increases) and in simulation studies (to document its Type 1 error and power rates). The last step is the explication of a procedure for how to carry out DIF analysis with real data. This last step is the focus of this chapter. To describe the implementation of DIF analysis procedures clearly requires some reference to and discussion of the DIF parameters and the DIF statistics, but the reader will be referred to appropriate references for detailed discussion of these parameters and statistics.

6.2. DIF Analysis Procedures

In our description of DIF analysis procedures, we will differentiate between two settings: a "stand-alone" test and a "linked" test. We use the term *stand-alone* to refer to a test that is developed, administered, and scored without any formal statistical connection to any other test. All the items on such a test are intended to be scored. A pilot study may or may not be carried out before the test is first used for scoring purposes. After the test has been administered once, any further administrations use the same items as in the first administration, with the notable exception of any items found to be faulty in the first administration.

We use the term *linked* to refer to a test that is linked through pretest items to other tests in a chain of statistically equated tests. Specifically, when such a test is administered, it is composed of two types of items: operational items and pretest items. The operational items are ones that have already been pretested with earlier administered tests in the chain and have been found to be high-quality items. The operational items are the ones that the test takers' scores will be based on. The pretest items are ones that have not been previously administered and are being tested out to see if they are of high enough quality to be used in a future administration. Test taker performance on the pretest items does not contribute to the test taker's reported test score. The pretest items are also used to ensure that every test has some common items with at least one other test, and these common items can be used to maintain a common scale across the chain of tests by employing IRT equating methodology. (See Lord, 1980, for an introduction to IRT equating.)

For purposes of detecting DIF, the use of linked tests is preferred over stand-alone tests because items can be tested for DIF before they are presented operationally. The use of linked tests is a common practice in the development of standardized tests by major testing companies. However, in many situations, pretesting items is not practical or feasible, but it is still important that DIF analysis be conducted.

6.2.1. Traditional DIF Analysis Procedure

6.2.1.1. Stand-Alone Tests

This DIF analysis is conducted the first time the test is administered, whether in a pilot study or in an operational setting. The traditional approach is often referred to as a "one-item-at-a-time" DIF analysis because each item is individually tested for DIF, with the matching criterion being the remaining items on the test.

The approach can be summarized by the following steps:

1. Calculate a DIF statistic and its standard error for each item on the test.

2. If the magnitude of the DIF is large and statistically significantly larger than a prescribed negligible level, then the item is flagged as displaying unacceptably high DIF.

The traditional method may also include a "purification" step in which the above analysis is redone using a matching criterion consisting only of those items that were not flagged in the initial analysis:

3. Remove from the matching criterion all items flagged in Step 2. This creates a "purified" matching criterion.

4. Repeat Steps 1 and 2 above. One might restrict the studied items to be only those flagged in Step 2 above, or one may choose to use all the items as studied items.

Finally, a decision is then made as to whether or which flagged items will be discarded:

5. The practitioner either automatically discards all the flagged items, or the flagged items are investigated by an item review committee and discarded only if the committee can agree on a substantive explanation for why the DIF occurred. This latter approach is favored when the cost of item replacement is expensive.

Conceivably, any established DIF statistic could be used to carry out this traditional DIF analysis; however, the Mantel-Haenszel (MH) DIF statistic (Holland & Thayer, 1988), denoted by $\hat{\Delta}$, has been the one most commonly used. Note that when the MH DIF statistic is used, the score on the studied item is included in the matching criterion. Because the MH statistic has been commonly used, standard criteria (see, e.g., Zieky, 1993) have been developed for Step 2, although, in truth, the criteria were not intended to be used outside of the testing programs (at Educational Testing Service) that they were originally developed for. The criterion that would be used in Step 2 above is as follows: An item is flagged for DIF when $|\hat{\Delta}| \geq 1.5$ and $|\hat{\Delta}|$ is significantly greater than 1.0 in the sense of statistical hypothesis testing. In other words, for this particular setting, a value of $|\hat{\Delta}|$ greater than 1.5 is interpreted to indicate a large amount of DIF, and a value less than 1.0 is interpreted to be negligible. So, if $|\hat{\Delta}|$ is large and significantly greater than a negligible amount, the item is flagged for unacceptably large DIF.

Similar rules could be established for other DIF statistics. Probably the most notable example would be that of the SIBTEST statistic (MH and SIBTEST are probably the two most thoroughly tested statistics), whose DIF estimate is denoted by $\hat{\beta}$. Based on Dorans (1989), a reasonable rule in this context would be the following: An item is flagged for DIF when $|\hat{\beta}| \geq 0.100$ and $|\hat{\beta}|$ is significantly greater than 0.050 in terms of hypothesis testing.

These rules for MH and SIBTEST are rather arbitrary. A quite fertile area for future research is the development of methods to help practitioners come up with more appropriate guidelines for particular testing situations. For ease of exposition, in the rest of this chapter, we will employ the above arbitrary rules (and corresponding rules for moderate DIF, as will be seen below) while also reminding the reader of the need for further research in this area.

6.2.1.2. Linked Tests

In this setting, for a given test administration, DIF analysis is conducted on both the pretest and operational items, although the process differs in significant ways for the two types of items. However, for both types of items, the analysis still follows the same general one-item-at-a-time approach.

6.2.1.2.1. Operational items. For the operational items, the approach is almost the same as described above for the stand-alone test. The main differences are that the purification process is generally not used and DIF items are not automatically discarded—they are reviewed by a committee, and an item is discarded only if the committee agrees on an identified problem with the item. The raising of the bar for throwing out an operational item is introduced because the operational items have already passed a DIF test as pretest items, and the cost of throwing out items once they reach the operational stage is quite high.

6.2.1.2.2. Pretest items. Thus, in this setting, it is very important to flag DIF items at the pretest stage, when the cost of throwing out an item is much lower than at the operational stage. At the pretest stage, it is desirable to have a more liberal approach to flagging DIF items so as to ensure that few large DIF items ever make it to the operational stage.

Although the general procedure is again similar to that described above for the stand-alone test, there are a number of notable differences. One major difference is that the matching criterion is the operational items. Having an external (i.e., external to the pretest studied items) matching criterion is a major advantage because every studied item has the same matching criterion, which makes for a more valid statistical analysis. When the matching criterion is simply the other studied items (an internal matching criterion), the DIF estimates are artificially constrained to approximately sum to zero,

no matter what the true levels of DIF are in the items. Another advantage of using the operational items as the matching criterion is that the operational items have already been tested previously for DIF and are thus likely to have only negligible values of DIF associated with them.

A second difference is that no purification of the matching criterion is used. This is because the matching criterion is the operational items that have already been screened for DIF when they were pretest items.

A third difference is that an item that displays large DIF and is significantly larger than negligible DIF would be *automatically* discarded.

Moreover, an additional rule that is sometimes used is that if an item displays moderate DIF and is significantly larger than zero DIF, then, if possible, the item is replaced by another item having a smaller DIF estimate.

Because there are so many differences between the DIF procedure for the pretest items as compared with the DIF procedure for a stand-alone test, we provide a separate summary of the step-by-step procedure for pretest items:

1. Calculate a DIF statistic and its standard error for each pretest item on the test using the operational items as the matching criterion.
2. If the magnitude of the DIF is large and statistically significantly larger than a prescribed negligible level, then the item is flagged as displaying unacceptably high DIF. If the magnitude of the DIF is moderate and statistically significantly larger than zero, the item is flagged as displaying moderately high DIF.
3. The practitioner automatically discards all the flagged high DIF items. The flagged moderately high DIF items are replaced, when possible, by items with lower DIF estimates.

The standard criterion (see Zieky, 1993) that has been developed for flagging moderate DIF items in Step 2 using the MH statistic is as follows: An item is flagged for moderately high DIF when $|\hat{\Delta}| \geq 1.0$ and $|\hat{\Delta}|$ is significantly greater than 0.0 in the sense of statistical hypothesis testing. Similarly, for SIBTEST, an item is flagged for moderately high DIF when $|\hat{\beta}| \geq 0.050$ and $|\hat{\beta}|$ is significantly greater than 0.0 in terms of hypothesis testing.

6.2.1.3. Strengths and Limitations of the Traditional Approach

The major strength of the traditional approach described above is that it provided the first statistically rigorous DIF analysis procedure. The traditional approach gave DIF analysis a strong statistical foundation. Second, even though the DIF effect size rules for the MH statistic were not intended for general consumption, the specification of one set of rules is a major accomplishment that provides an important benchmark for the many situations where no rules have yet to be established. Furthermore, the introduction of some substantive review of flagged DIF items made it possible for some understanding of the root causes of DIF to begin to be formed and potentially provide feedback to the test development process.

As DIF research progressed over the years since the introduction of the MH statistic and the concurrent development of the traditional DIF analysis procedure, a number of important limitations of the procedure have become apparent. First, the procedure is restricted to analyzing only one item at a time (an inherent limitation of the MH statistic). Statistically greater power can be gained by analyzing bundles of items, *if the bundles are carefully selected.* Hence, merely introducing the analysis of item bundles would be a meaningless adjustment without addressing an even more important limitation: The traditional procedure focuses on testing items for DIF rather than testing for the root causes of DIF. Even though the root causes of DIF are known to flow from the presence of multidimensionality, dimensionality considerations come into the traditional approach in only a very limited role—when substantive analyses of flagged items are conducted to try to determine the cause of DIF in individual items.

The one-item-at-a-time statistical analysis that dominates the traditional DIF analysis procedure is an essential component in DIF analysis, but by incorporating both statistical and substantive dimensionality analysis considerations into DIF analysis, it can be transformed from merely testing items for DIF to testing secondary dimensions for DIF/DBF.

6.2.2. Latest Developments in DIF Analysis Procedures

Roussos and Stout (1996) introduced a new multidimensionality-based DIF analysis procedure that integrates dimensionality analysis with DIF analysis at both the substantive and statistical levels. The resulting DIF analysis procedure focuses on finding the root causes of DIF by testing secondary dimensions through item bundles. Thus, this DIF analysis procedure can be described by two simple steps:

1. Development of DIF hypotheses
2. Testing of DIF hypotheses

6.2.2.1. Developing DIF Hypotheses

The first step is the development of hypotheses about whether particular substantive item characteristics will cause DIF. One natural way to accomplish this is by first identifying substantive item characteristics and then determining, if possible, if either the reference group or one of the corresponding focal groups would be expected to be favored based on the Shealy-Stout multidimensional model for DIF (Shealy & Stout, 1993), as described in Roussos and Stout (1996).

Another way to develop DIF hypotheses is from purely theoretical substantive considerations. A good example of this is a study by Gierl, Bisanz, Bisanz, and Boughton (2002), who reference a cognitive theory that implies that certain substantive item characteristics should result in DIF favoring females, whereas other characteristics should result in DIF favoring males.

The identification of item characteristics that may represent potentially DIF-causing secondary dimensions can come about in at least three general ways:

1. by item-writing specialists reading test items and using their expert judgment,
2. by flagging DIF items in a traditional DIF analysis implementation procedure and having item-writing specialists inspect the wordings of these items,
3. by conducting exploratory statistical dimensionality analyses and substantively inspecting the results. (See Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996, for a review of the latest advancements in nonparametric dimensionality analyses.)

A fuller discussion of methods that can be used to develop DIF hypotheses can be found in Roussos and Stout (1996). Furthermore, these methods have now been extensively demonstrated in research articles and papers such as the following: Douglas, Roussos, and Stout (1996); Walker and Beretvas (2001); Stout et al. (2003); Bolt (2000, 2002); Gierl and Kaliq (2001); McCarty, Oshima, and Raju (2002); Ryan and Chiu (2001); and Gierl et al. (2002).

Once these secondary dimensions have been identified, the multidimensional model for DIF described in Roussos and Stout (1996) is used to see if a directional DIF hypothesis can be conjectured. According to this model, when the mean proficiency on the secondary dimension, conditional on proficiency on the primary dimension, is greater for one group as compared to another, then the potential exists for DIF favoring the first group. Identifying potentially DIF-causing secondary dimensions is clearly easier to carry out than conjecturing about whether one group has a greater mean proficiency on a conditional distribution than another group, so that two-tail DIF hypotheses (a secondary dimension is identified, but it is not known which group would be favored by it) are the most typical ones developed. Still, Bolt (2002) has begun investigating a promising line of research in this regard employing estimation of a parametric multidimensional IRT model.

6.2.2.2. Testing of DIF Hypotheses

Once item characteristics that represent potentially DIF-causing secondary dimensions have been identified, item bundles are formed in which all the items in a given bundle share a common characteristic suspected of causing DIF. Because an item may contain several characteristics that it may share with other items, these bundles may exhibit some overlap in terms of the items they contain. These item bundles are then tested for DBF using an appropriate DBF statistic, such as SIBTEST (the MH statistic cannot be applied to item bundles).

The procedure by which these DIF hypotheses are tested does vary depending on whether the test is a stand-alone test or a linked test, similar to how the traditional procedure is varied. Note that the development of the DIF hypotheses is not restricted by whether the test is stand-alone or linked. DIF hypotheses are theoretical constructs that may arise from any test or situation and then be applied to other tests.

For stand-alone tests, a general framework for the testing of DIF hypotheses is as follows:

1. Read the items and label them according to whether they exhibit each of the hypothesized secondary dimensions of interest from the DIF hypotheses.
2. Form item bundles according to these potentially DIF-causing secondary dimensions.
3. The matching criterion will be the items that are not identified as measuring a potential DIF-causing secondary dimension. If the number of these items is too small (e.g., fewer than 20 or perhaps fewer than 15), then the matching criterion for each item bundle would be simply the remaining items on the test.
4. Calculate the DBF estimate and standard error for each item bundle and test if it is statistically significant (relative to zero DIF).
5. Item bundles having DBF estimates that are statistically significant represent secondary dimensions that exhibit strong evidence of causing DIF. The

DBF estimates, divided by the number of items in the bundle, provide a rough estimate of the average amount of DIF per item that this secondary dimension causes when it seems to be present in an item. More sophisticated analyses of the DIF indices based on analysis of variance can also be used to more accurately account for the overlapping bundles in estimating the DIF effect size for a secondary dimension—the reader is referred to Stout et al. (2003) and Bolt (2000) for two detailed examples.

6. If a DIF secondary dimension is an auxiliary dimension, the DIF is labeled *benign,* whereas if the secondary dimension is a nuisance dimension, the DIF is labeled *adverse.* Depending on how large the DIF effect size is and the type of DIF, different actions would be considered in response to the identified DIF dimensions. If the DIF is statistically significant but small, then probably no action should be taken other than to document the finding for future reference. Significant large adverse DIF would clearly call for item reviewers and item writers to be alerted so that they can ensure that items measuring this DIF dimension will be avoided on future tests. For significant large benign DIF, test development staff should be alerted to the finding so that they can keep this in mind during the test assembly process and perhaps develop new methods or test specifications that can minimize the use of such auxiliary dimensions. Indeed, automatic test assembly programs could include auxiliary dimension DBF as a variable that is constrained to be below a certain value.

For linked tests, the above stand-alone procedure might be carried out on the operational items. For the pretest items, the above procedure would be carried out on the pretest items using the operational items as the matching criterion. It should be noted here that DBF effect size criteria for item bundles are still a research area that needs further study.

6.2.2.3. Examples of Progress in Identifying Causes of DIF

Here we present brief examples of how the latest advancements in DIF analysis procedures focusing on developing and testing DIF hypotheses have resulted in significant progress in identifying the causes of DIF on standardized tests.

1. Bolt (2000) analyzed pretest items that were specially designed to test preformed DIF hypotheses about gender DIF on an SAT math test. In particular, he discovered that when items are presented in multiple-choice format as opposed to an open-ended format, the items exhibit DIF in favor of males. Another DIF hypothesis he was able to confirm was that DIF in favor of males also occurred for concrete-type items as opposed to abstract-type items. In both cases, however, the DIF effect sizes were clearly small enough to not be of concern.

2. Walker and Beretvas (2001) analyzed fourth-grade and seventh-grade math tests and confirmed another preformed DIF hypothesis that an item bundle consisting of open-ended math items that require students to communicate in writing about their solution would favor proficient writers over nonproficient writers. Their findings led to concrete recommendations for improving the fairness of the scoring of the open-ended items and also improving the communication of the test results to the teachers (and, hence, leading to improved instruction).

3. Stout et al. (2003) analyzed Graduate Record Examination (GRE) math pretest and operational data and identified 15 secondary dimensions (mostly auxiliary dimensions) using a combination of substantive and statistical dimensionality analyses. Using two different very large sets of pretest data, they were able to test the DIF hypotheses for consistency in a cross-validation study. Their results showed a remarkably high consistency for both item bundles that exhibited statistical rejection and those that exhibited nonrejection.

4. At the 2002 annual meeting of the National Council on Measurement in Education in New Orleans, Louisiana, an entire symposium was devoted to "New Approaches for Identifying and Interpreting Differential Bundle Functioning." As part of this symposium, a paper presented by Gierl et al. (2002) investigated preformed DIF hypotheses based on a cognitive theory about gender differences in mathematical problem solving. Their DBF analyses indicated strong support for the hypothesis that items requiring significant spatial processing show substantial DIF in favor of males. Moreover, the existence of this secondary dimension was supported by both DBF *and* dimensionality analyses.

5. At this same symposium, another paper by McCarty et al. (2002) analyzed item bundles on a survey instrument for rater DIF between parents and teachers on particular secondary dimensions. They found that teachers are more strict than parents in their ratings of assertive behaviors of children, whereas teachers are more lenient than parents in rating cooperative and self-control behaviors of children. Test developers can use this information to better tailor

survey questions for specific types of raters so as to minimize the manifest DIF.

Clearly, when DIF analysis procedures incorporate the Roussos and Stout (1996) multidimensionality-based DIF analysis paradigm focusing on the testing of secondary dimensions in the form of item bundles, significant progress can be achieved in identifying DIF-causing secondary dimensions and estimating the amount of DIF they may cause.

6.3. Summing Up: A More Complete Approach to DIF Analysis

The general purpose for conducting a DIF analysis is to help ensure test equity. The statistical flagging of items that exhibit evidence of DIF represents an essential contribution toward the achievement of this objective. Because tests are inherently multidimensional and multidimensionality is the basic cause of DIF, increased understanding of test multidimensionality and the effects of these dimensions on DIF hold the potential for a more accurate interpretation of the test score, more control over the influence of relevant auxiliary dimensions, and the reduction of influence by unintended and irrelevant nuisance dimensions.

Thus, the optimal approach for a DIF analysis procedure would seem to be one that incorporates the immediate critical goal of detecting DIF items, which is the focus of the traditional DIF analysis approach, and the longer range goal of identifying the DIF secondary dimensions, which is the focus of the more recent advancements that have been accomplished in DIF research.

It is important to note that the process of test design and development already involves consideration of a wide variety of substantive item characteristics through the item review processes (including the review of flagged DIF items and the sensitivity review of items for offensive language that could cause DIF) and the creation and implementation of test specifications, and these identified characteristics provide a ready source of secondary dimensions for DIF hypotheses. Also, in the case of linked tests, the large number of pretest items that are typically tested provides a more than adequate pool for forming item bundles for these hypotheses. Moreover, pretest items are already frequently used for research purposes so that some of these pretest slots can be reserved for controlled testing of DIF hypotheses (e.g., see Bolt, 2000).

Thus, the advantages of increased understanding of DIF secondary dimensions by augmenting the traditional DIF analysis implementation procedure with the developing and testing of DIF hypotheses do not necessarily involve any significant increase in expense. The inclusion of the developing and testing of DIF hypotheses in a DIF analysis implementation procedure often involves merely increased awareness that the hypotheses already exist and can be easily tested.

References

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67–91.

Bolt, D. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement, 37,* 307–327.

Bolt, D. (2002, April). *Studying the DIF potential of nuisance dimensions using bundle DIF and multidimensional IRT analyses.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2,* 217–233.

Douglas, J., Roussos, L. A., & Stout, W. F. (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement, 33,* 465–485.

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2002, April). *Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the DIF analysis framework.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Gierl, M. J., & Kaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement, 38,* 164–187.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 349–364). Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

McCarty, F. A., Oshima, T. C., & Raju, N. (2002, April). *Identifying possible sources of differential bundle functioning with polytomously scored data.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24,* 293–322.

Roussos, L. A., & Stout, W. F. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355–371.

Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education, 14,* 73–90.

Shealy, R., & Stout, W. F. (1993). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–239). Hillsdale, NJ: Lawrence Erlbaum.

Stout, W. F., Bolt, D., Froelich, A. G., Habing, B., Hartz, S. M., & Roussos, L. A. (2003). *Development of a SIBTEST bundle methodology for improving test equity with applications for GRE test development* (GRE Board Professional Rep. No. 98–15P, ETS Research Rep. 03–06). Princeton, NJ: Educational Testing Service.

Stout, W. F., Habing, B., Douglas, J., Kim, H. R., Roussos, L. A., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20,* 331–354.

U.S. Department of Education. (2002). *Draft regulations to implement Part A of Title I of the Elementary Secondary Education Act of 1965 as amended by the No Child Left Behind Act of 2001.* Washington, DC: Author.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement, 38,* 147–163.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Lawrence Erlbaum.

Chapter 7

UNDERSTANDING COMPUTERIZED ADAPTIVE TESTING

From Robbins-Monro to Lord and Beyond

HUA-HUA CHANG

7.1. OVERVIEW

Computerized adaptive testing (CAT) has become a popular mode of educational assessment in the United States. Examples of large-scale CATs include the Graduate Record Examination (GRE), the Graduate Management Admission Test (GMAT), the National Council of State Boards of Nursing, and the Armed Services Vocational Aptitude Battery (ASVAB).

A CAT test differs profoundly from a paper-and-pencil (P&P) test. In the former, different examinees are tested with different sets of items. In the latter, all examinees are tested with an identical set of items. The major goal of CAT is to measure the trait levels of examinees (θs) with greater precision than conventional P&P tests by building an individualized test for each examinee. Each examinee's latent trait level is fit precisely by selecting test items sequentially from a large item pool according to the current performance of an examinee. In other words, the test is tailored to each examinee's θ level, thus matching the difficulties of the items to the examinee being measured. Clever examinees can avoid responding to too many easy items, and less clever examinees can avoid being exposed to too many difficult items. So, the examinees are always challenged during the entire course of the testing. The major advantage of CAT is that it provides more efficient latent trait estimates (θ) with fewer items than would be required in conventional tests (e.g., Weiss, 1982).

Although the implementation of CATs has led to many advantages, such as new question formats, new types of skills that can be measured, easier and faster data analysis, and faster score reporting, many issues related to CATs are not well understood. One of them is the compatibility between CAT and P&P tests. It has been widely speculated that some examinees may get much lower scores than they would normally do if an alternative P&P version were given. According to Carlson (2000), in 2000, Educational Testing Service (ETS) found that the GRE CAT system does not produce reliable scores for about half of 1% of test takers. ETS offered them a chance to retake the test at no charge. However, examinees currently required to take the GRE are not given a choice between the standard P&P version of the tests and the CAT versions. Since the late 1990s, the GRE testing program has made a complete transition from P&P to CAT in the United States. Thus, without effective remedial

AUTHOR'S NOTE: The writing of this chapter is partially supported by the NSF PJ SES-0241020 "Improving Computerized Adaptive Testing in the U.S."

measures, this could significantly undermine the credibility of CAT.

Another important issue in the development and implementation of CAT is about test security and item pool usage. Wainer et al. (2000) noted that the basic notion of an adaptive test is to mimic automatically what a wise examiner would do. In doing so, certain types of items tend to be always selected by the computers, and many items are not selected at all, thereby making item exposure rates quite uneven. Because CATs are usually administered to small groups of examinees at frequent time intervals, examinees who take tests earlier may share information with examinees who will take tests later, escalating the risk that many items may become known.

In 1994, Kaplan Educational Centers sent its employees several times to take the GRE to memorize as many items as possible and to report those items back to Kaplan. Within a short period of time, Kaplan discovered that most of the items its employees collected were already on the list of compromised items. Kaplan notified ETS about the incident. Due to the large portion of the item pool made known to Kaplan, ETS temporarily shut down testing while new items were developed (Davey & Nering, 2002).

As the Kaplan-GRE event so clearly indicated, the major security weakness of CAT lies with *continuous testing.* Today, the CAT GRE is administered more than 100 days each year, whereas the conversional P&P version is administered only three times per year. Indeed, it has been nearly 10 years since the Kaplan-ETS incident, and people have just started to realize how vulnerable the item pools could be to organized item thievery during the period in which those pools are being used. On August 6, 2002, following an investigation that uncovered a number of Asian-language Web sites offering questions from live versions of the computer-based GRE General Test, ETS suspended the CAT GRE General Test and reintroduced P&P-based versions in China, Hong Kong, Taiwan, and Korea (www.ets.org, August 20, 2002).

In this chapter, our major goal is to address issues related to CAT test compatibility and security. To find the root of the problems, we need to understand some general principals and fundamental assumptions of *sequential design* from which the theoretical development CAT is based on. Likewise, we need to understand how it works for today's most commonly used Fisher information procedure, which has been adopted by some major testing programs, including the GRE and GMAT. Moreover, as already noted, our discussions here will focus exclusively on *issues* and *problems* instead of *advantages* and *achievements* of CAT, which have been reported extensively elsewhere.

7.2. Item Selections in CAT

The most important component in CAT is the item selection procedure that is used to select items during the course of the test. Suppose θ is the latent trait to be measured for a specific examinee. According to Lord (1970), an examinee is measured most effectively when test items are neither too difficult nor too easy. The dilemma is how to select such n test items from an item pool so that the examinee's corresponding responses will enable us to estimate θ as efficiently as possible. Heuristically, if the examinee answers an item correctly, the next item selected should be more difficult; if the answer is incorrect, the next item should be easier. This is referred to as the *branching rule* (see Lord, 1970). However, to carry out the branching rule, one must precalibrate all items in the pool according to their psychometric characteristics, such as *item difficulty, item discrimination,* and *guessing probability.*

7.2.1. Models for CAT

7.2.1.1. The Three-Parameter Logistic Model

The most commonly used model in CAT application is the three-parameter logistic model (1) described below. Let X_j be the score for a randomly selected examinee on the jth item, with $X_j = 1$ if the answer is correct and $X_j = 0$ if incorrect, and let $X_j = 1$ with probability $P_j(\theta)$ and $X_j = 0$ with probability $1 - P_j(\theta)$, where $P_j(\theta)$ denotes the probability of a correct response for a randomly chosen examinee of latent trait θ that is,

$$P_j(\theta) = P\{X_j = 1|\theta\},$$

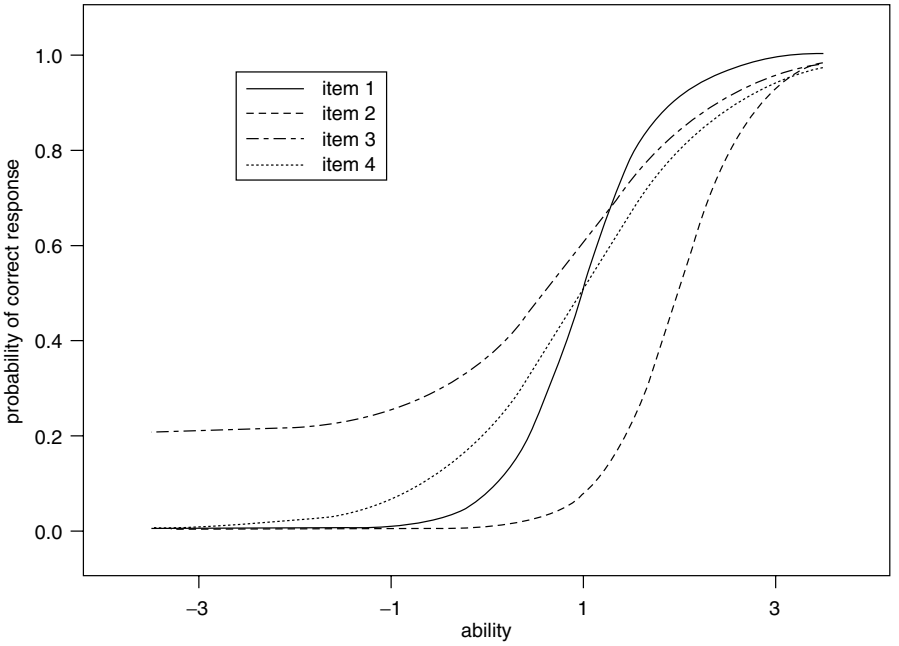
where θ is unknown and has the domain $(-\infty, \infty)$ or some subinterval on $(-\infty, \infty)$. When the three-parameter logistic model (3PL) is used, the probability becomes

$$P_j(\theta) = c_j + (1 - c_j)\frac{1}{1 + e^{-a_j(\theta - b_j)}}, \qquad (1)$$

where

a_j is the item discrimination parameter,
b_j is the difficulty parameter,
c_j is the guessing parameter.

Figure 7.1 Four Items With Item Parameters



NOTE: Item 1: $a = 1.5$, $b = 1.0$, $c = 0.0$; Item 2: $a = 1.5$, $b = 2.0$, $c = 0.0$; Item 3: $a = 0.5$, $b = 1.0$, $c = 0.2$; Item 4: $a = 0.8$, $b = 1.0$, $c = 0.0$.

Figure 7.1 shows item response functions for four hypothetical items. The horizontal axis presents the latent trait scale (θ), and the vertical axis corresponds to $P_i(\theta)$. Items 1 and 2 have the largest discrimination parameters, and their shapes are "steeper." Items 3 and 4 have smaller discrimination parameters, and their curves increase more slowly. Item 2 is the most difficult because it has the largest b-value, and Item 3 is the easiest since it has the smallest b-value. Item 3 has a guessing parameter $c = 0.2$, indicating the probability that a correct response occurs by guessing for low–ability level examinees.

There are two special cases. One is the two-parameter logistic model (2PL), in which $c_i \equiv 0$. The other is the one-parameter logistic model (1PL), in which $c_i \equiv 0$ and a_i is a fixed constant for all items.

According to the probability model, a difficult item will have large b-value, and an easy item will have small b-value. Knowing the difficulty levels of all the items in the pool, one can possibly develop an item selection algorithm based on branching. For instance, if the examinee answers an item incorrectly, the next item to be selected should have a lower b-value. By the same token, if he or she answers correctly, the next item should have a higher b-value. However, two fundamental questions need to be addressed to explain how the algorithm works: (a) how much the b-value should be varied from item to item and (b) how to score the responses after the items have been administered.

Let $b_1, b_2, \ldots, b_n$ be a sequence of the difficulty parameters after administering n items to the examinee. The new items should be selected such that b_n approaches to a constant b_0 as n is indefinitely large, where b_0 represents the difficulty level of an item that the examinee has about a 50% chance of answering correctly. Because for the 1PL and 2PL models, the probability for a randomly sampled examinee with θ to answer an item correctly is 0.5 (given that $\theta = b_0$), knowing b_0 is equivalent to knowing θ. Mathematically speaking,

$$b_n \to b_0, \text{ as } n \to \infty \tag{2}$$

where b_0 is the item difficulty level such that $P\{X = 1|\theta = b_0\} = 0.5$. If this happens, the item selection strategy will allow us to pinpoint the difficulty level at which the examinee answers half the items correctly. The convergence of b_n to b_0 in (2) indicates that at the beginning of the test, differences in bs may vary greatly from item to item, and these differences will be gradually diminished to reach a level of approximately equal difficulty. This implies that b_0 is a reasonable guess for θ, and thus we can

characterize θ in terms of the item difficulty level. Because our goal is to estimate θ, we can use b_0 as the score for the examinee's responses. Notice that b_0 can be linearly transformed to any meaningful score scale, which makes it convenient for us to score the examinee's test responses by a function of b_0.

The above process is called a Robbins-Monro process, and Lord is the first person who introduced the Robbins-Monro process in application in adaptive testing.

7.2.2. Robbins-Monro Process

The stochastic approximation of Robbins and Monro (1951) is a sequential design scheme for locating a point of interest, which is usually formulized as the zero of an unknown regression function. Let b denote the design point and x the corresponding response, and m is the mean of x, which is a function of b. Robbins and Monro proposed using b_n, which is generated from the following recursion,

$$b_{n+1} = b_n - \delta_n x_n, \tag{3}$$

to approximate the root m, where δ_n is a sequence of preassigned constants. Robbins and Monro showed that, with δ_n properly chosen, the sequentially determined b_n converges to the root of m. Numerous further refinements have been developed since the pioneering work, and this simple stochastic approximation in (3) has inspired many important applications, including those in engineering (Goodwin, Ramage, & Caines, 1980), biomedical science (Finney, 1978), and education (Lord, 1970).

Lord (1970) proposed several procedures as applications of the Robbins and Monro process, and one of them is described in the following equation:

$$b_{n+1} = b_n + d_n(x_n - m), \tag{4}$$

where x_n is the item response on the nth item ($x_n = 1$ if the answer is correct, $x_n = 0$ if the answer is incorrect); $d_1, d_2, \ldots$ is a decreasing sequence of positive numbers chosen before the testing; and m is a predetermined constant, say, $m = 0.5$. Assume the item pool is so rich that we can select any b-value from the range of $(-\infty, +\infty)$. Equation (4) indicates that the difficulty level of the $(n + 1)$th item to be selected is determined from that of the nth item plus $d_n/2$ if the answer is correct or minus $d_n/2$ otherwise. If d_1 is not too small, according to Hodges and Lehmann (1956), the sequence of d can be chosen as

$$d_i = d_1/i, \ i = 2, 3, \ldots \tag{5}$$

A point to be made here is that the sequence of $b_1, b_2, b_3, \ldots$ constructed from (4) will converge to b_0, where b_0 can be interpreted as the difficulty level of an item that the examinee will have a 50% chance to answer correctly.

In application of the Robbins-Monro process, it is essential to know the b-values for all the items in the pool, and the pool should be rich in b to such an extent that with any given b_{n+1} defined in (4), there is a corresponding item with the difficulty level b_{n+1}. Interestingly, the convergence of b_n to b_0 does not require strong assumptions, including that of the local independence or the assumption that the exact shapes of the item characteristic curves be known as described by (1). Because the design point is only b_n for the Robbins-Monro process, the guessing parameters and discrimination parameters defined in the 3PL model are not needed.

7.2.3. Lord's Maximum-Information Process

Let $\hat{\theta}_n$ be an estimator of θ based on n responses. $\hat{\theta}_n$ is called a consistent estimator if it converges to θ as n goes to ∞. Consistency is the most important property in our adaptive design because our objective is to identify the unknown θ. To this end, the Robbins-Monro process is handy and can be adequately used to construct a consistent estimator to the desired point for θ. As demonstrated by Lord (1970), the conditions for convergence can be approximated in practice. However, the size of the item pool must be large so that it contains various values of b. On the other hand, the speed of convergence may not be fast. In other words, it may need many items for $\hat{\theta}_n$ to be close to θ. Another problem frequently encountered in CAT design is efficiency. In addition to consistency, we would like to know whether our estimator has the smallest sample variance among other consistent estimators. In doing so, we need to compare the efficiency of different methods for estimating an examinee's ability, and hence it becomes necessary to include more information in our adaptive design, such as the exact shapes of the item response functions and the information functions as well.

7.2.3.1. Some Preliminary Conditions

One of the most important assumptions in item response theory (IRT) is *local independence,* which is defined in the following.

Definition 1. A test $X = (X_1, X_2, \ldots, X_n)$ is said to be locally independent with respect to a latent variable Θ if, for all $x = (x_1, x_2, \ldots, x_n)$ and θ,

$$P\{X = x|\theta\} = \prod_{i=1}^{n} P\{X_i = x_i|\Theta = \theta\}.$$

Suppose that an examinee with a fixed θ is given n items $X_2, X_2, \ldots, X_n$. According to the local independence assumption, the likelihood function can be expressed as

$$L_n(\theta) = \prod_{i=1}^{n} P_i(\theta)^{X_i} Q_i(\theta)^{1-X_i}, \tag{6}$$

where $Q_i(\theta) = 1 - P_i(\theta)$. Then, θ can be estimated by maximizing the likelihood function. Let $\hat{\theta}_n$ denote the resulting estimator. It is clear that $\hat{\theta}_n$ also solves the following maximum likelihood estimating equation:

$$\begin{aligned} U_n(\theta) &= \frac{\partial}{\partial\theta} \log L_n(\theta) \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log \frac{P_i(\theta)}{Q_i(\theta)} [X_i - P_i(\theta)] = 0. \end{aligned} \tag{7}$$

For conventional paper-and-pencil tests, it is well known that, under suitable regularity conditions, including the local independence condition, $\hat{\theta}_n$ is consistent and asymptotically normal, centered at the true θ, and with variance approximated by $I^{-1}(\hat{\theta}_n)$, where $I(\theta)$ is the Fisher test information function. Under the local independence condition, an important feature of $I(\theta)$ is that the contribution of each item to the total information is additive:

$$I(\theta) = \sum_{i=1}^{n} I_j(\theta), \tag{8}$$

where $I_j(\theta)$ is Fisher item information for item j, which is defined as

$$I_j(\theta) = \left[\frac{\partial P_j(\theta)}{\partial\theta}\right]^2 / P_j(\theta)[1 - P_j(\theta)].$$

Thus, under the local independence assumption, the total amount of information for a test can be readily determined. This feature is highly desirable in CATs because it enables test developers to separately calculate the information for each item and combine them to form updated test information at each stage. To make the sample variance of $\hat{\theta}_n$ small, we can sequentially select n items so that their information at $\hat{\theta}_j$, $j = 1, 2, \ldots, n$, is as large as possible.

7.2.3.2. The Maximum-Information Approach

Lord (1970) proposed a standard approach to item selection in CAT, which is to select the item with the maximum Fisher item information as the next item. Note that the original motivation for adaptive testing is to match items with the examinee's trait level θ (Lord, 1970). Under the 3PL model, maximizing Fisher information means intuitively matching item difficulty parameter values with the latent trait level of an examinee. Because the latent trait is unknown, the optimal item selection rule cannot be implemented but may be approximated using the updated estimate $\hat{\theta}$ each time a new item is to be selected. This is essentially the basic design behind Lord's original proposal of adaptive testing. Under the maximum-information-based design, items with high a-parameters will be preferentially selected.

The motivation for maximizing the Fisher information is to make $\hat{\theta}_n$ the most efficient. This can be achieved by recursively estimating θ with current available data and assigning further items adaptively. Note that in IRT, the large sample properties of $\hat{\theta}_n$, such as consistency and asymptotic normality, were established under the local independence assumption. In adaptive design, the selection of the next item is dependent on the basis of the examinee's responses to the items previously administered. Thus, the likelihood function may not be expressed as equation (6) (see Mislevy & Chang, 2000, for a detailed discussion about the local independence assumption in adaptive testing). Therefore, it is necessary to establish the corresponding large sample properties for the maximum-information approach. Chang and Ying (in press) showed that the asymptotic normality and the validity of using Fisher information to estimate variance continue to hold under the adaptive item allocation of CAT. Their result indicates that, for the 1PL model ($c_j \equiv 0$ and $a_j \equiv 1$ in equation (1)) and an infinitely large item pool, the maximum likelihood estimator of θ with maximum-information item selection is strongly consistent. For the 2PL ($c_j \equiv 0$ in equation (1)), consistency holds under the realistic assumption that the discrimination parameters are bounded. For the 3PL model, the same results hold under some reasonable regularity conditions, such as a bound on the guessing parameter being met and the likelihood equations not having a multiple solution.

The maximum-information method is more efficient than the Robbins-Monro process. For large n, the Fisher information measures the effectiveness of the estimator because the reciprocal of the Fisher information is the asymptotic lower bound of the sample

variance of $\hat{\theta}$. $\mathrm{Var}(\hat{\theta}_n) \rightarrow 1/I(\theta)$ as $n \rightarrow \infty$ under suitable conditions.

7.3. Limitations of the Maximum-Information Approach

The assumption of an "infinitely large item pool" never holds in reality. An operational item pool usually consists of several hundred items. Furthermore, the set of items selected for each examinee must satisfy certain nonstatistical constraints such as content balance. The more constraints one has to impose, the fewer degrees of freedom one can have in a design. To design a good CAT algorithm, we need further analytical study.

7.3.1. Constraint of Item Exposure Control

Under the maximum-information-based design, items with high a-parameters will be preferentially selected. In the simple case when all items follow $c = 0$, Fisher information becomes

$$I_j(\theta) = \frac{a_j^2 e^{a_j(\theta - b_j)}}{[1 + e^{a_j(\theta - b_j)}]^2}. \tag{9}$$

Suppose the examinee's true ability is θ_0. For a fixed a_j, Fisher information reached the maximum $a^2/4$ at $b_j = \theta_0$. Thus, if the true ability is known, the information approach tends to select an item with b_j close to θ_0 and a_j as large as possible. The rationale is that this leads to a substantial efficiency gain (Hau & Chang, 2001).

If information-based item selection methods are used, items with high a-parameters might be frequently exposed, whereas others might never be exposed. The exposure rate of an item is defined as the ratio between the number of times the item is administered and the total number of examinees. Because CATs are administered frequently to small groups of examinees, there is a risk that items with high exposure rates might become known to examinees. Thus, item exposure rates must be controlled (e.g., Hau & Chang, 2001; Mills & Stocking, 1996).

Remedies to restrain overexposure of high a-items have been proposed by McBride and Martin (1983), Sympson and Hetter (1985), Stocking and Lewis (1995), Davey and Parshall (1995), Thomasson (1995), and others. The most common method for controlling exposure rate was developed by Sympson and Hetter (SH), whose general idea is to put a "filter" between selection and administration—an item that is selected by the maximum-information criterion is evaluated to determine whether it will be administered. In so doing, the exposure rate can be kept within a certain prescribed value. Let us name the item selection method that maximizes the Fisher information while imposing that the SH exposure control the FSH method. Obviously, by restraining the actual use of the frequently chosen items, the FSH method puts a cap on the exposure rates of "popular" items and effectively keeps them within some desirable thresholds.

The SH control method suppresses the usage of the most overexposed items and spreads their usage over the next tier of overexposed items (e.g., see Chang & Ying, 1999). According to Hau and Chang (2001), the weakness of the SH control method is that it does not proactively raise the exposure of the least exposed items. The FSH method has a number of limitations. The selection rule guides the computer to choose items with certain specific characteristics (e.g., high discrimination); however, the FSH method devises a mechanism to suppress the chance that these items would be chosen so that they will not be overused. Because of these contradictory guidelines, the ability estimation efficiency of FSH is lower than the original Fisher method.

7.3.2. Should Low-Discrimination Items Ever Be Used?

If the computer algorithm only selects high a-items, we may have to force item writers to generate only high a-items. However, when item writers produce items, the items will follow certain distribution characteristics. Item writers may control some of the characteristics such as item content and item difficulty level, but it is extremely challenging to produce only highly discriminating items. As Mills and Stocking (1996) indicate, current testing programs are under greater pressure to produce the "best" items for CAT at a faster rate than for traditional P&P tests. The common practice to generate more relatively high a-items is to discard items whose a-parameter values are lower than a given threshold.

On the other hand, many items in the item pool still will never be selected by the computer. Once items are included in the pool, they have already undergone certain rigorous review processes and shown no problems. Items with relatively lower discrimination parameters are still of good quality and should be used. In practice, however, most item exposure control procedures currently available have failed to

yield more balanced item usage within a pool. Wainer (2000) examined the item usage within the GRE CAT pools and found that as few as 12% of the available items could account for as much as 50% of the functional pool (those items were actually administered). Obviously, one way to resolve this problem is to increase the usage of lower a-items.

7.3.3. When Should Low-Discriminating Items Be Used?

Ideally, all items in a pool should have similar exposure rates to meet the requirements of test security and efficiency of item usage as well (Hau & Chang, 2001; Mills & Stocking, 1996). When should lower a-items be used? The study of Hau and Chang (2001) confirmed that FSH administrated items with larger a-values at the earlier stages of testing. As testing progressed, more items with smaller a-values were selected. This practice follows the philosophy of maximizing Fisher information. Hau and Chang referred to this as the *descending a-method.* It is crucial to know when a low a-item should be used. For the purpose of accuracy, low a-items should be used first because estimation of θ could be inaccurate early in the test.

According to (9), for the 2PL model, items with high a-values and b-values close to the examinee's true θ provide the most information. This is true also for the 3PL model. Thus, more accurate $\hat{\theta}$s allow items with high a-values to provide more information. Using the highest a-items at the beginning of the testing may cause the underestimation problem. One major factor of uncertainty comes from inaccurate estimation of the latent trait in the initial stages when the number of administered items is small. This could result in grossly underestimating θ at early stages. To educate researchers in the field of CAT research and development, one must demonstrate certain strong evidence from analytical derivation.

7.3.4. Is Item Information Always Maximized When $\theta \approx b$?

To build a CAT system, one must specify a certain mathematical model for item response functions. The logistic and normal ogive models are the two most commonly used models in CAT research and implementation. Lord (1980) proved that for the logistic model, the corresponding item Fisher information is unimodal. Actually, $I_j(\theta)$ reaches the maximum value at $\theta = b_j$ for the 1PL and 2PL models and $\theta = b_i + \frac{1}{a_i}\ln(\frac{1}{2} + \frac{\sqrt{1+8c_i}}{2})$ for the 3PL model, respectively. According to Bickel, Buyske, Chang, and Ying (2001), this property also holds for the normal ogive model. Thus, for the two models, the item information function reaches the maximum value when θ is close to the difficulty parameter. Therefore, the maximum-information approach is equivalent to the basic design behind Lord's original proposal of adaptive testing. The fundamental assumption about the equivalence between matching ability with difficulty and maximizing information means that $I_j(\theta)$ reaches the maximum value when $\theta \approx b$.

However, logistic and normal ogive are just two convenient mathematical models for the true underlying item response functions, so it is important to check whether this fundamental assumption holds for a more general class of IRT models, which includes the logistic and the normal ogive models as special cases. This CAT delivery modeling issue should be addressed in light of Lord's (1970) maximum-information approach. Bickel et al. (2001) studied the sensitivity of the maximum-information item selection strategy to the assumed item response function modeling family. They show that two item response functional families that are similar in shape can in fact have different information optimizing strategies. If the IRT model uses one type of functional family, one obtains the usual optimal item selection rule of choosing an item with difficulty close to current estimated examinee ability. But if the IRT model uses the other type of functional family, one obtains the counterintuitive optimal item selection rule of choosing an item with difficulty as far away from the estimated ability as possible. Although we do not understand this study from a practical perspective, it suggests possible overreliance on the maximum-information item selection approach.

7.4. Revealing the Cause for Underestimation

Chang and Ying (2002) made an attempt to quantitatively reveal the cause that is most likely to account for the underestimation phenomenon for the CAT GRE exam reported by the *Chronicle of Higher Education* (Carlson, 2000). One major factor of uncertainty comes from inaccurate estimation of the latent trait in the initial stages when the number of administered items is small. Because the

maximum-information-based methods rely too heavily on the items administered at the initial stages, this could result in grossly underestimating the latent trait at early stages. Chang and Ying's analytical derivation shows that, under the current item selection strategy adopted by ETS, if an examinee failed a few items at the beginning of the test, easy (but more discriminating) items are likely to be administered, and such items are ineffective at bringing the estimate close to the true θ, unless the test is sufficiently long or a variable-length test is used. Their derivations show that a certain weighting mechanism is necessary to make the algorithm rely less on the items administered at the beginning of the test.

7.4.1. Information and MLE

Suppose that an examinee with a fixed θ is given n items $X_1, X_2, \ldots, X_n$. Then, θ can be estimated by maximizing the likelihood function specified in (6). Let $\hat{\theta}_n$ denote the resulting estimator. It is clear that $\hat{\theta}_n$ also solves the following maximum likelihood estimating equation (7).

It is well known that, under suitable regularity conditions, $\hat{\theta}_n$ is asymptotically normal, centered at the true θ, and with variance approximated by $I_n^{-1}(\hat{\theta}_n)$, where $I_n(\theta)$ is the Fisher information function.

An original motivation for CAT is to maximize the Fisher information so as to make $\hat{\theta}_n$ the most accurate. This can be achieved by recursively estimating θ with current available data and assigning further items adaptively.

7.4.2. Sensitivity of $\hat{\theta}_n$ for Small n

Chang and Ying (2002) proposed a way to illustrate the sensitivity of $\hat{\theta}_n$ on initial items in CAT. Their goal is to motivate remedies for the underestimation problem while promoting learning to CAT developers. Let us only demonstrate the case of the 2PL model. Although the cases for the 1PL and 3PL models are similar to that of the 2PL model, the 2PL model is more convenient to show where the problem is. See Chang and Ying for the discussions of 1PL and 3PL models. For the 2PL model, the Fisher test information function becomes

$$I_n(\theta) = \sum_{i=1}^{n} a_i^2 \frac{e^{a_i(\theta - b_i)}}{[1 + e^{a_i(\theta - b_i)}]^2}, \tag{10}$$

and the likelihood estimation function takes the form

$$U_n(\theta) = \sum_{i=1}^{n} a_i \left(X_i - \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \right) \tag{11}$$

after n items were administered. For the maximum likelihood estimator, $\hat{\theta}_n$, $U_n(\hat{\theta}_n) = 0$. Chang and Ying proved

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \frac{a_{n+1}}{I_{n+1}(\theta^*_{n+1})} \left(X_{n+1} - \frac{e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}}{1 + e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}} \right), \tag{12}$$

where $\hat{\theta}_n$ is the current estimator, $\hat{\theta}_{n+1}$ is the next estimator, b_{n+1} is the b-parameter of the $(n + 1)$th item, and θ^* is a point between $\hat{\theta}_n$ and $\hat{\theta}_{n+1}$. If the item pool is sufficiently rich, allowing each given θ to match a difficulty parameter b with the same value, then $b_{n+1} \approx \hat{\theta}_n$ or $e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}/(1 + e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}) \approx \frac{1}{2}$. This entails that the one-step update from $\hat{\theta}_n$ to $\hat{\theta}_{n+1}$ is $\pm\frac{1}{2}$ multiplied by $a_{n+1} I_{n+1}^{-1}(\theta^*_{n+1})$, which indicates that the size of the step may be determined by the value of a for small n. Consequently, the larger the n is, the smaller the one-step adjustment it gets. As indicated earlier, the maximum-information approach would select the items with the highest a-values, which may cause a big step size at the beginning of the test. Therefore, it is plausible that if the examinee misses a number of initial items and the test length is short to moderate, then he or she may not be able to regain a score (estimate) comparable (close) to the true θ, even though he or she responds well to the rest of the items.

Through their analytic derivations and simulation studies, Chang and Ying (1999) argued that, provided the necessary constraints are met, the a-parameter should be selected in an ascending order. Their motivations come from the considerations of efficiency improvement and item exposure balance. In view of (12), an additional benefit of the a-stratified approach of Chang and Ying is that it automatically adjusts step sizes in updating the current estimation of θ. Specifically, it shrinks weights at early stages, making it less likely to have extreme values in estimating θ. It also inflates weights at final stages, counteracting the effect of the multiplier $I_{n+1}^{-1}(\theta^*_{n+1})$ and making it more likely to adjust the final estimator of θ.

7.4.3. Overestimation Is Also Possible

In the 2000 GRE incident, even though ETS refused to comment on whether the examinees who were offered to retake the GRE were scored lower or higher (Carlson, 2000), our speculation is that they most likely received extremely low scores. One explanation for such conjecture is that examinees who received high scores would most likely not take the offer to retake

the test; in other words, it does not make sense to ask them to retake the test. If this speculation stands, then in 2000, about the same number of examinees as those who received the ETS offer were scored higher than what they deserved. According to equation (12), it is possible that a person who guesses correctly early in the test could have overestimated. More specifically, under the current design, large a-items are used first, allowing for big moves in estimates of ability. Thus, a person who guesses correctly at early stages of the test could obtain a high ability-level estimate, even though he or she did not do well for the remainder of the test. Actually, during the period from 2000 to 2002, "Never miss the first five items" was advised by several GRE preparation Web sites in China (e.g., www.taisha.org).

7.5. Alternative Approaches

An alternative to the maximum-information approach is the Bayesian method (e.g., Owen, 1975). Instead of using item information at $\hat{\theta}$, the Bayesian approach uses the posterior variance as the criterion for item selection. At the initial stages, posterior distributions depend heavily on the choice of prior distribution for θ, but the dependency diminishes at the later stages. Furthermore, according to Chang and Stout (1993), the posterior variance approaches the reciprocal of the test information when the number of items becomes large. For those who are interested in other selection models, see Folk and Smith (2002) for details.

7.5.1. Procedures to Deal With Early Stage Estimation

Several procedures have been proposed that deal with large estimation error at the beginning of the test. Chang and Ying (1996) suggested replacing Fisher information by Kullback-Leibler information. Generally, Kullback-Leibler information measures the "distance" between two likelihoods. The larger the Kullback-Leibler information, the easier it is to discriminate between two likelihoods. Veerkamp and Berger (1997) suggest using weighted Fisher's information with the likelihood function and selecting the kth item according to the maximum integrated information. van der Linden (1998) recommends using a Bayesian criteria for item selection that involves some form of weighting based on the posterior distribution of θ. Because the posterior distribution is a combination of the likelihood function and a prior distribution, the basic difference with the previous criterion is the assumption of a prior distribution.

7.5.2. The a-Stratified Method

On the basis of global information theory (Chang & Ying, 1996), Chang and Ying (1999) propose the ascending a-stratified item selection method. A simple version of the a-stratified method can be described as follows:

1. Partition the item pool into K levels according to item a-values.
2. Partition the test into K stages.
3. In the kth stage, select n_K items from the kth level based on the similarity between b and $\hat{\theta}$, then administer the items (note that $n_1 + n_2 +, \ldots, + n_K$ equals the test length).
4. Repeat Step 3 from $k = 1, 2, \ldots, K$.

The rationale behind the a-stratified method is that, because the accuracy of $\hat{\theta}$ generally becomes greater as the test progresses, one effective testing strategy is to stratify the item bank into levels according to item a-values and then partition the test into corresponding stages. That is, items from the lowest $a-$level would be administered at the early stages of the test, and those from the highest level would be administered at the last stage of the test. At each stage, only items from the corresponding level are selected.

Item pool stratification also affects item exposure rates. As indicated by Chang and Ying (1999), one major cause of unevenly distributed item exposure rates is that when using maximum-information item selection, items with large a-values are more likely to be selected than those with small a-values. By grouping items with similar a-values together and selecting within a group at each stage, exposure rates would be more evenly distributed because items with all a-values would be selected with equal frequency. Stratification would, therefore, both decrease exposure rates of high a-items and increase exposure rates of low a-items.

The a-stratified design has received positive remarks from many researchers. Davey and Nering (2002) indicate the following:

> Highly discriminating items are like a tightly focused spotlight that shines intensely but casts little light outside a narrow bean. Less discriminating items are more like floodlights that illuminate a wide area but not too brightly. The idea of Chang and Ying is to use the floodlights early on to search out and roughly locate

> the examinee, then switch to spotlights to inspect things more closely. (p. 181)

However, this method also received criticism. Stocking (1998) indicates,

> If the suggested approach worked well on real pools constructed for CATs, then the time-consuming iterations required to develop stable exposure control parameters in such exposure control approaches as Hetter and Sympson (1997), Stocking and Lewis (1998), and Davey and Nering (1998) could be eliminated. This might also be accompanied by more efficient pool usage in that items that are seldom or never used might be used with greater frequency.

Stocking's (1998) criticisms are basically from the following three aspects:

1. There is concern regarding the correlation between the item difficulty and item discrimination parameters, and this relationship may interfere with the predicted operating characteristic of CAT test designs, which depends on stratification of the pool on item discrimination (Stocking, 1998).
2. The *a*-stratified design did not incorporate the ability to handle item content (Stocking, 1998).
3. Other criticisms include the lack of guidelines regarding the number of strata to use as well as the number of items to administer from each stratum (Stocking, 1998).

The *a*-stratified method proposed in Chang and Ying (1999) is solely a prototype version. Their initial studies were too simplistic, and they did not address such designing issues as what the best set of stratification properties might be or whether these characteristics are general or dependent on the structure of the item pool and population distribution. Further research has taken place and yielded numerous refinements. Chang, Qian, and Ying (2001) have developed the *a*-stratified design with *b*-blocking to overcome the first problem by balancing the distributions of *b*-values among all strata. This method first prepartitions the item pool according to *b*-values and then implements the *a*-stratification. Their simulation study showed that the blocking method performs significantly better than the original stratified method in a sense that it improves item exposure rate control, reduces mean squared errors (MSE), and increases test reliability. Chang and van der Linden (2003) and van der Linden and Chang (2003) propose using 0-1 mathematical programming models, together with the *a*-stratified method, to balance contents and improve accuracy. Yi and Chang (2003) and Leung, Chang, and Hau (2003) proposed solutions to incorporating the ability to handle item content.

Application of the *a*-stratified method may be taken one step further to overcome the underestimation problem. As demonstrated analytically by Chang and Ying (2002), items with high-discrimination parameters tend to cause "big jumps" for the latent trait estimator at the very early stage of the test. To successfully maintain a normal pace at the beginning of the test, items to be administrated must possess the characteristic of low discrimination. Chang and Ying's simulation study revealed that the proposed ascending *a*-methodology is pivotal to overcoming the underestimation problem.

The theoretical results derived by Chang and Ying (2002) show that, for the 3PL model, the use of ascending order of a_n in item selection, as advocated in Chang and Ying (1999), plays a pivotal role in overcoming the underestimation problem encountered in current large-scale administrations of CATs. Some obvious benefits include the following:

- robustness, reducing fluctuation due to initial item response irregularity;
- effectiveness, offsetting initial item influence by the test performance based on the responses to later items;
- more balanced exposure rates, improving item pool usage and increasing test security;
- high reliability, increasing the score consistency between test and retest; and
- a higher level of efficiency, maintaining the high quality of the latent trait estimation by using high discriminating items when they can be most effectively used.

7.6. Assessing CAT Test Security Breaches

Around 10 years have passed since the Kaplan-ETS incident. However, unlike many other aspects in CAT, there is a lack of theoretical developments in assessing test security breaches. Way (1998) pointed out that there is no common understanding as of yet about issues such as what represents acceptable item exposure rates and how long CAT item pools should be used. Many rules currently used in large-scale CAT programs were derived essentially from simulation

studies, which may not be sufficient in assessing test security breaches caused by organized item thievery.

On August 6, 2002, ETS announced that it would temporarily suspend the CAT GRE General Test and reintroduce paper-based versions in several foreign countries. The news release is as follows (www.ets.org, August 20, 2002):

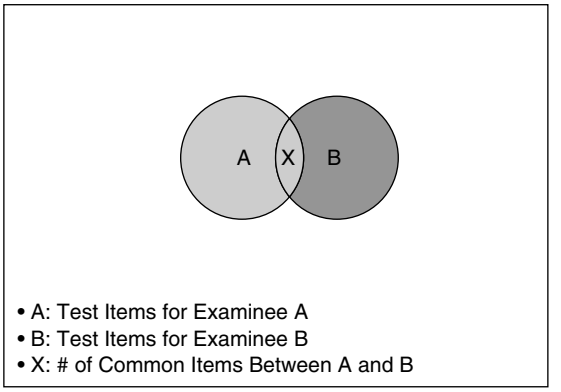
> ETS is undertaking the change at the request of the GRE Board, the policy setting body of the examination, following an investigation that uncovered a number of Asian-language Web sites offering questions from live versions of the computer-based GRE General Test. The Web sites included both questions and answers illegally obtained by test takers who memorize and reconstruct questions and share them with other test takers. The Web sites are located in China and Korea, and easily accessed in Hong Kong and Taiwan.

Clearly, CAT test security must be studied in a broad context, and certain theoretical justifications must be developed. The new emphasis should be on organized item theft. More specifically, for a given GRE item pool, if each examinee can memorize β items (say, $\beta = 10$), how many thieves are needed to steal the large-enough portion of the pool? Because different item selection strategies may yield different stealing rates, the objective of this kind of research is to develop a theoretical upper bound for the expected number of thieves under various CAT settings. An assessment that test developers typically need might very well be like the following: Assume the test length is 30 and the item pool size is 700; if every thief can remember 20 items, at most 20 thieves are needed to steal about 60% of items in the pool.

7.6.1. Chang and Zhang's Item Pooling Index

An *item overlap rate* for a group of examinees was originally defined as the ratio of the expected number of overlapping items encountered by two randomly sampled examinees from the group over the test length. The Venn diagram in Figure 7.2 shows two sets of items for Examinees A and B, respectively. The intersection indicates that the common items can be seen by both examinees. The item overlap rates can be estimated by calculating the percentage of the items that are shared by each pair of examinees and then averaging across all the pairs of examinees from the group. The estimated item overlap rate is also referred to as the *average item overlap rate* (Way, 1998). Ideally, the number of overlapping items within any group of examinees should be kept to a minimum. According to Chang and Zhang (2002), higher item overlap rates are evidence that item exposure rates are heavily skewed. If every item in the pool has an equal possibility of being selected, the number of common items among examinees will be kept to a minimum.

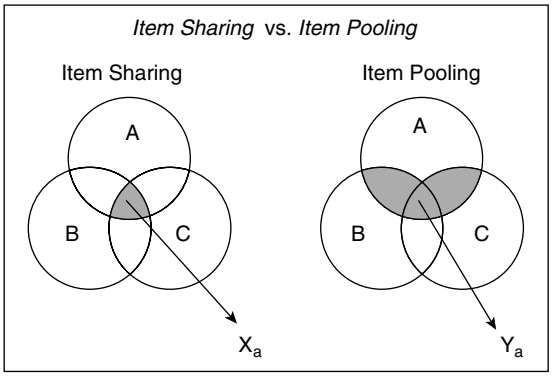
Figure 7.2 Items Can Be Shared by Two Examinees



There are two limitations in the original definition. First, it only considers two examinees in the rate calculation instead of a group of α examinees. In reality, it is often the case that one examinee pools information from several examinees who have taken the test. Second, it does not distinguish *beneficiary* from *nonbeneficiary*. An examinee who will take the test is a beneficiary, whereas examinees who have taken the test are nonbeneficiaries. So, it is necessary to extend the definition to overcome the limitations.

Let X_α denote the number of common items encountered by a group of test takers, where α is the number of test takers. For example, X_3 represents the number of common items encountered by three test takers; X_α is a random variable, and its randomness comes from both the item selection algorithm and test taker sampling. In this chapter, we only consider a fixed group of test takers so that the randomness of X_α uniquely comes from the item selection algorithm. Suppose α sets of items are assembled for α test takers; the common items should be the intersection of these α sets. See Figure 7.3 for the case of $\alpha = 3$, where the shaded area represents the common items encountered by these three test takers. We call this *item sharing* because the information in the intersection can be *shared* by all α examinees. Ideally, we would like to know the distribution of X_α so that its expected value can be calculated. A large value of $E[X_\alpha]$ indicates that the test overlap rate (for α test takers) is high.

Figure 7.3 Item Sharing Versus Item Pooling for Three Examinees



Chang and Zhang (2002) generalize the definition of item overlap rates from two examinees to a group of α examinees:

Definition 2. Let X_α be the number of common items shared by a group of randomly sampled α examinees; then $E[X_\alpha]$, which is the expected value of X_α, is called the *item-sharing index.*

When $\alpha = 2$, $E[X_\alpha]$ has an intuitive meaning that indicates how many common items we would expect that a student could get from a friend who just took the test. However, when $\alpha \geq 3$, although $E[X_\alpha]$ is still a good indicator of test security, it may be less intuitive in interpretation. Note that the information about the common items is beneficial to those who will take the test but not to those who have already taken the test. However, when $\alpha \geq 3$, $E[X_\alpha]$ does not distinguish the former from the latter.

Suppose test taker A, who will take the test, seeks help from two friends, B and C, who have taken the test. Let's call this *information pooling,* meaning that one beneficiary pools information from several nonbeneficiaries. Let Y_α be the number of overlapping items encountered by a test taker with other α test takers who have already taken the test. See Figure 7.3 for the graphical presentation for $\alpha = 2$, where A, B, and C represent the items taken by test takers A, B, and C, respectively. Clearly, $A \cap (B \cup C)$ are the items that test taker A can pool from B and C. Apparently, $Y_1 \equiv X_2$. But Y_α is different from $X_{\alpha+1}$ for $\alpha \geq 2$. Chang and Zhang proposed using a new definition to distinguish *item sharing* from *item pooling.*

Definition 3. Let Y_α be the number of common items one examinee can pool from randomly sampled α examinees, and then $E[Y_\alpha]$, which is the expected value of Y_α, is called the item-pooling index.

7.6.2. Lower Bounds of $E[X_\alpha]$ and $E[Y_\alpha]$

One aspect of test security control should be to keep test overlap rates to below a reasonable threshold. This raises an interesting question: What is the criterion for a small overlap rate? In item exposure rate control, it is common to set a threshold as an upper bound and to require that no item usage rates should exceed such a bound. Analogous to this, a straightforward way to set a criterion for test overlap control would be to base it on a minimum value of overlap rates. However, test overlap rates are highly sensitive to methods used in item selection, ability estimation, and exposure control. To make comparisons across all possible methods, one must search for a promising candidate across all possible item selection rules, ability estimation methods, and exposure control strategies. If such minimum value exists, it can serve as a lower bound. A test security panel may evaluate the discrepancy between the theoretical lower bound and the observed test overlap rate generated by the item selection algorithm under investigation. A big difference indicates that the algorithm needs to be further improved by lowering the overlap rate, and a small difference indicates that there is not much to improve.

Because different procedures may yield different overlap rates, the distributions of X_α and Y_α rely on the item selection procedure built in the CAT test, and so, in general, it may not be possible to derive theoretical distributions for the two random variables. However, for the randomized item selection procedure (i.e., we just randomly select n items to each examinee), the theoretical distributions of X_α and Y_α can be derived. The need for *randomization* is important in the mathematical derivation of the theoretical distributions of X_α and Y_α, but it may not be apparent to some practitioners because none of the CAT programs endorses the randomized item selection method. In terms of the consequence of the randomization assumption, as pointed out by Wainer (2000), when every item has an equal possibility to be administered to the examinees, test security will reach the maximum. As a result, the expectations defined in Definition 2 and Definition 3 can serve as two theoretical lower bounds for the item overlap rates.

7.6.3. About Theoretical Derivations

One purpose of test security control in a CAT is to lower the test overlap rate. To this end, it is desirable to find distributions of random variables X_α and Y_α, $\alpha = 2, 3, \ldots, n$, so that the expected values X_α and Y_α can be calculated. These values may serve

as theoretical lower bounds of test overlap rates for test developers, who can assess their CAT selection designs by comparing the observed or simulated overlap rates with the lower bounds. Under the assumption that every item has equal possibility of being selected, Chang and Zhang (2002) derived the theoretical distributions for the sharing variable X_α and the pooling variable Y_α for any given number α.

Although their derivations are mathematically rigorous, it becomes extremely simple for X_2 or, equivalently, Y_1. Let's consider the simplest case in which only two examinees are encountered, say, A and B. For A, we randomly selected m items from an item pool containing N items. After A finished the test, we put the m items back into the pool. Now we may consider the m items as "bad" items because they were used by A. Now, for test taker B, a set of m items is to be drawn from the same item pool with m "bad" items. It is interesting that this process is equivalent to the experiment in which one randomly selects m units from a set of N units with m defected units. Obviously, the number of the defected units that would be found in the m draws follows the hypergeometric distribution (Bickel & Doksum, 1977). Recall that X_α is the number of commonly overlapping items encountered by α examinees. Then, X_2 has a hypergeometric distribution, that is,

$$\text{Prob}\{X_2 = k\} = \frac{\binom{m}{k}\binom{N-m}{m-k}}{\binom{N}{m}},$$

$$k = 0, 1, 2, \ldots, m,$$

where N is item pool size, m is the test length, and k is the number of common items. Following any of elementary statistics textbooks, we have

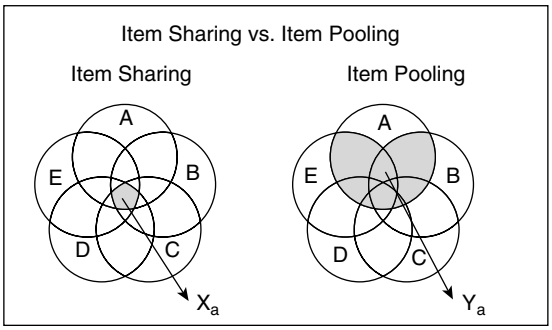
$$E[X_2] = \frac{m^2}{N}.$$

See Chang and Zhang (2002) for the derivations for $\alpha \geq 2$.

7.6.4. How to Use the Lower Bounds

Now, under the *best* test security consideration, the expected item overlap rate between one examinee and a group of α examinees can be precisely calculated. Based on the calculation, a table of lower bounds of item overlap rates for various combinations of test settings can be readily constructed. Such a table is useful as a benchmark for practitioners in evaluations of test security and item selection algorithms.

Figure 7.4 Item Sharing Versus Item Pooling for Five Examinees



Note that the results derived by Chang and Zhang (2002) are the lower bounds, and the observed overlap rates should be higher than those lower bounds. The discrepancy between the theoretical lower bound and the observed rate based on a particular item selection method provides information about the security perspective of this test design. Clearly, the use of the item-pooling rate may allow for better assessing test security for the methods used in item selections.

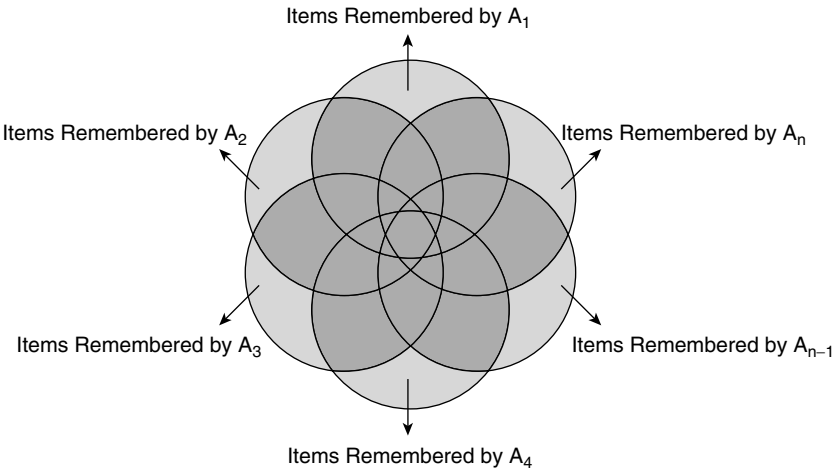
7.6.5. Limitations in Chang and Zhang's Original Derivation

The item-sharing index is the lower bound of the expected number of the common items encountered by a group of α examinees. Although the derivation is theoretically interesting, it may not be practically informative to practitioners when the number of examinees is large. According to Figure 7.4, as α gets large, the number of common items in the intersection part will become very small. On the other hand, the item-pooling index, which calculates the lower bound of the expected number of items one examinee can pool from a group of α examinees, should be a much more useful index. However, the derivation is based on the assumption that every examinee can memorize all the items in the test, which seems unlikely in a real situation. To what extent can the severity of item-pooling activity be assessed? Chang and Zhang's result provides a partial answer for CAT tests with a short test length. For CAT with a moderate to long test length, the assumption to memorize all the items in the test is unrealistic.

7.6.6. Extension of Chang and Zhang's Item-Pooling Index

Chang and Zhang (2003) found that for the application of the item-pooling index, one can simply assume

Figure 7.5 Items Can Be Compromised by N Thieves



that each examinee only memorizes β items, where $1 \leq \beta \leq n$ and where n is the test length. Note that the result of Chang and Zhang (2002) is a special case of $\beta = n$.

An interesting property in Chang and Zhang (2002) is that the test length can vary in the calculation of the item-pooling rate, which is the most concrete and convenient condition that can be employed in the extension. Let us assume that each thief can only remember β items out of the total n items administered to him or her; because those $n - \beta$ items that the thief cannot remember do not cause test security breach, one can simply remove them from the test so that the "authentic" test length is β. Even though the test length is reduced from n items to β items, all the derivations remain the same. As a consequence, the result of Chang and Zhang can be readily used under the assumption that each thief only remembers β items, and such generalization under the weakened condition is fairly straightforward.

7.6.7. Assess Organized Item Theft

For years, research efforts to defend CAT test security have been concentrated on item-sharing activities among examinees. The indexes discussed in this chapter so far are not specifically designed to measure security breaches caused by organized thievery activity. The Kaplan-ETS incident demonstrates that the organized item theft may cause more severe damage than sharing information among friends. To propose a quantitative model for organized item thievery, we might best think of the Kaplan-ETS example. Suppose a group of thieves takes a CAT sequentially. For convenience, these thieves are ordered by the sequence of their test times. To test administrators, the items in the union of $A_1, A_2, \ldots, A_\alpha$ can be considered bad items, where A_i is the set of n items that the ith examinee takes. See Figure 7.5 for a demonstration.

Definition 4. Let A_α be the set of n items that the αth thief takes, and $\bigcup_{i=1}^{\alpha} A_i$ are "bad" items that can be compromised by the α thieves. Let Z_α be the number of items in $\bigcup_{i=1}^{\alpha} A_i$.

Under the randomized item selection assumption, Chang and Zhang (2002) derived the theoretical distribution of Z_α so that $E[Z_\alpha]$ can be analytically calculated, which can be interpreted as the expected value of the *upper bound* of the number of items compromised by the α thieves. Again, let us consider that A_i only contains β items that the ith thief can remember; then, one should be able to predict how many thieves are needed at most to recover $\gamma\%$ ($0 \leq \gamma \leq 100$) of the item pool.

7.6.8. How Many Thieves Are Needed to Compromise a CAT GRE Item Pool?

To what extent can a high-stakes CAT test, such as the GRE, be compromised? Based on the findings of Chang and Zhang (2003), a table can be constructed of theoretical upper bounds of expected values of number thieves, which are needed for various combinations of test settings. The settings will include both the GRE

and the GMAT as special cases. Such tables are useful as an assessment of test security breaches in locations where organized crimes of stealing and sharing CAT items tend to take place.

Stocking (1994) proposed that the item pool size should be approximately 12 times the length of CAT exam, which Way (1998) referred to as a rule of thumb. According to the CAT GRE test length setting (www.gre.org, June 18, 2003), the Verbal test consists of 30 items, and the Quantitative test consists of 28 items. In line with the rule of thumb, the pool sizes should be 360 for the Verbal test and 336 for the Quantitative test. However, these numbers may seem too small. Let's double the sizes by assuming that each subtest consists of 2 item pools with a random chance to be assigned to each examinee. Thus, there are about 700 items in each of the subpools. In computing Chang and Zhang's (2003) $E[Z_\alpha]$, β can be any fixed number bounded by n, where n is the test length. Let $\beta = 10$; in other words, each thief can remember 10 items. Chang and Zhang (2003) calculated $E[Z_\alpha]$ for $\alpha = 2, 3, \ldots, 100$, where α stands for the number of thieves who have already taken the test. They found disturbing results for the above GRE setting—if every thief can remember 10 items, at most 50 thieves are needed to compromise about 55% of the items in the pool. However, if every thief can remember 20 items, at most 20 thieves are needed to steal the same amount of items.

Illegal actions of stealing and sharing CAT items will inflate test scores for some test takers and hurt honest test takers. Without effective measures, this could significantly undermine the credibility of CATs. The findings of Chang and Zhang (2003) are both encouraging and disturbing to CAT researchers. The proposed theoretical indices will allow assessing test security severity under various situations. Meanwhile, calculations based on these indices can help CAT developers to improve their CAT designs for better item pool usage and test security. However, their numerical result indicates that the current practice to form a high-stakes CAT with only several hundred items may not be suitable for areas where the prevention of stealing and sharing CAT items cannot be guaranteed.

As indicated earlier, Chang and Zhang's (2003) results are based on randomized item selection, and randomized item selection equalizes item exposure rates and hence yields the best test security control. Even under the best security design, one can get 385 items compromised from 700 items by talking to 50 test takers who can only remember 10 items after the test. In reality, ETS uses a constrained maximum-information item selection method that yields a greatly skewed item exposure distribution. So, the number of thieves needed to compromise 55% items in the pool could be much smaller than 50. It is interesting to note that Kaplan only sent 20 "thieves" in the 1994 incident, which resulted in ETS's temporary suspension of the CAT GRE test.

7.7. Conclusions

Computerized adaptive testing has become fashionable in many high-stakes testing programs. The principal component in CAT is the item selection procedure built into the CAT system, which selects the next item for the examinee on the basis of his or her responses to the items previously administered. For the past two decades, the most commonly used item selection procedure has been based on maximizing item information. More specifically, an item is selected that has maximum information at the currently estimated θ level ($\hat{\theta}$), which is estimated from the available responses at that time.

However, the original item selection algorithm developed by Lord (1970) is based on the Robbins-Monro process, and hence it can be considered as a non-IRT scoring approach. As noted by Lord, the aim of an adaptive test is to tailor the difficulty levels of the items administered to the latent trait θ of the examinee being tested. So, the items chosen for administration should have b-values that match the examinee's $\hat{\theta}$. When certain mathematical models are used for the item response functions, such as the logistic and normal ogive models, the item selected by maximizing item information at $\hat{\theta}$ should have its b-value close to $\hat{\theta}$. Bickel et al. (2001) show that the maximum-information approach is model sensitive. Many reasonable models that have shapes similar to that of the logistic model have different shapes of the information functions and hence different optimizing strategies. Although the practical perspective of this study is difficult to understand, it suggests that the maximum-information approach may overly depend on specific IRT mathematical models. Recently, several practitioners have emphasized that some alternatives for scoring computer-based testing should be considered (e.g., Dodd & Fitzpatrick, 2002; Plake, 2002).

Yes, the maximum-information approach yields more efficient estimates. However, the efficiency is under the assumption of an "infinitely large item pool" that never holds in reality. An operational item pool usually consists of several hundred items.

Furthermore, the set items selected for each examinee must satisfy nonstatistical constraints such as content balance. The more constraints one has to impose, the fewer degrees of freedom one can have in a design. To design a CAT algorithm that works reasonably well, one should consider some sampling strategy, including stratification.

One of the main purposes of this chapter is to intuitively reveal the cause for the underestimation/overestimation phenomenon. Based on the theoretical results of Chang and Ying (2002), these problems may be caused by heavy reliance on high-discrimination items at the beginning of the test, resulting in a lack of stability and consistency that are essential in every CAT administration, especially in high-stakes test situations. To this end, Chang and Ying propose modifying the statistical procedure used for CAT item selection by incorporating some analytic techniques. Their results show that weighting the likelihood score is a possibility in alleviating the problem of underestimation because the true θ will be closer to its ongoing estimator $\hat{\theta}$ after more CAT items have been administered.

CAT was originally developed for assessing the unidimensional latent trait, θ. Recently, Tatsouka (2002) and Xu, Chang, and Douglas (2003) proposed several promising item selection methods for cognitive diagnostic applications, which incorporate the ability of diagnostic assessment to provide helpful diagnostic information to examinees. One innovative future application of this research would be to use the computer adaptive approach to cognitive diagnosis in the realm of Web-based learning. A computer program could be developed to use the information provided by diagnostic testing as an online tutor. This research could be applied to the realm of Web-based instruction to produce a program that uses the diagnostic information from an individual's knowledge state estimate to provide additional instruction via the Internet. This specific instruction could be individualized to provide information for all of the nonmastered attributes, but not the mastered attributes, for the individual.

Finally, what conclusions may be drawn about organized item thievery in the context of test security? The analytical results discussed by Chang and Zhang (2003) clearly indicate that structuring an operational CAT exam with only several hundred items should be considered willful negligence. A high-stakes CAT exam must have, among many other things, a large item pool. This can be accomplished partly by including many items that have never been selected by the current maximum-item selection algorithms. Chang and Zhang show that test security can be strengthened greatly by increasing item pool size from several hundred items to a few thousand. This may also be accomplished by including many items that have been used in the past (maybe 20,000 used items for the GRE). According to Green (2000),

> If the item pool is sufficiently large, an examinee who has studied the pool has relatively little special advantage studying the pool amounts to reviewing the knowledge domain. But if the pool is small, a person who has studied the items may have an advantage. One possibility is to have two or more distinct item pools, or test forms. (p. 33)

Moreover, even more important, the structure of the organization should include on-the-fly item generation (e.g., see Bennett, 2003, for item generation) from schemas that allow, for some item types, indefinitely many items. In such a test, learning the principles behind the items becomes the most efficacious strategy (Robert Mislevy, personal communication, June 7, 2003).

REFERENCES

Bennet, R. E. (2003). An electronic infrastructure for a future generation of tests. In H. F. O'Neil & R. Perez (Eds.), *Technology applications in education: A learning view* (pp. 267–281). Hillsdale, NJ: Lawrence Erlbaum.

Bickel, P., Buyske, S., Chang, H., & Ying, Z. (2001). On maximizing item information and matching ability with item difficulty. *Psychometrika, 66,* 69–77.

Bickel, P. J., & Doksum, K. A. (1977). *Mathematical statistics.* San Francisco: Holden-Day.

Carlson, S. (2000, October 20). ETS finds flaws in the way online GRE rates some students. *Chronicle of Higher Education, 47*(8), A47.

Chang, H., Qian, J., & Ying, Z. (2001). *a*-Stratified multistage computerized adaptive testing with *b* blocking. *Applied Psychological Measurement, 25,* 333–341.

Chang, H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika, 58,* 37–52.

Chang, H., & van der Linden, W. J. (2003). Optimal stratification of item pools in *a*-stratified adaptive testing. *Applied Psychological Measurement, 27,* 262–274.

Chang, H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20,* 213–229.

Chang, H., & Ying, Z. (1999). *a*-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 211–222.

Chang, H., & Ying, Z. (2002, April). *To weight or not to weight? Balancing influence of initial and later items in CAT.* Paper presented at the annual meeting of National Council on Measurement in Education, New Orleans, LA.

Chang, H., & Ying, Z. (in press). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *Annals of Statistics.*

Chang, H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika, 67,* 387–398.

Chang, H., & Zhang, J. (2003, April). *Assessing CAT security breaches by the item pooling index—to compromise a CAT item bank, how many thieves are needed?* Paper presented at the annual meeting of National Council on Measurement in Education, Chicago.

Davey, T., & Nering, N. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward. (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 165–191). Mahwah, NJ: Lawrence Erlbaum.

Davey, T., & Parshall, C. G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Dodd, B. G., & Fitzpatrick, S. J. (2002) *Alternative for scoring CBTs.* In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 215–236). Mahwah, NJ: Lawrence Erlbaum.

Finney, D. J. (1978). *Statistical method in biological assay.* New York: Academic Press.

Folk, V. G., & Smith, R. L. (2002). Models for delivery of CBTs. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 41–66). Mahwah, NJ: Lawrence Erlbaum.

Goodwin, G. C., Ramage, P. J., & Caines, P. E. (1980). Discrete time multivariable adaptive control. *IEEE Transactions on Automatic Control, 25,* 449–456.

Green, B. F. (2000). System design and operation. In H. Wainer, N. Dorans, D. Eignor, R. Flaugher, B. Green, L. Steinberg, et al. (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 23–35). Hillsdale, NJ: Lawrence Erlbaum.

Hau, K.-T., & Chang, H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement, 38,* 249–266.

Hodges, J. L., & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t-test. *Annual of Mathematical Statistics, 27,* 324–335.

Leung, K., Chang, H., & Hau, K. T. (2003). Incorporation of content balancing requirements in stratification designs for computerized adaptive testing. *Educational and Psychological Measurement, 63,* 257–270.

Lord, M. F. (1970). Some test theory for tailored testing. In W. H. Holzman (Ed.), *Computer assisted instruction, testing, and guidance* (pp. 139–183). New York: Harper & Row.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223–236). New York: Academic Press.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9,* 287–304.

Mislevy, R., & Chang, H. (2000). Does adaptive testing violate local independence? *Psychometrika, 65,* 149–165.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351–356.

Plake, B. S. (2002). Alternatives for scoring CBTs and analyzing examinee behavior. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 267–274). Mahwah, NJ: Lawrence Erlbaum.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annual of Mathematical Statistics, 22,* 400–407.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Rep. No. 94-5). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1998). *A framework for comparing adaptive test designs.* Manuscript under review.

Stocking, M. L., & Lewis, C. (1995). *A new method of controlling item exposure in computerized adaptive testing* (Research Rep. No. 95-25). Princeton, NJ: Educational Testing Service.

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th Annual Meeting of the Military Testing Association* (pp. 973–977). San Diego: Navy Personnel Research and Development Center.

Tatsouka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of Royal Statistical Society, 51,* 337–350.

Thomasson, G. L. (1995, June). *New item exposure control algorithms for computerized adaptive testing.* Paper presented at the annual meeting of the Psychometric Society, Minneapolis, MN.

van der Linden, W. J. (1998). Bayesian item-selection criteria for adaptive testing. *Psychometrika, 63,* 201–216.

van der Linden, W. J., & Chang, H. (2003). a-Stratified adaptive testing with large number of content constraints. *Applied Psychological Measurement, 27,* 107–120.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Item-selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22,* 203–226.

Wainer, H. (2000). Rescuing computerized adaptive testing by breaking Zipf's law. *Journal of Educational and Behavioral Statistics, 25,* 203–224.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Steinberg, L., et al. (Eds.). (2000). *Computerized adaptive testing: A primer* (2nd ed). Hillsdale, NJ: Lawrence Erlbaum.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17,* 17–27.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473–492.

Xu, X., Chang, H., & Douglas, J. (2003, April). *A simulation study to compare CAT strategies for cognitive diagnosis.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Yi, Q., & Chang, H. (2003). a-Stratified multistage CAT design with content-blocking. *British Journal of Mathematical and Statistical Psychology, 56,* 359–378.

Section III

Models for Categorical Data

Chapter 8

Trends in Categorical Data Analysis

New, Semi-New, and Recycled Ideas

David Rindskopf

Kierkegaard probably knew nothing about statistics, but nonetheless he aptly summarized the plight of statisticians when he wrote that "life can only be understood backward, but it must be lived forward" (quoted in Smith, 1990, p. 6). In data analysis, we to try to predict the future by understanding the past.

Until recently, those of us who deal with categorical data have had to use a very restricted set of tools to attain understanding. And whether by design or by force of circumstance, most researchers end up with many variables in categorical form. For many years, the analysis of such data was, by and large, restricted to a simple chi-square test of independence in a two-way table of categorical variables, and perhaps the calculation of an appropriate measure of association. The field is now flooded with new techniques, and problems that once seemed insoluble are now either solved or on the brink of being solved.

In this chapter, I discuss some of these new methods. I begin with the broader task of presenting an overview of recent major advances in applied statistics. These advances will be illustrated using examples of research, both of my own and others, in the area of categorical data analysis. Some advances involve completely new ideas, whereas others represent either the resurrection or recycling of old ideas.

The independent rediscovery of what really are old ideas is more common now, because the literature is so extensive that no one could know every idea in it. But not all ideas that have roots in the past are the same as they were when originally proposed. I call an idea recycled if it was developed long ago in its basic form, but has been put in the context of modern theory and methods. In applied statistics, recycled ideas are put in a more general context than the original idea; they are put in a sound statistical framework, often using maximum likelihood estimation or Bayesian methods; and they are implemented in a computer program that makes their application feasible for the typical researcher. Later, I will discuss the partitioning of chi-square in contingency tables, which is both a good old idea and a good idea for recycling.

8.1. Balance of Emphasis in Applied Statistics

Broad trends in the development of applied statistics can be understood most easily by contemplating the rise and fall of emphasis on its three main components: description, exploration, and inference. The primary question addressed by *description* is, "What's there?";

its fundamental purpose is to summarize information, either numerically or graphically. *Exploration* deals with hypothesis generation and answers the question, "What might the data mean?" *Inference* is intended to settle the matter as much as possible; hypothesis testing, confidence intervals, prediction, and related methods respond to the desire to quantify the amount of evidence in the data. The evolution of mathematics, probability, statistics, and computational methods has led to changes in the emphasis given to description, exploration, and inference.

Many years ago, before the field of applied statistics existed, people primarily made qualitative descriptions and used informal (or no) inference. Gradually, the need for quantitative description was recognized, but inference was still informal. Eventually, statistical theory began to develop, and formal inference was possible. Unfortunately, formal inference began to monopolize the field, either replacing or dominating the descriptive function.

Only recently have we seen a move toward the balancing of description and inference, an emphasis on hypothesis generation as well as hypothesis testing, and the emergence of exploration as a primary purpose of applied statistics. For too long, hypothesis generation was ignored, not only in statistics courses but also more generally in methodology courses. Some considered it too hard to formalize these methods, or perhaps impossible because there was too much of the "human element" involved. Others (mistakenly) considered it unnecessary because hypotheses seemed so simple and obvious, such as "Treatment A has the same effect as Treatment B." Many students have thereby erroneously inferred that hypothesis generation is less important than hypothesis testing.

Luckily, exploratory methods are now in vogue, probably because they were advocated by John Tukey, who was respected by the "tough-minded" statistical theorists. Exploratory methods will not usually generate hypotheses by themselves, but they certainly help in the process by highlighting important features of the data.

Although the best-known exploratory methods deal with quantitative variables, some progress has been made for analyzing qualitative data. Correspondence analysis is fast becoming a popular technique in this area; it is a good example of an old idea that has been resurrected and recycled with additions such as graphical representation of the results. The technique was first proposed by, among others, R. A. Fisher (1930), and is actually just canonical correlation of frequencies in a two-way cross-tabulation. The corresponding inferential techniques using maximum likelihood estimation have recently been developed by Leo Goodman (1978).

Graphical methods for quantitative variables have become more widely known and used in the past few years, but most of those analyzing categorical data have had few tools to work with. That situation is changing, and many promising new methods are being developed. The state of the art is demonstrated in the excellent new book of graphical methods for categorical data by Friendly (2000).

8.2. Mathematics Versus Data Analysis Focus

The cause of the conflict between hypothesis generation and hypothesis testing is intriguing. It arose primarily because of a difference in focus between mathematical statisticians and data analysts. Statisticians typically want to get exactly the right answer, even if it is not the answer to the right question. Data analysts typically do not care if the answer is an approximate one, as long as it is an answer to the right question. This caricature might be a slight exaggeration, but not by much.

Theoretical statisticians are basically mathematicians; they place a high value on exactness. Data analysts are guided by research questions to be answered as well as possible, regardless of whether or not the answers are exact. Data analysts often develop ad hoc methods to attack important problems for which no methods based on statistical theory exist. Statisticians, repelled by the ad hoc nature of these methods, either dismiss them or try to develop and apply appropriate statistical theory. Thus, statisticians and data analysts sometimes differ in their method of approach to a problem, as well as in what they consider to be acceptable criteria for a satisfactory solution.

In architecture, "form follows function" was once the rule. But too often, form replaced or superceded function, so buildings looked interesting but did not work. In statistics, too often sophisticated mathematics replaces (instead of building on) thoughtful conceptualization. Happily, at least in some areas, there is a call to strike a balance between the goals of answering the right question and being rigorous (Wilkinson & Task Force on Statistical Inference, 1999). In the area of categorical data, partitioning of chi-square has practically disappeared, even though the usual hierarchical log-linear models cannot replace its function.

As we will see in some detail in the discussion of partitioning chi-square and nonstandard log-linear

models, current statistical methodology is moving toward allowing almost everyone to be satisfied: The right questions can be answered, and they can be answered in a sufficiently rigorous way to please most statisticians.

8.3. Realism, Complexity, Computability, and Generality

The past three decades have seen monumental changes in the realism of statistical models. Among the realities that can now be accommodated are the following:

- Data are often missing.
- Measurements almost always are made with error, and many constructs in psychology are latent variables, which can only be imperfectly observed.
- In many studies, people assign themselves to groups or drop out of the study.
- People live and work in groups, such as families, neighborhoods, and classrooms, necessitating multilevel models.
- The normal distribution does not adequately approximate the behavior of all continuous variables.
- Linear models often need to allow curves or interactions among predictors, and sometimes even this is not enough to avoid nonlinear statistical models.
- Important substantive hypotheses cannot always be expressed merely in terms of which main effects or interactions are significant.

Of necessity, greater realism leads to greater complexity of the theoretical and computational methods used for analysis. Unfortunately, greater complexity often leads to incomprehensible or uninterpretable results. We do not always know whether our results are correct, because sometimes there is no easy way to ask whether they seem reasonable.

Complex statistical models have existed in their theoretical form for quite a long time. Many good ideas languished for decades because they were impractical when first proposed. For example, Fisher invented maximum likelihood estimation (including the case with missing data) in the 1920s, and Lawley developed the theory of maximum likelihood factor analysis in the 1940s, but until computer power and numerical methods became available in the 1960s, no one could make use of this knowledge. Not only has rapid progress in the development of computer power made some of these methods practical, but, interestingly enough, it has also spawned many new methods (such as exact methods, data mining, the bootstrap and Monte Carlo techniques) that no one even considered until recently. When the new tools became available, people suddenly discovered many uses for them.

The initial breakthrough was the ability to solve a large number of linear equations simultaneously, so that multiple regression with large numbers of variables was simple to do. Then, techniques for finding eigenvalues and eigenvectors made many multivariate techniques possible. Techniques for solving nonlinear systems have been greatly improved. These evolved in part from the iterative solution to linear systems and, in part, from specialized methods such as the EM algorithm. Methods for numerical integration have made many Bayesian techniques practical. Finally, Monte Carlo methods, the bootstrap (and related resampling techniques), and other computational techniques have made it possible to test hypotheses without making assumptions about the underlying probability distribution involved, and to see how robust the usual techniques are when their assumptions are violated.

One previously mentioned disadvantage of computational progress is that many researchers have lost the close contact with their data that was the hallmark of previous work, because it is often difficult to know whether these complicated techniques are giving reasonable results. On the other hand, it is becoming easier to work with more general statistical models that can be used to analyze a variety of types of data from different research designs that used to require separate methods of analysis. One simple example is the use of multiple regression to do t-tests, ANOVA, ANCOVA, and so on.

Many current ideas about the analysis of categorical data have their origins in developments made decades ago. Some of these old ideas can be used as they are, "right out of the box," so to speak. Others have needed some "recycling," with changes to put them within the context of modern statistical theory and methodology. Yet other ideas are completely new, often representing a major extension or generalization of previous research. The rest of this chapter has two purposes: First, I will outline some of the major areas in which progress has been made, and illustrate these ideas and trends in applied statistics using some examples from research on categorical data analysis. Second, I will provide a context in which the other chapters in this section can be placed.

8.4. PARTITIONING CHI-SQUARE

Partitioning of chi-square is an old idea about testing very specific hypotheses for frequency data, rather than just testing a general hypothesis such as independence of rows and columns in a contingency table. As with many ideas in statistics, the idea for partitioning chi-square can be traced to Fisher. The simplest example is found in his book *Statistical Methods for Research Workers* (Fisher, 1930). Fisher described data from a genetics experiment in which corn was classified as either starchy or sugary, and as to whether the base leaf was either green or white. According to one genetic theory, the frequencies in the four cells of the contingency table should be in a 9:3:3:1 ratio. A chi-square test showed that the frequencies were not as the theory predicted, so what should be concluded? Obviously, the theory is wrong, but can more be said?

Fisher (1930) noted that the theory being tested could be wrong in any of its three assumptions: The expected 3:1 ratio of starchy to sugary might not hold, the expected 3:1 ratio of green to white might not hold, or the traits might not be independent. To test these assumptions, he partitioned the total chi-square with 3 degrees of freedom into three components, each with 1 degree of freedom, to test these three assumptions. He found that the 3:1 ratio held for each factor but that they were not independent.

Fisher's method did not catch on, perhaps because he discussed only genetics examples, and perhaps because he did not indicate a general method for the partitioning (although he did show how to test linear combinations of cell frequencies). In discussing another example, Fisher (1930) did, however, point to the need for subject matter theory to dictate which hypothesis tests would be performed:

> Mathematically the subdivision may be carried out in more than one way, but the only way which appears to be of biological interest is that which separates the parts due to inequality of the allelomorphs of the three factors, and the three possible linkage connections. (p. 93)

A simple example using a two-way table will show how partitioning chi-square can allow researchers to address the questions they consider important, rather than being limited to the usual global hypothesis tests. Consider the cross-tabulation shown in Table 8.1, adapted from Goleman (1985), which shows how well breast cancer patients with various psychological attitudes survive 10 years after treatment. Almost every researcher would know to do a test of independence for the data in this table; the familiar Pearson chi-square statistic is 8.01 with 3 degrees of freedom. Here, $p < .05$, so there is a relationship between the two variables: Attitude is related to survival. From the traditional point of view, that is that; there is nothing else to say. The issue of where the relationship lies is considered in only a few textbooks, most of which were written before 1980. (A pleasant exception is Wickens, 1989.)

Table 8.1 Ten-Year Survival of Breast Cancer Patients With Various Psychological Attitudes

	Response	
Attitude	*Alive*	*Dead*
Denial	5	5
Fighting	7	3
Stoic	8	24
Helpless	1	4

NOTE: LR = 7.95; P = 8.01. LR is the likelihood ratio goodness-of-fit statistic; P is the Pearson goodness-of-fit statistic. Each test has 3 degrees of freedom.

In this example, the researchers had a theory that active responses to cancer, such as fighting and denial, would be beneficial compared to passive responses such as stoicism and helplessness. They were not sure whether patients with different active modes of response would differ in survival rate, or whether patients with different passive modes would have different survival rates.

The theory immediately suggests that instead of a single overall test of independence, three tests should be done. The first should test whether fighters and deniers differ, the second whether stoics and the helpless differ, and the third whether the active responders differ from the passive responders. Each of these tests is displayed in Table 8.2, along with both Pearson and likelihood ratio chi-square tests. (The likelihood ratio chi-square, denoted LR in the tables here, differs little from the usual Pearson chi-square, but we will see that it is more useful for what we will do here. The actual formula appears below in the section on log-linear models.) Each of the three tests has 1 degree of freedom, and the results are as the researchers had hypothesized: Fighters and deniers do not differ in survival rate, nor do stoics and the helpless differ, but those with active modes of responding survive better than those with passive modes. These results would never have been tested without the availability of a technique such as partitioning chi-square.

Notice that the likelihood ratio statistics for the three tests of the specific hypotheses sum to the value of the

Table 8.2 Partition of Chi-Square for Attitude and Cancer Survival Data

	Response	
Attitude	*Alive*	*Dead*
Denial	5	5
Fighting	7	3

(LR = .84, P = .83.)

	Response	
Attitude	*Alive*	*Dead*
Stoic	8	24
Helpless	1	4

(LR = .06, P = .06.)

	Response	
Attitude	*Alive*	*Dead*
Denial + Fighting	12	8
Stoic + Helpless	9	28

(LR = 7.05, P = 7.10.)
NOTE: LR is the likelihood ratio goodness-of-fit statistic, and P is the Pearson goodness-of-fit statistic for the 2 × 2 table that precedes them. Each test has 1 degree of freedom.

test of the overall hypothesis of independence. That is, the overall chi-square has been partitioned into three components, each of which tests a specific hypothesis about comparisons among the groups. (The Pearson test statistics, although not partitioning exactly, still are valid tests of the same hypotheses.)

Although traditionally applied to studies with only two variables, partitioning of chi-square can be extended to test hypotheses in tables involving three or more variables; some examples are shown in Rindskopf (1990). However, the method also has its limitations because not all hypotheses one might wish to test can be specified using partitioning. Furthermore, several possible problems can arise regarding the proper use of partitioning; most important are whether post hoc use is justifiable and what should be done to control the Type I error rate of post hoc partitioned tests.

Partitioning chi-square is a simple technique; it can be taught in a short period of time to anyone familiar with the usual test of independence in a contingency table. Software is readily available: Everyone has access to a program that will produce chi-square statistics. Most important, it allows researchers to test hypotheses that are important to them, rather than hypotheses that statisticians tell them to test. Partitioning chi-square could be called a *context-dependent statistical technique* because the exact way it is implemented in the analysis of a specific data set depends on the context in which it is used. Statisticians can show some of the possibilities for the technique, but the research hypotheses suggested by the subject matter determine how the technique is used in any specific case. The importance of testing such focused contrasts is discussed for continuous variables by Aiken and West (1991) and Rosenthal and Rosnow (1985), among others, and for categorical variables by Rindskopf (1990, 1999).

8.5. Log-Linear and Logit Models

The development of statistical methods for categorical data has long lagged behind the development of techniques for continuous data. When faced with multivariate data sets consisting of continuous data, researchers could choose from a variety of tools, including regression, principal components and factor analysis, discriminant analysis, cluster analysis, and canonical correlation. When faced with multivariate categorical data, most researchers could do little but collapse over all but two variables and use the usual test of independence on these remaining two variables. The end result would be a set of tests of independence for all pairs of variables.

This methodology is inadequate for many reasons. Most important is the problem that the overall relationship between two variables, ignoring (i.e., collapsing over) other variables, can be very different from the relationship between those two variables at each level of other variables (i.e., conditional on the others). By now, most researchers have seen examples of this in the form of Simpson's paradox. For instance, in a random sample of people, there is a strong relationship between whether they get medical treatment and whether they die: Those getting medical treatment are more likely to die. Of course, we have collapsed over an important variable: Were these people seriously ill? If we look at those who are seriously ill, the relationship is the reverse of the overall relationship: Those who are treated are less likely to die.

8.5.1. Log-Linear Models

To deal with the problem of analyzing multivariate categorical data, we needed new approaches; the solution came with the development of log-linear models. From one viewpoint, there is a strong analogy between log-linear models and analysis of variance (ANOVA). The main emphasis in ANOVA is testing hypotheses

about main effects and interactions. The same is true of log-linear models, but the dependent variable for log-linear models is the logarithm of the cell frequency. For example, consider a situation with three categorical variables A, B, and C, with levels denoted by subscripts i, j, and k, respectively. A log-linear model with only main effects would be represented as

$$\ln(F_{ijk}) = \mu + a_i + b_j + c_k,$$

where F_{ijk} is the expected cell frequency for $A = i$, $B = j$, and $C = k$, and ln(.) means the natural logarithm. Except for the logarithm, the form is identical to an ANOVA model. Because this model contains no interaction terms, which would allow relationships among variables, this is the model for complete independence among the three variables. The model is usually specified by notation such as $[A]$ $[B]$ $[C]$, $\{A, B, C\}$, or simply A, B, C to indicate which terms are included.

Just as with the model for independence in two-way tables, expected frequencies can be calculated for this model. These can then be used to assess whether the model is consistent with the data by comparing the expected with the observed frequencies. This can be done using the usual Pearson goodness-of-fit statistic,

$$X^2 = \sum_t \{(O_t - E_t)^2 / E_t\},$$

where t has been used to index the cells of the cross-tabulated data, O represents the observed frequency, and E represents the expected frequency in a cell. The symbol $\sum_t$ means to sum over all the cells of the table. (The use of a single subscript t makes it possible to use this formula to easily represent tables of any dimension and also data sets that are not rectangular.) Conceptually, for each cell of the table, a number is calculated that measures how close the observed cell frequency is to the value that would be expected if the model were true. These numbers are then summed to give X^2. If the model is true, we would anticipate the value of X^2 to be small, but if the model is not true, we would anticipate a large value of X^2.

As discussed in the section on partitioning chi-square, an alternative fit statistic is the likelihood-ratio statistic,

$$G^2 = 2 \sum_t O_t \ln(O_t / E_t).$$

Although it is not so obvious why this is a reasonable measure of fit of a model to the data, notice what would happen if the model fit the data perfectly: Each observed frequency would equal the expected frequency, so O_t / E_t would equal 1 for each cell. Because the logarithm of 1 is 0, the value of G^2 would be zero, indicating perfect fit.

How large a value of X^2 or G^2 is necessary to reject a model as being inadequate to account for the observed pattern of frequencies? As with any statistic that follows a chi-square distribution, the number of degrees of freedom must be counted to find the critical value in a table. The total number of degrees of freedom in the data is the number of cells in the cross-tabulated table. The number of parameters in the model is subtracted from this total to give the number of degrees of freedom for the goodness-of-fit statistic.

Finding the number of parameters in the model is easy, because it is the same as in ANOVA models. There is 1 degree of freedom for the constant (intercept). For each main effect, the number of degrees of freedom is 1 less than the number of levels of that variable. For interactions (discussed further below), multiply the degrees of freedom for each variable involved in the interaction.

For example, consider a table with two variables, and suppose that one variable has three levels, the other four. The table thus has $3 \times 4 = 12$ cells. The log-linear model corresponding to the usual test of independence would have an intercept, $3 - 1 = 2$ parameters for one main effect, and $4 - 1 = 3$ parameters for the other main effect. In all, six parameters are estimated, so the goodness-of-fit test has $12 - 6 = 6$ degrees of freedom. (Notice that the usual rule for testing independence would also give $(2)(3) = 6$ degrees of freedom.)

As another example, consider the independence model for the three-way table described above, where there are three, four, and five levels of variables A, B, and C, respectively. Then there would be $1 + 2 + 3 + 4 = 10$ parameters in the model, and $3 \times 4 \times 5 = 60$ cells in the table. The goodness-of-fit test would have $60 - 10 = 50$ degrees of freedom. (Notice that trying to extend the usual rule would fail here: $(2)(3)(4) = 24$, which is incorrect.)

Of course, models of complete independence are not only too simple to explain most multivariate data, but researchers would be devastated if they did fit; after all, no one examines variables because they think that all of them will be unrelated to each other. Instead, we expect that there will be relationships, and we want to find the simplest model that accounts for these relationships. To do this, we start adding to the model what would be called interactions in the context of ANOVA.

To illustrate the general procedure, we will reanalyze a famous data set on ulcer and blood type, originally reported in Woolf (1955) and reproduced in Table 8.3. This data set has three variables: city

Table 8.3 Relationship Between Ulcer and Blood Type

		Ulcer?		
City	*Blood Type*	*Yes*	*No*	*% Ulcer*
London	O	911	4,578	16.6
	A	579	4,219	12.1
Manchester	O	361	4,532	7.4
	A	246	3,775	6.1
Newcastle	O	396	6,598	5.7
	A	219	5,261	4.0

Table 8.4 Fit of Log-Linear Models to Ulcer and Blood Type Data

Model	G^2	*df*	*p*
U, B, C	754.47	7	.000
BU, C	700.97	6	.000
CU, B	83.59	5	.000
BC, U	737.74	5	.000
BU, BC	684.25	4	.000
BU, CU	30.10	4	.000
BC, CU	66.87	3	.000
BC, BU, CU	2.96	2	.227
BCU	0.0	0	1

NOTE: U = ulcer; B = blood type; C = city.

(London, Manchester, and Newcastle), blood type (only O and A are included here), and ulcer (whether or not the person has an ulcer). The table has $3 \times 2 \times 2 = 12$ cells. Table 8.4 contains the fit of several log-linear models for this data set. A common abbreviated notation is used: If an interaction is listed, then all lower order interactions and main effects of those variables are also in the model. For example, if an *AB* term is in the model (indicating that an $A \times B$ interaction is included), then *A* and *B* main effects are also assumed to be present. This is called the hierarchy principle; most applications of log-linear models follow this principle.

As can be seen in Table 8.4, no simple model fits the data. The last model, called the saturated model, has no degrees of freedom left to test the model: It fits the data exactly because it uses all of the information in each cell. Because this model represents no simplification over the frequencies themselves, one would hope that other models would fit the data. In this case, the model [*BC*] [*BU*] [*CU*], with all main effects and three two-way relationships (*BC*, *BU*, and *CU*) but no three-way relationship, fits well. So blood type is related to city, blood type is related to ulcers, and city is related to ulcers, but the relationship between any two variables is the same at each level of the third variable (no three-way relationship).

Researchers often have one or more ordered variables (e.g., no symptoms, mild symptoms, severe symptoms). The most frequently used strategy in the past has been to treat ordered variables as if they were continuous. Now there are many methods for more adequately analyzing such data; these methods are discussed by Johnson and Albert (Chapter 9, this volume).

8.5.2. Logit Models

Frequently, a researcher considers one variable to be an outcome variable, and the others to be control or predictor variables. For this data set, ulcer (U) might be considered an outcome, blood type (B) a predictor, and city (C) a control or possible moderator of the effect of blood type on ulcer. In the usual ANOVA terminology, we would be interested in the main effect of blood type (on likelihood of ulcer), the main effect of city, and the interaction between blood type and city. The most obvious approach would be to model the probability of ulcer as a function of blood type and city. But using probability as an outcome is problematic: It can only vary between 0 and 1, whereas in ANOVA models, the outcome variable can have any value. The solution is to use the logit of the probability as the outcome, where the logit is defined as the logarithm of the odds:

$$\text{logit}(p) = \ln\{p/(1-p)\}.$$

The logit can take on any real value and is therefore appropriate as an outcome variable.

Although logit models can be represented in different ways, one useful approach is to note a correspondence between logit models and log-linear models: Each logit model is equivalent to a log-linear model (but not all log-linear models are logit models). To understand the equivalence, consider each logit model as if it were a regression model. In regression models, no constraint is placed on relationships among the predictors; the predictors might be independent, but more likely they are related. Similarly, the log-linear version of a logit model contains (i.e., allows) all possible relationships among predictor variables. This is done because we are not concerned with relationships among predictors but with relationships between predictors and the outcome variable.

In the ulcer data, this means that any logit model would include a *BC* term (and, because of the hierarchy principle, *B* and *C* terms by implication). Furthermore, all logit models include a term involving the dependent variable U. Any log-linear model that includes these components is a logit model also.

For example, the log-linear model [*BC*] [*BU*] can be interpreted as a logit model in which blood type is related to ulcers (*BU*), but city is not (no *CU* term). Note that *BU* is an interaction in a log-linear model but a main effect (of *B* on *U*) in the corresponding logit model. Furthermore, there is no interaction (*BCU* term), so the effect of blood type on ulcers is the same in each city.

The log-linear model that actually fit the data was [*BC*] [*BU*] [*CU*]. This can be interpreted as a logit model in which blood type affects ulcers and city affects ulcers, but there is no interaction between the blood type and city effects on ulcers. (As an example of a log-linear model that is not a logit model, consider the independence model: [*B*][*C*][*U*]. Because there is no *BC* term, this is not a logit model.)

8.5.3. Logistic Regression

In some cases, the outcome variable is dichotomous, but one or more predictors are continuous. In this case, an analysis is desired that is similar to multiple regression, but that takes into account the categorical nature of the dependent variable. Logistic regression is such a procedure; the outcome is the logit of the probability of the outcome event occurring. That is, the model is the same as a logit model, except that one or more predictors are continuous.

Because one or more predictors are continuous, the data are not easily summarized in a contingency table; such a table would have a large number of cells. Many of the cells would be empty, and few would contain more than one observation. This means that the G^2 and X^2 statistics are not good approximations to the chi-square distribution and cannot be used to assess the goodness of fit of the model. The utility of the predictor variables must be assessed by examining either the ratio of parameters to their standard errors (commonly denoted as t or z in computer output), or the difference in G^2 statistics for models with and without a set of parameters. The first method is similar to what is done in testing individual parameters in a regression model; the second method is comparable to testing the increase in R^2 when a set of predictors is added to a regression model.

As with log-linear models, extensions of logit and logistic regression models allow polytomous (more than two-category) dependent variables. Some polytomous variables are unordered (e.g., race), whereas others are ordered; both types of situations are handled by more complex versions of the models discussed previously.

8.6. Nonstandard Log-Linear and Logit Models

Partitioning of chi-square is a simple technique to learn and use, and it can go a long way toward testing hypotheses that are important to researchers. But because it cannot test all important hypotheses, a more general method is needed.

To provide a context, consider the data on admissions to graduate school at UC Berkeley that have become well publicized (see, e.g., Freedman, Pisani, & Purves, 1978, pp. 12–15). For six major areas of study, Table 8.5 shows data on what proportion of each gender were admitted in each of six major areas of study. The three variables will be called *Major, Gender,* and *Admission* (or *M*, *G*, and *A* for brevity).

We would presume that there might be a relationship between gender and major (a $G \times M$ effect) because males and females might tend to apply to different major areas at different rates. We would also presume that there might be a relationship between major area and admissions (an $M \times A$ effect) because some major areas get many more applicants per opening than others do.

If there is no bias in admissions, however, we would hope to find no relationship between gender and admission in any major area. (We oversimplify here and ignore the possibility of other confounding variables, such as prior achievement or aptitude.) The usual log-linear model described by this situation would be specified as [*GM*] [*MA*], to show inclusion of both $G \times M$ and $M \times A$ effects in the model. If there is bias, the additional term (*GA*) would be added to the model to show that gender is related to admissions.

If there is bias, and if that bias differs across major areas, then there would be a three-way Gender × Major × Admission interaction (*GMA*) in the model. This is the saturated log-linear model, with zero degrees of freedom; it will fit the data exactly but provides no simple interpretation of the data.

In fact, this occurs for the Berkeley data: The model of no three-way interaction does not fit the data and is rejected, leaving the conclusion that there is bias, and it differs across major areas. For those who limit themselves to the usual hierarchical log-linear models, there is not much else to say here, but inspection of the data in Table 8.5 shows something very interesting. For major area A, it appears that males are admitted at a lower rate than females. For each of the other major areas, there is no apparent difference in rates of admission.

Table 8.5 Graduate Admissions Data at UC Berkeley

Major Area	*Gender*	*% Admitted*
A	M	62
	F	82
B	M	63
	F	68
C	M	37
	F	34
D	M	33
	F	35
E	M	28
	F	24
F	M	6
	F	7

This description does not correspond to any standard log-linear model; therefore, a nonstandard log-linear model is needed. (In this instance, we could use partitioning chi-square, but that will not be possible for all nonstandard models.) A model of no bias in major areas B through F, but possible bias in major area A, has a likelihood ratio chi-square of 2.33 with 5 degrees of freedom, and thus fits the data quite well. The simplest way to represent the model is as a logit model, with admission (A) as the dependent variable. The model matrix (sometimes called a design matrix) for the logit form of the model is presented as follows:

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The rows of this matrix correspond to the 12 groups in the study, as shown in Table 8.5 (i.e., six major areas by two genders). Those who are used to looking at such matrices will notice that the first column represents the intercept term, and the next five columns represent the main effect of major area. There is no column for the main effect of gender, but there is one column for the interaction of gender and major. Normally, there would be five such columns; they would be the product of the gender effect with each of the five major area effects. Here, however, we are including such an interaction only for major area A. The omitted main effect for gender and the four omitted interaction terms produce the 5 degrees of freedom mentioned above for testing the model. Of course, the hypothesis tested here is post hoc, and the results must be considered tentative.

Nonstandard log-linear models illustrate one of the main trends in applied statistics discussed previously in the section on partitioning chi-square, the testing of context-dependent models. Nonstandard models also illustrate another trend, the increasing generality of statistical models. They provide a framework that includes as special cases many situations that were previously dealt with separately by other researchers. Most obviously, the usual hierarchical log-linear models and partitioning chi-square can be put in this framework. In addition, the nonstandard log-linear approach includes models for data with structural zeros, incomplete designs, models for symmetry and quasi-symmetry, models with linear restrictions on parameters, polynomial models, and many of Goodman's models for association with ordered variables (details can be found in Rindskopf, 1990). One general framework and one computer program can deal with this wide variety of problems.

8.7. Methods for Rates (Survival Analysis)

Data from studies with dichotomous outcome variables are not always best analyzed using logit or log-linear models. Some outcomes, such as marriage, divorce, contraction of an illness, or job termination, occur after different lengths of time for different people; these time or duration differences should be used in the data analysis. Furthermore, the event does not occur for every person in the study. If we were to try using time until occurrence of the event as an outcome variable, what length of time should be used when the event does not occur (at least during the period during which the study is conducted)? These issues are dealt with by a set of statistical procedures known collectively as *survival analysis.* This topic is treated in detail by Willett and Singer (Chapter 11, this volume); here I discuss one particular analytic method for survival analysis that is related to log-linear models.

The example I will use is a well-known data set to those doing survival analysis (Laird & Olivier, 1981). It is also a small enough data set to analyze by hand. Table 8.6 lists the number of deaths for old and young males who had either aortic or mitral heart valve

Table 8.6 Data on Heart Valve Replacement Operations

Type	*Age*	*Deaths*	*Exposure*	*Death Rate*
A	Y	4	1,259	3.177
M	Y	1	2,082	0.480
A	O	7	1,417	4.940
M	O	9	1,647	5.464

NOTE: For type of operation, A = aortic, M = mitral; for age, Y = young, O = old; exposure is in patient-months; death rate (= $1,000 \times$ Deaths/Exposure) is per 1,000 patient-months.

replacements. We could, of course, list the number of people in each group who did not die, and fit a logit model for the probability of dying as a function of age and type of valve. But this would ignore the fact that we are able to observe some people for longer periods of time than others. Furthermore, even if the same proportion of people eventually died in each group, those in some groups may live longer than those in other groups.

To deal with these problems, instead of looking at the number of deaths, we will calculate the death rate per unit of time that people are observed. To illustrate, if five people are observed for 2, 3, 3, 5, and 7 months, the total period of observation is 20 months. If two of these people die during our study, the death rate is $2/20 = .10$ deaths per person-month. Table 8.6 lists 1,000 times the death rate for each of the four groups. For example, the young subjects who received aortic valves have a death rate of 3.177 per 1,000 person-months.

Symbolically, the expected rate can be represented as F_i/z_i, where F_i is the expected number of deaths in group i, and z_i is the total exposure time of all people in group i. To extend the log-linear model so that rates instead of frequencies are modeled, we can write

$$\ln(F_i/z_i) = b_0 + b_1X_1 + b_2X_2 + \ldots,$$

where X_1, X_2, and so forth are predictor variables, and b_0, b_1, b_2, and so forth are the parameters of the model (like regression coefficients). The only difference between this model and the usual log-linear model is the denominator z_i.

Fitting the model with main effects of age and type of operation to the heart valve data, we find that the likelihood ratio chi-square is 3.223 with 1 degree of freedom. Although this model fits well, it can be simplified; tests of the statistical significance of the two parameters indicate that only the effect of age, not type of operation, is significantly different from zero. Fitting a model with only an age main effect gives a likelihood ratio chi-square of 3.790 with 2 degrees of freedom. This model also fits well. Examining the difference in the fit of the two models ($3.790 - 3.223 = 0.567$, with $2 - 1 = 1$ degree of freedom) shows that the second model fits no worse than the first model; the second is therefore preferred because it is more parsimonious.

Standard model tests would stop at this point, but examination of the death rates leads to consideration of another model. The death rates for three of the groups seem similar; only the young subjects who received mitral valves seem to have a lower death rate. Using nonstandard log-linear models, we can test the hypothesis that the other three groups have equal death rates. The likelihood ratio chi-square for this model is .909 with 2 degrees of freedom, which provides support for this model. (Distinguishing between this model and the model with only the main effect of age would require a greater amount of data.)

8.8. Latent Class Analysis

Latent class analysis provides an example of a recycled technique that is bringing a radical change in applied statistics for categorical data. Most quantitative psychologists are acquainted with factor analysis; in many ways, latent class analysis (LCA) is the categorical variable analog of factor analysis. As with factor analysis, LCA models presume that relationships among a number of observed variables can be explained by a smaller number of unobserved or latent variables. In these models, the variables we observe are presumed to be measured with error; we would rather observe the latent variable directly but cannot do so.

LCA is closely related to an area with which most researchers are at least somewhat familiar—genetic models for discrete traits such as blood type, eye color, and certain diseases. These genetic models presume that observed characteristics (phenotypes) are determined by unobserved characteristics (genotypes).

An interesting psychological example arises in considering Piaget's theories. In a group of children, there should be two types: Those who can conserve number and those who cannot. If we were to administer to children a four-item test to assess conservation and if there were no errors of responding, then children who can conserve should get each item right, but those who cannot should get each item wrong. According to the theory, no one should get some items right and others wrong.

Of course, no matter how well we write items and how well we develop scoring schemes to assess reasoning behind answers, through sheer perversity children will not accommodate us by responding perfectly. Would this mean that Piaget's theory was wrong? Or is it possible that a model for two kinds of children may still be right if we allow errors of responding?

Latent class models allow us to test such hypotheses. The simplest such model would include the two kinds (i.e., classes) of children specified by the original model. One type of child would be those who can conserve, and the other type would be children who cannot conserve. These types would be the two latent classes. For any item testing conservation, one type of child should have a high probability (though not necessarily perfect) of answering the item correctly, whereas the other type of child should have a low probability (though not necessarily zero) of answering correctly. If we have enough items (four in this case), the theory that there are only two types of children can be tested.

If this simple model is wrong, we can test other models, such as those that include a transition class for children who are "on their way" toward acquiring conservation. More complicated models can test theories about the sequence of acquisition of various types of conservation. Rindskopf (1987) discusses a variety of such models in the context of developmental psychology, as well as the work of others in this area. Magidson and Vermunt's chapter in this volume (Chapter 10) illustrates many extensions of the basic latent class model. The original articles that put LCA on a firm statistical basis are reprinted in Goodman (1978).

Latent class models illustrate a variety of trends in data analysis. First, they involve latent variables and therefore are more realistic than models that do not. Consequently, they (like factor analysis models) can be complex computationally and often involve various subtleties not encountered in most models that involve only observed variables.

Second, many latent class models are context dependent. This is especially true for those models of learning and development that hypothesize specific sequences in which skills should develop. Many special cases of latent class models can be devised to test specific theories and hypotheses.

Finally, latent class models also illustrate the trend toward generality in statistical models because many apparently different models fit within the latent class framework. One example is a model for a dichotomous outcome variable with error of classification; such a model can be tested using latent class analysis. However, the next example, which concerns missing categorical data, demonstrates that there are even more general frameworks that include latent class models as special cases.

8.9. Missing Data Problems

Missing data is a problem that plagues most researchers, and yet only recently have computational and theoretical advances enabled such problems to be treated appropriately. For categorical data, little progress was made for decades after Fisher (1930) used maximum likelihood to estimate the parameter of a genetic model with missing data.

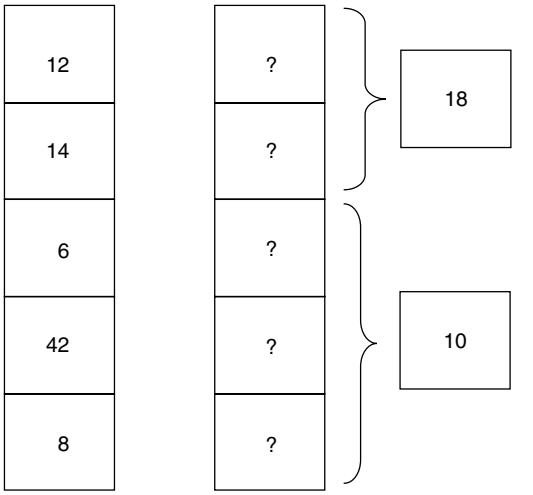
We can now realistically treat many missing data problems, at some cost in computational complexity. The approach I will use as an illustration results in a very general framework for analysis with categorical missing data, into which many special cases fit.

The general principles are very simple, although it is not always obvious how a particular case fits within the framework. First, one constructs a model expressing the relationships that would be observed if there were no missing data. The usual type of linear or log-linear model can often be used to represent these relationships. Then, another part of the model specifies how the hypothetical complete data are collapsed (i.e., summed) to form the observed data.

To see how this framework is implemented, consider the following example: A study has been done in which an ordered variable has been more finely classified for some subjects than others, perhaps because of cost considerations. In a study of psychotherapy, some patients might be rated as either improved or not, whereas others might have each of those categories more finely divided, such as much improved or slightly improved; and stable, slightly worse, or much worse.

Figure 8.1 illustrates some hypothetical data for such a study. In a real study, other variables would be included such as predictors of improvement and control variables, but for clarity, these are omitted from the figure. The numbers in the cells indicate observed frequencies; a question mark indicates that the frequency is not observed. The five frequencies on the left-hand side of the figure represent people who were actually assessed on the 5-point scale. The numbers 18 and 10 are observed for people who are classified only as either improved or not. Each of these two frequencies is the sum of cells that we would like to observe directly but cannot. We do not know how many of the 18 who were rated only as improved were really much improved and how many were slightly improved.

Figure 8.1 Data From a Coarsely Categorized Variable Conceptualized as Incompletely Observed Data



A statistical model would be specified for the complete data that would be observed if everyone were measured on the more finely categorized variable. The remaining part of the model would specify that certain cells are not observed; only their sums, such as those indicated in the figure, are observed. In this case, the second part of the model would show that for the second group of people, two unobserved cells would be merged (i.e., summed) to create the observed category "improved," and three unobserved cells would be collapsed to create the observed category "not improved." Researchers familiar with confirmatory factor analysis and structural equation models will probably guess (correctly) that some missing data models have unidentified parameters, so the analysis is sometimes complicated.

A wide variety of missing data cases can be analyzed using this framework. These include estimating frequencies when some people are missing data on some variables; fitting log-linear models when there are missing data; fitting latent class models and, more generally, models with fused cells (such as genetics models); fitting latent class models when some observed variables have missing data; fitting models when some variables are more finely categorized than others; and fitting models with various assumptions about the missing data process. Some of the above models were not previously conceptualized as missing data problems, so it was not realized how many situations could be treated within one general framework. Rindskopf (1992) describes these models in detail.

Even more general models for missing data can be estimated using the Bayesian program BUGS (Spiegelhalter, Thomas, & Best, 1999; Spiegelhalter, Thomas, Best, & Gilks, 1996). BUGS was primarily developed for Bayesian analysis with missing data, but it has been applied to a wide variety of statistical models. The categorical data models for which it has been used include logistic regression, Poisson regression, item response theory, latent class analysis, multilevel (nested) models, and log-linear models.

One trend illustrated by this example is obviously the move toward a comprehensive, general model that includes many special cases. This approach also requires numerical methods involving heavy computation, especially for large problems. Even relatively large problems can now be analyzed using a microcomputer.

8.10. Summary and Implications

Categorical data analysis, like most of applied statistics, has become more realistic, more general, more comprehensive, and more complex. In fact, there are models even more general than some discussed here. For example, generalized linear models (McCullagh & Nelder, 1989) have been developed that include regression, ANOVA, logistic regression, and log-linear models (among others) as special cases. Computer hardware and software (e.g., BUGS, Mplus, LEM, SPlus) that did not previously exist have made many of these new methods possible, and have stimulated the development of more statistical methods.

Many other areas of recent research have expanded the set of tools for analyzing categorical data. Some of these are too specialized for discussion here (e.g., exact methods, meta-analysis, and data-mining methods such as CHAID, CART, and neural networks). Others are covered in separate sections of this volume (e.g., multilevel models, longitudinal models, item response theory, and structural equation models).

Researchers must also keep in mind that analysis has implications for design; a badly designed study cannot be rescued by a brilliant analysis. Complex statistical methods require additional design considerations beyond those encountered with more traditional designs. In particular, latent variable models and multilevel models cannot be used without a properly designed study.

I hope that the examples presented here have provided a taste for the exciting new developments in

categorical data analysis. We have not only exciting new methods but also exciting old methods; what could be better?

REFERENCES

Aiken, L. C., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Fisher, R. A. (1930). *Statistical methods for research workers* (3rd ed.). Edinburgh, UK: Oliver & Boyd.

Freedman, D., Pisani, R., & Purves, R. (1978). *Statistics*. New York: W. W. Norton.

Friendly, M. (2000). *Visualizing categorical data*. Cary, NC: SAS Publishing.

Goleman, D. (1985, October 22). Strong emotional response to disease may bolster patient's immune system. *New York Times*, p. C1.

Goodman, L. A. (1978). *Analyzing qualitative/categorical data: Log-linear models and latent structure analysis*. Cambridge, MA: Abt Books.

Laird, N. M., & Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association, 76*, 231–240.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.

Rindskopf, D. (1987). Using latent class analysis to test developmental models. *Developmental Review, 7*, 66–85.

Rindskopf, D. (1990). Nonstandard loglinear models. *Psychological Bulletin, 108*, 150–162.

Rindskopf, D. (1992). A general approach to categorical data analysis with missing data using generalized linear models with composite links. *Psychometrika, 57*, 29–42.

Rindskopf, D. (1999). Some hazards of using nonstandard log-linear models, and how to avoid them. *Psychological Methods, 4*, 339–347.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. New York: Cambridge University Press.

Smith, J. (1990, October 7). Take my advice. *Los Angeles Times Magazine*, p. 6.

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (1999). *WinBUGS Version 1.2 user manual*. Cambridge, UK: MRC Biostatistics Unit.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1996). *BUGS: Bayesian inference using Gibbs sampling, Version 0.5 (Version ii).* Cambridge, UK: MRC Biostatistics Unit.

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics, 19*, 251–253.

Chapter 9

ORDINAL REGRESSION MODELS

VALEN E. JOHNSON

JAMES H. ALBERT

9.1. REGRESSION MODELS FOR ORDINAL DATA

Ordinal data are the most frequently encountered type of data in the social sciences. Survey data, in which respondents are asked to characterize their opinions on scales ranging from *strongly disagree* to *strongly agree,* are a common example of such data. For our purposes, the defining property of ordinal data is that there exists a clear ordering of the response categories but no underlying interval scale between them. For example, it is generally reasonable to assume an ordering of the form

$$\textit{strongly disagree} < \textit{disagree} < \textit{don't know} < \textit{agree} < \textit{strongly agree},$$

but it usually does not make sense to assign integer values to these categories. Thus, statements of the type

$$\textit{disagree} - \textit{strongly disagree} = \textit{agree} - \textit{don't know}$$

are not assumed to be valid.

9.2. ORDINAL DATA VIA LATENT VARIABLES

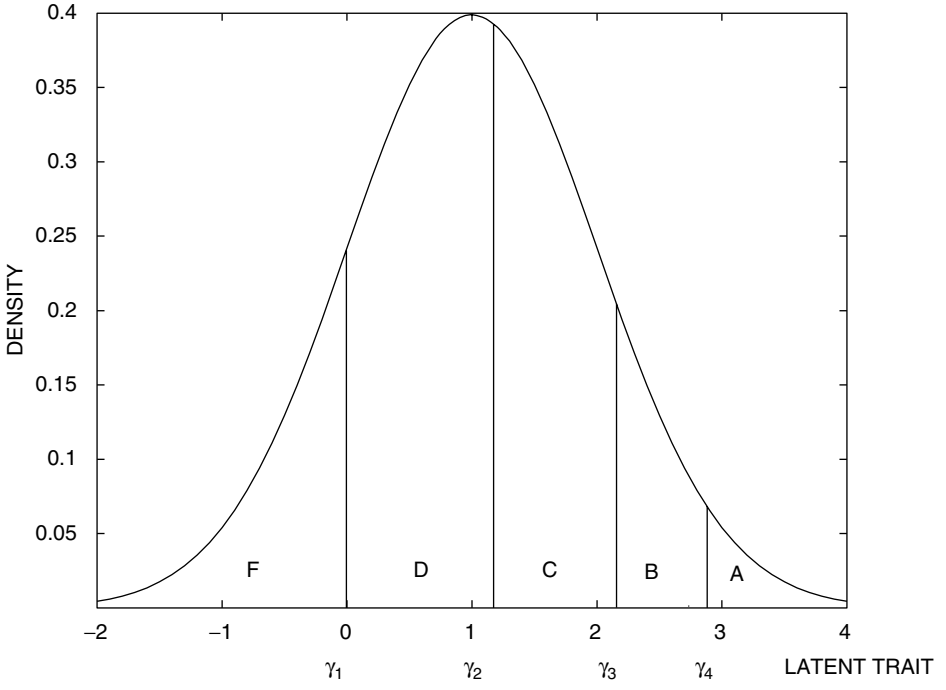
The most natural way to view ordinal data is to assume the existence of an underlying latent (unobserved) variable associated with each response. Such variables are often assumed to be drawn from a continuous distribution centered on a mean value that varies from individual to individual. Often, this mean value is modeled as linear function of the respondent's covariate vector.

To illustrate this concept, suppose that we are interested in estimating the effects of, say, SAT scores on the performance of students in an introductory college statistics course. Assuming that the course is graded on an A to F scale, the latent variable approach to this problem can be defined by assuming the existence of four category cutoffs on the latent scale that separate the observed values of the latent variables into the observed grade categories. Also, because the response categories are ordered, we must impose a corresponding constraint on the grade cutoffs. Letting the upper grade cutoff for an F be denoted by γ_1, the upper grade cutoff for a D be denoted by γ_2, and so on, this ordering constraint may be stated mathematically as

$$-\infty < \gamma_1 \equiv 0 \leq \gamma_2 \leq \gamma_3 \leq \gamma_4 \leq \gamma_5 \equiv \infty.$$

AUTHORS' NOTE: Much of the material in this chapter was adapted from Johnson, V. E. & Albert, J. H., *Ordinal Data Modeling.*

Figure 9.1 Latent Trait Interpretation of Ordinal Classification



NOTE: In this plot, the logistic density represents the distribution of latent traits for a particular individual. It is assumed that a random variable is drawn from this density, and the value of this random variable determines an individual's classification. For example, if a deviate of 0.5 is drawn, the individual receives a D grade.

Note that the upper cutoff for an A, γ_5, is assumed to be unbounded. For notational convenience, we define $\gamma_0 = -\infty$.

Geometrically, Figure 9.1 illustrates the way a latent variable formulation can be used to define a model for the probability that students in the statistics class receive grades A through F, assuming grade cutoffs $\gamma_1, \ldots, \gamma_4$. In this figure, we imagine a latent variable—say, Z—that underlies the generation of the ordinal data. In extending this framework to the regression setting, we further assume that the variable Z may be expressed

$$Z = \mathbf{x}'\beta + \varepsilon, \tag{1}$$

where ε is a random variable drawn from the standard logistic distribution. When Z falls between the grade cutoffs γ_{c-1} and γ_c, the observation is classified into category c. To link this model for the data generation to the probability that an individual receives a particular grade, let f denote the density of the standard logistic distribution, and let F denote the logistic distribution function. Denote by p_c the probability that an individual receives a grade of c. Then from (1), it follows that

$$\begin{aligned} p_c &= \int_{\gamma_{c-1-x'\beta}}^{\gamma_{c-x'\beta}} f(z)dz \\ &= \Pr(\gamma_{c-1} < Z < \gamma_c) \\ &= F(\gamma_c - \mathbf{x}'\beta) - F(\gamma_{c-1} - \mathbf{x}'\beta). \end{aligned}$$

The latent variable formulation of the problem thus provides a model for the probability that a student receives a particular grade in the course or, in the more general case, that a response is recorded in a particular category. If we also assume that the responses or grades for a sample of n individuals are independent of one another given these probabilities, the sampling distribution for the observed data is given by a multinomial distribution.

To specify this multinomial distribution, let us assume that there are C possible grades, denoted by $1, \ldots, C$. Also, suppose that n items are observed and that the grades or categories assigned to these n items are denoted by $y_1, \ldots, y_n$; y_i denotes the

grade observed for the ith individual.[1] Associated with the ith individual's response, we define a continuous latent variable Z_i, and, as above, we assume that $Z_i = x_i'\beta + \varepsilon_i$, where x_i is the vector of covariates associated with the ith individual, and ε_i is distributed according to the distribution F. We observe the grade $y_i = c$ if the latent variable Z_i falls in the interval (γ_{c-1}, γ_c). If p_{ic} denotes the probability that a single response from the ith respondent falls into category c, we may write this probability as

$$\begin{aligned} p_{ic} &= \Pr(\gamma_{c-1} < Z_i < \gamma_c) \\ &= F(\gamma_c - \mathbf{x}_i'\beta) - F(\gamma_{c-1} - \mathbf{x}_i'\beta). \end{aligned} \tag{2}$$

In addition, let $\mathbf{p}_i$ denote the vector of probabilities associated with assignment of the ith item into the categories $1, \ldots, C$; that is, $\mathbf{p}_i = (p_{i1}, \ldots, p_{iC})$. Let $\mathbf{y} = (y_1, \ldots, y_n)$ denote the observed vector of responses for all individuals. It then follows that the probability of observing the data $\mathbf{y}$, for a fixed value of the probability vectors $\{\mathbf{p}_i\}$, is given by a multinomial density proportional to

$$\Pr[\mathbf{y}|\{\mathbf{p}_i\}] \propto \prod_{i=1}^{n} p_{i_{y_i}}. \tag{3}$$

Substituting the value of p_{ic} from (2) leads to the following expression for the likelihood function for β:

$$\begin{aligned} L(\beta, \gamma) = \prod_{i=1}^{n} [&F(\gamma_{y_i} - \mathbf{x}_i'\beta) \\ &- F(\gamma_{y_i - 1} - \mathbf{x}_i'\beta)]. \end{aligned} \tag{4}$$

In terms of the latent variables $\mathbf{Z}$, the likelihood function may be reexpressed as

$$L(\beta, \gamma, \mathbf{Z}) = \prod_{i=1}^{n} f(Z_i - \mathbf{x}_i'\beta) I(\gamma_{y_i - 1} \le Z_i < \gamma_{y_i}), \tag{5}$$

where $I(\cdot)$ indicates the indicator function. Note that the latent variables Z_i may be integrated out of (5) to obtain (4).

1. In defining the multinomial sampling density for an ordinal response, we assume that the multinomial denominator associated with each response is 1. For the more general case in which the ordinal responses are grouped by covariate, so that the multinomial denominator (say, m_i) for the ith individual is greater than 1, this simply means that the m_i observations associated with the ith individual are considered independently in our model description. Because a multinomial observation with a denominator greater than $m_i > 1$ can always be reexpressed as m_i multinomial observations with denominator 1, this distinction is irrelevant for most of the theoretical development discussed in this chapter, and it somewhat simplifies notation and exposition. Of course, the likelihood function is unaffected by this change. The distinction only becomes important in defining the deviance statistic and individual deviance contributions, but further comments on this point are delayed until these quantities are introduced in Section 9.4.

9.2.1. Cumulative Probabilities and Model Interpretation

Ordinal regression models are often specified in terms of cumulative probabilities rather than individual category probabilities. If we define

$$\theta_{ic} = p_{i1} + p_{i2} + \cdots + p_{ic},$$

then the regression component of an ordinal model of the form (2) may be rewritten as

$$\theta_{ic} = F(\gamma_{ic} - \mathbf{x}_i'\beta). \tag{6}$$

For example, if a logistic link function is assumed, equation (6) becomes

$$\log\left(\frac{\theta_{ic}}{1 - \theta_{ic}}\right) = \gamma_c - \mathbf{x}_i'\beta. \tag{7}$$

Note that the sign of the coefficient of the linear predictor is negative, as opposed to the positive sign of this term in the usual binary regression setting.

An interesting feature of model (7) is that the ratio of the odds for the event $y_1 \le c$ to the event $y_2 \le c$ is

$$\frac{\theta_{1c}/(1 - \theta_{1c})}{\theta_{2c}/(1 - \theta_{2c})} \exp[-(\mathbf{x}_1 - \mathbf{x}_2)'\beta], \tag{8}$$

independently of the category of response, c. For this reason, (7) is often called the *proportional odds model* (see, e.g., McCullagh, 1980).

Another common regression model for ordinal data, the proportional hazards model, may be obtained by assuming a complementary log-log link in (2). In this case,

$$\log[-\log(1 - \theta_{ic})] = \gamma_c - \mathbf{x}_i'\beta.$$

If one interprets $1 - \theta_{ic}$ as the probability of survival beyond (time) category c, this model may be considered a discrete version of the *proportional hazards model* proposed by Cox (1972). Further details concerning the connection between this model and the proportional hazards model may be found in McCullagh (1980).

Another link function often used to model cumulative probabilities of success is the standard normal distribution. With such a link, (2) becomes

$$\Phi(\theta_{ic})^{-1} = \gamma_c - \mathbf{x}_i'\beta.$$

This model is referred to as the *ordinal probit model.* The ordinal probit model produces predicted probabilities similar to those obtained from the proportional odds model, just as predictions from a probit model for binary data produce predictions similar to those

obtained using a logistic model. However, the ordinal probit model possesses a property that makes sampling from its posterior distribution particularly efficient. For that reason, it may be preferred over other model links (at least in preliminary studies) if a Bayesian analysis is to be performed.

9.3. Parameter Constraints and Prior Models

An ordinal regression model with C categories and $C-1$ unknown cutoff parameters $\gamma_1, \ldots, \gamma_{C-1}$ is overparameterized if an intercept is included in the regression function. To see this, note that if we add a constant to every cutoff value and subtract the same constant from the intercept in the regression function, the values of $\gamma_c - \mathbf{x}_i'\beta$ used to define the category probabilities are unchanged. Two approaches might be taken toward resolving this identifiability problem. The first is to simply fix the value of one cutoff, usually the first. In other words, we might assume that γ_1, the upper cutoff for the lowest category of response, is 0. A second approach that can be taken for establishing identifiability of parameters is to specify a proper prior distribution on the vector of category cutoffs, γ. Of course, for ordinal data containing more than three categories, a Bayesian approach toward inference requires that a prior distribution be specified for at least one category cutoff, regardless of which approach is taken. For that reason, we now turn our discussion to prior specifications for ordinal regression models.

9.3.1. Noninformative Priors

In situations in which little prior information is available, the simplest approach toward constructing a prior distribution over the category cutoffs and regression parameter begins by fixing the value of one cutoff, usually γ_1, at 0. The values of the remaining cutoffs are then defined relative to the first, and posterior variances of category cutoffs represent the variances of the contrasts $\gamma_c - \gamma_1$. After fixing the value of one cutoff, a uniform prior can then be assumed for the remaining cutoffs, subject, of course, to the constraint that

$$\gamma_1 \leq \cdots \leq \gamma_{C-1},$$

Normally, the components of the category cutoff vector and the regression parameter are assumed a priori to be independent, and a uniform prior is also taken for β.

This choice of prior results in a maximum a posteriori (MAP) estimate of the parameter values that is identical to the maximum likelihood estimation (MLE). In general, these point estimators provide satisfactory estimates of the multinomial cell probabilities when moderate counts are observed in all C categories. However, if there are categories in which no counts are observed or in which the number of observations is small, the MLE and MAP estimates will differ significantly from the posterior mean. Furthermore, the bias and other properties of estimators of the extreme category cutoffs may differ substantially from the corresponding properties of estimators of the interior category cutoffs.

9.3.2. Informative Priors

As in the case of binary regression, informative priors on the components of γ and β may be specified using the conditional means approach of Bedrick, Christensen, and Johnson (1996). However, in addition to the specification of an independent assessment for each component of the regression parameter β, an independent assessment must also be specified for each random component of γ. If the dimension of the regression parameter is a and the dimension of the random component of γ is b, then $a+b$ independent assessments are needed for the specification of a proper prior. For example, if an intercept term is included in the regression parameter, so that the total dimension of β is b, and γ_1 is set to 0 so that there are $C-2$ random components of γ, then $b+C-2$ independent assessments must be solicited to set the joint prior on γ and β. The precision of each assessment must also be specified.

In setting the prior using the conditional means approach, it is often easier to specify prior estimates of cumulative success probabilities than it is to specify estimates of the probability of observing specific categories of response. Also, to establish identifiability of parameters in the prior, we need to estimate at least one cumulative probability for each random component of the vector γ. In other words, if there are four categories of response and $\gamma_1 = 0$, at least one prior assessment must be made of the cumulative probability that a response is observed to be less than or equal to the second category ($\gamma_i \leq 2$), and at least one prior assessment must be made of the probability of observing at least one response less than or equal to the third category. In addition, the design matrix selected for the covariate values (including category cutoffs) should be invertible.

Suppose then that there are a unknown components of the cutoff vector γ and b unknown components of the regression vector β. To construct a conditional means prior, we must examine $M = a + b$ values of the covariate vector x—call these covariate vectors $x_2, \ldots, x_M$. For each of the covariate vectors x_j, we specify a prior estimate and prior precision of our estimate of the corresponding cumulative cutoff probability $\theta_{(j)}$. Thus, for each covariate value, two items are specified:

1. An assessment at the cumulative probability $\theta_{(j)}$—call this assessment g_j.
2. A statement about the precision of this assessment in terms of the number of "prior observations." Denote this prior sample size by K_j.

This prior information about $\theta_{(j)}$ can be incorporated into the model specification using a Dirichlet density with parameters $K_j g_j$ and $K_j(1 - g_j)$ if the prior distributions of the cumulative probabilities $\theta_{(1)}, \ldots, \theta_{(M)}$ are assumed to be independent. In that case, it follows that the joint prior density is given by the product

$$g(\theta_{(1)}, \ldots, \theta_{(M)}) \propto \prod_{j=1}^{M} \theta_{(j)}^{K_j g_j - 1} (1 - \theta_{(j)})^{K_j(1-g_j)-1}.$$

By transforming this prior on the cumulative probabilities back to (β, γ), the induced conditional means prior may be written

$$g(\beta, \gamma) \propto \prod_{j=1}^{M} F(\gamma_{(j)} - \mathbf{x}_j'\beta)^{K_j g_j} \cdot [1 - F(\gamma_{(j)} - \mathbf{x}_j'\beta)]^{K_j(1-g_j)} f(\gamma_{(j)} - \mathbf{x}_j'\beta), \quad (9)$$

subject to $\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_{C-1}$. As before, $F(\cdot)$ denotes the link distribution function, and $f(\cdot)$ is the link density.

9.4. Residual Analysis and Goodness of Fit

Associated with every multinomial observation are C categories, and an individual's response (or absence of a response) in each of these categories can be used to define a residual. For binomial data ($C = 2$), two such residuals are $y_i - n_i p_i$ and $y_i - n_i(1 - p_i)$. Of course, if you know the value of the first residual—that is, if you know p_i—you can figure out the value of the second, which depends only on $(1 - p_i)$ (because y_i and n_i are assumed known). The same is true for ordinal data with C categories; if you know the values of p_{ic} for $C - 1$ of the categories, you can figure out the probability for the last because the probabilities have to sum to 1. Thus, for ordinal data, we potentially have $C - 1$ residuals for each multinomial observation.

This increase in dimensionality, from 1 to $C - 1$, complicates residual analyses. Not only are there more residual values to be examined, but the $C - 1$ residuals from each observation are also correlated. It is therefore not clear how classical residuals (e.g., Pearson, deviance, and adjusted deviance residuals) should be displayed and analyzed. In the case of Bayesian residual analyses, the standard Bayesian residual and posterior-predictive residuals both involve $(C - 1)$-dimensional distributions, which again complicates model criticism. One possible solution to this problem is to create a sequence of binary residuals by collapsing response categories. For example, we might redefine a "success" as exceeding the first, second, ..., or $(C - 1)$st category. The resulting binary residuals can then be analyzed using the procedures described in, for example, Chapter 3 of Johnson and Albert (1999), keeping in mind that the residuals defined for each success threshold are highly correlated. From a practical viewpoint, the binary residuals formed using exceedance of the extreme categories (Categories 1 and $(C - 1)$) are often the most informative in identifying outliers, and so attention might be focused first on these residuals.

In contrast, residuals based on the vector of latent variables $\mathbf{Z}$ do not suffer from the problem of dimensionality because only a single latent variable is defined for each individual. The latent residual for the ith observation is defined as

$$r_{i,L} = Z_i - x_i'\beta.$$

Nominally, the residuals $r_{1,L}, \ldots, r_{n,L}$ are independently distributed as draws from the distribution F. Deviations from the model structure should therefore be reflected as deviations of the observed values of these quantities from typical samples drawn from F. For this reason, case analyses are generally easier to perform and interpret using the scalar-valued latent residuals.

To judge the overall goodness of fit of an ordinal regression model, we can use the deviance statistic, defined as

$$D = 2 \sum_{i=1}^{n} \sum_{j=1}^{C} I(y_i = j) \log(I(y_i = j)/\hat{p}_{ij}),$$

where $\hat{p}_{ij}$ denotes the maximum likelihood estimate of the cell probability p_{ij}, and I is the indicator function. In this expression, the term $I() \log(I()/\hat{p}_{ij})$ is

assumed to be 0 whenever the indicator function is 0. The degrees of freedom associated with the deviance statistic is $n - k - (C - 1)$, where k is the number of regression parameters in the model, including the intercept. Asymptotically, the deviance statistic for ordinal regression models has a chi-square distribution only when observations are grouped according to covariate values and the expected counts in each cell become large. When only one observation is observed at each covariate value, the deviance statistic is not well approximated by a chi-square distribution.[2]

Besides its role as a goodness-of-fit statistic, the deviance statistic can also be used for model selection. Perhaps surprisingly, the distribution of differences in deviance statistics for nested models is often remarkably close to a chi-square random variable, even for data in which the expected cell counts are relatively small. The degrees of freedom of the chi-square random variable that approximates the distribution of the difference in deviances is equal to the number of covariates deleted from the larger model to obtain the smaller model.

Related to the model deviance are the contributions to the deviance that accrue from individual observations. In the case of binary residuals, the signed square root of these terms was used to define the deviance residuals. However, for ordinal data, it is preferable to examine the values of the deviance contribution from individual observations directly, or

$$d_i = 2 \sum_{j=1}^{C} I(y_i = j) \log(I(y_i = j)/\hat{p}_{ij}).$$

Observations that contribute disproportionately to the overall model deviance should be regarded with suspicion.[3]

Turning to Bayesian case analyses, posterior-predictive residuals provide a generally applicable tool by which model adequacy can be judged and outlying observations can be identified. As in the case of binary regression, observations for which the residual posterior-predictive distributions are concentrated away from zero represent possible outliers.

9.5. EXAMPLES

9.5.1. Grades in a Statistics Class

For a simple application of this methodology, we first consider the grades received by students in an advanced statistics class. Interest in this example focuses on predicting the grades of the students in this class using their SAT math scores and their grades in a prerequisite class. The data for this example are depicted in Table 9.1. We begin by illustrating maximum likelihood estimation for a proportional odds model. After discussing classical model-checking procedures, we then discuss Bayesian analyses using both informative and noninformative priors.

9.5.1.1. Maximum Likelihood Analysis

As a first step in the analysis, we assume that the logit of the probability that a student receives a grade in category c or worse is a linear function of his or her SAT-M score. That is, we assume a proportional odds model of the form

$$\log\left(\frac{\theta_{ic}}{1 - \theta_{ic}}\right) = \gamma_c - \beta_0 - \beta_1 \times \text{SAT} - \text{M}_i. \quad (10)$$

Because an intercept is included in this relation, to establish identifiability, we fix $\gamma_1 = 0$.

The maximum likelihood estimates and associated standard errors for the parameters γ and β are displayed in Table 9.2. These estimates were obtained using MATLAB routines described in Johnson and Albert (1999) and are available from that publication's Web site. The corresponding estimates of the fitted probabilities that a student receives each of the five possible grades are plotted as a function of SAT-M score in Figure 9.2. In this figure, the white area reflects the probability that a student with a given SAT-M received an A, the lightly shaded area the probability of a B, and so on. From the plot, we see that the probability that a student with a 460 SAT-M score receives a D or F is about 57%, that a student scoring 560 on the SAT-M has approximately a 50% chance of receiving a B, and

2. For grouped ordinal data, a more general definition of the deviance is needed. Letting y_{ij} denote the observed counts in category j for the ith observation, the deviance statistic can be redefined as

$$2 \sum_{i=1}^{n} \sum_{j=1}^{C} y_{ij} \log(y_{ij}/\hat{y}_{ij}),$$

where $\hat{y}_{ij}$ denotes the maximum likelihood estimate of the expected cell counts y_{ij}. As the expected number of counts in each cell of every observation approaches infinity (i.e., > 5), the distribution of this more general form of the deviance statistic does approach a chi-square distribution. Whenever it is possible to group observations, this form of the deviance function should therefore be used when assessing goodness of fit and for model selection.

3. For grouped ordinal data, an alternative definition of the deviance contribution from an individual observation is

$$\frac{2}{m_i} \sum_{j=1}^{C} y_{ij} \log(y_{ij}/\hat{y}_{ij}),$$

where $m_i = \sum_j y_{ij}$.

Table 9.1 Grades for a Class of Statistics Students

Student #	*Grade*	*SAT-M Score*	*Grade in Previous Statistics Course*
1	D	525	B
2	D	533	C
3	B	545	B
4	D	582	A
5	C	581	C
6	B	576	D
7	C	572	B
8	A	609	A
9	C	559	C
10	C	543	D
11	B	576	B
12	B	525	A
13	C	574	F
14	C	582	D
15	B	574	C
16	D	471	B
17	B	595	B
18	D	557	C
19	F	557	A
20	B	584	A
21	A	599	B
22	D	517	C
23	A	649	A
24	B	584	C
25	F	463	D
26	C	591	B
27	D	488	C
28	B	563	B
29	B	553	B
30	A	549	A

NOTE: The first column is student number. The second column lists the grade received in the class by the student, and the third and fourth columns provide the SAT-math score and grade for a prerequisite statistics course.

Table 9.2 Maximum Likelihood Estimates and Standard Errors for Proportional Odds Model for Statistics Class Grades Example

Parameter	*Estimate*	*Standard Error*
γ_2	2.22	0.64
γ_3	3.65	0.78
γ_4	6.51	1.33
β_0	–20.08	6.98
β_1	.0430	0.012

that a student who scored 660 on his or her SAT-M has a better than 80% chance of earning an A in the course.

An important property of the ordinal regression model that underlies the model for these data is that the interpretation of regression parameters is invariant with respect to the number of classification categories used. In the present case, the regression parameter β in the proportional odds model has the same interpretation as would the regression parameter appearing in a logistic model, in which grade categories were collapsed into a pass/fail system (i.e., if Ds and Fs were considered failing and As to Cs were considered passing). Further discussion of this point within the context of this example may be found in Johnson and Albert (1999).

As a cursory check for model fit, we plotted the contributions to the deviance from individual observations against observation number in Figure 9.3. The most extreme observation in the proportional odds model appears to be Student 19, who received an F in the course while having an above-average SAT score of 559. It is also interesting to note that Student 30's grade resulted in the second highest deviance contribution; this student had a slightly below-average SAT-M score but received an A in the course.

For purposes of comparison, we next fit the ordinal probit model to the same data. In this case, the ordinal probit model takes the form

$$\theta_{ic} = \Phi(\gamma_c - \beta_0 - \beta_1 \times \text{SAT} - \text{M}_i). \qquad (11)$$

As before, an intercept was included in this model because γ_1 was assigned the value 0.

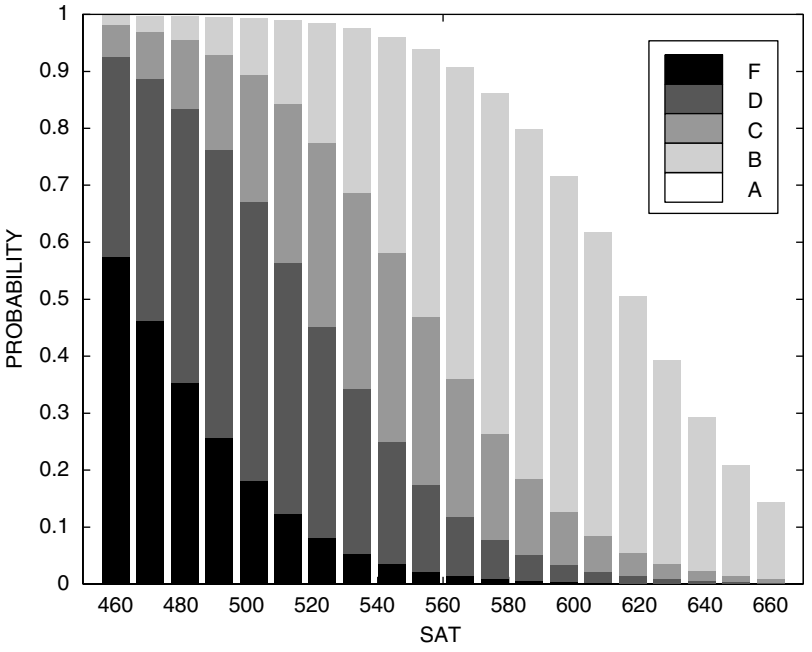
The maximum likelihood estimates for the probit model appear in Table 9.3 and were obtained using MATLAB functions described in Johnson and Albert (1999).

As in the proportional odds model, one can plot the deviance contributions from each observation. The appearance of this plot was almost identical to Figure 9.3, and so comments regarding the fit of the proportional odds model to individual student marks apply to the ordinal probit model as well. The similarity of the two deviance plots is a consequence of the fact that the fitted values under each model are nearly identical. This point is illustrated in Figure 9.4, in which the predicted cell probabilities under the two models are plotted against one another. The deviance under the ordinal probit model was 73.5, but it was 72.7 under the proportional odds model.

9.5.1.2. Bayesian Analysis With a Noninformative Prior

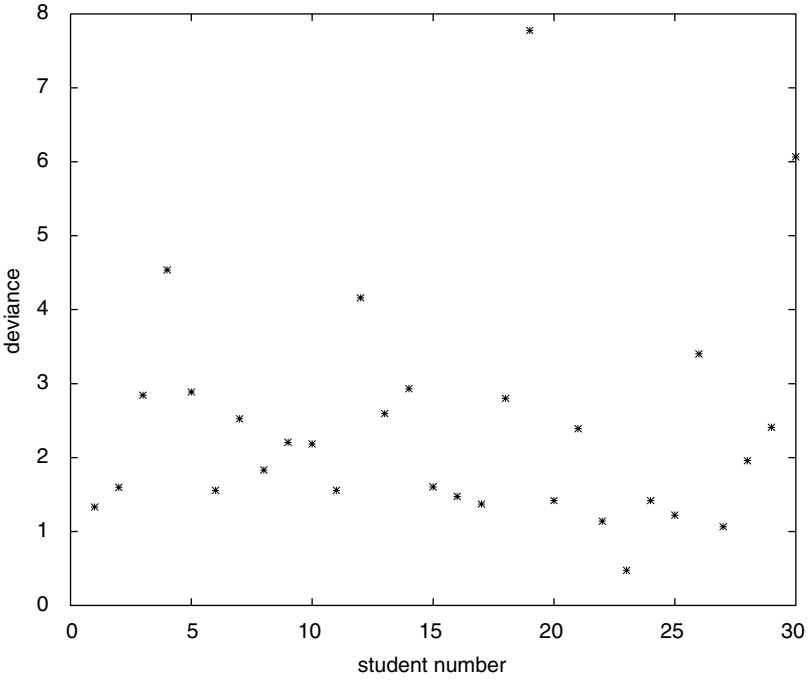
To further investigate the relationship between the student grades and SAT-M score, we next considered a Bayesian model using a vague prior on the parameters γ and β. Because of the similarity of fitted values obtained under the ordinal probit and proportional hazards model, as well as the computational simplicity of sampling from the ordinal probit model using Cowles's

Figure 9.2 Fitted Multinomial Probabilities From the Maximum Likelihood Fit of the Proportional Odds Model



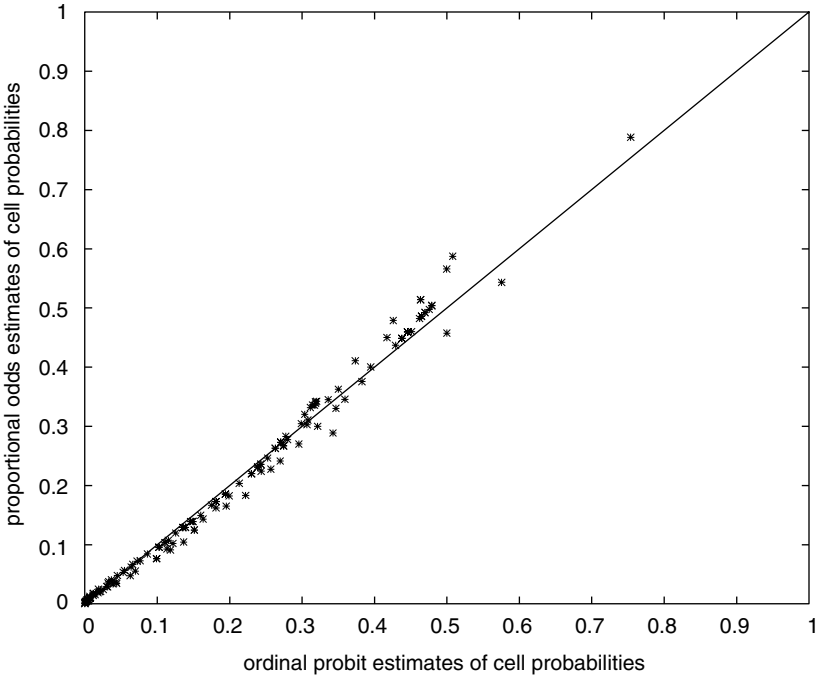
NOTE: For each SAT value, the five shaded areas of the stacked bar chart represent the fitted probabilities of the five grades.

Figure 9.3 Deviance Contributions in the Proportional Odds Model for Student Grades



NOTE: This plot does not depict deviance residuals, as the square root of the deviance contributions was not taken (nor was there a natural way to attribute a sign to each observation).

Figure 9.4 Fitted Probabilities Under the Ordinal Probit Model Versus Fitted Probabilities for the Proportional Odds Model



NOTE: All 150 predicted cell probabilities from the 30 observations are shown.

Table 9.3 Maximum Likelihood Estimates and Standard Errors for Ordinal Probit Model

Parameter	*Estimate*	*Asymptotic Standard Deviation*
γ_2	1.29	0.35
γ_3	2.11	0.41
γ_4	3.56	0.63
β_0	–11.22	3.64
β_1	.0238	0.0063

Table 9.4 Simulation Estimates of the Posterior Means and Standard Deviations for the Ordinal Probit Model Using Vague Priors

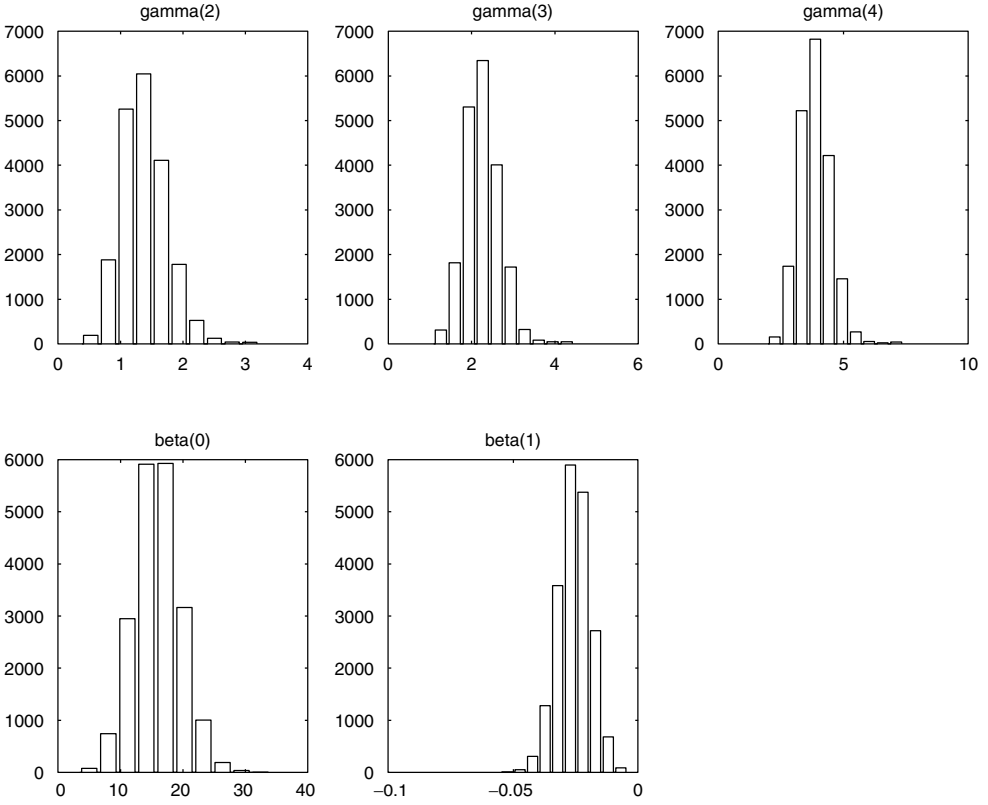
Parameter	*Posterior Mean*	*Posterior Standard Deviation*
γ_2	1.38	0.37
γ_3	2.26	0.42
γ_4	3.86	0.63
β_0	–12.05	3.73
β_1	.0257	0.0065

algorithm (Cowles, 1996), we restrict attention to the probit link.

In applying Cowles's algorithm to these data, we initialized the parameter vectors with the maximum likelihood values. We then performed 20,000 Monte Carlo Markov chain (MCMC) iterations. The MCMC sample estimates of the posterior means of the parameter values are displayed in Table 9.4 and indicate that the posterior means agree well with the maximum likelihood estimates provided in Table 9.3. This fact suggests that the posterior distribution of the parameter estimates is approximately normal. The histogram estimates of the marginal posterior distributions displayed in Figure 9.5 support this conclusion.

A by-product of the MCMC algorithm used to estimate the posterior means of the parameter estimates is the vector of latent variables $\mathbf{Z}$. As discussed at the end of Section 9.4, these variables provide a convenient diagnostic for detecting outliers and assessing goodness of fit. A priori, the latent residuals $Z_1 - x_1'\beta, \ldots, Z_n - x_n'\beta$ are a random sample from a $N(0, 1)$ distribution. Thus, deviations in the values of the latent residuals from an independent sample

Figure 9.5 Histogram Estimates of the Marginal Posterior Distributions of the Regression and Category Cutoff Parameters in the Statistics Grades Example



of standard normal deviates are symptomatic of violations of model assumptions.

A normal scores plot of the posterior means of the latent residuals is depicted in Figure 9.6. As might be predicted from the deviance plots presented above, the most extreme latent residuals correspond to Students 19 and 30. The value of Student 19's latent residual, −2.57, is smaller than would be expected in a sample of this size, and thus this observation might be regarded as an outlier. Student 30's latent residual value is 2.27 and is somewhat less suspicious.

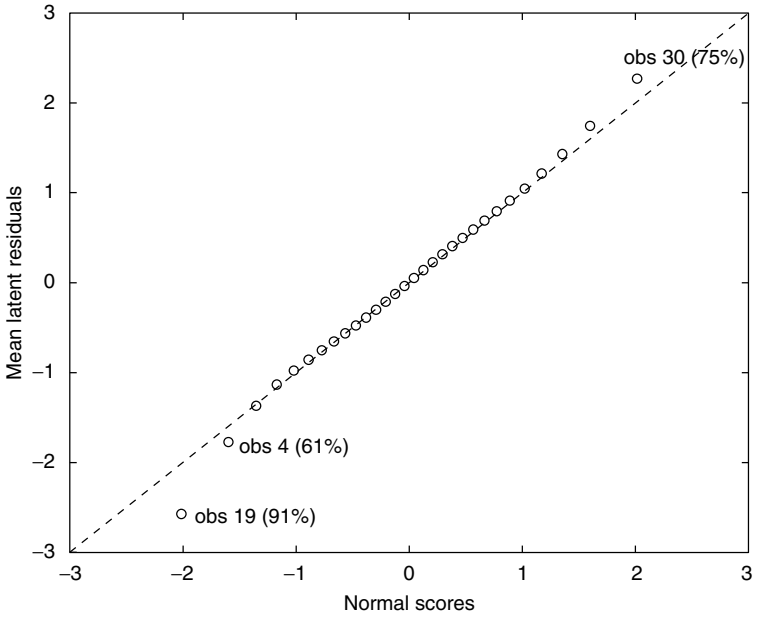
Overall, the normal scores plot does not suggest serious violations of model assumptions.

9.5.2. Prediction of Essay Scores From Grammar Attributes

A problem faced by large educational testing companies (e.g., ETS, ACT) involves grading thousands of student essays. As a result, there is great interest in automating the grading of student essays or—failing this—determining easily measurable qualities of essays that are associated with their ranking. The purpose of this example is to study the relationships between essay grades and essay attributes. The data in this example consist of grades assigned to 198 essays by five experts, each of whom rated all essays on a 10-point scale. A score of 10 indicates an excellent essay. Similar data have also been analyzed by, for example, Page (1994) and Johnson (1996). For present purposes, we examine only the grades assigned by the first expert grader and the essay characteristics of average word and sentence length, number of words, and the number of prepositions, commas, and spelling errors.

Following a preliminary graphical analysis of the data, we chose to examine the predictive relationships between an expert's grade of an essay and the variable's square root of the number of words in the essay (SqW), average word length (WL), percentage of prepositions (PP), number of commas × 100 over the number of words in the essay (PC), the percentage of spelling

Figure 9.6 Normal Scores Plot of the Posterior Means of the Sorted Latent Residuals From Grades Example



errors (PS), and the average sentence length (SL). Plots of each of these variables versus the essay grades are displayed in Figure 9.7.

On the basis of the plots in Figure 9.7, we posited a baseline model of the form

$$\Phi^{-1}(\theta_{ic}) = \gamma_c + \beta_0 + \beta_1 \mathrm{WL} + \beta_2 \mathrm{SqW} + \beta_3 \mathrm{PC} + \beta_4 \mathrm{PS} + \beta_5 \mathrm{PP} + \beta_6 \mathrm{SL}, \qquad (12)$$

where, as before, θ_{ic} denotes the cumulative probability that an essay received a score of c or below, and Φ denotes the standard normal distribution function. The maximum likelihood estimates for this model are displayed in Table 9.5.

The deviance of model (12) was 748.7 on $198 - 15 = 183$ degrees of freedom, using the usual convention that the number of degrees of freedom in a generalized linear model is equal to the number of observations less the number of estimated parameters. The deviance statistic is much larger than the degrees of freedom, suggesting some overdispersion in the model. This confirms our prior intuition that the six explanatory variables in the model cannot accurately predict the grades assigned by any particular human expert. (In fact, we might expect considerable variation between the grades assigned by different experts to the same essay.) Thus, it is probably prudent to apply a correction for overdispersion for us to interpret the standard errors in the table. Because the usual estimate of overdispersion for ordinal regression models is deviance/degrees of freedom (in this case, 4.09), each of the standard errors in Table 9.5 should be multiplied by the square root of the estimated overdispersion ($\approx$ 2.0) to obtain a more realistic estimate of the sampling uncertainty associated with each parameter.

To investigate the source of overdispersion, we need to examine the deviance contribution from each essay grade. To this end, a plot of deviance contribution versus the square root of the number of words is provided in Figure 9.8. As this figure illustrates, there are several observations for which the deviance exceeds 8 and two observations for which the deviance exceeds 14. The values 8 and 14 correspond approximately to the 0.995 and 0.9998 points of a χ^2_1 random variable, although it is unlikely that the asymptotic distribution of either the total deviance or the deviance of individual observations is well approximated by a chi-square random variable. However, the large values of the deviance associated with these observations provide further evidence that the grammatical variables included in the model do not capture all features used by the grader in evaluating the essays.

From Table 9.5 and the preliminary plots of the essay grades versus explanatory variables, it is clear that several of the variables included in the baseline model were not significant in predicting essay

Figure 9.7 Plots of Essay Grades Obtained From the First Expert Grader Versus Six Explanatory Variables

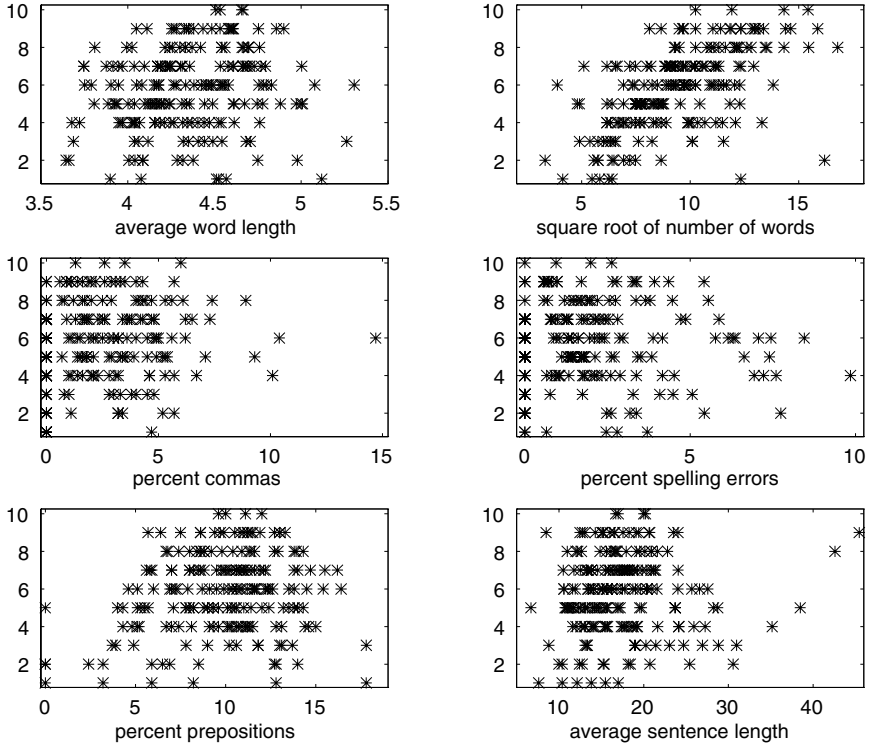


Table 9.5 Maximum Likelihood Estimates and Asymptotic Standard Deviations for the Baseline Regression Model for Essay Grades

Parameter	*Maximum Likelihood Estimate*	*Asymptotic Standard Deviation*
γ_2	0.632	0.18
γ_3	1.05	0.20
γ_4	1.63	0.21
γ_5	2.19	0.22
γ_6	2.71	0.23
γ_7	3.39	0.24
γ_8	3.96	0.26
γ_9	5.09	0.35
β_0	–3.74	1.08
β_1	0.656	0.23
β_2	0.296	0.032
β_3	0.0273	0.032
β_4	–0.0509	0.038
β_5	0.0461	0.023
β_6	0.00449	0.013

grade. To explore which of the variables should be retained in the regression function, we used a backward selection procedure in which variables were excluded sequentially from the model. The results of this procedure are summarized in the analysis of deviance table displayed in Table 9.6. Both the reduction in deviance associated with the deletion of each model variable and the reduction in the deviance corrected for overdispersion are provided.

By comparing the corrected changes in deviance displayed in the table to the corresponding tail probabilities of a χ^2_1 random variable, it appears that the variables SL, PC, and PS (average sentence length and percentage of commas and spelling errors) were not important in predicting the essay scores assigned by this grader. Likewise, the variable PP (percentage of prepositions) appears to be only marginally significant as a predictor, whereas the variables WL and SqW (word length and square root of number of words) are significant or highly significant. These results suggest that the variables SL, PC, and PS might be excluded from the model, leaving a predictive model of the form

$$\Phi^{-1}(\theta_{ic}) = \gamma_c + \beta_0 + \beta_1 \mathrm{WL} + \beta_2 \mathrm{SqW} + \beta_3 \mathrm{PP}. \tag{13}$$

Turning next to a default Bayesian analysis of these data, if we assume a vague prior on all model parameters, we can use an MCMC algorithm similar

Figure 9.8 Deviance Contribution From Individual Essay Grades Versus the Square Root of the Number of Words Contained in Each Essay

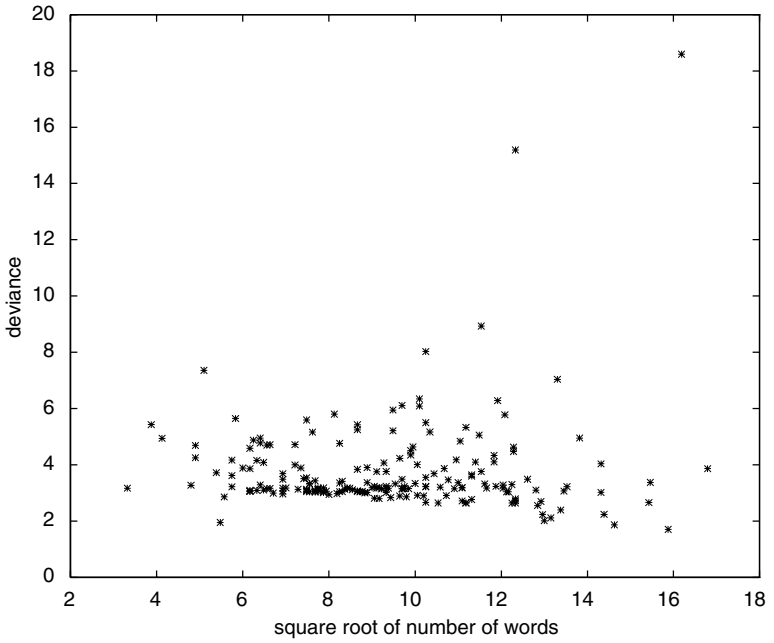


Table 9.6 Analysis of Deviance Table for Essay Grades

Model	*Change in Deviance*	*Corrected Change in Deviance*
Full model	—	—
SL	0.12	0.03
PC	0.71	0.17
PS	1.28	0.31
PP	3.93	0.96
WL	8.84	2.16
SqW	86.42	21.13

NOTE: The entries in the second column represent the increase in deviance resulting from deletion of the variable indicated in the first column as compared to the model on the previous row. The entries in the third column represent the entries in the second column divided by 4.09, the estimate of the model overdispersion from the full model.

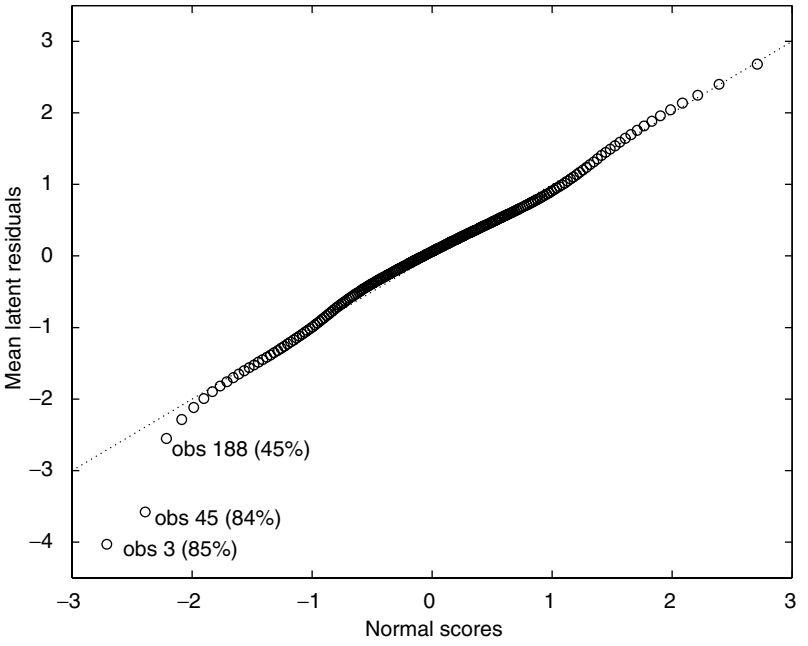
Table 9.7 Posterior Means of Parameter Estimates and Standard Deviation for the Full Regression Model for the Essay Grades

Parameter	*Posterior Mean*	*Posterior Standard Deviation*
γ_2	0.736	0.16
γ_3	1.19	0.18
γ_4	1.79	0.21
γ_5	2.35	0.21
γ_6	2.88	0.22
γ_7	3.59	0.22
γ_8	4.18	0.24
γ_9	5.30	0.30
β_0	–3.76	1.12
β_1	0.670	0.24
β_2	0.305	0.033
β_3	0.0297	0.033
β_4	–0.0520	0.038
β_5	0.0489	0.024
β_6	0.00463	0.013

to that described in Chapter 4 of Johnson and Albert (1999) to sample from the posterior distribution on the parameters appearing in either the full model (12) or the reduced model (13). For purposes of illustration, we generated 5,000 iterates from the full model and used these sampled values to estimate the posterior means of the regression parameters. These estimates are provided in Table 9.7 and are quite similar to the maximum likelihood (and, in this case, maximum a posteriori) estimates listed in Table 9.5.

Bayesian case analyses based on output from the MCMC algorithm proceed as in the previous example. By saving the latent variables values generated in the MCMC scheme, we can easily construct a normal scores plot of the latent residuals, as depicted in Figure 9.9. Like the deviance plot, this figure suggests that at least two observations did not conform

Figure 9.9 Normal Scores Plot of the Posterior Means of the Sorted Latent Residuals for the Essay Grading Example



to model assumptions. There is also evidence that the distribution of the latent residuals is non-Gaussian, another form of model misspecification.

In addition to the latent residuals, we can also examine the posterior-predictive residuals to further investigate the overdispersion detected in the likelihood-based analysis. If we let y_i^* denote the posterior distribution of the simulated essay grade for the ith essay and let y_i denote the observed grade of the ith essay, then the posterior-predictive residual distribution for the ith observation is defined as the distribution of $y_i - y_i^*$.

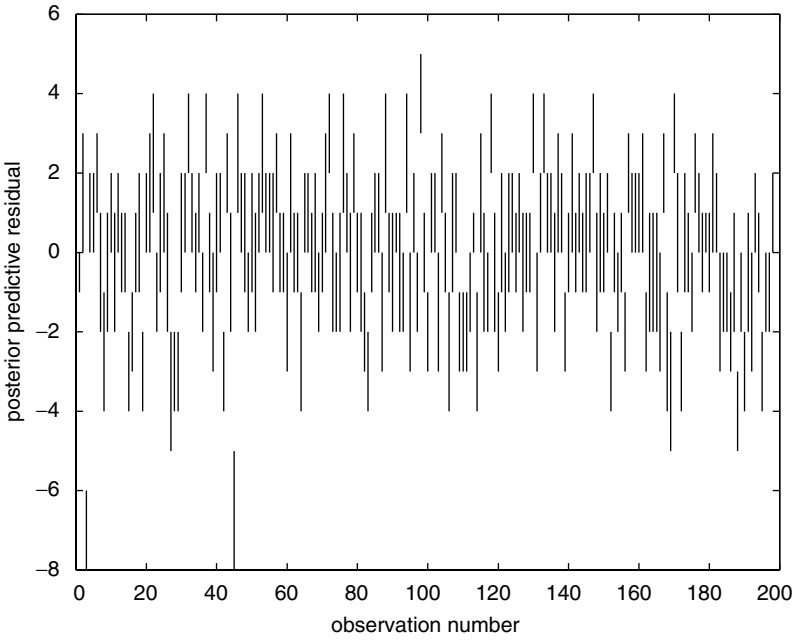
A plot of the estimated interquartile ranges of the posterior-predictive residuals is provided in Figure 9.10. The appearance of this plot indicates model lack of fit. To more formally quantify this lack of fit, we might again posit a random-effects model, but in this case, there are at least two distinct sources of error that we would like to model. The first is the inability of the regression model to fully explain the nuances of human graders; the regression model clearly cannot account for all of the essay attributes used by the expert in arriving at a grade for an essay. The second is the variability between experts in assigning grades to essays. As we mentioned at the beginning of this example, there were four other experts who also assigned grades to these same essays, and there was considerable disagreement among the experts on the appropriate grade for any particular essay. Thus, a simple random-effects model is unlikely to capture both sources of overdispersion, which suggests that a more comprehensive model is needed. We investigate such models in the next section.

9.6. Analyzing Data From Multiple Raters

In the example of the last section, we examined the relationship between an expert's ratings of a set of high school students' essays and several easily quantifiable attributes measured from these essays. The particular essay grades that we examined happened to be the grades from the first of five experts who graded the essays. However, with more than one expert grader, an obvious question becomes the following: How would our analysis change if we used another expert's ratings or if we somehow combined the grades from all experts?

In the previous section, we assumed that the "true grade" of each essay was known, and then we analyzed

Figure 9.10 Interquartile Ranges of the Posterior-Predictive Residuals



NOTE: The fact that a high proportion of these ranges does not cover 0 is an indication of overdispersion, or other lack of fit.

the essays to assess the relationship between these grades and various grammatical attributes. Unfortunately, when we examine ratings from several raters, it generally happens that the classifications assigned to individuals by different raters are not consistent. We must therefore decide how to combine the information gathered from different raters.

Numerous approaches have been proposed for analyzing ordinal data collected from multiple raters. Often, emphasis in such analyses focuses on modeling the agreement between raters. Among the more commonly used indices of multirater agreement in social sciences and medicine is the κ-statistic (Cohen, 1960). Assuming that all judges employ the same number of rating categories, the κ-statistic can be estimated by constructing a contingency table in which each judge is treated as a factor having K levels. The κ-statistic is then defined by

$$\kappa = \frac{p_0 - p_c}{1 - p_c},$$

where p_0 represents the sum of the observed proportions in the diagonal elements of the table, and p_c represents the sum of the expected proportions under the null hypothesis of independence. Large positive values of κ may be interpreted as indicating systematic agreement between raters. This statistic has been developed and extended by a number of authors, including Fleiss (1971), Light (1971), and Landis and Koch (1977a, 1977b). A related index has been proposed by Jolayemi (1991a, 1991b).

A more recent, model-based approach toward measuring rater agreement was proposed by Tanner and Young (1985). In their paradigm, the contingency table used in the construction of the κ-statistic was analyzed in the context of a log-linear model. Indicator variables corresponding to subsets of diagonal cells in subtables were used to model agreement between different judges. An advantage of this approach over the κ-statistic is that specific patterns of rater agreement can be investigated. Both methodologies are applicable to nominal and ordinal categorical data. Further work in this direction was proposed by Uebersax (1992) and Uebersax and Grove (1993).

In contrast to these approaches, the approach that we advocate emphasizes the tasks of evaluating rater precision, estimating the relative rankings of individuals, and predicting rankings from observed covariates. Unlike the approaches mentioned above, we assume a priori that all judges essentially agree on the merit of various individuals and that an underlying trait (or trait vector) determines the "true" ranking of an individual in relation to all others. Generally, we assume that this trait is scalar valued.

9.7. Essay Scores From Five Raters

To illustrate our modeling approach, let us again consider the essay grade data that we encountered at the end of Section 9.5.

Figure 9.11 depicts the marginal distribution of the grades assigned by each of the five judges to the 198 essays. From this figure, we see that the proportion of essays assigned to each grade category varies substantially from judge to judge. The raters vary with respect to their average ratings and also with respect to the spread of their ratings. For example, Rater 1 appears to assign higher ratings than Rater 2. Rater 3 seems unusual with respect to the relatively large variation of his or her ratings. Of course, the variation between ratings that we see in Figure 9.11 does not necessarily mean that the *rankings* of the essays were not consistent across judges; it might mean only that the grade cutoffs employed by the judges were different. To examine the consistency of the rankings, we can plot the essay grades assigned by the judges against one another. Such a plot is provided in Figure 9.12. To make this plot easier to interpret visually, we have plotted the elements in the cross-tabulation tables as a gray-scale image, with darker squares corresponding to higher counts in the bivariate histogram. The extent to which raters agree is indicated by the concentration of dark squares along a line with positive slope. When raters agree both in their rankings of individuals and also employ similar definitions of the category cutoffs, the slope of this line is approximately 1.

From Figure 9.12, we see that the variability of the third rater is comparatively large in comparison to the other four raters. It also appears that the second, fourth, and fifth raters produced rankings that were largely consistent with one another and that the second and fifth raters used similar category definitions.

9.8. The Multiple-Rater Model

9.8.1. The Likelihood Function

As in preceding sections, we denote the "true" value of the ith individual's latent trait on a suitably chosen scale by Z_i. The vector of latent traits for all individuals is denoted by $\mathbf{Z} = \{Z_i\}$.

We assume that the available data have the following general form. There are n individuals rated, and each individual is rated by at most J judges or raters. In many cases, every judge rates every individual. In the situation when all individuals are not rated by all judges, we assume that the decision for a judge to rate an individual is made independently of the qualities of both the judge and individual. We further assume that judge j classifies each individual into one of K_j ordered categories. Typically, all judges use the same number of categories, in which case we drop the subscript j and let K denote the common number of ordered categories. We let $\mathbf{y} = \{y_{ij}\}$ denote the data array, with y_{ij} denoting the rating assigned by judge j to individual i. The matrix of covariates relevant for predicting the relative rankings of the individuals is denoted by $\mathbf{X}$.

In assigning a category or grade to the ith object, we assume that judge j observes the value of Z_i with an error denoted by e_{ij}. The quantity $t_{ij} = Z_i + e_{ij}$ then denotes judge j's estimate of the latent trait for individual i on the underlying trait scale.

The error term e_{ij} incorporates both the observational error of judge j in assessing individual i and the bias of judge j in assessing the true value of Z_i. In some cases, it might be sensible to model e_{ij} as a function of individual covariates, although in what follows, we assume that the expectation of e_{ij}, averaged over all individuals in the population that have the same covariate values as individual i, is zero.

As in the single-rater setting, we assume that individual i is assigned to category c by judge j if

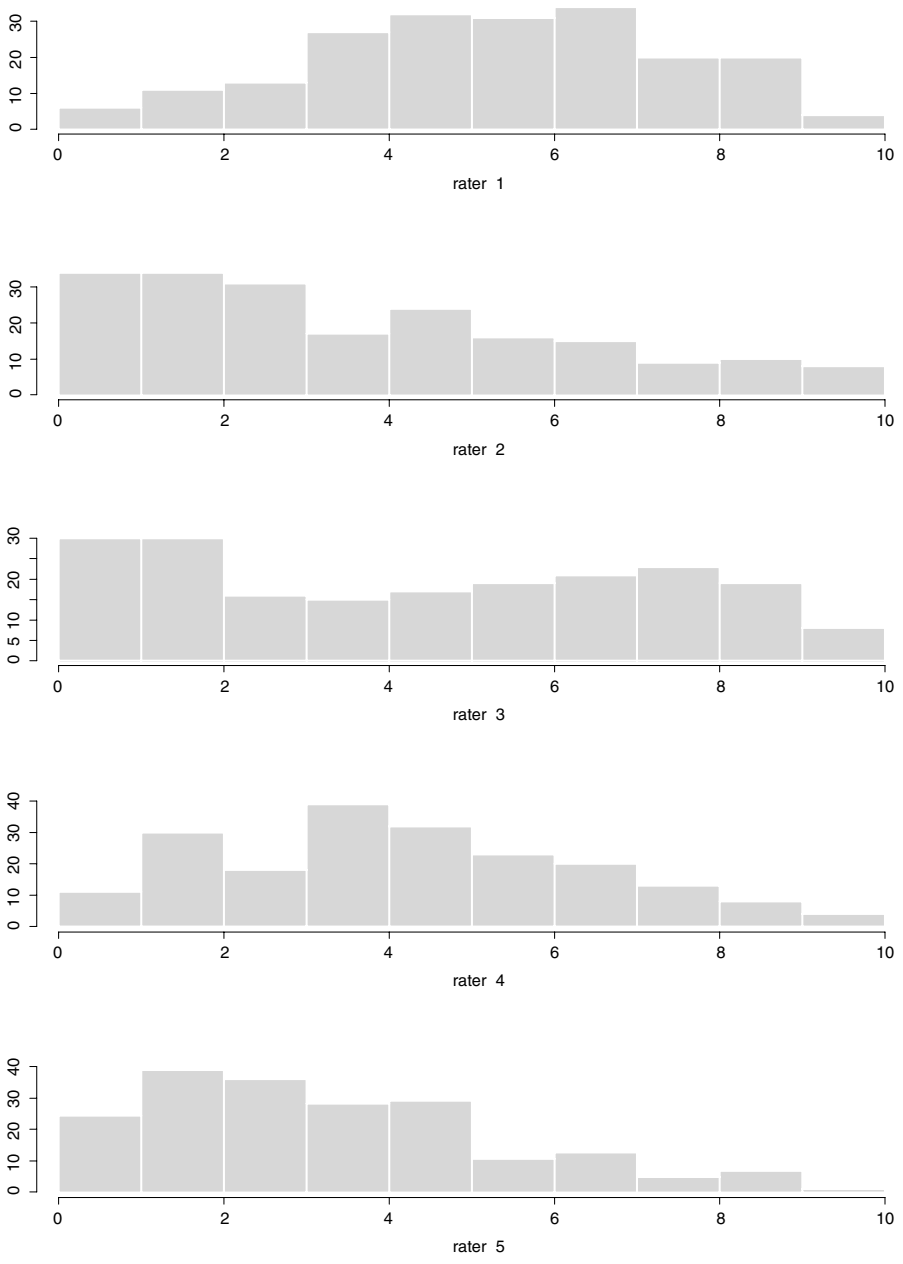
$$\gamma_{j,c-1} < t_{ij} \leq \gamma_{j,c} \tag{14}$$

for judge-specific category cutoffs $\gamma_{j,c-1}$ and $\gamma_{j,c}$. As in the single-rater case, we define $\gamma_{j,0} = -\infty$, $\gamma_{j,K} = \infty$ and let $\gamma_j = (\gamma_{j,1}, \ldots, \gamma_{j,K-1})$ denote the vector of cutoffs for the jth judge. Let $\gamma = \{\gamma_1, \ldots, \gamma_J\}$ denote the array of category cutoffs for all judges.

To this point, the model for the multirater ordinal data generation is entirely analogous to the single-rater case. However, in specifying the distribution of the error terms e_{ij}, we must decide whether we wish to assume that all judges rank individuals with equal precision or whether some judges provide rankings that are more accurate than others.

In either case, it is convenient to assume a common distributional form for the error terms e_{ij} across judges. We therefore assume that e_{ij}, the error of the jth judge in rating the ith individual, has a distribution with mean 0 and variance σ_j^2. We write the distribution function of e_{ij} as $F(e_{ij}/\sigma_j)$ for a known distribution function F. We denote the density function corresponding to F by f.

Figure 9.11 Histogram Estimate of the Marginal Distribution of the Grades Assigned by Each Expert Rater to the Essays

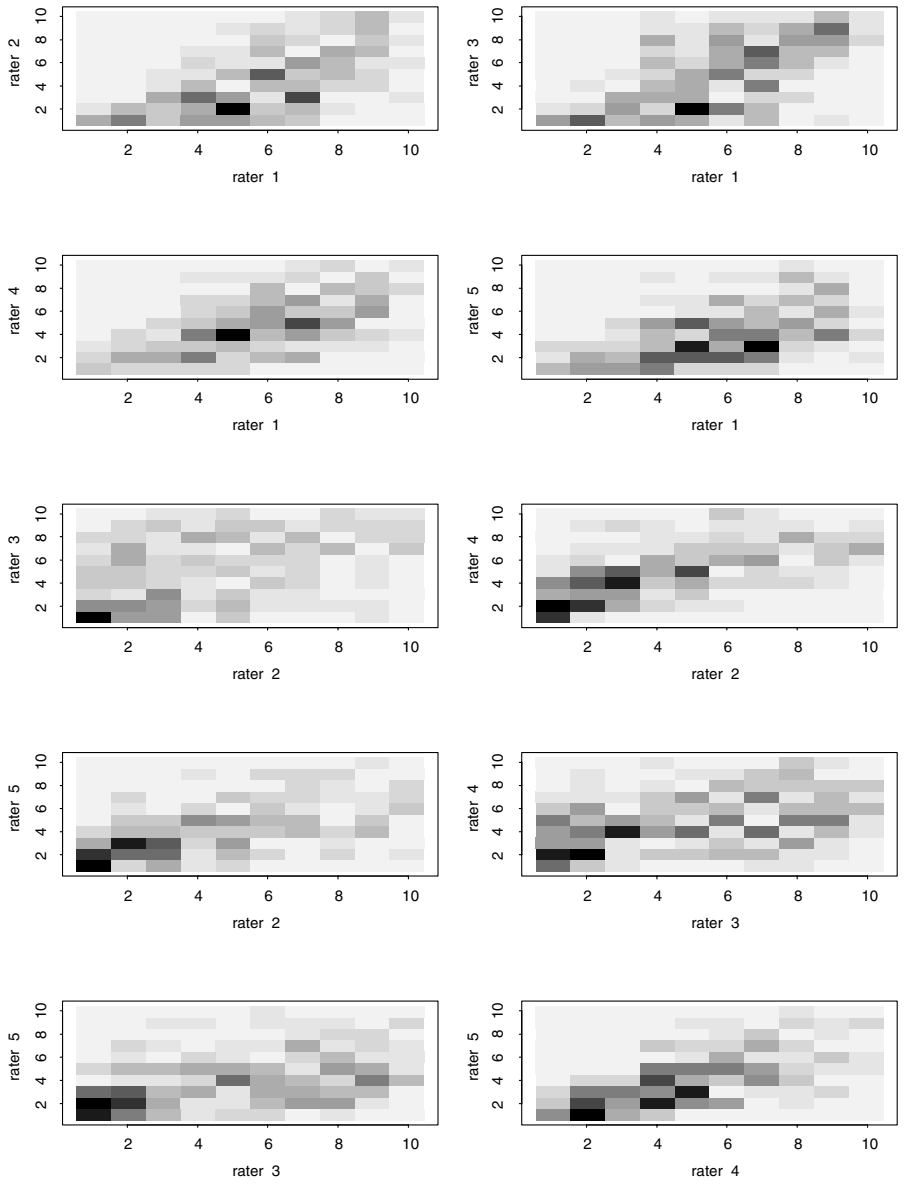


By taking $\sigma_j^2 = \sigma^2$ for all j, we impose the constraint that all judges rank individuals with similar precision. In practice, however, this assumption is seldom supported by data, and so unless explicitly stated otherwise, we assume distinct scale parameters for each judge.

Under these assumptions, it follows that the likelihood function for the observed data **y** (ignoring, for the moment, regression of the latent traits **Z** on explanatory variables **X**) may be written as

$$L(\mathbf{Z}, \gamma, \{\sigma_j^2\}) = \prod_{i=1}^{n} \prod_{j \in C_i} \left[F\left(\frac{\gamma_{j,y_{ij}} - Z_i}{\sigma_j} \right) - F\left(\frac{\gamma_{j,y_{ij}-1} - Z_i}{\sigma_j} \right) \right], \quad (15)$$

Figure 9.12 Bivariate Histogram Representations of Joint Marginal Distributions of the Grades Assigned by Each Pair of Expert Raters to the Essays



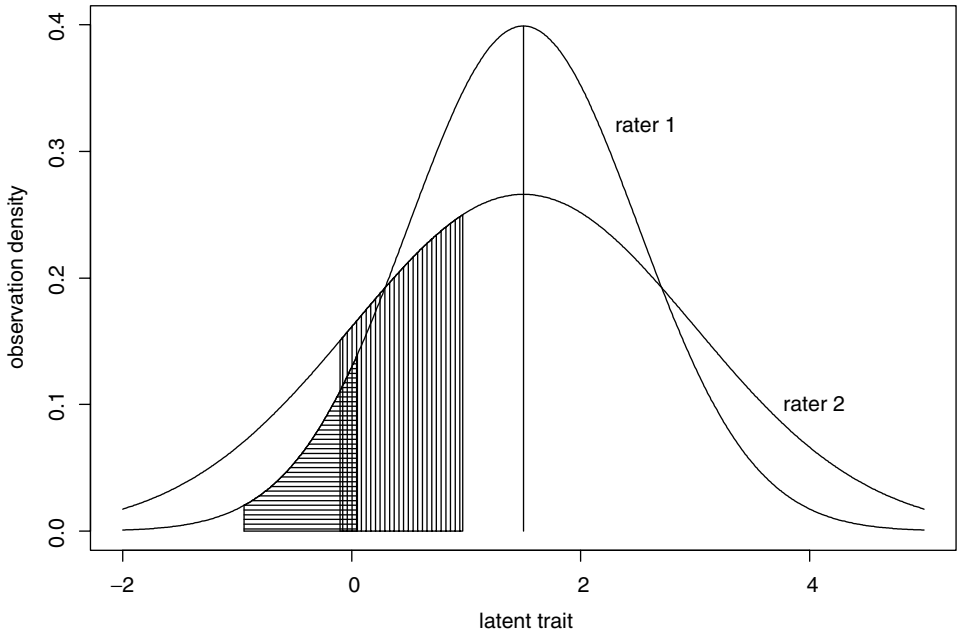
NOTE: Darker squares represent higher counts.

where C_i denotes the set of raters who classified individual i. If we introduce the latent trait estimates t_{ij} into the estimation procedure, the augmented likelihood function can be expressed as

$$L(\mathbf{Z}, \{t_{ij}\}, \gamma, \{\sigma_j^2\}) = \prod_{i=1}^{n} \prod_{j \in C_i} \frac{I}{\sigma_j} f\left(\frac{t_{ij} - Z_i}{\sigma_j}\right) \cdot I(\gamma_{j,y_{ij}-1} < t_{ij} \leq \gamma_{j,y_{ij}}). \quad (16)$$

As before, $I(\cdot)$ denotes the indicator function. Graphically, this model for the likelihood function is illustrated in Figure 9.13. In this plot, two raters classify an individual with true trait 1.5, indicated by the isolated vertical line. The distribution of their observations of this individual's trait is depicted by the two normal densities, from which it is clear that the second rater is less precise. The horizontally shaded region represents the probability that the first rater classifies

Figure 9.13 Depiction of Multirater Ordinal Data Model



this individual as "2," supposing that the lower and upper category cutoffs for the first rater's second category were $(\gamma_{1,1}, \gamma_{1,2}) = (-1.0, 0.1)$. Similarly, the vertically shaded region depicts the probability that the second rater classified this individual in the second category, given that the corresponding category cutoffs were $(\gamma_{2,1}, \gamma_{2,2}) = (-0.2, 1.0)$.

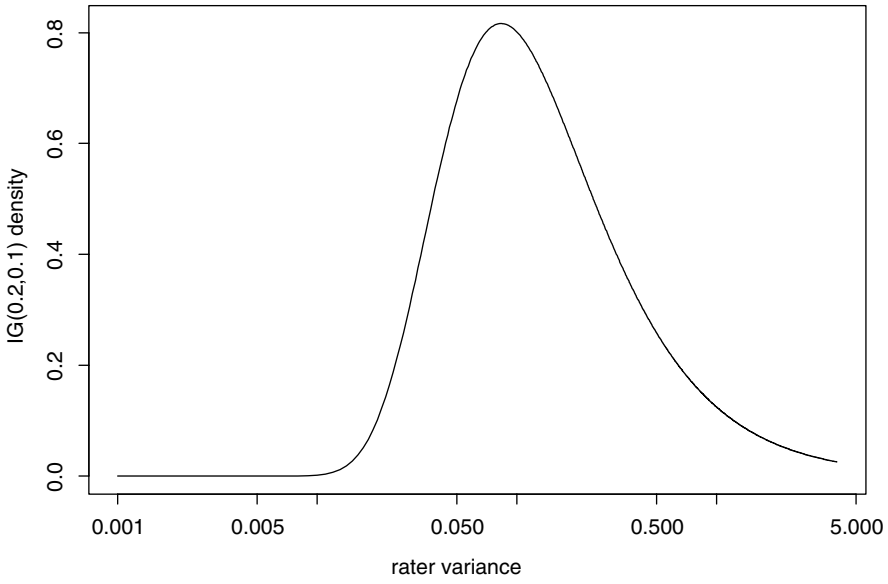
9.8.2. The Prior

Upon careful examination of the likelihood function (15 or 16), it is clear that the model parameters are not identifiable. That is, for any constants a and $b > 0$, we may replace $\mathbf{Z}$ with $b(\mathbf{Z} - a)$, t_{ij} with $b(t_{ij} - a)$, γ with $b(\gamma - a)$, and σ_j with $b\sigma_j$ without changing the value of the likelihood. We faced a less severe identifiability problem when we composed a model for single-rater ordinal data. To solve that problem, we imposed a constraint on the value of the first category cutoff to make γ and the regression intercept identifiable. In this case, the problem is exacerbated because it is generally unreasonable to assume that the upper cutoff for the lowest category is the same for all raters. Furthermore, when data from only one rater are available, the value of the rater variance σ^2 can be assigned the fixed value of 1. This constraint on σ eliminates the scaling problem (i.e., multiplying all model parameters by a positive constant b) but, as stated above, is generally not appropriate for multirater data because different raters exhibit different levels of precision in their rankings.

The identifiability problem can be overcome by imposing proper prior distributions on some or all model parameters. The location of the trait distribution can be fixed by specifying a proper prior for the latent traits. For convenience, we assume throughout the remainder of this chapter that the proper prior chosen for the latent traits is a Gaussian distribution or, in other words, that the latent traits $Z_1, \ldots, Z_n$ are distributed a priori from independent standard normal distributions.

In addition to specifying a proper prior distribution on the latent trait vector $\mathbf{Z}$, we also assume a specific distributional form for the rater variance parameters σ. In particular, we assume that rescaled versions of the rater variances are distributed according to a known distribution F, and then we assign a proper prior distribution to the scaling factors applied to each rater variance so that they have distribution F. In other words, we assume that the rater error terms e_{ij}/σ_j are distributed according to F and take an informative prior on σ_j. A convenient choice for $F(\cdot)$ is a standard normal distribution. If we combine this assumption with the assumptions made above, it follows that the conditional distribution of the rater-observed latent traits $\{t_{ij}\}$, given Z_i, are independent and normally distributed with mean Z_i and variance σ_j^2.

Figure 9.14 Inverse Gamma Density With Parameters $\lambda\psi = 0.2$ and $\lambda\psi = 0.1$



NOTE: The x-axis is plotted on the logarithmic scale to better illustrate the behavior of the density near 0.

The assumption of normality of the judge error terms can be at least partially justified by noting that the errors in a judge's perception of an individual's attributes usually result from a large number of small effects. By the central limit theorem, we might therefore expect that the rater errors are approximately Gaussian. Furthermore, it should be noted that predictions obtained under a model that assumes normally distributed rater errors generally produces predictions that are quite similar to predictions obtained under other common error models. Thus, the final conclusions drawn from this class of models tend to be relatively insensitive to the particular distributional form assumed for the components of e_{ij}.

A priori, we also assume that the rater variances $\sigma_1^2, \ldots, \sigma_J^2$ are independent. If F is chosen to be a standard normal distribution, then the conjugate prior for the variance parameters σ_j^2 is an inverse gamma density, expressible in the form

$$\pi(\sigma_j^2; \lambda, \alpha) = \frac{\lambda^\alpha}{\Gamma(\alpha)}(\sigma_j^2)^{-\alpha-1}\exp\left(-\frac{\lambda}{\sigma_j^2}\right), \quad \alpha, \lambda > 0. \tag{17}$$

We denote the inverse gamma distribution corresponding to this density by $IG(\alpha, \lambda)$; the mean and mode of the distribution are $\lambda/(\alpha - 1)$ (assuming $\alpha > 1$) and $\lambda/(\alpha + 1)$, respectively. A density plot for an $IG(0.2, 0.1)$ random variable is depicted in Figure 9.14. The parameters α and λ can be chosen so that the prior density on the rater variances concentrates its mass in the interval (0.01, 4.0). It is important to assign a positive value to λ to prevent singularities in the posterior distribution that would occur if the components of σ^2 were allowed to become arbitrarily small.

To complete the prior model, we need to specify a distribution on the vector of category cutoffs γ. For present purposes, we assign independent uniform priors on the category cutoff vectors γ_j, subject to the constraint that

$$\gamma_{j,1} \leq \cdots \leq \gamma_{j,K-1}.$$

Combining all of these assumptions, the joint prior density on $(\mathbf{Z}, \gamma, \{\sigma_j^2\})$ is given by

$$g(\mathbf{Z}, \gamma, \{\sigma_j^2\}) = \prod_{i=1}^{n}\varphi(Z_i; 0, 1)\prod_{j=1}^{J}\pi(\sigma_j^2; \lambda, \alpha) \tag{18}$$

where $\varphi(x; \mu, \sigma)$ denotes a normal density with mean μ and standard deviation σ. Taken together, this set of assumptions defines what we refer to as the multirater ordinal probit model. As we demonstrate below, this model provides a useful framework for analyzing a wide variety of ordinal data sets.

9.8.3. Analysis of Essay Scores From Five Raters (Without Regression)

To quantify the qualitative conclusions drawn from Figures 9.12 and 9.13, we apply the model described in the last section to obtain the posterior distributions on each rater's variance parameter. As a by-product of this model-fitting procedure, we also obtain the posterior distribution on the underlying trait for each essay's grade.

With the introduction of the latent trait estimates t_{ij} into the estimation problem, the joint posterior density of all unknown parameters is given by

$$g(\mathbf{Z}, \{t_{ij}\}, \gamma, \{\sigma_j^2\}) \propto L(\mathbf{Z}, \{t_j\}, \gamma, \{\sigma_j^2\}) g(\mathbf{Z}, \gamma, \{\sigma_j^2\}),$$

where the likelihood function is given by (15) and the prior density by (18). To obtain samples from this posterior distribution, we modify the MCMC algorithm described for single-rater data to accommodate additional raters. After initializing model parameters, we begin the MCMC algorithm by sampling from the conditional distribution of $\mathbf{Z}$. From (16), we see that the conditional distributional of the component Z_i, given the array $\{t_{ij}\}$ and σ_j^2, is normally distributed with mean s/r and variance $1/r$, where

$$r = 1 + \sum_{j \in C_i} \frac{1}{\sigma_j^2} \quad \text{and} \quad s = \sum_{j \in C_i} \frac{t_{ij}}{\sigma_j^2}. \tag{19}$$

Given the value of $\mathbf{Z}$, updating the components of γ_j proceeds as in the single-rater case. Similarly, rater-specific trait values t_{ij} can be sampled from a truncated Gaussian density with mean Z_i and variance σ_j^2, truncated to the interval $(\gamma_{j,y_{ij}-1}, \gamma_{j,y_{ij}})$.

Finally, the conjugate prior structure specified for the variances $\sigma_1^2, \ldots, \sigma_J^2$ makes sampling from the conditional distributions of these parameters straightforward. If we let D_j denote the set of individuals rated by the jth judge and take n_j to be the number of elements of D_j, then the conditional distribution of σ_j^2 is

$$\sigma_j^2 \sim IG\left(\frac{n_j}{2} + \alpha, \frac{S}{2} + \lambda\right), \quad \text{where}$$
$$S = \sum_{i \in D_j} (t_{ij} - Z_i)^2. \tag{20}$$

These conditional distributions can be used to sample from the joint posterior distribution on all model parameters, as described in more detail in Johnson and Albert (1999). After this MCMC algorithm is implemented, we can estimate the posterior means and variances of the rater variance parameters from the MCMC output. Estimates obtained in this way are depicted in Table 9.8. Note that the values displayed in this table agree qualitatively with the graphical analysis of the data presented in Figure 9.12. As predicted, the third rater tended to assign essay grades that were not consistent with the grades assigned by the other raters.

Table 9.8 Posterior Means and Posterior Standard Deviations of Rater Variance Parameters

	Rater				
	1	2	3	4	5
Posterior mean	0.91	0.53	2.05	0.61	0.89
Standard Deviation	0.22	0.14	0.54	0.14	0.24

The values obtained from the MCMC algorithm can also be used to perform residual analyses in ways similar to those described for single-rater data. For example, the simulated values of t_{ij} and $\mathbf{Z}$ obtained from the MCMC algorithm can be used to define standardized residuals of the form

$$r_{ij} = \frac{t_{ij} - Z_i}{\sigma_j}.$$

A normal scores plot of the posterior means of the standardized residuals for the first rater's grades is shown in Figure 9.15. In this case, none of the grades assigned by this rater appears unusually large.

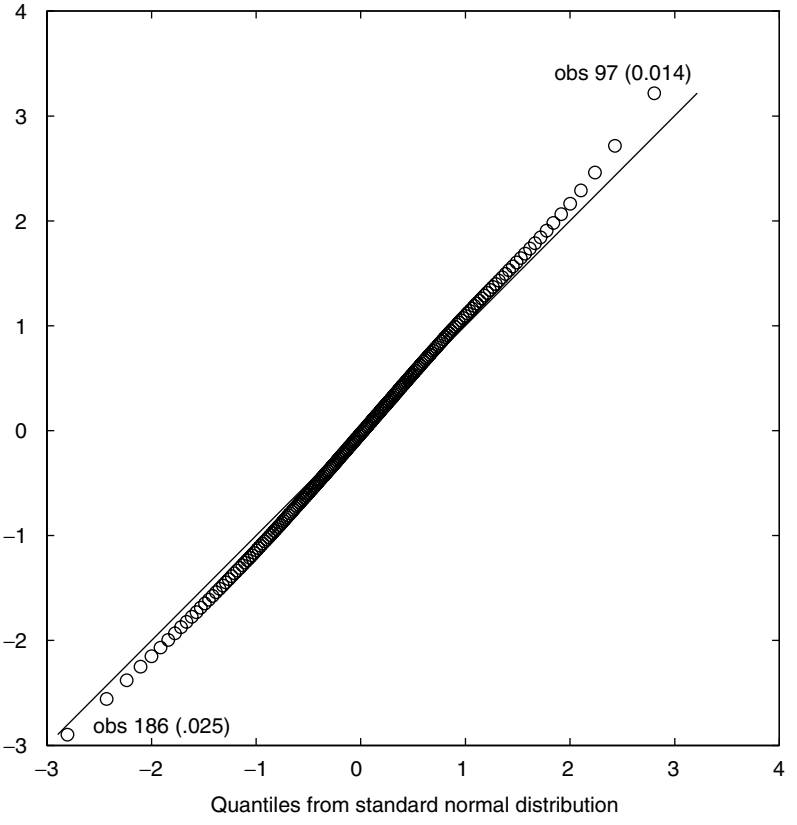
9.9. Incorporating Regression Functions Into Multirater Data

If we compare the model framework outlined above for multirater ordinal data to the standard model for single-rater ordinal data described, we find two basic differences. First, in the case of single-rater data, there is no loss of generality incurred by fixing the value of the first category cutoff at zero, provided that we include an intercept in the linear regression of the latent trait variables $\{Z_i\}$ on the matrix of explanatory variables X. Second, and more important, is the fact that we implicitly assumed a value of 0 for the rater variance parameter in the single-rater case. Coupled with the assumption that

$$Z_i = \mathbf{x}_i'\beta + \eta_i, \quad \eta_i \sim F(\cdot), \tag{21}$$

where $F(\cdot)$ denoted the link function for the regression model, this allowed us to define a scale of measurement for both the latent variables and the regression

Figure 9.15 Normal Scores Plot for the Posterior Means of the Standardized Latent Residuals for the First Rater's Grades



parameter. Of course, with data from only one rater, we really had no choice but to make the assumption that the rater correctly categorized each observation, that is, that the rater's error variance was exactly 0. Indeed, in many instances, this assumption might actually be justified from substantive considerations. In testing mechanical parts for failure, for example, the binary classification of tested parts as either a success or failure might be completely objective.

However, for multirater data, the situation changes. The very fact that data in an experiment or study were collected from multiple raters implies that the classification of individuals into categories was subjective. That is, different raters are *expected* to have different opinions on the relative merit of each individual. Substantively, the subjectiveness of the observed data means that one must question the validity of the regression assumption.

To illustrate the importance of this point, recall that in the case of single-rater data, the latent traits Z_i were assumed to follow the regression relation (21). If we assume that $F(.)$ is a standard normal distribution function, equation (21) can be combined with the model assumptions of the previous section to obtain the following expression for the value of the latent trait observed by a single rater:

$$t_{ij} = \mathbf{x}_i'\beta + \eta_i + \varepsilon_{ij}, \quad \text{where} \quad \eta_i \sim N(0, 1) \quad \text{and} \quad \varepsilon_{ij} \sim N(0, \sigma_j^2). \tag{22}$$

It follows that the rater-observed latent trait t_{ij} has a normal distribution with mean $\mathbf{x}_j'\beta$ and variance $1 + \sigma_j^2$. This implies that the estimated variance for those raters whose classifications most closely follow the regression function will be the smallest. Data obtained from these raters will consequently be given more weight than data from the other raters in estimating the true ranking of individuals.

These ideas are well illustrated in terms of our example involving the essay grades collected independently from five judges. In the previous section, we posited a linear model for the latent performance variable

associated with the grade acquired from the first judge. Assuming that a similar model can be used to predict the grade of any of the five judges participating in the study, we obtain the following regression equation for the prediction of t_{ij}:

$$t_{ij} = \beta_0 + \beta_1 \mathrm{WL}_i + \beta_2 \mathrm{SqW}_i + \beta_3 \mathrm{PP}_i + \eta_i + \varepsilon_{ij}. \tag{23}$$

As before, WL, SqW, and PP represent the average word length, the square root of the number of words in the essay, and the percentage of prepositions used, respectively. The variance of e_{ij}, σ_j^2 measures the agreement of the jth judge's ratings with the explanatory variables. From a substantive viewpoint, the critical question is the following: Do we wish to give more weight to the rankings of those judges whose grades were most linear in these explanatory variables? In this example, the answer is probably not. Our primary interest in performing this regression was to investigate the extent to which an expert's grade might be modeled using easily quantified grammatical variables. We did not anticipate that these variables would ideally predict the relative merit of the essays.

To overcome this difficulty, we need to specify our regression model so that lack of fit of the regression function can be accommodated. One way to do this is to assume that for given values of the parameters β and τ^2,

$$z_i = \mathbf{x}_i'\beta + \zeta_i, \quad \text{where} \quad \zeta_i \sim N(0, \tau^2). \tag{24}$$

That is, we put the regression equation on equal footing with the ratings obtained from a single judge. The error ζ_i term accounts for the "lack-of-fit" error associated with the regression equation. This term is completely analogous to the term ε_{ij} associated with the observation of the latent trait by a single rater.

The precision of the regression relationship depends on the value of τ^2, which might also be estimated from within the model framework.

The difficulty with this formulation (24) is that it is inconsistent with the assumptions made in the last section concerning the marginal distribution on $\mathbf{Z}$—that $\mathbf{Z} \sim N(0, \mathbf{I})$. Recall that this assumption was needed to make parameters in the likelihood identifiable. Unfortunately, if a vague prior is specified for β, constraint (24) does not establish a scale of measurement for the latent traits.

To summarize our discussion to this point, we have reached two conclusions. First, in many applications, it is not reasonable to assume that the "true" rating of an individual is exactly predicted by the regression function. For this reason, the assumptions implied by model (22) are often inappropriate for modeling ordinal data. Second, we cannot assume that the values of the latent trait vector $\mathbf{Z}$ are governed by a relationship of the type expressed in (24) without assuming a proper prior on the components of β. Doing so leads to an improper prior on $\mathbf{Z}$ and nonidentifiability of all model parameters.

We can solve each of these problems by building our ordinal regression model on top of the multiple-rater ordinal model described. Essentially, this means that we must specify the conditional distribution of β given $\mathbf{Z}$, rather than reversing the conditionality relationships and specifying the distribution of $\mathbf{Z}$ given β. In this way, we can preserve the prior assumption that $\mathbf{Z} \sim N(0, \mathbf{I})$ while coupling the scale of β to the scale used to model the rater variances.

To specify the conditional distribution of β given $\mathbf{Z}$, we use standard results from the Bayesian analysis of the normal linear model. In the normal setting, if we let $\mathbf{W}$ denote the data vector and assume that $\mathbf{W} \sim N(\mathbf{Xb}, a^2\mathbf{I})$ for an unknown regression parameter $\mathbf{b}$ and known variance a^2, then the posterior distribution of $\mathbf{b}$ is

$$\mathbf{b} \sim N((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}, a^2(\mathbf{X}'\mathbf{X})^{-1}),$$

provided a uniform prior is assumed for $\mathbf{b}$. If the prior on $\mathbf{b}$ is $N(d, D)$, then the posterior distribution for $\mathbf{b}$ is

$$\mathbf{b} \sim N(f, a^2 F),$$

where

$$f = [(\mathbf{X}'\mathbf{X}) + D^{-1}]^{-1}(\mathbf{X}'\mathbf{W} + D^{-1}d)$$

and

$$F = a^2(\mathbf{X}'\mathbf{X} + D^{-1})^{-1}.$$

Applying these normal theory results to our problem, we might therefore assume that the *prior* distribution for β, *given* $\mathbf{Z}$ and τ^2, can be expressed as

$$\beta|\mathbf{Z}, \mathbf{X}, \tau^2 \sim N(c, C). \tag{25}$$

In the absence of specific prior information regarding the prior density for β, we take c and C to be the least squares estimates of β and C:

$$c = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z} \tag{26}$$

and

$$C = \tau^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{27}$$

Alternatively, when prior information concerning the regression parameter β is available, the parameters c and C might be chosen as

$$c = [(\mathbf{X}^t\mathbf{X}) + D^{-1}]^{-1}(\mathbf{X}^t\mathbf{Z} + D^{-1}d),$$
$$C = \tau^2(\mathbf{X}^t\mathbf{X} + D^{-1})^{-1},$$

where d and $\tau^2 D$ are the prior mean and covariance of β.

Continuing the analogy with the normal theory models, we complete our specification of the prior model by taking the prior for τ^2, given $\mathbf{Z}$, to be an inverse gamma density of the form

$$g(\tau^2|\mathbf{Z}) = \frac{(RSS/2 + \lambda_r)^{(n-p)/2+\alpha_r}}{\Gamma[(n-p)/2 + \alpha_r]}(\tau^2)^{-(n-p)/2-\alpha_r-1} \times \exp\left(-\frac{RSS/2 + \lambda_r}{\tau^2}\right). \quad (28)$$

In (28), RSS denotes the residual sum of squares of the regression of $\mathbf{Z}$ on $\mathbf{X}\beta$; that is, $RSS = \mathbf{Z}^t(\mathbf{I} - \mathbf{X}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X})\mathbf{Z}$, and α_r and λ_r are prior hyperparameters.

From a technical standpoint, it is interesting to note that this specification of the prior density for β and τ^2 (given $\mathbf{Z}$) differs from the prior density implied by model (24) by a factor of

$$(RSS/2 + \lambda_r)^{(n-p)/2+\alpha_r}.$$

In the case of a vague prior for β and τ^2 (i.e., when $\lambda_r = \alpha_r = 0$), this factor approaches zero as the residual sum of squares approaches zero. Multiplication by this factor in the revised model prevents the posterior distribution from becoming arbitrarily large in a region near the value $\tau^2 = 0$ because the residual sum of squares approaches zero as $\mathbf{Z}$ becomes small. Without this factor, the prior specified in (24) leads to a posterior distribution that is unbounded for small values of $\mathbf{Z}$ and τ^2.

Finally, we must specify values for the hyperparameters α_r and λ_r. A natural choice for these parameters is to assign them the values α and λ used in the specification of the prior on the rater variances. Doing so facilitates the comparison of the posterior distribution on the rater variances and the variance of the regression relation. Alternatively, the prior on τ^2 might be taken to be a scaled version of the prior assumed for σ^2, with a scaling constant determined from prior knowledge of the relative reliability of a single judge's scores relative to predictions obtained from the regression relation. Further details, examples, and software implementation for multirater ordinal regression models may be found in Johnson and Albert (1999).

REFERENCES

Bedrick, E. J., Christensen, R., & Johnson, W. (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association, 91*, 1450–1460.

Cohen, J. (1960). A coefficient of agreement for nominal tables. *Educational and Psychological Measurement, 20*, 37–46.

Cowles, M. K. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing, 6*, 101–111.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B, 34*, 187–220.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 88*, 322–328.

Johnson, V. E. (1996). On Bayesian analysis of multirater ordinal data. *Journal of the American Statistical Association, 91*, 42–51.

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. New York: Springer-Verlag.

Jolayemi, E. T. (1990a). A multirater's agreement index for ordinal classification. *Biometrics Journal, 33*, 485–492.

Jolayemi, E. T. (1990b). On the measurement of agreement between two raters. *Biometrics Journal, 32*, 87–93.

Landis, J. R., & Koch, G. G. (1977a). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics Journal, 33*, 363–374.

Landis, J. R., & Koch, G. G. (1977b). The measurement of observer agreement for categorical data. *Biometrics Journal, 33*, 159–174.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin, 5*, 365–377.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B, 42*, 109–142.

Page, E. (1994). New computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62*(2), 127–142.

Tanner, M., & Young, M. (1985). Modeling agreement among raters. *Journal of the American Statistical Association, 80*, 175–180.

Uebersax, J. (1992). A review of modeling approaches for the analysis of observer agreement. *Investigative Radiology, 17*, 738–743.

Uebersax, J., & Grove, W. M. (1993). A latent trait finite mixture model for the analysis of rating agreement. *Biometrics, 49*, 823–835.

Chapter 10

LATENT CLASS MODELS

JAY MAGIDSON

JEROEN K. VERMUNT

10.1. INTRODUCTION

Latent class (LC) modeling was initially introduced by Lazarsfeld and Henry (1968) as a way of formulating latent attitudinal variables from dichotomous survey items. In contrast to factor analysis, which posits continuous latent variables, LC models assume that the latent variable is categorical, and areas of application are more wide-ranging. The methodology was formalized and extended to nominal variables by Goodman (1974a, 1974b), who also developed the maximum likelihood (ML) algorithm that serves as the basis for many of today's LC software programs. In recent years, LC models have been extended to include observable variables of mixed scale type (nominal, ordinal, continuous, and counts) and covariates, as well as deal with sparse data, boundary solutions, and other problem areas.

In this chapter, we describe three important special cases of LC models for applications in cluster, factor, and regression analyses. We begin by introducing the LC cluster model as applied to nominal variables (the traditional LC model), discuss some limitations of this model, and show how recent extensions can be used to overcome them. We then turn to a formal treatment of the LC factor model and an extensive introduction to LC regression models before returning to show how the LC cluster model, as applied to continuous variables, can be used to improve on the K-means approach to cluster analysis. We use the Latent GOLD computer program (Vermunt & Magidson, 2003) to illustrate the use of these models as applied to several data sets.

10.2. TRADITIONAL LATENT CLASS MODELING

Traditional LC analysis (i.e., Goodman, 1974b) assumes that each observation is a member of one and only one of T latent (unobservable) classes and that *local independence* exists between the manifest variables. That is, conditional on latent class membership, the manifest variables are mutually independent of each other. This model can be expressed using (unconditional) probabilities of belonging to each latent class and conditional response probabilities as parameters. For example, in the case of four nominal manifest variables A, B, C, and D, we have

$$\pi_{ijklt} = \pi_t^X \pi_{it}^{A|X} \pi_{jt}^{B|X} \pi_{kt}^{C|X} \pi_{lt}^{D|X}, \qquad (1)$$

where π_t^X denotes the probability of being in latent class $t = 1, 2, \ldots, T$ of latent variable X; $\pi_{it}^{A|X}$ denotes the conditional probability of obtaining the ith response to item A, from members of class t, $i = 1, 2, \ldots, I$; and $\pi_{jt}^{B|X}$, $\pi_{kt}^{C|X}$, $\pi_{lt}^{D|X}$, $j = 1, 2, \ldots, J$, $k = 1, 2, \ldots, K$, $l = 1, 2, \ldots, L$, denote the corresponding conditional probabilities for items B, C, and D, respectively.

Model 1 can be described graphically in terms of a path diagram (or a graphical model) in which manifest variables are not connected to each other directly but indirectly through the common source X. The latent variable is assumed to explain all of the associations among the manifest variables. A goal of traditional LC analysis is to determine the smallest number of latent classes T that is sufficient to explain away (account for) the associations (relationships) observed among the manifest variables.

The analysis typically begins by fitting the $T = 1$ class baseline model (H_0), which specifies mutual independence among the variables. Model H_0:

$$\pi_{ijkl} = \pi_i^A \pi_j^B \pi_k^C \pi_l^D.$$

Assuming that this *null* model does not provide an adequate fit to the data, a one-dimensional LC model with $T = 2$ classes is then fitted to the data. This process continues by fitting successive LC models to the data, each time adding another dimension by incrementing the number of classes by 1, until the simplest model is found that provides an adequate fit.

10.2.1. Assessing Model Fit

Several complementary approaches are available for assessing the fit of LC models. The most widely used approach uses the likelihood ratio chi-squared statistic L^2 to assess the extent to which maximum likelihood (ML) estimates for the expected cell frequencies, $\hat{F}_{ijkl}$ differ from the corresponding observed frequencies, f_{ijkl}:

$$L^2 = 2 \sum_{ijkl} f_{ijkl} \ln(\hat{F}_{ijkl}/f_{ijkl}).$$

A model fits the data if the value of L^2 is sufficiently low to be attributable to chance (within normal statistical error limits—generally, the .05 level).

The $\hat{F}_{ijkl}$ are obtained using the following two-step process. First, ML estimates for the model parameters are obtained and substituted into the right side of equation (1) to obtain ML estimates of the probabilities $\hat{\pi}_{ijklt}$. These probability estimates are then summed over the latent classes to obtain estimated probabilities for each cell in the observed table and multiplied by the sample size N to obtain the ML estimates for the expected frequencies:

$$\hat{F}_{ijkl} = N \sum_{t=1}^{T} \hat{\pi}_{ijklt}.$$

In the case that $\hat{F}_{ijkl} = f_{ijkl}$ for each cell (i, j, k, l), the model fit will be perfect and L^2 equals zero. To the extent that the value for L^2 *exceeds* 0, the L^2 measures lack of model fit, quantifying the amount of association (nonindependence) that remains unexplained by that model. When N is sufficiently large, L^2 follows a chi-square distribution, and as a general rule,[1] the number of degrees of freedom (df) equals the number of cells in the full multiway table minus the number of distinct parameters M minus 1. For example, in the case of four categorical variables, the number of cells equals $IJKL$, and the number of parameters is the following:

$$M = T - 1 + T[(I - 1) + (J - 1) + (K - 1) + (L - 1)].$$

M is obtained by counting the $T - 1$ distinct LC probabilities and, for each latent class, the $I - 1$ distinct conditional probabilities associated with the categories of variable A, the $J - 1$ distinct conditional probabilities associated with B, and so on. Because probabilities sum to 1, the probability associated with one category of each variable is redundant (and hence not counted as a *distinct* parameter): It can be obtained as 1 minus the sum of the others.

In situations involving sparse data, the chi-squared distribution should not be used to compute the p-value because L^2 would not be well approximated. Instead, the bootstrap approach can be used to estimate p (Langeheine, Pannekoek, & Van de Pol, 1996). Sparse data often occur when the number of observed variables or the number of categories of these variables is large. In such cases, the total number of cells in the resulting multiway frequency table will be large relative to the sample size, resulting in many empty cells. This situation is illustrated below with a data example. Sparse data also result when LC models are extended to include continuous variables, which is illustrated in the last section.

An alternative approach to assessing model fit in the case of sparse data uses an information criterion weighting both model fit and parsimony. Such measures, such as Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), are especially useful in comparing models. The most widely used in LC analysis is the BIC statistic, which can be

1. According to the general rule, if it turns out that $df < 0$, the model is not identifiable, which means that unique estimates are not available for all parameters. For example, for $I = J = K = L = 2$, $df = -4$ for $T = 4$, which means that the four-class model is not identifiable. In some cases, however, this general counting rule may yield $df > 0$, yet the model may still not be identifiable. For example, Goodman (1974b) shows that in this situation of four dichotomous variables, the three-class model is also unidentifiable despite the fact that the counting rule yields $df = 1$. See also Note 3.

defined as follows: $\text{BIC}_{L^2} = L^2 - \ln(N)df$ (Raftery, 1986). A model with a lower BIC value is preferred over a model with a higher BIC value. A more general definition of BIC is based on the log-likelihood (LL) and the number of parameters (M) instead of L^2 and df; that is,

$$\text{BIC}_{\text{LL}} = -2\text{LL} + \ln(N)M.$$

Again, a model with a lower BIC value is preferred over a model with a higher BIC value.[2]

If the baseline model (H_0) provides an adequate fit to the data, no LC analysis is needed because there is no association among the variables to be explained. In most cases, however, H_0 will not fit the data, in which case $L^2(\text{H}_0)$ can serve as a baseline measure of the total amount of association in the data. This suggests a third approach for assessing the fit of LC models by comparing the L^2 associated with LC models for which $T > 1$, with the baseline value $L^2(\text{H}_0)$ to determine the percent reduction in L^2. Because the total association in the data may be quantified by $L^2(\text{H}_0)$, the percent reduction measure represents the total association explained by the model. This less formal approach can complement the more statistically precise L^2 and BIC approaches.

As an example of how these measures are used, suppose that the L^2 suggests that a three-class model falls short of providing an adequate fit to some data (say, $p = .04$) but explains 90% of the total association. Moreover, suppose a four-class model is the simplest model that fits according to the L^2 statistic, but this model only explains 91% of the association. In this case, it may be that, on practical grounds, the three-class model is preferable because it explains almost as much of the total association.

10.2.1.1. Example: Survey Respondent Types

We will now consider a first example that illustrates how these tools are used in practice. It is based on the analysis of four variables from the 1982 General Social Survey given by McCutcheon (1987) to illustrate how traditional LC modeling can be used to study the different types of survey respondents. Two of the variables ascertain the respondent's opinion regarding (A) the purpose of surveys and (B) how accurate they are, and the others are evaluations made by the interviewer of (C) the respondent's levels of understanding of the survey questions and (D) cooperation shown in answering the questions. McCutcheon initially assumed the existence of two latent classes corresponding to "ideal" and "less than ideal" types.

The study included separate samples of White and Black respondents. Beginning with an analysis of the White sample, McCutcheon (1987) later included data from the Black sample to illustrate a two-group LC analysis. We will use these data to introduce the basics of traditional LC modeling and to illustrate several recent developments that have been made over the past decade. These include allowing for specific local dependencies (Section 10.3.1), the usage of LC factor models (Section 10.3.2), and the inclusion of covariates as well as the methodology for making multigroup comparisons (Sections 10.3.3 and 10.3.4).

Traditional exploratory LC analysis begins by fitting the null model H_0 to the sample of White respondents. Because $L^2(\text{H}_0) = 257.3$ with $df = 29$ (see Table 10.1), the amount of association (nonindependence) that exists in these data is too large to be explained by chance, so the null model must be rejected ($p < .001$) in favor of $T > 1$ classes.

Next, we consider McCutcheon's (1987) two-class model (H_1). For this model, the L^2 is reduced to 79.5,[3] a 69.1% reduction from the baseline model, but still much too large to be acceptable with $df = 22$. Thus, we increment T by 1 and estimate model H_{2C}, the three-class model. This model provides a further substantial reduction in L^2 to 22.1 (a 91.5% reduction over the baseline) and also provides an adequate overall fit ($p > .05$). Table 10.1 shows that the four-class LC model provides some further improvement. However, the BIC statistic, which takes parsimony into account, suggests that the three-class model is preferred over the four-class model (see Table 10.1).

The parameter estimates obtained from the three-class model are given in the left-most portion of Table 10.2. The classes are ordered from largest to smallest. Overall, 62% are estimated to be in Class 1, 20% in Class 2, and the remaining 18% in Class 3. Analogous to factor analysis, in which names are assigned to the factors based on an examination of the "factor loadings," names may be assigned to the latent classes based on the estimated conditional probabilities. Like factor loadings, the conditional probabilities provide the measurement *structure* that defines the latent classes.

2. The two formulations of BIC differ only with respect to a constant. More precisely, BIC_L^2 equals BIC_{LL} minus the BIC_{LL} corresponding to the saturated model.

3. This value differs slightly from the value 79.3 reported in McCutcheon (1987) because our models include a Bayes constant set equal to 1 to prevent boundary solutions (estimated model probabilities equal to zero). For further information on Bayes constants, see the technical appendix of the Latent GOLD 3.0 manual (Vermunt & Magidson, 2003, or www.latentclass.com).

Table 10.1 Results From Various Latent Class Models Fit to the General Social Survey 1982 Data

Model		BIC_{LL}	L^2	df	*p-Value*	*% Reduction in* $L^2(H_0)$
Sample of White respondents						
Traditional						
H_0	One-class	5787.0	257.3	29	2.0×10^{-38}	0.0
H_{1C}	Two-class	5658.9	79.5	22	2.0×10^{-8}	69.1
H_{2C}	Three-class	5651.1	22.1	15	.11	91.4
H_{3C}	Four-class	5685.3	6.6	8	.58	97.4
Nontraditional						
H_{1C+}	Two-class + $\{CD\}$ direct effect	5606.1	12.6	20	.89	95.1
H_{2F}	Basic two-factor	5640.1	11.1	15	.75	95.7
Sample of Black respondents						
Traditional						
H'_0	One-class	2402.1	112.1	29	1.0×10^{-11}	0.0
H'_1	Two-class	2389.6	56.9	22	.00006	49.2
H'_{2C}	Three-class	2393.8	18.3	15	.25	83.7
H'_{3C}	Four-class	2427.6	9.4	8	.31	91.6
Nontraditional						
H'_{1C+}	Two-class + $\{CD\}$ direct effect	2360.2	15.2	20	.77	86.4
H'_{2F}	Basic two-factor	2387.0	11.5	15	.72	89.7
Full sample (multiple-group analysis)						
Traditional						
M_0	One-class	8185.1	400.0	64	4.3×10^{-50}	0
M_1	Two-class	8013.8	169.5	56	2.4×10^{-13}	57.6
M_{2C}	Three-class unrestricted (complete heterogeneity)	8077.4	40.4	30	.10	89.9
M_{2CR}	Three-class restricted (partial homogeneity)	7953.0	49.4	48	.42	87.7
M_{2CRR}	Three-class restricted (complete homogeneity)	7962.1	73.3	50	.02	81.7
M_{3CR}	Four-class restricted (partial homogeneity)	7989.8	27.0	40	.94	93.3
Nontraditional						
M_{2F}	Basic two-factor unrestricted	8059.6	22.6	30	.83	94.4
M_{2FR}	Basic two-factor restricted	7934.9	31.3	48	.97	92.2

NOTE: BIC = Bayesian information criterion.

McCutcheon (1987) assigned the name "ideal" to latent Class 1, reasoning as follows:

> The first class corresponds most closely to our anticipated ideal respondents.
>
> Nearly 9 of 10 in this class believed that surveys "usually serve a good purpose"; 3 of 5 expressed a belief that surveys are either "almost always right" or "right most of the time"; 19 of 20 were evaluated by the interviewer as "friendly and interested" during the interview; and nearly all were evaluated by the interviewer as having a good understanding of the survey questions. (p. 34)

He named the other classes "believers" and "skeptics" based on the interpretations of the corresponding conditional probabilities for those classes.

10.2.2. Testing the Significance of Effects

The next step in a traditional LC analysis is to delete from the model any variable that does not exhibit a significant difference between the classes. For example, to test whether to delete variable A from a T-class model, one would test the null hypothesis that the distribution over the I categories of A is identical within each class t:

$$\pi_{i1}^{A|X} = \pi_{i2}^{A|X} = \cdots = \pi_{iT}^{A|X} \text{ for } i = 1, 2, \ldots, I.$$

Table 10.2 Parameter Estimates for the Three-Class Latent Class (LC) Model by Sample

	White Sample			*Black Sample*		
	Class 1 Ideal	*Class 2 Believers*	*Class 3 Skeptics*	*Class 1 Ideal*	*Class 2 Believers*	*Class 3 Skeptics*
LC probabilities	0.62	0.20	0.18	0.49	0.33	0.18
Conditional probabilities						
(*A*) PURPOSE						
Good	0.89	0.92	0.16	0.87	0.91	0.19
Depends	0.05	0.07	0.22	0.08	0.04	0.17
Waste	0.06	0.01	0.62	0.05	0.05	0.65
(*B*) ACCURACY						
Mostly true	0.61	0.65	0.04	0.54	0.65	0.01
Not true	0.39	0.35	0.96	0.46	0.35	0.99
(*C*) UNDERSTANDING						
Good	1.00	0.32	0.75	0.95	0.37	0.68
Fair, poor	0.00	0.68	0.25	0.05	0.63	0.32
(*D*) COOPERATION						
Interested	0.95	0.69	0.64	0.98	0.56	0.64
Cooperative	0.05	0.26	0.26	0.01	0.37	0.25
Impatient/hostile	0.00	0.05	0.10	0.00	0.07	0.11

To implement this test, we make use of the relationship between the conditional response probabilities and the log-linear parameters (see, e.g., Formann, 1992; Haberman, 1979; Heinen, 1996):

$$\pi_{it}^{A|X} = \frac{\exp(\lambda_i^A + \lambda_{it}^{AX})}{\sum_{i'=1}^{I} \exp(\lambda_{i'}^A + \lambda_{i't}^{AX})}.$$

Standard log-linear modeling techniques can then be used to test the null hypothesis, reexpressed in terms of the log-linear parameters associated with the AX relationship:

$$\lambda_{i1}^{AX} = \lambda_{i2}^{AX} = \cdots = \lambda_{iT}^{AX} = 0 \text{ for } i = 1, 2, \ldots, I.$$

One way to test for significance of the four indicators in our three-class model is by means of an L^2 difference test, where ΔL^2 is computed as the difference between the L^2 statistics obtained under the *restricted* and *unrestricted* three-class models, respectively. The ΔL^2 values obtained by setting the association parameters corresponding to one of the indicators to zero were 145.3, 125.4, 61.3, and 101.1, for A, B, C, and D, respectively. These numbers are higher than of the corresponding Wald statistics, which took on the values 29.6, 8.4, 7.4, and 19.0. This is because the latter test is uniformly less powerful than the ΔL^2 statistic. Under the assumption that the unrestricted model is true, both statistics are distributed asymptotically as chi-square with $df = (I-1)\cdot(T-1)$, where I denotes the number of categories in the nominal variable. The encountered values show that each of the four indicators included in the model is significantly related to class membership.

10.2.3. Classification

The final step in a traditional LC analysis is to use the results of the model to classify cases into the appropriate latent classes. For any given response pattern $(i, j, k, 1)$, estimates for the posterior membership probabilities can be obtained using Bayes's theorem as follows:

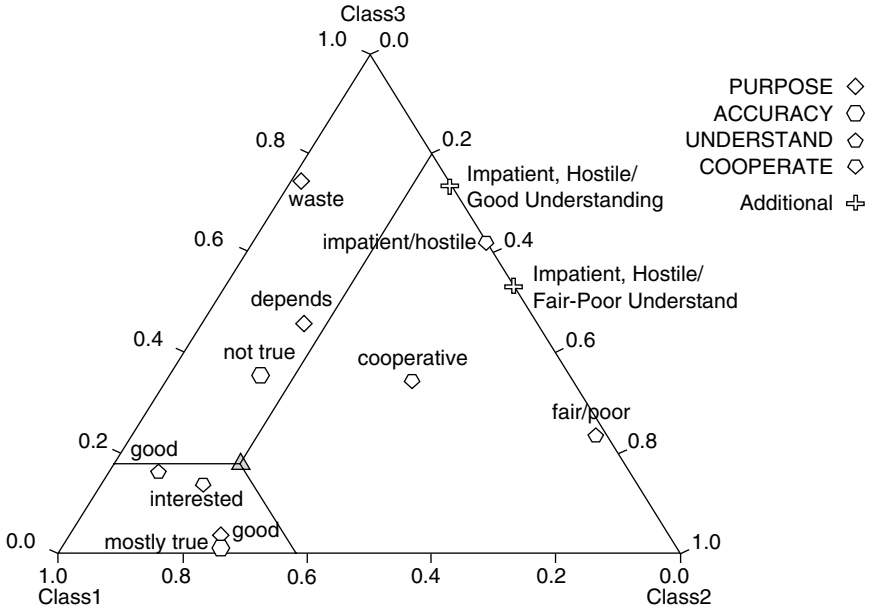
$$\hat{\pi}_{tijkl}^{X|ABCD} = \frac{\hat{\pi}_{ijklt}^{ABCDX}}{\sum_{t=1}^{T} \hat{\pi}_{ijklt}^{ABCDX}}, \quad t = 1, 2, \ldots, T \quad (2)$$

where the numerator and denominator in equation (2) are obtained by substituting the model parameter estimates in place of the corresponding parameters in equation (1).

Magidson and Vermunt (2001) and Vermunt and Magidson (2002) refer to this kind of model as a LC *cluster* model because the goal of classification into T homogeneous groups is identical to that of cluster analysis. In contrast to an ad hoc measure of distance used in cluster analysis to define homogeneity, LC analysis defines homogeneity in terms of probabilities. As indicated by equation (1), cases in the same latent class are similar to each other because their responses are generated by the same probability distribution.

Cases are then assigned to the class for which the posterior probability is highest (i.e., the modal

Figure 10.1 Barycentric Coordinate Display for Three-Class Model



class). For example, according to the three-class LC model, someone with response pattern $A = 1$ (PURPOSE = "good"), $B = 1$ (ACCURACY = "mostly true"), $C = 1$ (UNDERSTANDING = "good"), and $D = 1$ (COOPERATION = "interested") has posterior membership probabilities equal to 0.92, 0.08, and 0.00. This means that such a person is assigned to the first class.

10.2.4. Graphical Displays

Because for any given response pattern (i, j, k, l), the T-class membership probabilities sum to 1, only $T - 1$ such probabilities are required as the probability of belonging to the remaining class can be obtained from the others. Hence, the class membership probabilities $\hat{\pi}_{tijkl}^{X|ABCD}$ can be used to position each response pattern in $T - 1$ dimensional space, and for $T = 3$, various two-dimensional barycentric coordinate displays can be produced.

Rather than plotting every one of the many response patterns, instructive plots of the kind used in correspondence analysis can be produced, where points are plotted for each category of each variable as well as other meaningful aggregations of these probabilities (Magidson & Vermunt, 2001).

Figure 10.1 depicts the corresponding barycentric coordinate display under the three-class LC model. Points are plotted for each category of each of the four variables in our example. Because these points contain information equivalent to the LC parameter estimates (Van der Heijden, Gilula, & Van der Ark, 1999), this type of plot provides a graphical alternative to the traditional tabular display of parameter estimates and can yield new insights into data. Also displayed in Figure 10.1 are two additional aggregations associated with the response categories UNDERSTANDING = "good" and "fair, poor" ($k = 1, 2$) among those for whom COOPERATION = "hostile/impatient" ($1 = 3$).

The horizontal dimension of the plot corresponds to differences between McCutcheon's (1987) "ideal" and "believer" types (latent Classes 1 and 2). We see that the categories of the variable C tend to spread out along this dimension. Respondents showing "good" understanding are most likely to belong to the ideal class (the corresponding symbol is plotted closest to the lower left vertex that represents Class 1), whereas those showing only "fair or poor" understanding are plotted closest to the lower right vertex that represents Class 2.

Differences along the vertical dimension of the plot are best shown by the categories of A and B. For

Table 10.3 Descriptive Information and Parameter Estimates From Three-Class and Two-Factor Latent Class (LC) Models Obtained With the Landis and Koch (1977) Data

	Descriptive Information			*Three Classes*			*Two Factors (Joint Probabilities)*			
							Factor 1=1 (True Negative)		*Factor 1 = 2 (True Positive)*	
	% of Slides Rated Positive	*% of Ratings That Agree With*					*Factor 2 = 1 (Negative Bias)*	*Factor 2 = 2 (Positive Bias)*	*Factor 2 = 1 (Negative Bias)*	*Factor 2 = 1 (Positive Bias)*
		5+ Raters	*6+ Raters*	*Class 1*	*Class 2*	*Class 3*				
Class size				0.44	0.37	0.18	0.36	0.19	0.30	0.16
F	21	64	58	**0.47**	0.00	0.00	0.00	0.01	**0.23**	0.86
D	27	70	62	**0.59**	0.00	0.06	0.00	0.05	**0.37**	0.92
C	38	80	64	0.85	0.00	0.01	0.00	0.00	0.83	0.83
A	56	82	64	1.00	0.06	0.51	0.06	**0.47**	0.99	1.00
G	56	85	66	1.00	0.00	0.63	0.01	**0.58**	0.99	1.00
E	60	80	64	1.00	0.06	0.76	0.06	**0.72**	0.99	1.00
B	67	75	61	0.98	0.15	0.99	0.13	**0.99**	0.97	1.00

NOTE: False-negative and false-positive rates are highlighted in bold.

example, respondents agreeing that the purpose of surveys is "good" are plotted close to the lower left (Class 1) vertex. Those who say "it depends" are plotted somewhat midway between the Class 1 and Class 3 (top) vertex. Those who say "it's a waste of time and money" are most likely to be in Class 3 and are positioned near the top vertex. The fact that the positioning of categories for both *A* and *B* spreads out over the vertical dimension suggests a high degree of association between these variables. In contrast, the categories of *C* are spread over the horizontal dimension, suggesting that the association between *C* and the two variables *A* and *B* is close to nil.

The categories of the variable *D* form an interesting diagonal pattern. Respondents showing they are "interested" in the questions are most likely to be in Class 1 ("ideal"), whereas those who are only "cooperative" or exhibit "impatience/hostility" are plotted closer to Classes 2 and 3. This suggests the hypothesis that impatience and hostility may arise for either of two different reasons: (a) disagreement that surveys are accurate and serve a good purpose (indicated by the vertical dimension of the plot) and/or (b) lack of understanding (indicated by the horizontal dimension).

The additional points plotted deal with the relationship between variables *C* and *D*. The positioning of these points suggest that among impatient/hostile respondents, those who show good understanding of the questions tend to be more in Class 3, whereas those whose understanding is fair/poor tend to be about equally likely to be in Class 2 or 3.

We will revisit these data and obtain further insights later when we examine an alternative *nontraditional* two-dimensional LC model, the two-factor LC model.

10.2.4.1. Example: Sparse Multirater Agreement Data

We next consider an example with sparse data in which seven pathologists each classified 118 slides as to the presence or absence of carcinoma in the uterine cervix (Landis & Koch, 1977) that was also analyzed by Agresti (2002). LC modeling will be used here to estimate the false-positive and false-negative rates for each pathologist and to use multiple ratings to distinguish between slides that indicate carcinoma and those that do not (for similar medical applications, see Rindskopf & Rindskopf, 1986; Uebersax & Grove, 1990). The second column of Table 10.3 shows that the raters vary from classifying only about 1 of every 5 slides as positive (Rater D) to classifying more than 2 of every 3 as positive (Rater B). The next two columns indicate for which percentage of slides the ratings agree among five or more and six or more raters. This information shows that agreement is highest among Raters C, A, G, and E.

As a starting point, Agresti (2002) formulated a model containing two latent classes, in an attempt to confirm the hypothesis that slides are either "true positive" or "true negative." The assumption of local independence in the two-class model means that rater agreement is caused solely by the differing

Table 10.4 Results From Various Latent Class (LC) Models Fit to Landis and Koch (1977) Data

Model		BIC_{LL}	L^2	*Bootstrap p-Value*	*% Reduction in* $L^2(H_0)$
Traditional					
H_0	One-class	1082.3	476.8	.00	0.0
H_1	Two-class	707.9	64.2	.00	86.5
H_{2C}	Three-class	699.6	17.7	.49	96.3
H_{3C}	Four-class	729.4	9.3	.79	98.0
Nontraditional					
H_{2C+}	Three-class + {*DF*} direct effect	698.0	11.3	.83	97.6
H_{2FR}	Restricted basic two-factor	688.4	11.3	.90	97.6

NOTE: BIC = Bayesian information criterion.

characteristics between these two types of slides. That is, given that a slide is in the class of "true positive" ("true negative"), any similarities and differences between raters represent pure error. However, in his analysis of these data, he found that three classes were necessary to obtain an acceptable fit.

Although there are $2^7 = 128$ possible response patterns, because of the large amount of inter-rater agreement, 108 of these patterns were not observed at all. As mentioned above, sparse data such as these cause a problem in testing model fit because the L^2 statistic does not follow a chi-square distribution. For this reason, Agresti (2002) simply alluded to the obvious discrepancy between the expected frequencies estimated under the two-class model and the observed frequencies and speculated that this model does not provide an adequate fit to these data. He then compared estimates obtained from the three-class model and suggested that the fit of this model was adequate.

We report the bootstrap p-value in Table 10.4, which confirms Agresti's (2002) speculation that the fit of the two-class model *is* poor and that of the three-class model is adequate. It also shows that the three-class model is preferred over the four-class model according to the BIC criteria.

The parameter estimates obtained with the three-class model are given in the middle portion of Table 10.3. The largest class (44%) refers to slides that all pathologists (except for D and F) almost always agree show carcinoma ("true positive"). Class 2 (37%) refers to slides that all pathologists (except occasionally B) agree show no carcinoma. The remaining class of slides (18%) shows considerable disagreement between pathologists—B, E, and G usually diagnose carcinoma, whereas C, D, and F rarely do, and A diagnoses carcinoma half the time.

If we assume that Class 1 represents cases of true carcinoma, the results reported in Table 10.3 show that those pathologists who rated the *fewest* slides as positive (D and F) have the highest false-*negative* rates (42% and 53%, respectively, highlighted in bold). Similarly, under the assumption that Class 2 represents cases free from carcinoma, the results show that the pathologist who rated the *most* slides as positive, Rater B, shows a false-positive rate (15%) that is substantially larger than the other pathologists.

The traditional model-fitting strategy requires us to reject our two-class hypothesis in favor of a three-class alternative in which the third latent class consists of slides that cannot be classified as either "true positive" or "true negative" for cancer. Next we consider some nontraditional LC models that provide classification of each slide according to its likelihood of carcinoma. In particular, we will show that a two-factor LC model provides an attractive alternative whereby Factor 1 classifies all slides as either "true positive" or "true negative," and Factor 2 classifies slides according to a tendency for ratings to be biased toward false-positive or false-negative error.

10.3. Nontraditional Latent Class Modeling

Rejection of a traditional T-class LC model for lack of fit means that the local independence assumption does not hold with T classes. In such cases, the traditional LC model-fitting strategy is to fit a $T + 1$ class model to the data. In both of our examples, theory supported a two-class model but because this model failed to provide an adequate fit, we formulated a three-class model. In this section, we consider some alternative

strategies for modifying a model. In both cases, we will see the nontraditional alternatives lead to models that are more parsimonious than traditional models, as well as models that are more congruent with our initial hypotheses. The alternatives considered are as follows:

1. adding one or more direct effects,
2. deleting one or more items,
3. increasing the number of latent variables.

Alternative 1 is to include "direct-effect" parameters in the model (Hagenaars, 1988) that account for the residual association between the observed variables that are responsible for the local dependence. This approach is particularly useful when some external factor, unrelated to the latent variable, creates an irrelevant association between two variables. Examples of such external factors include similar question wording used in two survey items, as well as two raters using the same incorrect criterion in evaluating slides.

Alternative 2 also deals with the situation in which two variables are responsible for some local dependency. In such cases, rather than add a direct effect between two variables, it may make more sense to eliminate the dependency by simply deleting one of the two items. This variable reduction strategy is especially useful in situations in which there are *many* redundant variables.

Alternative 3 is especially useful when a group of several variables accounts for a local dependency. Magidson and Vermunt (2001) show that by increasing the dimensionality through the addition of latent variables rather than latent classes, the resulting LC *factor* model often fits data substantially better than the traditional LC cluster models having the same number of parameters. In addition, LC factor models are identified in some situations when the traditional LC model is not.[4]

In the next section, we introduce a diagnostic statistic called the *bivariate residual* (BVR) and illustrate its use to develop some nontraditional alternative models for our two data examples. The BVR helps pinpoint those *bivariate* relationships[5] that fail to be adequately explained by the LC model and can help determine which of the three alternative strategies to employ. We will see that even in situations when the L^2 statistic reports that the model provides an adequate *overall* fit, the fit in one or more two-way tables may not be adequate and may indicate a flaw or weakness in the model.

Table 10.5 Values for Bivariate Residuals Obtained Under Various Models for the Sample of White Respondents

Two-Way Table	*Model*					
	H_0	H_1	H_{2C}	H_{3C}	H_{2C+}	H_{2F}
$\{AB\}$	61.6	0.1	0.1	0.0	0.0	0.0
$\{AC\}$	0.5	0.7	0.1	0.0	0.2	0.0
$\{AD\}$	10.6	0.0	0.1	0.0	0.2	0.1
$\{BC\}$	0.3	1.1	0.0	0.0	0.0	0.0
$\{BD\}$	8.6	0.4	0.3	0.2	0.2	0.4
$\{CD\}$	43.4	32.3	2.4	0.0	0.0	0.2

10.3.1. Bivariate Residuals and Direct Effects

A formal measure of the extent to which the observed association between two variables is reproduced by a model is given by the BVR statistic (Vermunt & Magidson, 2003). Each BVR corresponds to a Pearson chi-square statistic (divided by the degrees of freedom) in which the observed frequencies in a two-way cross-tabulation of the variables are contrasted with those expected counts estimated under the corresponding LC model.[6] A BVR value substantially larger than 1 suggests that the model falls somewhat short of explaining the association in the corresponding two-way table.

10.3.1.1. Example: Survey Respondent Types (Continued)

Table 10.5 reports BVRs for each variable pair under each of several models estimated in our first example. Because model H_0 corresponds to the model of mutual independence, each BVR for this model provides a measure of the overall association in the corresponding observed two-way table; that is, each BVR equals the usual Pearson chi-square statistic used to test for independence in the corresponding two-way table divided by the degrees of freedom. The results show that except for the nonsignificant relationships in

4. For example, with four dichotomous variables, an LC two-factor model (composed of four latent classes) is identified, whereas a traditional three-class model is not (Goodman, 1974b).

5. Traditional factor analysis, through the assumption of multivariate normality, limits its focus to bivariate relationships (i.e., the correlations) because higher order relationships are assumed not to exist. In contrast, LC models do not make strict distributional assumptions and hence attempt to explain higher order associations as well. Nevertheless, the two-way (bivariate) associations are generally the most prominent, and the ability to pinpoint specific two-way tables in which lack of fit may be concentrated can be useful in suggesting alternative models.

6. These residuals are similar to Lagrangian statistics. A difference is that they are limited-information fit measures: Dependencies with parameters corresponding to other items are not taken into account.

the $\{AC\}$ and $\{BC\}$ tables, all of the remaining BVRs are quite large, attesting to several significant associations (local dependencies) that exist among these variables. The BVR is especially large for $\{AB\}$ and for $\{CD\}$. For example, in Table $\{CD\}$, a Pearson chi-square test confirms that the observed relationship is highly significant ($\chi^2 = 86.8$, $df = 2$, $p < .001$; BVR $= 86.8/2 = 43.4$).

Under the two-class model (H_1), note that the BVRs are all near or less than 1, except for one very large value of 32.3 for $\{CD\}$. This suggests that the overall lack of fit for this model can be traced to this single large BVR. The traditional way to account for the lack of fit is by adding another latent class. However, Table 10.5 shows that even after the addition of a third class, the BVR for $\{CD\}$ under the three-class model H_{2C} remains unacceptably high (BVR $= 2.4$). Although the inclusion of the third class does add a second dimension that causes the *overall* fit to be adequate, it is not until we add a fourth class (model H_{3C}) that *all* BVRs are at acceptable levels.

Below, we consider the alternative approach of adding a "direct effect" to the model to account for the residual correlation. In addition, we consider use of the two-factor LC model and further explore the differences between the three- and four-class models.

10.3.1.2. Example: Sparse Multirater Agreement Data (Continued)

Turning now to our second example, Table 10.6 shows that all of the BVRs under the one-class model of mutual independence (model H_0) are very large,[7] indicating that the amount of agreement between each pair of raters is highly significant. Under the two-class model, many BVRs remain large. Although the three-class model provides an acceptable overall fit to these data, again we see that there is a single BVR that remains unacceptably large—BVR $= 4.5$ for Raters D and F, the two pathologists who rated the fewest slides positive (recall Table 10.3). This large BVR suggests that Raters D and F may be using some rating criterion not shared by the other raters.

To account for this large residual association, we will use nontraditional Alternative 1 and modify the three-class model by adding the D through F direct-effect parameter λ^{DF} into the model (Hagenaars, 1988; for a slightly different formulation, see Uebersax, 1999). Formally, this new model, H_{2C+}, is expressed as

$$\pi_{ijklmpt} = \pi_{it}^{A|X}\pi_{jt}^{B|X}\pi_{kt}^{C|X}\pi_{mt}^{E|X}\pi_{lpt}^{DF|X},$$

7. The smallest BVR under model H_0 is 20.8, which occurs in table $\{EF\}$.

Table 10.6 Bivariate Residuals Obtained Under Various Models for Landis and Koch (1977) Data

	Model					
	Traditional			*Nontraditional*		
Two-Way Table[a]	H_0	H_1	H_{2C}	H_{2C+}	H_{2FR}	H_{2FRC}
$\{BE\}$	66.4	8.4	0.0	0.0	0.0	0.1
$\{DF\}$	38.0	7.2	4.5	0.0	0.0	0.0
$\{BG\}$	66.7	5.2	0.0	0.0	0.1	0.1
$\{EG\}$	77.2	3.3	0.1	0.1	0.2	0.2
$\{AB\}$	54.5	1.7	0.1	0.0	0.1	0.1
$\{CF\}$	28.0	1.3	0.0	0.0	0.0	0.0
$\{CE\}$	47.7	1.1	0.1	0.1	0.2	0.1
$\{DE\}$	24.5	0.0	0.7	0.6	0.6	1.2

a. These are the two-way tables for which the bivariate residuals were larger than 1 under any of the reported models (other than H_0).

where the probabilities $\pi_{lpt}^{DF|X}$ are constrained as follows:

$$\pi_{lpt}^{DF|X} = \frac{\exp(\lambda_l^D + \lambda_p^F + \lambda_{lp}^{DF} + \lambda_{lt}^{DX} + \lambda_{pt}^{FX})}{\sum_{l=1}^{L}\sum_{p=1}^{P}\exp(\lambda_l^D + \lambda_p^F + \lambda_{lp}^{DF} + \lambda_{lt}^{DX} + \lambda_{pt}^{FX})}.$$

By relaxing the local independence assumption between Raters D and F, model H_{2C+} is able to account for the excessive association between D and F that is not explainable by the latent classes. The ΔL^2 test shows that inclusion of the direct-effect parameter provides a significant improvement over the traditional model H_{2C} ($\Delta L^2 = 17.7 - 11.3 = 6.4$; $p = .01$).

From a practical perspective, models H_{2C} and H_{2C+} do not differ much as both models assign the 118 slides to the same classes under the modal assignment rule. This occurs despite the fact that model H_{2C+} gives D and F less weight than model H_{2C} during the computation of the posterior probabilities. The primary benefit of model H_{2C+} is to suggest the possibility that Raters D and F share a bias when evaluating Class 1 slides, those slides that D and F often rate negative but that the other pathologists almost always rate positive (recall Table 10.3). The implication of including the direct effect is that model H_{2C+} provides higher predictions of *agreement* between Raters D and F than model H_{2C} on Class 1 slides.[8]

8. Because model H_{2C} assumes local independence, the expected probability of both raters agreeing that a given Class 1 slide is free from cancer can be computed by multiplying the corresponding conditional probabilities. Using the estimates from Table 10.3, the probability of both agreeing that a Class 1 slide is negative is $.42 \times .53 = .22$, and similarly, the probability of both agreeing that it was positive is $.59 \times .47 = .28$. In contrast, model H_{2C+} predicts higher probabilities (.31 and .35, respectively) for Raters D and F agreeing in both cases. Under the assumption that Class 1 slides are "true positive," the results from model H_{2C+} mean that Raters D and F both tend to share a bias toward committing a false-negative error.

Returning to the first data example for a moment, we might now expect to find similar insights by the inclusion of the direct-effect parameters, λ_{kl}^{CD}, in the two-class model. Table 10.1 shows that this model (H_{1C+}) provides a good fit to the data. However, under this model, the parameter measuring the contribution of C to the latent classes is no longer significant, and therefore C can be deleted from the LC model completely. As this amounts to deleting an association simply because it could not be explained by a model with two latent classes, Alternative 1 does not provide a desirable solution here.

10.3.2. LC Factor Models

Next we consider Alternative 3, in which we use LC factor models to include more than one latent variable in the model. LC factor models were proposed as a general alternative to the traditional exploratory LC modeling by Magidson and Vermunt (2001). For both examples, the results (given in Table 10.1 and Table 10.4) show that a two-factor model is preferable to the other models. We shall see that the two-factor model is actually a restricted four-class model. In both cases, the fit is almost as good as the (unrestricted) four-class solution but is more parsimonious and parameterized in a manner that allows easier interpretation of the results.

LC factor models were initially proposed by Goodman (1974a) in the context of confirmatory latent class analysis. Certain traditional LC models containing four or more classes can be interpreted in terms of two or more component latent variables by treating those components as a joint variable (see, e.g., Hagenaars, 1990; McCutcheon, 1987). For example, a latent variable X consisting of $T = 4$ classes can be reexpressed in terms of two dichotomous latent variables $V = \{1, 2\}$ and $W = \{1, 2\}$ using the following correspondence:

	$W = 1$	$W = 2$
$V = 1$	$X = 1$	$X = 2$
$V = 2$	$X = 3$	$X = 4$

Thus, $X = 1$ corresponds with $V = 1$ and $W = 1$, $X = 2$ with $V = 1$ and $W = 2$, $X = 3$ with $V = 2$ and $W = 1$, and $X = 4$ with $V = 2$ and $W = 2$.

Formally, for four nominal variables, the four-class LC model can be reparameterized as a LC factor model with two dichotomous latent variables as follows:

$$\begin{aligned}\pi_{ijklrs} &= \pi_{rs}^{VW}\pi_{ijklrs}^{ABCD|VW} \\ &= \pi_{rs}^{VW}\pi_{irs}^{A|VW}\pi_{jrs}^{B|VW}\pi_{krs}^{C|VW}\pi_{lrs}^{D|VW}.\end{aligned}$$

Magidson and Vermunt (2001) consider various restricted factor models. They use the term *basic* LC factor models to refer to certain LC models that contain two or more dichotomous latent variables that are mutually independent of each other and that exclude higher order interactions from the conditional response probabilities. Such a model is analogous to the approach of traditional factor analysis in which multiple latent variables are used to model multidimensional relationships among manifest variables.

It turns out that by formulating the model in terms of R mutually independent, dichotomous latent factors, the basic LC factor model has the same number of distinct parameters as a traditional LC model with $R+1$ classes. That is, the LC factor parameterization allows specification of a 2^R-class model with the same number of parameters as a traditional LC model with only $R+1$ classes! This offers a great advantage in parsimony over the traditional T-class model as the number of parameters is greatly reduced by natural restrictions.

As mentioned previously, the basic two-factor model provides an excellent fit to both of our example data sets. For the first example, Table 10.1 shows that this model (model H_{2F}) is preferred over any of the LC cluster models according to the BIC. In addition, this model explains all bivariate relationships in the data (see Table 10.5). We will interpret the results from this model in the next section in conjunction with a more extensive analysis, including both the White and Black sample.

10.3.2.1. Example: Sparse Multirater Agreement Data (Continued)

Regarding our second example, Table 10.4 shows that the basic two-factor model is preferred over all the other models according to the BIC criteria. The rightmost portion of Table 10.3 provides the parameter estimates[9] that we used to name the factors. These are joint latent class and conditional response probabilities for combinations of factor levels. We assigned the names "true negative" and "true positive" to Levels 1 and 2 of Factor 1, respectively. Each of these levels is split again into two levels by Factor 2, which we

9. The two-factor model in Table 10.3 was further restricted by setting the effect of indicator C on Factor 2 to zero because this effect was not significant.

named "tendency toward ratings bias." We named the two levels of Factor 2 "tend to negative bias" and "tend to positive bias," respectively.

Comparing the four factor cells (right-most portion of Table 10.3) to the classes in the three-class model (middle portion of Table 10.3), we see the following similarities. First, note that Class 1 of the three-class solution (representing 44% of the slides mostly rated positive) corresponds primarily to Factor 1, Level 2 slides (those named "true positive"), which account for 46% of all slides. These "true-positive" slides are divided according to Factor 2 into cell (2, 1), accounting for 30% of all slides, and cell (2, 2), accounting for 16% of the slides. Note that the former slides show a clear tendency toward a false-negative error, especially among Raters D and F.

Next, notice the similarity between Class 2 of the three-class solution, representing 37% of the slides rated mostly negative, and factor cell (1, 1), accounting for 36% of the slides rated mostly negative. In addition, from Table 10.3, we can also see the strong similarity between Class 3 of the three-class solution and factor cell (1, 2), identified in the table as "true-negative" slides that are prone to false-positive error, especially by Raters A, B, E, and G.

In conclusion, we have shown that the two-factor LC model fits better than the traditional three-class model and offers two substantive advantages. First, it provides a clear way to classify slides as "true positive" or "true negative." Second, it provides a further grouping of slides that may be useful in pinpointing the reasons for rater disagreement. Of course, whether Factor 1 *actually* distinguishes between "true negative" and "true positive" and whether the error characterization given by Factor 2 is accurate are important questions that could be addressed in future research.

10.3.3. Multigroup Models

Multigroup LC models can be used to compare models across groups. A completely unrestricted multigroup LC model, referred to by Clogg and Goodman (1984) as the model of complete heterogeneity, is equivalent to the estimation of a separate T-class LC model for each group. The fit of such a model can be obtained by simply summing the L^2 values (and corresponding degrees of freedom) for the corresponding models in each group.

Let G denote a categorical variable representing membership in group g. The model of *complete heterogeneity* is expressed as (model $\mathrm{M}_{2\mathrm{C}}$)

$$\pi^{ABCDX|G}_{ijklt|g} = \pi^{X|G}_{t|g}\pi^{A|X,G}_{it|g}\pi^{B|X,G}_{jt|g}\pi^{C|X,G}_{kt|g}\pi^{D|X,G}_{lt|g}.$$

10.3.3.1. Example: Survey Respondent Types (Continued)

The second part of Table 10.1 provides the results of repeating our Example 1 analyses for the sample of *Black* respondents. These results turn out to be very similar to those obtained for the White respondents (see first part Table 10.1). As in our analysis for the White sample, we again reject the one- and two-class models in favor of three classes to obtain a model that provides an overall fit to the data that is adequate. The right-most portion of Table 10.2 presents the parameter estimates obtained from the three-class model (model $\mathrm{H}'_{2\mathrm{C}}$) as applied to the sample of Blacks. As in our earlier analysis, the classes are ordered from largest to smallest.

In comparing results across these two groups, it is important to be able to interpret the three classes obtained from the Black respondents as representing the same latent constructs ("ideal," "believers," and "skeptics") as in our analysis of the White respondents. Otherwise, any between-group comparisons would be like comparing apples with oranges. Although it is tempting to interpret Class 1 for both samples as representing the "ideal" respondents, this is not appropriate without first restricting the measurement portion of the models (the conditional probabilities) to be equal. These restrictions are accomplished using the model of *partial* homogeneity (model $\mathrm{M}_{2\mathrm{CR}}$):

$$\pi^{ABCDX|G}_{ijklt|g} = \pi^{X|G}_{t|g}\pi^{A|X}_{it}\pi^{B|X}_{jt}\pi^{C|X}_{kt}\pi^{D|X}_{lt}. \qquad (3)$$

Estimates from this model are given in the left-most portion of Table 10.7. The third part of Table 10.1 compares the fit of the unrestricted model $\mathrm{M}_{2\mathrm{C}}$ and restricted model $\mathrm{M}_{2\mathrm{CR}}$. The ΔL^2 statistic can be used to test the restrictions made under model $\mathrm{M}_{2\mathrm{CR}}$. Because $\Delta L^2 = 9.0$ with 18 *df* is *not* significant, we are free to use this restricted model for our group comparisons.

The model of complete homogeneity (model $\mathrm{M}_{2\mathrm{CRR}}$) imposes the further restriction that the latent class probabilities across the groups are identical: $\pi^{X|G}_{t|1} = \pi^{X|G}_{t|2}$, for $t = 1, 2, 3$. Because these restrictions yield a significant increase in L^2, we reject the model of complete homogeneity in favor of the model of partial homogeneity and conclude that there are significant differences in latent class membership between the White and Black samples.

Table 10.1 also includes results obtained from the LC factor model counterparts to the models of complete heterogeneity and partial heterogeneity. Because these models contain two dichotomous and independent factors, they contain the same number of parameters

Table 10.7 Parameter Estimates for the Three-Class Latent Class (LC) Model of Partial Homogeneity (Model M_{2CR}) and the Corresponding LC Two-Factor Model M_{2FR}

	Three Classes			*Two Factors (Marginal Probabilities)*			
	Class 1	*Class 2*	*Class 3*	*Factor V*		*Factor W*	
	Ideal	*Believers*	*Skeptics*	*Level 1*	*Level 2*	*Level 1*	*Level 2*
LC probabilities							
Whites	0.68	0.15	0.17	0.81	0.19	0.85	0.16
Blacks	0.51	0.30	0.19	0.79	0.21	0.70	0.31
Conditional probabilities							
PURPOSE							
Good	0.89	0.90	0.16	0.90	0.20	0.76	0.78
Depends	0.06	0.06	0.21	0.06	0.21	0.09	0.07
Waste	0.05	0.04	0.63	0.05	0.59	0.15	0.15
ACCURACY							
Mostly true	0.60	0.64	0.01	0.63	0.02	0.50	0.55
Not true	0.40	0.36	0.99	0.37	0.98	0.50	0.45
UNDERSTANDING							
Good	0.94	0.32	0.74	0.79	0.76	0.92	0.26
Fair, poor	0.06	0.68	0.26	0.21	0.24	0.08	0.74
COOPERATION							
Interested	0.95	0.57	0.65	0.86	0.66	0.90	0.50
Cooperative	0.05	0.35	0.25	0.12	0.24	0.09	0.38
Impatient/hostile	0.00	0.08	0.10	0.02	0.10	0.01	0.12

as the three-class models M_{2C} and M_{2CR}. The lower part of Table 10.1 shows that these models fit better than the corresponding LC cluster models according to the BIC criteria. Also, the smaller BVRs than the LC cluster counterpart confirm that the LC factor model fits the data better.

The parameter estimates from the two-factor model M_{2FR} are presented in the right-most portion of Table 10.7. These are marginal latent class and conditional response probabilities for factors V and W, which are obtained by summing over the other factor. Note that variable D is strongly related to both factors V and W. That is, respondents at Level 1 of each factor have a higher probability (.90 or .91) of being "interested" than those at Level 2. Variables A and B relate only to factor V, and variable C relates only to factor W. That is, for factor V, those at Level 1 are substantially more likely to agree that surveys serve a good purpose and are more accurate than those at Level 2, but the two levels are about equal in showing a good understanding of the questions. For factor W, Level 1 shows good understanding, but Level 2 does not.

Moreover, Table 10.7 shows that group differences exist primarily with respect to Factor 2 (observed group differences on factor V are not significant). Black respondents are twice as likely as Whites to be at Level 2 of Factor 2 (30% vs. 15%). These results allow us to formulate a more rigorous test of our earlier hypothesis that cooperation may be due to two separate factors—one associated with the belief that surveys serve a good purpose and are accurate (as assessed by LC Factor 1) and the second related to understanding the questions (as assessed by LC Factor 2).

Before concluding this section, we note that thus far, we have treated the trichotomous variables COOPERATE (A) and PURPOSE (C) as nominal. Alternatively, they can be treated as ordinal, which serves to simplify the model by reducing the number of parameters. The most straightforward approach is to restrict the log-linear parameters by using uniform scores v_i^A and v_k^C for the categories of A and C, implying the following constraints: $\lambda_{ir}^{AV} = \lambda_r^{AV} v_i^A$ and $\lambda_{is}^{AW} = \lambda_s^{AW} v_i^A$ (see, e.g., Formann, 1992; Heinen, 1996).

The use of these restrictions in our example increased the L^2 by very little, indicating that variables A and C may in fact be treated as ordinal. In the next section, we present the results of a modified two-factor model in which variables A and C are treated as ordinal.

10.3.4. Covariates

The parameters in the traditional LC model consist of unconditional and conditional probabilities. The conditional probabilities comprise the measurement

portion of the model. They characterize the distribution among the observed variables (indicators) conditional on the latent classes. The *unconditional* probabilities describe the distribution of the latent variable(s). To obtain improved description/prediction of the latent variable(s), we use a multinomial logit model to express these probabilities as a function of one or more exogenous variables **Z**, called covariates (Dayton & Macready, 1988).

The multigroup model described in the previous section is an example of the use of a single nominal covariate ($\mathbf{Z} = G$). For example, the term $\pi_{tg}^{X|G}$ in equation (3) can be expressed as

$$\pi_{tg}^{X|G} = \frac{\exp(\gamma_t^X + \gamma_{tg}^{XG})}{\sum_{t=1}^{T} \exp(\gamma_t^X + \gamma_{tg}^{XG})}.$$

Although the latent variable(s) explain all of the associations among the indicators, associations between the covariates are *not* explained by the latent variables. This is what distinguishes the indicators from the covariates.

10.3.4.1. Example: Survey Respondent Types (Continued)

Regarding the interpretation of the three-class solution, McCutcheon (1987) questioned whether some of the difference in latent class membership between Black and White respondents might be explained by education, a question that falls outside the scope of traditional LC modeling. We address this question below by including E: EDUCATION as a second covariate in the two-factor model—$\mathbf{Z} = (G, E)$.

The model provides a good fit to the data. The results indicate that the effect of education *does* explain most, but not all, of the group effect on factor W. The logit parameter estimates are given in Table 10.8, in which nonsignificant estimates were set to zero. The multinomial model used for the covariates was

$$\pi_{rsge}^{VW|GE} = \frac{\exp(\gamma_r^V + \gamma_s^W + \gamma_{gs}^{GW} + \gamma_{es}^{EW})}{\sum_{r=1,s=1}^{R,S} \exp(\gamma_r^V + \gamma_s^W + \gamma_{gs}^{GW} + \gamma_{es}^{EW})}.$$

The gamma parameters in Table 10.8 indicate that the higher the educational level, the lower the score on factor W. The race effect is very weak: Blacks have a slightly higher score on factor W than Whites.

The results for this two-factor restricted multigroup model are also displayed in the biplot display (Magidson & Vermunt, 2001) given in Figure 10.2. Like the barycentric coordinate display in Figure 10.1, we see that the horizontal axis, corresponding to factor W. is associated with UNDERSTANDING. Overall, respondents having a good understanding are highly likely to be at Level 1 of factor W,whereas those with a fair/poor understanding are highly likely to be at Level 2. The figure makes it clear that education is much more related to this factor than race. The vertical dimension is highly related to PURPOSE. Figure 10.2 shows more clearly than Figure 10.1 that COOPERATION is related to both factors. In particular, those rated as impatient/hostile tend to include two different types of respondents—those whose understanding is fair/poor, as well as those who view the purpose of surveys as a "waste of time and money."

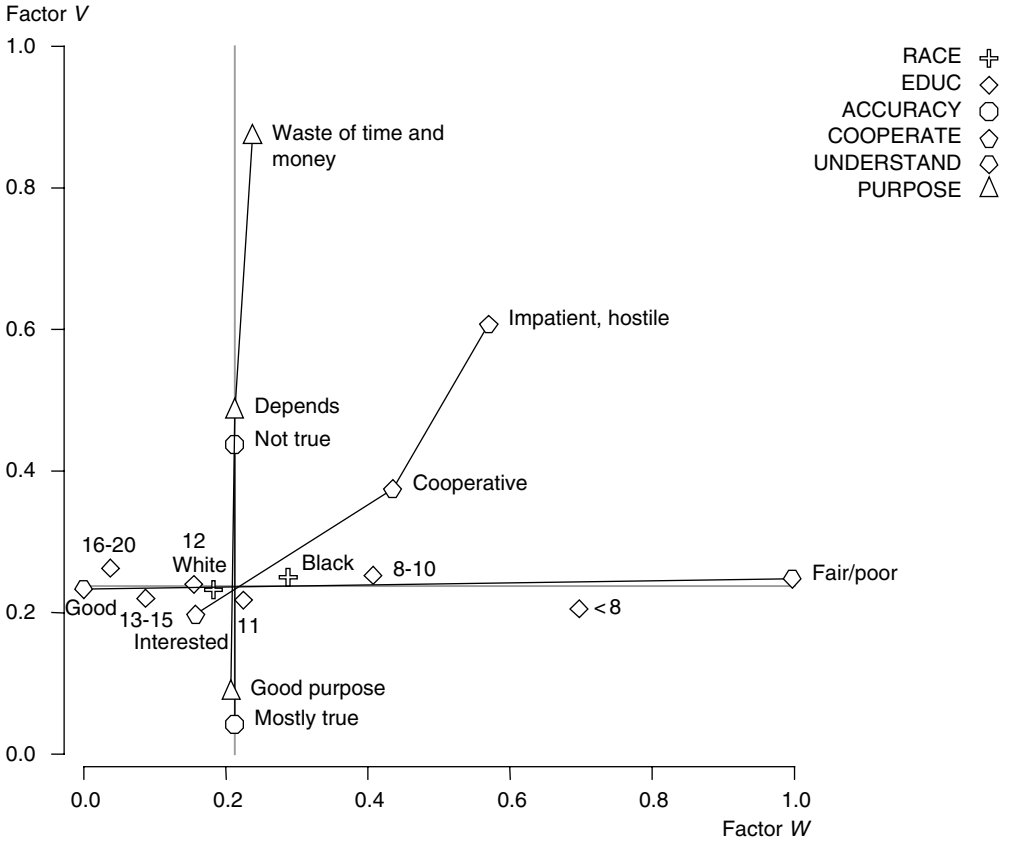
Table 10.8 Parameter Estimates for the Two-Factor Restricted Multigroup Latent Class (LC) Model With Covariates

	Factor	
	V	*W*
Covariates (gammas)		
G: Group		
WHITE	0	−0.20
BLACK	0	0.20
E: Years of education		
< 8	0	2.19
8–10	0	0.97
11	0	0.08
12	0	−0.34
13–15	0	−1.01
16–20	0	−1.89
Indicator variables (lambdas)		
A: PURPOSE	2.26	0
B: ACCURACY		
Mostly true	−1.34	0
Not true	1.34	0
C: UNDERSTANDING		
Good	0	−5.14
Fair/poor	0	5.14
D: COOPERATION	0.98	1.26

10.4. Other Types of Latent Class Models

Thus far, we have focused on the traditional LC modeling approach, including some important extensions such as covariates, several latent variables, and local dependencies. Some common characteristics of these models are that they serve as scaling methods or tools for dealing with measurement error, that indicators are nominal or ordinal, and that local

Figure 10.2 Biplot for Two-Factor Model With Covariates



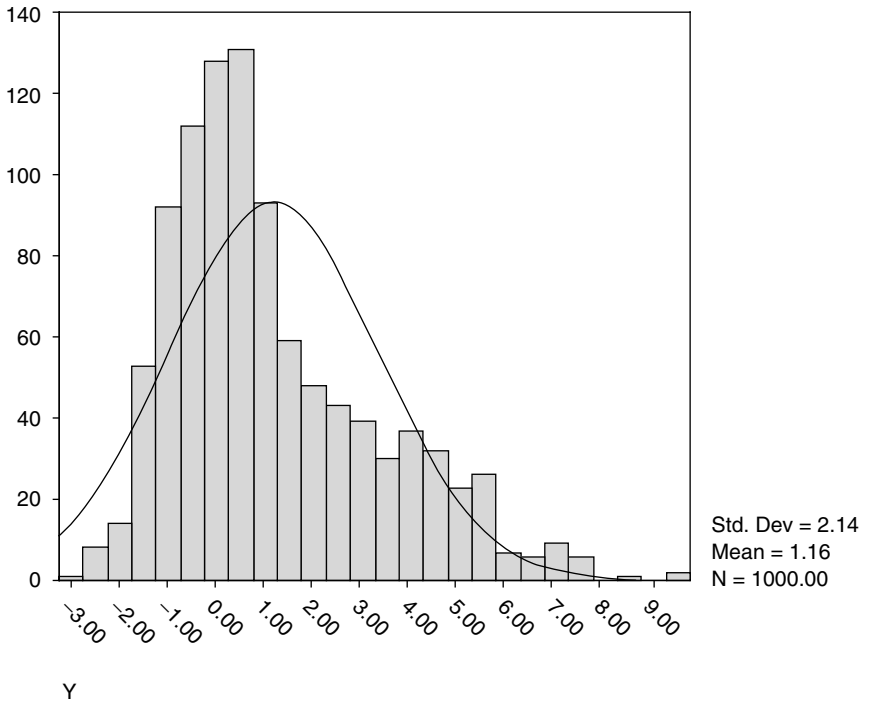
independence between indicators is the primary model assumption. In this section, we discuss other types of LC models. They are not used as scaling tools but as clustering methods, tools for dealing with unobserved heterogeneity, density estimation methods, or random-coefficients models (McLachlan & Peel, 2000). Moreover, indicators or dependent variables can be of scale types other than nominal or ordinal, and local independence is no longer the basic model assumption. As we will see, in some cases, there is only one indicator or dependent variable.

The next section presents simple mixture models for univariate distributions, with examples of mixtures of normals and mixtures of Poisson distributions. Then, we extend this basic model by including predictors, yielding what is called mixture regression or LC regression models. We present an example of a mixed linear regression model and show how the method can deal with various types of repeated measurements. Special attention is given to the relationship with hierarchical or multilevel models. Then, we present another extension of the simple mixture model, that is, a mixture model for multivariate distributions. As will be shown, the resulting LC model can be seen as a model-based alternative to standard hierarchical clustering methods such as K-means. We end with a short overview of LC methods that were not discussed in detail.

10.4.1. Simple Mixture Models

Consider the histogram depicted in Figure 10.3. This generated data set of 1,000 cases is obtained from a population consisting of a mixture of two normal distributions. For 60% of the population, the variable of interest follows a normal distribution with a mean of 0 and a variance of 1, $N(0, 1)$; for the other 40%, the mean equals 3 and the variance 4, $N(3, 4)$. The normal curve that is drawn through the histogram shows that the resulting mixture is clearly not normally distributed.

Figure 10.3 Simulated Distribution From a Mixture of Two Normals



A model that can be used to describe such a phenomenon is a finite mixture model (Everitt & Hand, 1981; McLachlan & Peel, 2000), which is a particular kind of LC model. The basic formula for a mixture of univariate distributions is

$$f(y|\boldsymbol{\vartheta}) = \sum_{t=1}^{T} \pi_t^X f(y|\boldsymbol{\varphi}_t). \quad (4)$$

The left-hand side of equation (4) indicates that we are interested in describing the distribution of a random variable y, which depends on a set of unknown parameters $\boldsymbol{\vartheta}$. The right-hand side contains two terms: π_t^X is the probability of belonging to latent class or mixture component t, and $f(y|\boldsymbol{\varphi}_t)$ is the distribution of y within latent class t, given some unknown parameters $\boldsymbol{\varphi}_t$. The class-specific distribution of y is assumed to belong to a particular parametric family. Depending on the scale type of y, this can, for instance, be a normal, Poisson, binomial, exponential, or gamma distribution. The summation on the right-hand side indicates that the distribution of y is a weighted mean of the class-specific distributions, where the latent class proportions serve as weights.

Mixture models such as these have two important types of applications. The first is density estimation: Complicated distributions can be approximated by a mixture of simple parametric distributions. Another important application type is clustering, in which case the class-specific parameters are used to define the clusters, and the posterior membership probabilities are used to classify cases into the most appropriate cluster.

Table 10.9 Test Results for Generated Mixture of Normals Data

Model	*Log-Likelihood*	BIC_{LL}	*Number of Parameters*
Equal variances			
One-class	−2177.75	4369.31	2
Two-class	−2066.99	4161.61	4
Three-class	−2050.78	4143.00	6
Four-class	−2046.25	4147.75	8
Unequal variances			
Two-class	−2048.14	4130.81	5
Three-class	−2047.78	4150.83	8
Four-class	−2045.41	4166.80	11

NOTE: BIC = Bayesian information criterion.

Table 10.9 presents test results for various models fitted to the data depicted in Figure 10.3. We estimated one- to four-class mixtures of normal distributions with equal and unequal within-cluster variances. As can be seen, the BIC measure identifies the correct model, the two-cluster model with unequal

Table 10.10 Observed and Estimated Frequency Distribution of Packs of Hard Candy Purchased During Past 7 Days Under the One-Class and Three-Class Poisson Model

	Frequencies		
Number of Packages	*Observed*	*One-Class Model*	*Three-Class Model*
0	102	8.43	101.67
1	54	33.63	54.63
2	49	67.11	50.03
3	62	89.28	53.89
4	44	89.09	47.25
5	25	71.11	34.14
6	26	47.30	22.00
7	15	26.97	14.37
8	15	13.46	11.02
9	10	5.97	10.18
10	10	2.38	10.17
11	10	0.86	9.97
12	10	0.29	9.20
13	3	0.09	7.90
14	3	0.03	6.32
15	5	0.01	4.72
16	5	0.00	3.30
17	4	0.00	2.18
18	1	0.00	1.36
19	2	0.00	0.80
20	1	0.00	0.45

within-cluster variances, as best. The three-class model with equal within-cluster variances fits almost as well, showing that a simpler parametric form can sometimes be compensated by a larger number of mixture components.

In the two-class model with unequal variances, the estimated probability of belonging to Class 1 is .64. This class has an estimated mean of -0.03 and a variance of 1.01. The mean and variance of the other class equal 3.24 and 3.95. Note that these estimates are close to the population values we used to generate this data set.

Table 10.10 provides a data set taken from Dillon and Kumar (1994) that we will use as a second example. It gives the observed frequency distribution of the number of packs of hard candy consumed by 456 respondents during the 7 days prior to the survey. Because the outcome variable is a count without a fixed maximum, it is most natural to assume that it follows a Poisson distribution. The table also reports the estimated frequency distribution obtained with a standard, or one-class, Poisson model, as well as with a three-class mixture Poisson model. As can be seen, the standard Poisson model does not fit the empirical distribution at all, whereas the three-class Poisson describes the data almost perfectly. This shows that a mixture of simple parametric distributions can be used to describe a quite complicated empirical distribution.

Test results obtained when applying mixture Poisson models to the hard candy data set show that models with two and three mixture components perform much better than the standard Poisson model. As is typical, there is a saturation point at which increasing the number of classes no longer increases the log-likelihood function: In this case, it occurs at four classes. The three-cluster solution is the one that is preferred according to the BIC criterion.

The estimated latent class proportions in the three-class model are 0.54, 0.28, and 0.18, and the Poisson rates are 3.48, 0.29, and 11.21. This means that we identified a small cluster of heavy users (more than 11 packs in 7 days), a cluster containing slightly more than a quarter of the respondents with almost no usage, and a large group of moderate users.

10.4.2. LC Regression Models

In the simple mixture models discussed above, it was assumed that the mean of the chosen parametric distribution differs across latent classes. This can also be expressed by specifying a linear regression model for the mean of the distribution of interest, μ_t, after applying some transformation or link function $g(..)$ that depends on the scale type of the y variable. For the mean of a binomial or multinomial distribution, we use a logit transformation; for a Poisson mean, a log transformation; and for a normal mean, no transformation or an identity link. The regression model has the form

$$g(\mu_t) = \beta_{0t}.$$

As can be seen, this regression model contains only an intercept, and this intercept is class specific.

Let w denote a set of predictors or explanatory variables. Suppose we are no longer interested in the unconditional distribution of y but in the conditional distribution of y given w, $f(y|\mathbf{w}, \boldsymbol{\varphi}_t)$. A natural way to express the dependency of y on w is by the inclusion of the set of predictors $\mathbf{w}$ on the right-hand side of the regression equation. In the case of a single predictor w, the resulting LC regression model (Wedel & DeSarbo, 1994) has the form

$$g(\mu_t) = \beta_{0t} + \beta_{1t} w,$$

where β_{0t} and β_{1t} are the class-specific regression coefficients.

Figure 10.4 Simulated Two-Class Latent Class (LC) Regression Model

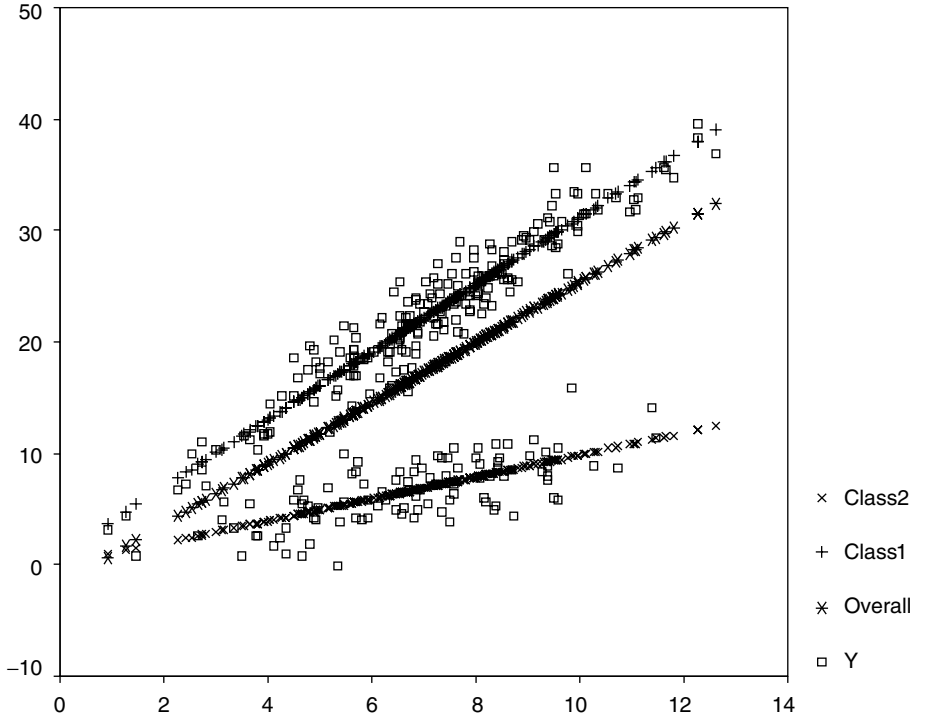


Figure 10.4 depicts a data set generated from a population consisting of two latent classes, with class-specific regression models equal to $\mu_1 = 1 + 3w$ and $\mu_2 = 0 + 1w$. It also compares the estimated y values for the two-class model (YLC2) with the standard one-class regression model (YLC1). As can be seen, the description given by the standard regression model is very poor compared to the two-class model. The LC regression modeling procedure has no problem identifying the two regression lines without preknowledge of class membership.

In an LC regression model, the latent variable is a predictor that interacts with the observed predictors, which means that it serves as a moderator variable. Compared to a standard regression model in which all predictors are observed, this basic LC regression model provides several useful functions. First, it can be used to weaken standard regression assumptions about the nature of the effects (linear, no interactions) and the error term (independent of predictors, particular distribution, homoskedastic). Second, it makes it possible to identify and correct for sources of unobserved heterogeneity. As explained below, this is especially useful in situations when there are repeated measurements or other types of dependent observations. Longitudinal data applications are sometimes referred to as LC or mixture growth models (each latent class has its own growth curve). Third, it can be used to detect outliers because these are cases for which the primary regression model does not hold.

An important application area for LC regression modeling is clustering or segmentation (Wedel & Kamakura, 1998). In particular, ratings- and choice-based conjoint studies are designed to identify subgroups (segments) that react differently to product characteristics, which is the same as saying that these groups have different regression coefficients. This type of application is illustrated in more detail below with an empirical example.

10.4.2.1. Example: Repeated Measurements or Clustered Observations

As explained below, the LC regression model can be viewed as a random-coefficients model that, similar to multilevel or hierarchical models, can take dependencies between observations into account. This extends the application of LC regression models to

Table 10.11 Parameter Estimates for the Abortion Example

Parameter	*Class 1*	*Class 2*	*Class 3*	*Class 4*	*Mean*	*Standard Deviation*
Class size	0.30	0.28	0.24	0.19		
Intercept	−0.34	0.60	3.33	1.59	1.16	1.38
Year						
1983	0.14	0.26	0.47	−0.58	0.12	0.35
1984	−0.12	−0.46	−0.35	−1.11	−0.45	0.34
1985	0.04	−0.44	−0.26	1.43	0.10	0.66
1986	−0.06	0.64	0.14	0.26	0.24	0.27
Religion						
Roman Catholic	−0.53	−0.53	−0.53	−0.53	−0.53	0.00
Protestant	0.20	0.20	0.20	0.20	0.20	0.00
Other	−0.10	−0.10	−0.10	−0.10	−0.10	0.00
No religion	0.42	0.42	0.42	0.42	0.42	0.00

situations with repeated measurements or other types of dependent observations.

We will illustrate LC regression with repeated measurements using an application to longitudinal survey data. This is, therefore, an example of an LC growth model. The data set consists of 264 participants in the 1983 to 1986 yearly waves of the British Social Attitudes Survey (McGrath & Waterton, 1986). The dependent variable is the number of yes responses on seven yes/no questions as to whether it is a woman's right to have an abortion under specific circumstances. Because this is a count variable with a fixed total, it is most natural to work with a logit link and binomial error function. The predictors that we used are the year of measurement (1 = 1983, 2 = 1984, 3 = 1985, 4 = 1986) and religion (1 = Roman Catholic, 2 = Protestant, 3 = other, 4 = no religion). The effect of year of measurement is assumed to be class dependent, and the effect of religion is assumed to be the same for all classes.

We estimated models with one to five classes, and the four-class model turned out to perform best in terms of the BIC criterion. We also estimated more restricted models in which the time effect is assumed to be linear and/or the time effect is assumed to be class independent. These models did not describe the data as well as our four-class model, which indicates that the time trend is nonlinear and heterogeneous.

The parameters obtained with the four-class model appear in Table 10.11. The parameter means across classes indicate that the attitudes are most positive at the last time point and most negative at the second time point. Furthermore, the effects of religion show that people without religion are most in favor and Roman Catholics and others are most against abortion. Protestants have a position that is close to the no-religion group.

The class-specific parameters indicate that the four latent classes have very different intercepts and time patterns. The largest Class 1 is most against abortion, and Class 3 is most in favor of abortion. Both latent classes are very stable over time. The overall level of latent Class 2 is somewhat higher than of Class 1, and it shows somewhat more change of the attitude over time. People belonging to latent Class 4 are very instable: At the first two time points they are similar to Class 2, at the third time point to Class 4, and at the last time point again to Class 2 (this can be seen by combining the intercepts with the time effects). Class 4 could therefore be labeled as random responders. It is interesting to note that in a three-class solution, the random-responder class and Class 2 are combined. Thus, by going from a three- to a four-class solution, one identifies the interesting group with less stable attitudes.

Vermunt and Van Dijk (2001) used the same empirical example to illustrate the similarity between LC regression models and random-coefficients, multilevel, or hierarchical models. Using terminology from multilevel modeling, the time variable is a Level 1 predictor and religion a Level 2 predictor. The effect of the Level 1 predictor time is allowed to vary across Level 2 units—in this case, individuals. The LC regression output can be transformed into the usual output produced by a standard multilevel or hierarchical model—means, variances, covariances of the intercept, and the three time effects—by elementary statistical operations. The most important part of this multilevel output is what appears in the last two columns of Table 10.11.

A difference between LC regression analysis and standard hierarchical models is that the former does not make strong assumptions about the distribution of the random coefficients. LC regression models can, therefore, be seen as nonparametric hierarchical models in which the distribution of the random coefficients is approximated by a limited number of mass points (= latent classes). As shown by Vermunt and Van Dijk (2001), the LC approach has the practical advantage of being much less computationally intensive than parametric models, and substantively easier to interpret results are often obtained.

10.4.2.2. Example: Application to Choice-Based Conjoint Studies

The LC regression model is a popular tool for the analysis of data from conjoint experiments in which individuals rate or choose between sets of products having different attributes (Wedel & Kamakura, 1998). The objective is to determine the effect of product characteristics on the rating or the choice probabilities. LC analysis is used to identify subgroups or market segments for which these effects differ.

For illustration of LC analysis of data obtained from choice-based conjoint experiments, we use a generated data set. The products are 10 pairs of shoes that differ on three attributes: fashion (0 = traditional, 1 = modern), quality (0 = low, 1 = high), and price (ranging from 1 to 5). Eight choice sets offer 3 of the 10 possible alternative products to 400 individuals. Each choice task consists of indicating which of the 3 alternatives they would purchase, with the response "none of the above" allowed as a fourth choice option.

The model that is used is a multinomial logit model with choice-specific predictors, also referred to as the conditional logit model. Let M be the number of choice sets, K the number of choices per set, and J the number of predictors. A particular set, choice, and predictor are denoted by m, k, and j, respectively. The regression model of interest is

$$\pi_{mkt} = \frac{\exp(\sum_{j=1}^{J} \beta_{jt} w_{mjk})}{\sum_{k=1}^{K} \exp(\sum_{j=1}^{J} \beta_{jt} w_{mjk})}.$$

Here, π_{mkt} denotes the probability that someone belonging to class t selects choice alternative k in choice set m. The predictors we use are the three product attributes (fashion, quality, and price), as well as a dummy variable for the "none" category.

The BIC values indicated that the three-class model is the model that should be preferred. The parameter estimates obtained with the three-class model are reported in Table 10.12. As can be seen, fashion has a major influence on choice for Class 1, quality for Class 2, and both fashion and quality for Class 3. The price effect is similar for all three classes. The Wald test for the equality of effects between classes indicates that the difference in price effects across classes is not significant. The price effects could, therefore, be assumed to be class independent.

Table 10.12 Parameter Estimates for Conditional Logit Model in Conjoint Study Example

	Class 1	*Class 2*	*Class 3*	*Wald for No Effect*	*Wald for Equal Effects*
Fashion	3.03	−0.17	1.20	494.74	216.37
Quality	−0.09	2.72	1.12	277.96	171.16
Price	−0.39	−0.36	−0.56	144.48	3.58
None	1.29	0.19	−0.43	82.39	59.26

Table 10.13 Parameter Estimates for the Latent Variable Regression for Conjoint Study Example

	Class 1	*Class 2*	*Class 3*	*Wald*
Intercept	0.37	0.00	−0.37	8.22
SEX				
Male	−0.66	−0.34	1.01	24.15
Female	0.66	0.34	−1.01	
AGE				
16–24	1.02	−0.15	−0.87	62.76
25–39	−0.59	−0.37	0.96	
40+	−0.43	0.52	−0.09	

In addition to the conditional logit model, which shows how the predictors affect the likelihood of choosing one alternative over another, differentially for each class, we specified a second logit model to describe the latent class variable as a function of the covariates sex and age. Table 10.13 shows that females turn out to belong more often to Class 1 and males to Class 3. Younger persons have a higher probability of belonging to Class 1 (emphasize fashion in choices), and older persons are most likely to belong to Class 2 (emphasize quality in choices).

In conclusion, the LC regression model offers computational and interpretive advantages over the more traditional hierarchical modeling approach that tends to overfit data (Andrews, Ansari, & Currim, 2002). In our example, we used the BIC criteria to select a parsimonious number of classes. However, researchers who prefer the results to show *higher* levels of individual variation in regression coefficients can obtain such

with LC regression models by simply increasing the number of latent classes to produce the desired amount of variation.

10.4.3. LC Analysis as an Alternative to K-Means Clustering

An important application of LC analysis is clustering (Banfield & Raftery, 1993; McLachlan & Peel, 2000; Vermunt & Magidson, 2002). Actually, we already saw several cluster-like applications. The traditional LC model was used to construct a typology of survey respondents using a set of categorical indicators. We also showed that simple mixture models such as mixtures of normals or mixtures of Poisson distributions could be used for clustering purposes.

In this section, we will concentrate on LC analysis as a tool for cluster analysis with *continuous* indicators. These LC models can be seen as multivariate extensions of the mixtures of univariate normals discussed above. Instead of assuming a univariate normal distribution, we assume multivariate normal distributions within latent classes. The most general form of the mixture model concerned assumes that each latent class has its own set of means, variances, and covariances. More formally,

$$f(\mathbf{y}|\boldsymbol{\vartheta}) = \sum_{t=1}^{T} \pi_t^X f(\mathbf{y}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t).$$

Here, $\boldsymbol{\mu}_t$ denotes the vector with class-specific means and $\boldsymbol{\Sigma}_t$ the class-specific variance-covariance matrix. Note that, contrary to traditional LC modeling, it is not necessary to assume local independence between the indicators.

The above LC cluster model is similar to the model used in discriminant analysis. An important difference is, of course, that in cluster analysis, group membership is unobserved or latent, which is the reason that LC cluster analysis is sometimes referred to as latent discriminant analysis.

The first part of Figure 10.5 depicts a data set that we will use to illustrate the LC cluster model for continuous variables. Three measures are available to diagnose diabetes: glucose, insulin, and steady-state plasma glucose (SSPG) (see Fraley & Raftery, 1998). In addition to these measures, we have information on the clinical diagnosis consisting of the three categories "normal," "chemical diabetes," and "overt diabetes." However, in practice, a gold standard is not available in cluster applications. Our objective here is to construct a mixture model that yields a classification that is close to the clinical diagnosis, without use of the information on the clinical diagnosis. We use this data set to demonstrate the flexibility of LC clustering compared to other clustering methods. The gold

Figure 10.5 Matrix Scatter Plot of Diabetes Data Set for the Clinical Classification, the K-Means-Like Five-Cluster Solution, and the Final Three-Cluster Solution

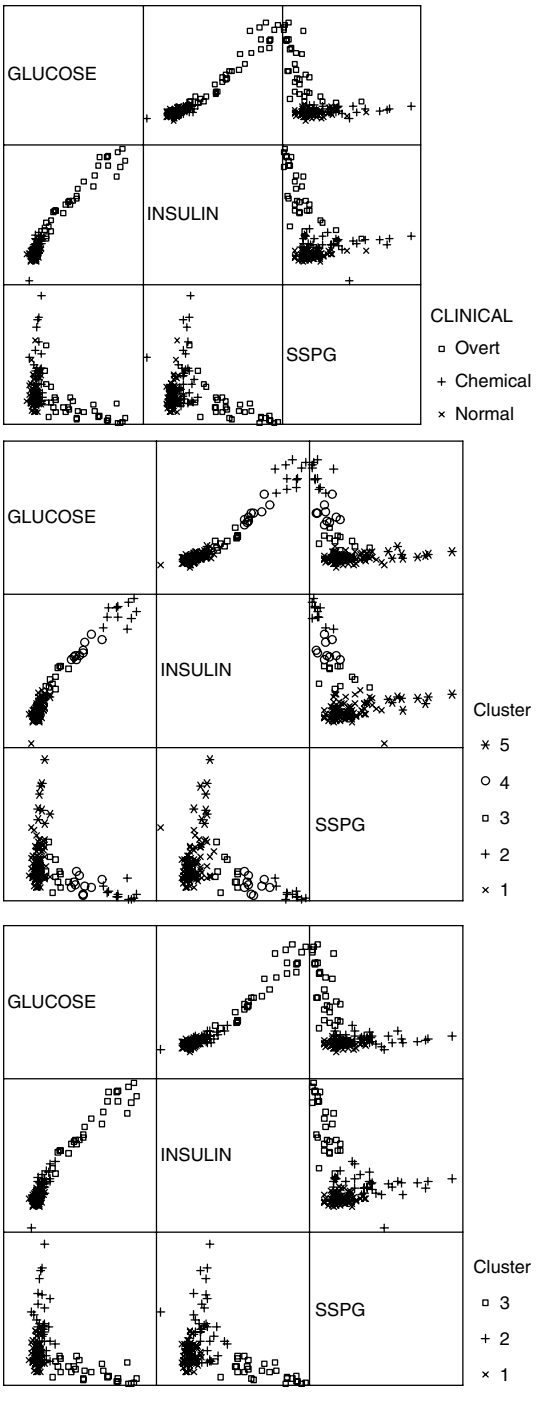


Table 10.14 Test Results for Diabetes Data

Model	*Log-Likelihood*	BIC_{LL}	*Number of Parameters*
Equal and diagonal			
One-cluster	−2750.13	5530.13	6
Two-cluster	−2559.88	5169.52	10
Three-cluster	−2464.78	4999.24	14
Four-cluster	−2424.46	4938.49	18
Five-cluster	−2392.56	**4894.60**	22
Unequal and diagonal			
One-cluster	−2750.13	5530.13	6
Two-cluster	−2446.12	4956.94	13
Three-cluster	−2366.92	4833.38	20
Four-cluster	−2335.38	**4805.13**	27
Five-cluster	−2323.13	4815.47	34
Unequal and full			
One-cluster	−2546.83	5138.46	9
Two-cluster	−2359.12	4812.80	19
Three-cluster	−2308.64	**4761.61**	29
Four-cluster	−2298.13	4790.34	39
Five-cluster	−2284.97	4813.79	49
Unequal and $y_1 - y_2$ free			
One-cluster	−2560.40	5155.64	7
Two-cluster	−2380.27	4835.19	15
Three-cluster	−2320.57	**4755.61**	23
Four-cluster	−2303.14	4760.56	31
Five-cluster	−2295.05	4784.19	39

NOTE: Bold numbers (minimum Bayesian information criterion [BIC]) indicate the model that would be selected according to the BIC criterion.

standard makes it possible to judge whether the methods do what we want them to do.

LC cluster analysis is a model-based clustering procedure. As such, it is a probabilistic and more flexible alternative to K-means clustering. K-means clustering performs well under very strict conditions—that is, if indicators are locally independent and if error variances are cluster invariant and equal across indicators ($\boldsymbol{\Sigma}_t = \sigma^2\mathbf{I}$). These implicit assumptions of K-means imply that in a three-dimensional scatter plot, each cluster has the form of a sphere with the same radius, and in each two-dimensional plot, each cluster will have the form of a sphere with the same radius. The assumption of equal error variances across indicators is the reason that in K-means clustering, it is advised to standardize the variables prior the analysis. Although standardization often improves the situation, it does not solve the problem because equating the variance in the total sample is not the same as equating the within-group variances (Magidson & Vermunt, 2002).

Having a closer look at Figure 10.5, it can easily be seen that it is impossible to describe the shape of the three diabetes clusters by a K-means model, that is, by three spheres with the same radius. The within-cluster variances are very different across clusters and across indicators. Moreover, the glucose and insulin indicators are strongly correlated within the group with overt diabetes. Nevertheless, because the clusters are well separated, a reliable cluster method should be able to yield a three-cluster solution that is similar to the clinical classification.

The problems associated with K-means are confirmed by the test results reported in Table 10.14. We estimated one- to five-cluster models, each with four different specifications of the variance-covariance matrix: diagonal (= local independence) and equal across classes, diagonal and unequal, glucose-insulin covariance and unequal, and all covariances and unequal. It can be seen that when the specifications are too restrictive, one needs five and four clusters, respectively. Actually, with the first K-means-like specification, even more than five clusters are needed.

Although the BIC values indicate that the two additional local dependencies ($y_1 - y_3$ and $y_2 - y_3$) in the full model are not needed (compare the three-cluster solutions for the last two specifications), the fit measures also show that both the model with the fully unrestricted covariance matrix and the model with only the glucose-insulin covariance detect the correct three-cluster solution. This means that working with a

model with insufficient restrictions does not harm in this example, but this is not always the case.

The middle part of Figure 10.5 shows the five nearly spherical clusters identified with the most restricted specification we used. Similar results would have been obtained with K-means. The lower part of Figure 10.5 depicts the three-cluster solution that turned out to be the best according to the BIC criterion. It can be seen that the three clusters identified by this model are very similar to the clinical classification. Our three-cluster solution is smoother in the sense that some of the overlap between the clinical classes disappears, which is, of course, what can be expected from a statistical model. The correspondence between the three-cluster and the clinical classification is 87%, which is only slightly lower than the 93% correct classifications of a quadratic discriminant analysis (in which cluster membership is treated as known).

The LC cluster model can be applied not only with continuous indicators but also with indicators of other scale types and different combinations of scale types. Depending on the scale type, one will specify the most appropriate within-cluster distribution for the indicator concerned. This yields a general cluster model for mixed-model data (Hunt & Jorgensen, 1999; Vermunt & Magidson, 2002). Note that the traditional LC model is the special case in which all indicators are categorical variables.

10.4.4. Other Developments in LC Modeling

In this chapter, we presented what we believe to be the most important types of LC models. We did not discuss LC models for transition, survival, or event history data (Vermunt, 1997). Most of these models are mixture regression models and can, therefore, be handled within the LC regression framework. Another important class of models for transition data are latent or hidden Markov models that can be used to separate true change from measurement error in the outcome variable of interest (see, e.g., Langeheine & Van de Pol, 1994). The structure of latent Markov models is similar to the LC models with several latent variables discussed in the previous section.

In the previous section, we presented LC models that can be used for scaling. There also exist more sophisticated LC scaling models, which can be obtained by imposing certain constraints on the parameters of the traditional LC model. Examples are LC models for probabilistic Guttman scaling, LC models with order constraints, LC Rasch models, LC models for preference data, and LC models for distance data (see Böckenholt, 2002; Croon, 2002; Dayton, 1998; Heinen, 1996).

Another more advanced type of LC model we would like to mention is the LISREL-type framework for categorical variables developed by Hagenaars (1990) and extended by Vermunt (1997). Any type of LC models with categorical indicators, including LC models for transition data and sophisticated LC scaling models, are special cases of this general model. A limitation of this approach is that it is restricted to categorical indicators.

A final recent development that we would like to mention is the development of more sophisticated restricted mixtures of multivariate normals than that discussed above. LC models have been proposed in which the class-specific covariance matrices are constrained by means of principal component (Fraley & Raftery, 1998) or factor-analytic (Yung, 1997) structures or by structural equation models (Jedidi, Jagpal, & DeSarbo, 1997).

REFERENCES

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: John Wiley.

Andrews, R. L., Ansari, A., & Currim, I. S. (2002). Hierarchical Bayes versus finite mixture conjoint analysis models: A comparison of fit, prediction, and partworth recovery. *Journal of Marketing Research, 39*, 87–98.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics, 49*, 803–821.

Böckenholt, U. (2002). Comparison and choice: Analyzing discrete preference data by latent class scaling models. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 163–182). Cambridge, UK: Cambridge University Press.

Clogg, C. C., & Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association, 79*, 762–771.

Croon, M. A. (2002). Ordering the classes. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 137–162). Cambridge, UK: Cambridge University Press.

Dayton, C. M. (1998). *Latent class scaling models*. Thousand Oaks, CA: Sage.

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association, 83*, 173–178.

Dillon, W. R., & Kumar, A. (1994). Latent structure and other mixture models in marketing: An integrative survey and overview. In R. P. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 352–388). Cambridge, UK: Blackwell.

Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions*. London: Chapman & Hall.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association, 87*, 476–486.

Fraley, C., & Raftery, A. E. (1998). *How many clusters? Which clustering method? Answers via model-based cluster analysis* (Tech. Rep. No. 239). Seattle: Department of Statistics, University of Washington.

Goodman, L. A. (1974a). The analysis of systems of qualitative variables when some of the variables are unobservable: Part I. A modified latent structure approach. *American Journal of Sociology, 79*, 1179–1259.

Goodman, L. A. (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215–231.

Haberman, S. J. (1979). *Analysis of qualitative data: Vol. 2. New developments.* New York: Academic Press.

Hagenaars, J. A. (1988). Latent structure models with direct effects between indicators: Local dependence models. *Sociological Methods & Research, 16*, 379–405.

Hagenaars, J. A. (1990). *Categorical longitudinal data–loglinear analysis of panel, trend and cohort data.* Newbury Park, CA: Sage.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences.* Thousand Oaks, CA: Sage.

Hunt, L., & Jorgensen, M. (1999). Mixture model clustering using the MULTIMIX program. *Australian and New Zealand Journal of Statistics, 41*, 153–172.

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science, 16*, 39–59.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Langeheine, R., Pannekoek, J., & Van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research, 24*, 492–516.

Langeheine, R., & Van de Pol, F. (1994). Discrete-time mixed Markov latent class models. In A. Dale & R. B. Davies (Eds.), *Analyzing social and political change: A casebook of methods* (pp. 171–197). London: Sage.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots and related graphical displays. *Sociological Methodology, 31*, 223–264.

Magidson, J., & Vermunt, J. K. (2002). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research, 20*, 37–44.

McCutcheon, A. L. (1987). *Latent class analysis.* Newbury Park, CA: Sage.

McGrath, K., & Waterton, J. (1986). *British social attitudes, 1983–1986 panel survey* (Technical report). London: Social and Community Planning Research.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models.* New York: John Wiley.

Raftery, A. E. (1986). Choosing models for cross-classifications. *American Sociological Review, 51*, 145–146.

Rindskopf, R., & Rindskopf, W. (1986). The value of latent class analysis in medical diagnosis. *Statistics in Medicine, 5*, 21–27.

Uebersax, J. S. (1999). Probit latent class analysis with dichotomous or ordered category measures: Conditional independence/dependence models. *Applied Psychological Measurement, 23*, 283–297.

Uebersax, J. S., & Grove, W. M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine, 9*, 559–572.

Van der Heijden, P. G. M., Gilula, Z., & Van der Ark, L. A. (1999). On a relationship between joint correspondence analysis and latent class analysis. *Sociological Methodology, 29*, 147–186.

Vermunt, J. K. (1997). *Log-linear models for event histories.* Thousand Oaks, CA: Sage.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge, UK: Cambridge University Press.

Vermunt, J. K., & Magidson, J. (2003). *Latent GOLD 3.0 user's guide*. Belmont, MA: Statistical Innovations, Inc.

Vermunt, J. K., & Van Dijk, L. (2001). A nonparametric random-coefficients approach: The latent class regression model. *Multilevel Modelling Newsletter, 13*, 6–13.

Wedel, M., & DeSarbo, W. S. (1994). A review of recent developments in latent class regression models. In R. P. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 352–388). Cambridge, UK: Blackwell.

Wedel, M., & Kamakura, W. A. (1998). *Market segmentation: Concepts and methodological foundations*. Boston: Kluwer Academic.

Yung, Y. F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika, 62*, 297–330.

Chapter 11

Discrete-Time Survival Analysis

John B. Willett

Judith D. Singer

An important class of research questions asks whether and, if so, when a variety of events occur (Singer & Willett, 1991; Willett & Singer, 1991). Researchers investigating the consequences of childhood traumas on later well-being, for instance, ask *whether* an individual ever experiences depression and, if so, *when* onset first occurs (Wheaton, Roszell, & Hall, 1997). Other researchers ask questions about whether and when street children return to their homes (Hagan & McCarthy, 1997), whether and when college students drop out of school (DesJardins, Ahlburg, & McCall, 1999), whether and when recently married couples get divorced (South, 2001), and whether and when adolescent boys first have sexual intercourse (Capaldi, Crosby, & Stoolmiller, 1996).

Familiar statistical techniques, such as regression and analysis of variance, and their more sophisticated cousins, such as structural equation modeling, are ill-suited for addressing questions about the timing and occurrence of events. These usually versatile methods fail because they are unable to handle situations in which the value of the outcome—whether and when the event occurs—is unknown for some people under study. When studying event occurrence, this type of information shortfall is inevitable. No matter how long a researcher collects data, some people in the sample will not experience the target event while he or she watches—some adults will not have a depressive episode, some street children will not return to their homes, some college students will not drop out of school, some recently married couples will not divorce, and some boys will remain virgins. Statisticians say that such observations are *censored.*

Censoring creates an analytic dilemma. Although the researcher knows something about individuals with censored event times—if they ever experience the event, they will do so *after* data collection ends—this knowledge is imprecise. If an adolescent boy does not have sexual intercourse by 12th grade, for example, we would not want to conclude that he will *never* do so. All we can say is that by the end of 12th grade, the individual was still a virgin. Yet the need to analyze simultaneously the data from individuals with censored and noncensored event times is apparent because the former are a key group of people—those *least likely* to experience the event.

Sound investigation of event occurrence requires an analytic method that deals consistently and evenhandedly with noncensored and censored

AUTHORS' NOTE: The order of the authors was determined by randomization.

observations. Biostatisticians modeling human lifetimes (time to death) initially developed a class of appropriate methods because they were faced with a related problem, in which some of the individuals in their studies (thankfully) did not die by the end of data collection (Cox, 1972; Kalbfleisch & Prentice, 2002). Despite the foreboding appellations of these techniques—known variously as survival analysis, event history analysis, and hazard modeling—these tools are invaluable for social scientists because they provide a sound mathematical basis for exploring the "whether" and "when" of *any* type of event.

In this chapter, we provide a conceptual introduction to survival methods, focusing specifically on the principles of discrete-time survival analysis. After distinguishing between discrete-time and continuous-time survival methods and describing why we encourage first-time learners to begin with the former approach, we use data describing the age of first onset of depression to introduce the fundamental building blocks of the methods—the hazard and survivor functions. We then describe the statistical models that can be used to link the pattern of temporal risk to predictors, commenting on the types of predictors that can be included in these models and how to interpret the results of statistical modeling. Finally, we show how researchers can be misled if they use traditional analytic techniques instead of survival methods. Our presentation here is nontechnical and conceptual. Readers seeking practical information and data-analytic advice should consult Singer and Willett (2003) before using survival analysis in their research.

11.1. How Do You Measure Time and Record Event Occurrence?

To study event occurrence and its predictors, a researcher must record how long it takes, from some common starting time, for each individual in a sample to experience the target event. Researchers have a great deal of flexibility in identifying the "beginning of time." Because birth is both handy and meaningful across a wide variety of contexts, most researchers choose to use it as the "beginning of time," using an individual's *age* (time since birth) as the marker of when the event occurred (see, e.g., Wheaton et al., 1997). But researchers need not restrict themselves to the metric of chronological age. Another common way of setting the beginning of time is to tie it to the occurrence of a precipitating event—one that places all individuals in the population *at risk* of experiencing the target event. When modeling street children's return to a parental home, for example, the "beginning of time" may be defined as the time when the child first left the parental home (making "time on the street" the metric for analysis).

Once a common start time is defined, the researcher follows individuals (either prospectively on a periodic basis or through retrospective event history reconstruction) to record whether and, if so, when the target event occurs. All individuals who experience the target event during data collection are assigned event times equal to the value of time when they actually experience the event. Individuals who do not experience the target event during data collection are assigned *censored event times,* set equal to the value of time when data collection ended or when the individual was no longer at risk of experiencing the event. This censored event time, although seemingly imprecise, tells us a great deal about event occurrence: It tells us that the individual *did not experience the target event at any earlier time.*

Some researchers can record event occurrence data very precisely. When studying the relationship between experiences of childhood adversity and death, for example, Friedman, Tucker, Schwartz, and Tomlinson-Keasey (1995) used public records of vital statistics to determine the precise time (year, month, and even day) when each individual who had died had actually passed away. Other researchers can record only that the target event occurred within some finite time *interval.* A researcher might know, for example, the *year* when a person first experienced depressive symptoms, the *month* when an individual began a new job, or the *grade* when a youngster transitioned from adult-supervised care to self-care. We distinguish between these two scales of measurement (very precise and somewhat coarser) by calling the former *continuous-time* data and the latter *discrete-time* data.

In this chapter, we focus on statistical methods for analyzing data recorded in discrete time (Singer & Willett, 1993; Willett & Singer, 1993). We have six reasons for this emphasis. First, we have found that discrete-time methods are intuitively more comprehensible than their continuous-time cousins, facilitating initial mastery and later transition to continuous-time methods (if required). Second, we believe that these methods are very appropriate for much of the event history data collected by social scientists because, for logistical and financial reasons, data are often recorded only in terms of intervals (see Lin, Ensel, & Lai, 1997). Third, this approach facilitates inclusion of both *time-invariant* and *time-varying* predictors, whereas inclusion of the latter is more difficult under the continuous-time approach. Thus, with

discrete-time models, researchers can easily examine the effects of predictors whose values fluctuate naturally over the life course such as family structure and employment status. Fourth, discrete-time survival analysis fosters inspection of how the pattern of risk shapes up over time. The most popular continuous-time survival analysis strategy ("Cox regression"; Cox, 1972) ignores the shape of the temporal risk profile entirely in favor of estimating the influence of predictors on risk, under a restrictive assumption of "proportionality." Fifth, under the discrete-time approach, the proportionality assumption is easily checked and "non-proportional" models fitted. Finally, in discrete-time survival analysis, all estimation can be conducted using standard statistical software packages that fit logistic regression models. This avoids reliance on the dedicated computer software required for continuous-time survival analyses.

11.2. Describing Survival Data

The *hazard function* and the *survivor function* are the two fundamental tools for describing the occurrence and timing of events. Estimates of these functions provide answers to the two key descriptive questions: "When is the target event most likely to occur?" and "How much time passes before people are likely to experience the event?"

11.2.1. The Hazard Function

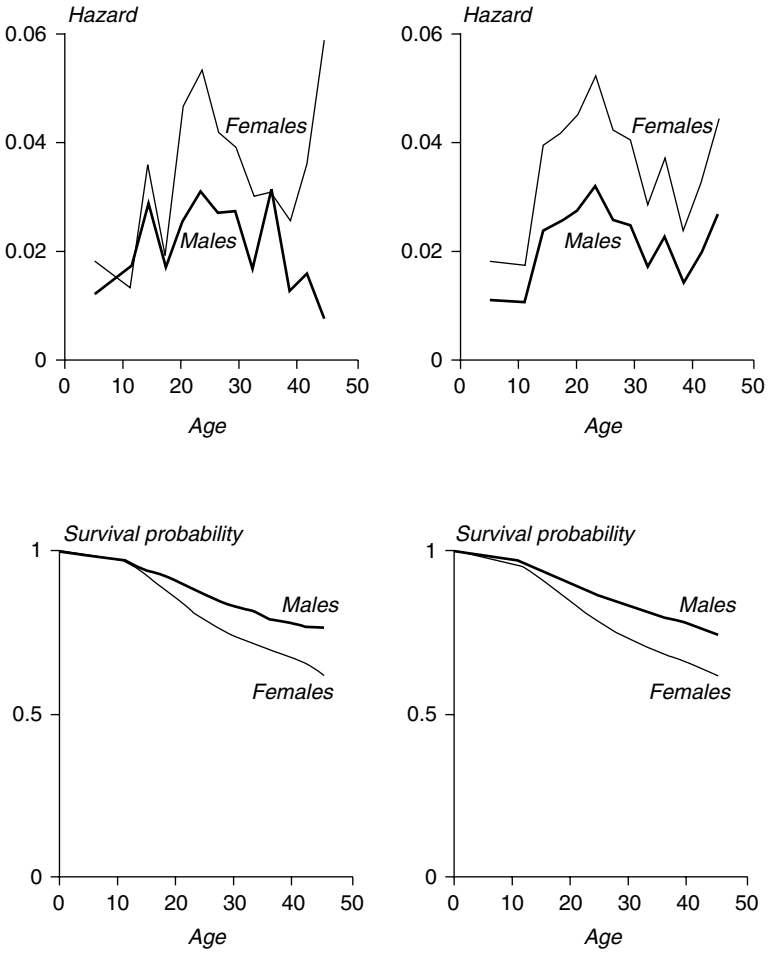
When examining the occurrence of an event—such as "experiencing an initial episode of depression"—for a random sample of individuals, we begin by asking about the pattern of event occurrence over time. We might ask, for example, the following: When are individuals at greatest risk of first experiencing a depressive episode—during childhood, during their teens, or during their 20s, 30s, or 40s? When we pose such questions, we are implicitly asking about the "risk" of event occurrence across time periods. Knowing how the risk of experiencing a depressive episode fluctuates over time provides answers to questions about the "whether" and "when" of event occurrence.

How can we summarize the risk of event occurrence among individuals in a sample, especially if some of these people have censored event times—that is, by the end of data collection, they had never been clinically depressed? In discrete-time survival analysis, the fundamental quantity that represents the risk of event occurrence in each time period is called the *hazard probability.* Its computation in the sample is straightforward: In each time period, identify the pool of people still "at risk" of experiencing the event (those who have reached this time period without experiencing the event, the so-called "risk set") and compute the proportion of this group that experiences the event during the time period. Notice that this definition is inherently conditional; once someone experiences the event (or is censored) in one time period, he or she no longer is a member of the risk set in a future time period. The plot of the set of hazard probabilities against time yields the *hazard function,* a chronological summary of the risk of event occurrence.

In the top panel of the left-hand side of Figure 11.1, we present an illustrative hazard function using retrospective data gathered from a probability sample of 1,393 adults in metropolitan Toronto who were asked whether and, if so, when they first experienced a depressive episode (for a complete description of these data, see Wheaton et al., 1997). The panel presents two sample hazard functions, computed separately for men and women, describing the "risk" of initially experiencing a depressive episode in each of 13 successive time periods—age 9 or younger, 10 to 12, 13 to 15, 16 to 18, and so on in 3-year increments until the time intervals 40 to 42 and age 43 and older. Inspection of the sample hazard function helps pinpoint when events are most likely and least likely to occur. Examining these two hazard functions, we see that for both males and females, the risk of experiencing an initial episode of depression is relatively low in childhood, increases during adolescence, and then peaks in the early 20s. After this point in time, the risk of initial onset of depression, *among those individuals who have not yet had a depressive episode,* is much lower, and by the early 40s, it declines back to preadolescent levels for men, although it rises again for women. Beyond this overall pattern of risk, also notice that in all but two time periods, there is a sex differential: In general, women are at greater risk of experiencing a depressive episode than are men.

The "conditionality" inherent in the definition of *hazard* is critical. It ensures that all individuals remain in the risk set until the last time period when they are eligible to experience the event (at which point, they are either censored by the end of data collection or experience the target event). For example, the hazard probability for initial onset of depression during the age period 31 to 33 years is estimated conditionally using the data from all those individuals (852 of the initial sample of 1,393) who were at least age 31 when data were collected but who *had not yet had a depressive episode during any earlier time period.* Individuals

Figure 11.1 Hazard and Survivor Functions Describing Age at First Onset of Depression for 1,393 Adults in Toronto, by Gender



NOTE: The left panel presents sample functions; the right panel presents fitted functions.

who were not yet in their early 30s ($n = 227$) or who had already experienced a depressive episode ($n = 314$) are no longer "at risk" and are therefore excluded from the calculation of hazard in this time period and all subsequent time periods. This conditionality is crucial, for it ensures that the sample hazard probability deals evenhandedly with censoring—using *all* the information available in the sample event histories but not overextending this knowledge beyond the time when the researcher has data.

11.2.2. The Survivor Function

In addition to using the hazard function to explore the conditional risk of event occurrence in each time period, it is useful to cumulate these period-by-period risks to display the proportion of the sample that "survives" through each time period—that *does not experience the event.* The term *survival probability* refers to this proportion, and the term *survivor function* refers to plots arraying the survival probabilities against time. Sample survivor functions summarize aggregate event histories. They are easily computed by cumulating the entries in the sample hazard function over time (see Singer & Willett, 2003).

In the bottom panel of the left-hand side of Figure 11.1, we display the sample survivor functions for men and women corresponding to the sample hazard functions displayed in the top panel. These survivor functions indicate the proportion of adults who "survived"—*did not experience an initial*

depressive episode—through each successive time period—ages 1 to 9, 10 to 12, 13 to 15, and so on. Notice that the curves remain high at the beginning of time and then drop more sharply as time passes. At birth, all individuals are "surviving"—none of them has experienced a depressive episode—and so the survival probabilities are 1.00. Over time, as individuals experience depressive episodes, the survivor functions drop. Because most adults do not experience a depressive episode *at any time* in their lives, the curves do not reach zero, ending in this sample at .77 for men and .62 for women. These proportions indicate that by the end of their late 50s, an estimated 77% of men and 62% of women had not yet experienced a depressive disorder. By subtraction, we estimate that 23% of men and 38% of women *have* experienced a depressive episode *at some point before their 60s.*

All sample survivor functions have a similar shape—a monotonically nonincreasing function of time. The rate of decline, however, can differ across groups. For example, although the two sample survivor functions in Figure 11.1 have similar shapes, the sharper decline among women suggests that, in comparison to men, they are at greater risk of experiencing a depressive episode.

11.3. Detecting Predictors of Event Occurrence Using a Discrete-Time Hazard Model

Estimated hazard functions and survivor functions describe when (and whether) a group of individuals is likely to experience a target event. These descriptive statistics can also be used to answer questions about differences between groups. Are maltreated children more likely than nonmaltreated children to repeat a grade in school (Rowe & Eckenrode, 1999)? Are children of divorced parents more likely than children of intact families to experience a divorce themselves? Are individuals from larger families less likely to experience a depressive episode than individuals from smaller families?

Each of these examples implicitly uses individual characteristics—child maltreatment, parental divorce, and family size—to predict the risk of event occurrence. When we examine the pair of sample hazard and survivor functions displayed in the left-hand side of Figure 11.1, we, too, are implicitly treating gender as a predictor of age at first onset of depression. But implicit comparisons such as these are limited. How can we examine the effects of continuous predictors using such plots? How can we examine the effects of several predictors simultaneously or explore statistical interactions among predictors? How can we make inferences about the population from which the sample was drawn? With survival analysis, we achieve these goals by postulating and fitting statistical models of the hazard function and by conducting hypothesis tests about the values of population parameters in these models.

Statistical models of hazard express hypothesized population relationships between entire hazard profiles and predictors. To motivate our representation of these models, examine the two sample hazard functions in the top panel of the left side of Figure 11.1 and imagine that we have created a dummy variable, FEMALE, which takes on two values (0 for males, 1 for females). In this formulation, we are making the entire hazard function the conceptual "outcome" and the dummy variable FEMALE the potential "predictor."

What is the relationship between the predictor and the outcome? Ignoring differences in the shapes of the profiles for the moment, when FEMALE $= 1$, the sample hazard function is generally "higher" relative to its location when FEMALE $= 0$, indicating that in virtually every time period, women are more likely to experience an initial depressive episode. So conceptually, at least, the effect of the predictor FEMALE is to "shift" one sample hazard profile vertically relative to the other. A population hazard model formalizes this conceptualization by ascribing the vertical displacement in hazard profiles to variation in predictors in much the same way as an ordinary linear regression model ascribes differences in mean levels of a continuous noncensored outcome to variation in predictors.

The difference between a hazard model and a linear regression model, of course, is that the entire hazard profile is no ordinary continuous outcome. The discrete-time hazard profile is a set of conditional probabilities, each bounded by 0 and 1. Statisticians modeling a bounded outcome as a function of predictors generally do not use a linear function to express this relationship but rather use *a nonlinear link function* that has the net effect of transforming the outcome so that it is unbounded. This prevents derivation of fitted values that fall outside the range of permissible values—in this case, between 0 and 1. When the outcome is a probability, the *logit* link function is especially popular (Hosmer & Lemeshow, 2000). If p represents a probability, then *logit* (p) is the natural logarithm ($\log_e$) of $p/(1-p)$ and, in the case of these data, can be interpreted as *the log-odds of initial onset of depression.*

Letting $h(t)$ represent the entire population hazard profile, then, a statistical model that relates the logit transform of $h(t)$ to the predictor FEMALE is

$$\text{logit } h(t) = \beta_0(t) + \beta_1 \text{FEMALE}. \qquad (1)$$

The parameter $\beta_0(t)$ is the *baseline logit-hazard profile*. It represents the value of the outcome (the entire logit-hazard profile) when the value of the predictor FEMALE is 0 (i.e., it specifies the profile for men). We write the baseline as $\beta_0(t)$, a function of time, and not as β_0, a single term unrelated to time (as in regression analysis), because the outcome (logit $h(t)$) is an entire temporal profile. The discrete-time hazard model in (1) specifies that differences in the value of the predictor "shift" the baseline logit-hazard profile up or down. The "slope" parameter β_1 captures the magnitude of this shift; it represents the vertical shift in logit-hazard associated with a one-unit difference in the predictor. Because the predictor here is a dichotomy, FEMALE, β_1 captures the differential risk of onset (measured in the logit-hazard scale) for women in comparison to men.

Discussion of methods for estimating the parameters of discrete-time hazard models, evaluating goodness of fit, and drawing inferences about the population is beyond the scope of this chapter. All of these goals are easily achieved using standard software for fitting logistic regression models (see Singer & Willett, 2003). Without delving into details, suffice it to say that once a discrete-time hazard model has been fit, its parameters can be reported along with standard errors and goodness-of-fit statistics in much the same way that the results of familiar regression analyses are reported. And just as fitted lines can be used to illustrate the influence of important predictors in the context of multiple regression, so, too, can fitted hazard functions (and survivor functions) be displayed for prototypical people—those who share substantively important values of statistically significant predictors.

We illustrate the results of this estimation process in the right-hand panel of Figure 11.1, which presents *fitted* hazard and survivor functions for the model presented in (1). Comparing the right and left panels, notice that the *fitted* plots on the right side are far smoother without the crossing and zigzagging characteristic of the *sample* plots on the left side. This smoothness results from the constraints inherent in the population hazard model stipulated in (1), which forces the vertical separation between the two hazard functions to be identical (in logit-hazard scale) in every time period. Just as we do not expect a fitted regression line to touch every data point in a scatter plot, we do not expect a fitted hazard function in survival analysis to match every sample value of hazard. Indeed, analyses using procedures described in our companion papers reveal that the discrepancies between the sample and fitted plots presented in Figure 11.1 can be ascribed to nothing more than sampling variation.

What have we learned by fitting this statistical model to these data? First, we can see the more clearly articulated profile of risk across time that is revealed by pooling information across individuals and asking questions about the population that gave rise to these sample data. Here, this reveals a clear pattern of risk resembling that found by many researchers studying the initial onset of depressive disorders (e.g., Sorenson, Rutter, & Aneshensel, 1991): The risk of onset is relatively low in childhood, rises steadily through adolescence, and reaches a peak in the early 20s, at which point it declines, falling not back to zero but to moderate levels that never quite reach the peak risks of early adulthood.

Second, we can quantify the increased risk of initially becoming depressed among women in comparison to men, and we can conduct a hypothesis test of whether this gender differential may be a result of sampling variation. Our analyses yield a parameter estimate for β_1 of 0.52, which indicates that the vertical separation in the *logit-hazard scale* between the profiles of risk for men and women is 0.52. Conducting the appropriate hypothesis test (described elsewhere), we obtain a chi-square test statistic of 23.20 on 1 degree of freedom ($p < .0001$), indicating that we may reject the null hypothesis that the predictor FEMALE has no effect on the population hazard profile (i.e., we reject the null hypothesis H_0: $\beta_1 = 0$). Because few researchers have an intuitive understanding of the *logit-hazard scale,* we recommend using the same data-analytic practice used when fitting ordinary logistic regression models: *Antilog* the coefficient and interpret it in terms of *odds* and *odds ratios* (Hosmer & Lemeshow, 2000). Antilogging .52 (i.e., taking $e^{.52}$), we conclude that the estimated odds of experiencing a depressive episode in any given time period are 1.67 times higher for women in comparison to men.

The fitting of discrete-time hazard models provides a flexible approach to investigating predictors of event occurrence that appropriately includes data from both censored and noncensored individuals. Although hazard models may appear unusual, they actually resemble familiar multiple linear and logistic regression models. Like these familiar models, hazard models can incorporate several predictors simultaneously, simply through the inclusion of additional predictors. Inclusion of multiple predictors permits examination of the effect of one predictor while

controlling statistically for the effects of others. Similarly, we can examine the synergistic effect of several variables by including statistical interactions between predictors.

Rather than describe the *similarities* between hazard models and familiar regression models (for these are presented extensively elsewhere), let us turn now to the *unique* analytic possibilities offered by hazard models—possibilities unavailable with standard statistical methods. We do so because we believe that it is the *unique* features of hazard models (such as the ability to investigate time-varying effects) that make them so exciting for the empirical researcher.

11.4. What If the Values of Predictors Vary Over Time? Including Time-Varying Predictors

Hazard models can include two very different types of predictors: those that are time invariant and those that are time varying. As befits their label, the former describe immutable characteristics of people, such as their sex or race, whose values are stable across the lifetime, whereas the latter describe characteristics of people that may fluctuate with time, as might an individual's self-esteem, marital status, or income. For clarity, when writing statistical models that include time-varying predictors, we include a parenthetical t in the variable name to distinguish such predictors from their time-invariant cousins.

There are at least two reasons why we believe that the ability to include time-varying predictors represents an especially exciting analytic opportunity for researchers studying the predictors and consequences of events across the life course. First, researchers often find themselves studying behavior across extended periods of time, sometimes encompassing more than 20, 30, or even 40 years. Although researchers studying behavior across short periods of time may reasonably argue that the values of time-varying predictors will be relatively stable during the study period (enabling them to use time-invariant indicators of these time-varying features), the tenability of this assumption decreases as the length of time studied increases. Second, many research questions focus on the *links between the occurrence of several* different *events.* Researchers ask about whether the occurrence of one stressful event (e.g., parental divorce or death of a spouse) predicts the occurrence of another stressful event (e.g., one's own divorce or the onset of depression). Although it is possible to address such questions by comparing the trajectories of individuals who have had and who have not had the precipitating event *at any time during the interval covered by data collection,* this approach requires the researcher to set aside data on all individuals who experienced the precipitating event *during* the period of data collection. By coding the precipitating event using a time-varying predictor, data from *all individuals* may be analyzed simultaneously.

We illustrate the use of a time-varying predictor by considering the dummy variable PARDIV(t), which indicates whether the individual's parents had divorced by time t (0 = not yet divorced; 1 = divorced). We could investigate the effects of adding this time-varying predictor to model (1) by fitting the following model:

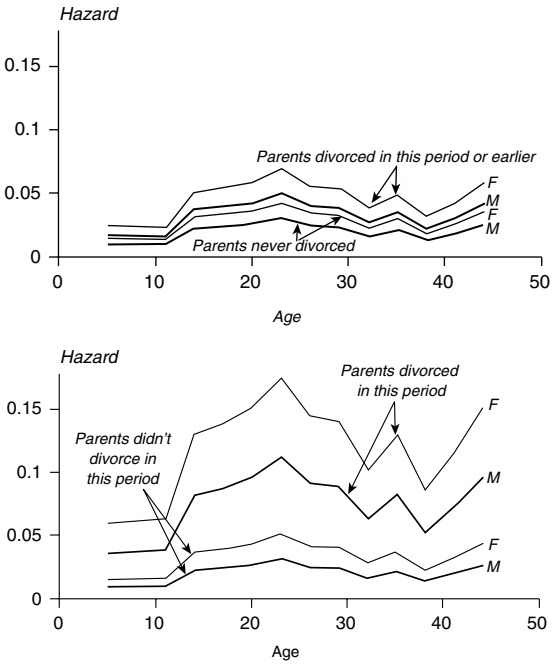
$$\text{logit } h(t) = \beta_0(t) + \beta_1 \text{FEMALE} + \beta_2 \text{PARDIV}(t). \quad (2)$$

This model allows the *values* of the dummy variable PARDIV(t) to vary over time (beginning at 0 among intact families and switching to 1 if, and when, the individual's parents divorce). However, it also stipulates that the *effect* of parental divorce on the risk of onset is constant over time, represented by the single parameter β_2. If β_2 is positive, individuals whose parents divorced are more likely to develop depressive symptoms (*after* the divorce occurs); if it is negative, they are less likely; and if it is zero, parental divorce has *no effect* on risk.

The top panel of Figure 11.2 presents the results of fitting the population discrete-time hazard model postulated in (2) to these sample data. We present the results of this "main effects model" because analyses (not presented here) confirmed that there was no statistical interaction between these predictors—in other words, the effect of parental divorce on risk was identical for men and women. Comparison of the four fitted hazard functions clearly illustrates the large and statistically significant effects of the two predictors: Women are at greater risk of experiencing an initial depression onset as are individuals whose parents divorced.

Because PARDIV(t) is a time-varying predictor, however, these fitted plots cannot be interpreted in the same way as the fitted plots presented in Figure 11.1. To learn how to interpret these plots, focus first on the bottom fitted hazard profile (in the top half of the figure), which depicts the risk of experiencing a depressive episode among men whose parents *never divorced.* This is the lowest of the four fitted hazard profiles because this group of individuals is at lowest risk of experiencing a depressive disorder. Now consider the profile that would result if a boy's (or man's)

Figure 11.2 Fitted Hazard Functions Describing Age at First Onset of Depression, by Gender, for Children Whose Parents *Had* and *Had Not* Divorced



NOTE: In the top panel, divorce effects are coded so that they *persist* throughout an individual's lifetime; in the bottom panel, divorce effects are coded so that they are *interval specific.*

parents divorce. While the parents were married, the boy's risk profile would still be represented by the lowest of the four hazard functions. When they divorce, however, the latter portion of *this boy's risk profile* (the portion occurring *after the divorce*) would be described by the *other* fitted hazard profile for males, which is substantially higher, capturing the increased risk of initial depression onset among males whose parents had divorced. In essence, then, the fitted hazard profiles presented in the top panel of Figure 11.2 provide an *envelope* of all possible hazard profiles corresponding to the many different possible times when parents may divorce. Individuals whose parents are not divorced remain on the lower profile (for their gender); if and when their parents' divorce, their risk of initial onset rises to the level represented by the higher hazard profile for their gender.

Another analytic opportunity made possible through hazard modeling is the option of exploring different ways of parameterizing the effects of time-varying predictors. In the model we have just fit for parental divorce, we have assumed that the effect of parental divorce on the risk of depression remains with a person throughout his or her lifetime. But consider an alternative possibility: Parental divorce may increase an individual's risk of depression, *but only during the time period when the parental divorce occurs.* Letting the dummy variable DIVNOW(t) indicate whether the individual's parents had divorced *at time t* (0 = not divorced in this time period; 1 = divorced in this time period), we could investigate the effects of this time-varying predictor by fitting the following model:

$$\text{logit } h(t) = \beta_0(t) + \beta_1\text{FEMALE} + \beta_2\text{DIVNOW}(t). \quad (3)$$

As with the model postulated in (2), the *values* of the dummy variable DIVNOW(t) may vary over time (being 0 among individuals whose parents did not divorce during this time period and being 1 if they do). So, too, we continue to hypothesize that the *effect* of parental divorce on the risk of initial onset of depression is constant over time and is represented by the single parameter β_2. The difference between the two models is that once the variable PARDIV(t) takes on the value 1 for an individual, it *stays* at the value 1 for the remainder of that individual's record; for the variable DIVNOW(t), in contrast, it would take on the value of 1 *only during the time period when the individual's parents actually divorced.* Thus, model (3) postulates that the effects of parental divorce are *interval specific.* If β_2 is positive, then individuals whose parents *divorced in this interval* are at greater risk of *depression in this interval.* This model does not allow the effects of parental divorce to carry over into any interval *after* the divorce occurs.

The bottom panel of Figure 11.2 presents the results of fitting this alternative model to these data. We have plotted these fitted functions on a scale identical to that used in the top panel so that the differential effect associated with parental divorce in the two models is apparent. Once again, begin with the bottom hazard profile, which represents the risk of onset of depression among males whose parents *did not divorce during the time period in question.* Our model specifies that a male whose parents remained married would have this profile of risk over time. If and when that male's parents divorce, however, his risk of onset during that time period would skyrocket, jumping up to the upper hazard profile for men. The difference between this model and the previous model, however, is that here, *after the time period in question,* the male's risk profile would return to the lower level represented by the bottom hazard function.

Why are these models so different? Both confirm that the effect of parental divorce is statistically significant, but the *magnitude* of the effect differs because they code the parental divorce variable in dramatically different ways. The first model allows the effect of parental divorce to persist throughout a person's lifetime. It yields an estimated coefficient of 0.34, which indicates that the odds that children of divorced parents become depressed are $e^{0.34} = 1.41$ times higher than the corresponding odds for children of nondivorced parents. The second model, in contrast, stipulates that the effect of parental divorce "kicks in" *only during the time period when the divorce occurred.* It yields a much larger coefficient (1.36), which implies that the effect of parental divorce on the risk of depression is much higher *at that particular time period.* Antilogging (exponentiating) this coefficient, we find that the odds of depression among children of divorced parents are 3.88 times higher at that point in time, but that afterwards, they revert back, and their risk profiles are indistinguishable from those whose parents had not been divorced. The first model essentially amortizes the dramatically elevated period-specific risk across an individual's postparental divorce life, whereas the latter focuses exclusively on what happens to an individual during the time period when his or her parents actually divorced.

The ease with which time-varying predictors can be incorporated into hazard models offers social scientists an innovative analytic opportunity. Many important predictors of trajectories and turning points fluctuate naturally with time: family and social structure, employment, opportunities for emotional fulfillment, and, perhaps most important, the occurrence and timing of other events. In traditional statistical analyses, temporal fluctuation in such predictors must be reduced to a single measure across time. With the advent of hazard modeling, this is no longer the case. Researchers can examine relationships between event occurrence and dynamically changing predictors.

11.5. What If the *Effects* of Predictors Vary Over Time? Including Interactions With Time

When processes evolve dynamically, the *effects* of both time-invariant and time-varying predictors may fluctuate over time. A predictor whose effect is constant over time has the same impact in all time periods. A predictor whose effect varies over time has a different impact on hazard in different time periods.

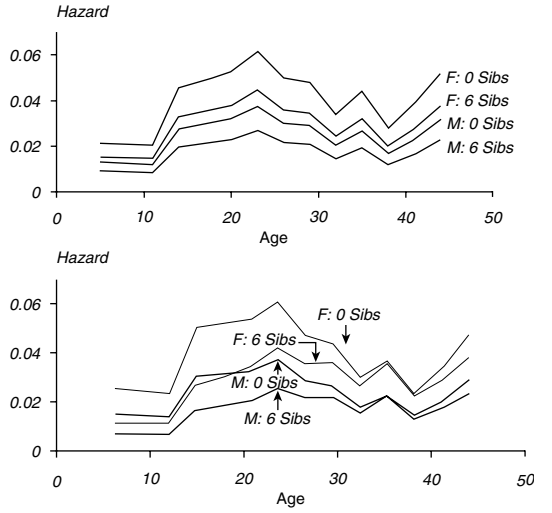
Both time-invariant and time-varying predictors can have time-varying effects. Consider the effects of parental divorce (as measured by the variable PARDIV(t)) on the risk of depression. PARDIV(t) is a time-varying predictor—its value goes from 0 to 1 if and when parents divorce—but its *effect* on hazard might be constant over time (as we have stipulated so far). If the effect is time invariant, this means that the effect of parental divorce on the risk of onset is the same regardless of whether the divorce takes place during childhood, adolescence, or adulthood. If the effect of parental divorce varies over time, in contrast, divorce might have a larger effect on the risk of depression among children who are still living at home than among adults who have already moved out of the house.

The discrete-time hazard models posited so far have not permitted a predictor's effect to vary with time; they are called *proportional-odds models.* Hazard profiles represented by such models have a special property: In every time period (t) under consideration, the effect of the predictor on the logit hazard is the same. In equation (1), for example, the vertical shift in the logit-hazard profile for women is always β_1, and consequently, the hypothesized logit-hazard profiles for women and men have identical *shapes* because their profiles are simply shifted versions of each other. Generally, in proportional-odds models, the entire family of logit-hazard profiles represented by all possible values of the predictors shares a common shape and is mutually parallel, differing only in its relative elevations. If the logit-hazard profiles are parallel and have the same shape, the corresponding *raw* hazard profiles are (approximate) magnifications and diminutions of each other—they are *proportional.*[1] Because the models presented so far include predictors with only time-constant effects, the fitted hazard functions displayed appear to have the required "proportionality."

But is it sensible to assume that the effects of all predictors are unilaterally time constant and that all hazard profiles are proportional in practice? In reality, many predictors will not only displace the logit-hazard profile but also alter its shape. If the effect of a predictor varies over time, we must specify a *nonproportional model* that allows the shapes of the logit-hazard profiles

1. For pedagogic reasons, we have taken some mathematical liberties here. In discrete-time models, the proportionality of the raw hazard profiles is only approximate because vertical shifts in logit hazard correspond to magnifications and diminutions of the untransformed hazard profile only when the magnitude of the hazard probability is small (say, less than .15 or .20). In much empirical research, as in the example we present here, discrete-time hazard is about this magnitude or less, and therefore the approximation tends to hold quite well in practice (see Singer & Willett, 2003, for further discussion of this issue).

Figure 11.3 Fitted Hazard Functions Describing the Age at First Onset of Depression, by Gender and the Individual's Number of Siblings, From Two Discrete-Time Hazard Models



NOTE: Top panel is a main effects model, in which the effect of number of siblings *is constant over time;* bottom panel is an interaction with time model, in which the effect of number of siblings *varies over time.*

to differ. When the effect of one predictor differs by the levels of another, we say that the two predictors *interact;* in this case, we say that the predictor *interacts with time.* To add such an effect into our hazard models, we include the cross-product of that predictor and time as an additional predictor.

Figure 11.3 illustrates the types of information that can be gleaned from determining whether a predictor interacts with time, presenting the results of fitting two discrete-time hazard models to the depression data using the time-invariant predictor NSIBS, which indicates the number of siblings for each respondent.[2] Because NSIBS is a continuous variable (its values vary from 0 to 26), we present fitted hazard profiles for two prototypical individuals: those who were only children (zero sibs) and those who came from larger families (six sibs). The figure presents fitted hazard profiles from two distinct models: a "main effects" model (top panel) and an "interaction with time" model (bottom panel). The main effects model suggests that siblings protect against depression: For both men and women, the greater the number of siblings, the lower the risk of onset. The four fitted hazard profiles appear proportional because the main effects model constrains the effect of NSIBS to be the same in each time period.

But a more accurate and complex story emerges from the "interaction with time" model displayed in the bottom panel of Figure 11.3, in which the effect of NSIBS is allowed to vary over time. Comparing the fitted hazard functions from the interaction with time model with those from the main effects model illustrates the untenability of the proportionality assumption due to the statistically significant interaction between NSIBS and time. The hazard functions in the bottom panel *are clearly not proportional.* In childhood, when individuals are still living at home, family size *does* have a protective effect: Boys and girls from larger families are at lower risk of having a depressive episode. Over time, however, the protective effect of family size diminishes, and by the time an individual reaches his or her early 30s, the effect is virtually nonexistent. Instead of having a constant vertical separation in logit-hazard space, the relative differences between the hazard functions differ, being larger in childhood and trivial in adulthood.

We believe that the ability to include and test the importance of interactions with time represents a major analytic opportunity for empirical researchers. When studying the behavior of individuals over very long periods of time, it seems reasonable to hypothesize that the effects of predictors will vary as people pass through different life stages. Although the effects of some predictors *will* remain with an individual throughout his or her lifetime, the effects of others may dissipate or increase over time. Our example of the changing effects of family size is but one of hundreds of reasonable possibilities that depression researchers might want to investigate. Were we to look for predictors of depression whose effects might *increase over time,* we might find that characteristics of the individual's own family in adulthood (say, number of children) might be an important predictor of depression in this phase of life.

We believe that it is not hyperbole to state that interactions with time are everywhere, if only researchers took the time to look for them. Present data-analytic practice (and the widespread availability of prepackaged computer programs) permits an almost unthinking (and often untested) adoption of proportional hazards models ("Cox" regression), in which the effects of predictors are constrained to be constant over time. Yet we have found, in a wide variety of

2. Because of data limitations, the values of this predictor are assumed to be constant during an individual's lifetime. If we had data indicating when the respondent's siblings were born, we could have coded this predictor as being time varying.

substantive applications, including not only our own work on employment duration (Murnane, Singer, & Willett, 1989; Singer, 1993a, 1993b) but also others' work on topics such as age at first suicide ideation (Bolger, Downey, Walker, & Steininger, 1989) and child mortality (Trussel & Hammerslough, 1983), that interactions with time seem to be the rule rather than the exception. We have every reason to believe that once researchers start looking for interactions with time, they will arise commonly. The key is to *test* the tenability of the assumption of a time-invariant effect. Although we have not outlined the statistical procedures for doing so here, we refer the interested reader to Singer and Willett (2003).

11.6. Is Survival Analysis Really Necessary?

In this chapter, we have introduced a class of statistical methods for analyzing longitudinal data on the occurrence and timing of events. Our presentation so far has encouraged researchers to learn more about these methods because they offer analytic capabilities that other methods do not. But now we turn to another reason for learning about survival methods: Failure to use them when appropriate can mislead a researcher alarmingly.

How can traditional methods for analyzing event occurrence deceive the investigator? The answer to this question depends on which traditional approach replaces the survival method and how that approach responds to the problem of censoring. Survival methods deal evenhandedly with censored cases—they contribute information to the analysis up until the time at which they are censored. Traditional analytic methods, in contrast, deal with censoring in an ad hoc way, which can create a series of problems we now describe.

One common way of "resolving" the censoring dilemma is to ignore the censored cases completely, treating them as if they were missing. Traditional statistical analyses can then be conducted in the subsample of noncensored individuals, with event time (or perhaps its logarithm) playing the role of the dependent variable. Descriptive statistics can be used to summarize subsample variability in event time. Correlation analysis, regression analysis, and analysis of variance can be used to investigate the relationship between event time and predictors. Unfortunately, this approach reduces statistical power (due to the smaller sample size) and leads to negatively biased estimates of aggregate event time (with a corresponding impact on estimates of the relationship between event time and predictors). When studying age at first divorce, for example, an investigator might be tempted to subsample only those individuals who had divorced by the end of data collection. But omitting individuals who remain married modifies the sample in an unfortunate way, reducing its size by eliminating the very individuals at lowest risk of divorce! After all, some of these individuals *will* divorce; they will just do so after data collection ends. The average time to divorce among the full population of "ever marrieds" must be longer than that found among the noncensored ("divorced") subsample analyzed.

An alternative to the "convert all censored cases into missing values" approach involves imputing their unknown event times and regressing its logarithm on predictors in the traditional fashion. The imputation allows the censored cases (for whom continuous duration data are unavailable) to be included in analyses with noncensored cases. The basic idea is well intentioned—there must be equivalent information available on both groups if they are to be included in the same traditional analysis. Although the approach maintains the sample at its original size (thereby apparently avoiding a loss of statistical power), it does not resolve the problem of bias. The censored event times are usually imputed arbitrarily or simply set equal to the length of data collection. Such a decision is not completely unreasonable because all the censored individuals did not experience the target event until that point in time. But many did not experience the target event for many more years to come. Full-sample summaries of event times based on such data necessarily underestimate the true length of time to event because, for an unknown proportion of the sample, the ultimate event times must be greater than the imputed value.

To avoid arbitrary data imputation while retaining both censored and noncensored cases, many investigators set aside the continuous event-time information (which is unknown for one of the subgroups) and focus on the categorical data that are known for both groups—data on whether each member of the sample experienced the target event by a particular point in time, usually the end of data collection. Dichotomization provides a new analytic outcome for which individuals who experienced the target event prior to the chosen cutoff are assigned a value of 1, and those who did not (the censored cases) are assigned the value 0. Descriptive statistics can summarize the proportion of cases experiencing the event prior to the chosen time, and logistic regression can be used to investigate the relationship between event occurrence and predictors.

Dichotomization can be viewed as the coarsest form of discrete-time survival analysis available. But its coarseness creates problems that can obscure knowledge about transitions. First, the approach destructs perfectly good continuous duration data to create the new dichotomous outcome. Consider what would happen, for example, if we followed a sample of individuals for 50 years and asked whether and, if so, when they first experienced a depressive episode. Dichotomization would eliminate known and potentially meaningful variation in event times by clustering together everyone who onset prior to the cutoff of age 50. Those whose initial depressive episodes occurred in early childhood would be pooled with those who did not become depressed until their late 40s, and yet such individuals undoubtedly differ enormously in the causes of their depression and in their ultimate prognosis.

A further problem is that any particular cutoff time—even one seemingly relevant to the process under study—is somewhat arbitrary. A researcher studying the predictors of obtaining employment after being laid off, for example, might follow a sample of people for 1 year to see whether they successfully secured a first job (as in Ginexi, Howe, & Caplan, 2000). But highly disparate temporal profiles of risk can lead to similar employment rates at a specific point in time. Just because individuals with high self-esteem and low self-esteem were equally likely to find jobs after 2 years does not mean that they got there by following similar trajectories. Perhaps most of the high self-esteem individuals obtained their jobs relatively quickly, soon after losing their jobs, whereas the low self-esteem individuals may have gotten their jobs only after months and months of searching. The 2-year cut point is convenient but not purposeful. By avoiding dichotomization and using survival analysis to disaggregate risk, we can better document variation in risk over time; by discovering what predicts variation in risk, we can better understand why some individuals find jobs early and others do not. Traditional methods disregard the temporal profile of risk; with survival methods, the risk profile becomes the primary analytic focus.

Disregard for the temporal variation in risk leads to yet another problem with the dichotomization approach; contradictory conclusions can result from nothing more than differences in the particular cutoff time adopted. Not only will the overall proportion of the sample experiencing the event differ as the cutoff is modified, but the relationship with predictors may also change. In our previous example, choosing cutoffs of 2 months, 1 year, and 2 years may lead to three entirely discrepant conclusions about the rate at which high school dropouts find jobs. The 2-month rates might erroneously indicate that *low* self-esteem individuals are more likely to find jobs (because they will take the first job that comes along), the 1-year rates might register no difference, and the 2-year rates might suggest that *high* self-esteem individuals are more likely to find jobs (because they persist and ultimately do secure employment). By using survival analysis to look at the relationship between hazard and self-esteem, the source of these cumulative differences in risk may be revealed as a statistically significant interaction between self-esteem and time. Researchers using traditional methods must constantly remind themselves that their conclusions can fluctuate as they modify their cutoff. Although such caveats usually appear in the "Results" section of an article, they often disappear by the "Discussion" section. In survival analysis, the time frame itself is an integral part of the answer; it highlights, rather than obscures, variation in risk over time.

The dichotomization "solution" is rendered further ineffective if censoring occurs at different times for different members of the sample. This occurs when sampled individuals are observed for different lengths of time, perhaps because of the research design (as when interviewing an age-heterogeneous sample and obtaining retrospective event history data) or because of the gradual onset of attrition (a common problem in longitudinal research). If censoring times differ across sample members, then both cutoff time and the opportunity for event occurrence differ as well. People followed for longer periods of time have greater opportunity to experience the target event than do those followed for shorter periods of time. So observed differences in cumulated risk might be attributable to nothing more than research design. Although it is possible to make risk periods equivalent across all sample members by discarding data describing behavior that occurred after the earliest possible censoring point for any member of the sample, this will eliminate large quantities of perfectly good data already collected. With survival analysis, a person who does not experience the event of interest is censored at the particular time that his or her data record ends; censoring times need not be identical for everyone under study.

Finally, traditional analytic methods offer few mechanisms for including predictors whose values vary over time or for permitting the effects of predictors to fluctuate over time. To overcome this limitation, researchers studying the effects of variables such as family functioning, socioeconomic status, or marital

status often use predictor values corresponding to a single point in time, the average of the several values over time, or perhaps a rate of change in values over time. Survival analysis makes this approach unnecessary. The analytic effort is identical whether including predictors that are static over time or predictors that change over time; so, too, it is easy to determine whether the effects of predictors are constant over time or whether they differ over time. Traditional methods force researchers to build static models of dynamic processes; survival methods allow researchers to model dynamic processes dynamically.

For all these reasons, we believe that empirical researchers should investigate the possibilities offered by survival methods. In the recent past, when these methods were in their infancy and statistical software was neither available nor user-friendly, researchers reasonably adopted other approaches. But these methods, originally developed to model an event seemingly beyond a person's control (i.e., death), lend themselves naturally to the study of individual behavior and development. The time has come for empirical researchers to fully exploit the utility of survival analysis. We are convinced that there is much that these methods can reveal.

REFERENCES

Bolger, N., Downey, G., Walker, E., & Steininger, P. (1989). The onset of suicide ideation in childhood and adolescence. *Journal of Youth and Adolescence, 18*, 175–189.

Capaldi, D. M., Crosby, L., & Stoolmiller, M. (1996). Predicting the timing of first sexual intercourse for at-risk adolescent males. *Child Development, 67*, 344–359.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B, 34*, 187–202.

DesJardins, S. L., Ahlburg, D. A., & McCall, B. P. (1999). An event history model of student departure. *Economics of Education Review, 18*, 375–390.

Friedman, H. S., Tucker, J. S., Schwartz, J. E., & Tomlinson-Keasey, C. (1995). Psychosocial and behavioral predictors of longevity: The aging and death of the "Termites." *American Psychologist, 50*, 69–78.

Ginexi, E. M., Howe, G. W., & Caplan, R. D. (2000). Depression and control beliefs in relation to reemployment: What are the directions of effect? *Journal of Occupational Health Psychology, 5*, 323–336.

Hagan, J., & McCarthy, B. (1997). Intergenerational sanction sequences and trajectories of street-crime amplification. In I. H. Gotlib & B. Wheaton (Eds.), *Stress and adversity over the life course: Trajectories and turning points* (pp. 212–232). New York: Cambridge University Press.

Hosmer, D. W., Jr., & Lemeshow, S. (2000). *Applied logistic regression.* New York: John Wiley.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). New York: John Wiley.

Lin, N., Ensel, W. M., & Lai, W. G. (1997). Construction and use of the life history calendar: Reliability and validity of recall data. In I. H. Gotlib & B. Wheaton (Eds.), *Stress and adversity over the life course: Trajectories and turning points* (pp. 343–354). New York: Cambridge University Press.

Murnane, R. J., Singer, J. D., & Willett, J. B. (1989). The influences of salaries and "opportunity costs" on teachers' career choices: Evidence from North Carolina. *Harvard Educational Review, 59*, 325–346.

Rowe, E., & Eckenrode, J. (1999). The timing of academic difficulties among maltreated and nonmaltreated children. *Child Abuse & Neglect, 23*(8), 813–832.

Singer, J. D. (1993a). Are special educators' career paths special? Results of a 13-year longitudinal study. *Exceptional Children, 59*, 262–279.

Singer, J. D. (1993b). Once is not enough: Special educators who return to teaching. *Exceptional Children, 60*, 58–73.

Singer, J. D., & Willett, J. B. (1991). Modeling the days of our lives: Using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events. *Psychological Bulletin, 110*, 268–298.

Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics, 18*, 155–195.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford University Press.

Sorenson, S. G., Rutter, C. M., & Aneshensel, C. S. (1991). Depression in the community: An investigation into age of onset. *Journal of Consulting and Clinical Psychology, 57*, 420–424.

South, S. J. (2001). Time-dependent effects of wives' employment on marital dissolution. *American Sociological Review, 66*, 226–245.

Trussel, J., & Hammerslough, C. (1983). A hazards-model analysis of the covariates of infant and child mortality in Sri Lanka. *Demography, 20*, 1–26.

Wheaton, B., Roszell, P., & Hall, K. (1997). The impact of twenty childhood and adult traumatic stressors on the risk of psychiatric disorder. In I. H. Gotlib & B. Wheaton (Eds.), *Stress and adversity over the life course: Trajectories and turning points.* New York: Cambridge University Press.

Willett, J. B., & Singer, J. D. (1991). From whether to when: New methods for studying student dropout and teacher attrition. *Review of Educational Research, 61*(4), 407–450.

Willett, J. B., & Singer, J. D. (1993). Investigating onset, cessation, relapse and recovery: Why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *Journal of Consulting and Clinical Psychology, 61*, 952–965.

Section IV

Models for Multilevel Data

Chapter 12

An Introduction to Growth Modeling

Donald Hedeker

12.1. Introduction

Longitudinal studies are increasingly common in social sciences research. In these studies, subjects are measured repeatedly across time, and interest often focuses on characterizing their growth across time. Traditional analysis of variance methods for such growth curve analysis are described in Bock (1975). However, these traditional methods are of limited use because of restrictive assumptions concerning missing data across time and the variance-covariance structure of the repeated measures. The univariate "mixed-model" analysis of variance assumes that the variances and covariances of the dependent variable across time are equal (i.e., compound symmetry). Alternatively, the multivariate analysis of variance for repeated measures only includes subjects with complete data across time. Also, these procedures focus on estimation of group trends across time and provide little help in understanding about how specific individuals change across time. For these and other reasons, hierarchical linear models (HLMs) (Bryk & Raudenbush, 1992) have become the method of choice for growth modeling of longitudinal data.

Variants of HLMs have been developed under a variety of names: random-effects models (Laird & Ware, 1982), variance component models (Dempster, Rubin, & Tsutakawa, 1981), multilevel models (Goldstein, 1995), two-stage models (Bock, 1989a), random-coefficient models (de Leeuw & Kreft, 1986), mixed models (Longford, 1987; Wolfinger, 1993), empirical Bayes models (Hui & Berger, 1983; Strenio, Weisberg, & Bryk, 1983), and random regression models (Bock, 1983a, 1983b; Gibbons, Hedeker, Waternaux, & Davis, 1988). A basic characteristic of these models is the inclusion of random subject effects into regression models to account for the influence of subjects on their repeated observations. These random subject effects thus describe each person's growth across time and explain the correlational structure of the longitudinal data. In addition, they indicate the degree of subject variation that exists in the population of subjects.

Several features make HLMs especially useful in longitudinal research. First, subjects are not assumed to be measured on the same number of time points; thus, subjects with incomplete data across time are included in the analysis. The ability to include subjects with incomplete data across time is an important advantage relative to procedures that require complete data across time because (a) by including all data, the analysis has increased statistical power, and

AUTHOR'S NOTE: The author thanks David Kaplan and Michael Seltzer for helpful and constructive comments on an earlier version of this chapter. Preparation of this chapter was supported by National Institutes of Mental Health (NIMH) Grant MH44826.

(b) complete case analysis may suffer from biases to the extent that subjects with complete data are not representative of the larger population of subjects. Because time is treated as a continuous variable in HLMs, subjects do not have to be measured at the same time points. This is useful for analysis of longitudinal studies in which follow-up times are not uniform across all subjects. Both time-invariant and time-varying covariates can be included in the model. Thus, changes in the outcome variable may be due to both stable characteristics of the subject (e.g., their gender or race) as well as characteristics that change across time (e.g., life events). Finally, whereas traditional approaches estimate average change (across time) in a population, HLMs can also estimate change for each subject. These estimates of individual change across time can be particularly useful in longitudinal studies in which a proportion of subjects exhibit change across time that deviates from the average trend.

As these methods have developed, several textbooks describing HLMs for longitudinal data analysis, to various degrees, have been published (Brown & Prescott, 1999; Bryk & Raudenbush, 1992; Davis, 2002; Diggle, Liang, & Zeger, 1994; Goldstein, 1995; Hand & Crowder, 1996; Hox, 2002; Longford, 1993; Raudenbush & Bryk, 2002; Singer & Willett, 2003; Verbeke & Molenberghs, 2000). Similarly, several collected editions are available (Bock, 1989b; Collins & Sayer, 2001; Leyland & Goldstein, 2001; Moskowitz & Hershberger, 2002) containing a variety of HLM developments. Also, review, comparison, and/or tutorial articles on longitudinal data analysis treating HLMs have proliferated (Albert, 1999; Burchinal, Bailey, & Snyder, 1994; Cnaan, Laird, & Slasor, 1997; Delucchi & Bostrom, 1999; Everitt, 1998; Gibbons et al., 1993; Gibbons & Hedeker, 2000; Keselman, Algina, Kowalchuk, & Wolfinger, 1999; Lesaffre, Asefa, & Verbeke, 1999; Manor & Kark, 1996; Omar, Wright, Turner, & Thompson, 1999; Sullivan, Dukes, & Losina, 1999). Most of these articles concern continuous response variables, although ones dealing specifically with categorical outcomes have also appeared (Agresti & Natarajan, 2001; Fitzmaurice, Laird, & Rotnitzky, 1993; Gibbons & Hedeker, 1994; Hedeker & Mermelstein, 1996, 2000; Pendergast et al., 1996; Zeger & Liang, 1992).

Applications of growth modeling are steadily increasing and can be found in many different fields, including studies on alcohol (Curran, Stice, & Chassin, 1997), smoking (Niaura et al., 2002), HIV/AIDS (Gallagher, Cottler, Compton, & Spitznagel, 1997), drug abuse (Carroll et al., 1994; Halikas, Crosby, Pearson, & Graves, 1997), psychiatry (Elkin et al., 1995; Serretti, Lattuada, Zanardi, Franchini, & Smeraldi, 2000), and child development (Campbell & Hedeker, 2001; Huttenlocher, Haight, Bryk, & Seltzer, 1991), to name a few. Not only do these articles illustrate the wide applicability of HLMs, but they also give a sense of how HLM results are typically reported in the various literatures. Thus, they can be very useful for investigators who are new to HLMs and their usage.

This chapter will focus on describing HLMs for continuous outcomes in a very practical way. We will first illustrate how HLMs can be seen as an extension of an ordinary linear regression model. Starting with a simple linear regression model, the model will slowly be extended and described to guide the reader going from familiar to less familiar territory. Following the descriptions of the statistical models, several HLM analyses will be presented using a longitudinal psychiatric data set. These analyses will illustrate many of the key features of HLMs for growth modeling. For further illustration, interested readers can download the data set and program files to replicate the analyses in this report from http://www.uic.edu/~hedeker/long.html.

12.2. HLMs for Longitudinal Data

To introduce HLMs, consider a simple linear regression model for the measurement y of individual i ($i = 1, 2, \ldots, N$ subjects) on occasion j ($j = 1, 2, \ldots, n_i$ occasions):

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij}. \tag{1}$$

Ignoring subscripts, this model represents the regression of the outcome variable y on the independent variable time (denoted t). The subscripts keep track of the particulars of the data—namely, whose observation it is (subscript i) and the relative order of the observation (the subscript j). The independent variable t gives a value to the level of time and may represent time in weeks, months, and so forth. Because y and t carry both i and j subscripts, both the outcome variable and the time variable are allowed to vary by individuals and occasions.

In linear regression models, such as (1), the errors ε_{ij} are assumed to be normally and *independently* distributed in the population with zero mean and common variance σ^2. This independence assumption makes the model given in equation (1) an unreasonable one for longitudinal data. This is because the outcomes y are observed repeatedly from the same individuals, and so it is much more likely to assume that errors within an individual are correlated to some degree. Furthermore,

the above model posits that the growth, or change across time, is the same for all individuals because the model parameters describing growth (β_0, the intercept or initial level, and β_1, the linear change across time) do not vary by individuals. For both of these reasons, it is useful to add individual-specific effects into the model that will account for the data dependency and describe differential growth for different individuals. This is precisely what HLMs do. Thus, a simple HLM is given by

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \upsilon_{0i} + \varepsilon_{ij}, \tag{2}$$

where υ_{0i} represents the influence of individual i on his or her repeated observations.

To better reflect how this model characterizes an individual's influence on his or her observations, we can represent the model in a hierarchical or multilevel form. For this, it is partitioned into the within-subjects (or Level 1) model,

$$y_{ij} = b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij}, \tag{3}$$

and the between-subjects (or Level 2) model,

$$\begin{aligned} b_{0i} &= \beta_0 + \upsilon_{0i}, \\ b_{1i} &= \beta_1. \end{aligned} \tag{4}$$

Here, the Level 1 model indicates that individual i's response at time j is influenced by his or her initial level b_{0i} and time trend, or slope, b_{1i}. The Level 2 model indicates that individual i's initial level is determined by the population initial level β_0, plus a unique contribution for that individual υ_{0i}. Thus, each individual has his or her own distinct initial level. Conversely, the present model indicates that each individual's slope is the same; all are equal to the population slope β_1. Another way to think about it is that each person's trend line is parallel to the population trend determined by β_0 and β_1. The difference between each individual's trend and the population trend is υ_{0i}, which is constant across time.

The between-subjects, or Level 2, model is sometimes referred to as a "slopes as outcomes" model (Burstein, 1980). The hierarchical representation shows that just as within-subjects (Level 1) covariates can be included in the model to explain variation in Level 1 outcomes (y_{ij}), between-subjects (Level 2) covariates can be included to explain variation in Level 2 outcomes (the subject's intercept b_{0i} and slope b_{1i}). Note that combining the within- and between–subjects models (3) and (4) yields the previous single-equation model (2).

Because individuals in a sample are typically thought to be representative of a larger population of individuals, the individual-specific effects υ_{0i} are treated as random effects. That is, υ_{0i} are considered to be representative of a distribution of individual effects in the population. The most common form for this population distribution is the normal distribution, with mean 0 and variance σ_υ^2. In the model given by equation (2), the errors ε_{ij} are now assumed to be normally and *conditionally independently* distributed in the population with zero mean and common variance σ^2. *Conditional independence* here means conditional on the random individual-specific effects υ_{0i}. Because the errors now have an influence due to individuals removed from them, this conditional independence assumption is much more reasonable than the ordinary independence assumption associated with (1). Because individuals deviate from the regression of y on t in a parallel manner (because there is only one subject effect υ_{0i}), this model is sometimes referred to as a random-intercepts model, with each υ_{0i} indicating how individual i deviates from the model. Figure 12.1 represents this model graphically.

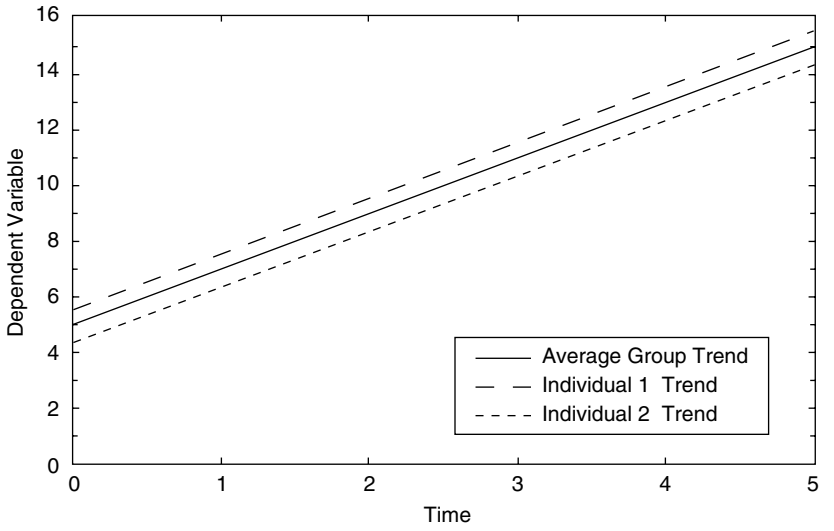
In this figure, the solid line represents the population average trend, which is based on β_0 and β_1. Also depicted are two individual trends, one below and one above the population (average) trend. For a given sample, there are N such lines, one for each individual. The variance term σ_υ^2 represents the spread of these lines. If σ_υ^2 is near zero, then the individual lines would not deviate much from the population trend. In this case, individuals do not exhibit much heterogeneity in growth. Alternatively, as individuals differ from the population trend, the lines move away from the population trend line and σ_υ^2 increases. In this case, there is more individual heterogeneity in growth.

For longitudinal data, the above random-intercepts model is often too simplistic for a number of reasons. First, it is unlikely that the rate of growth, or trend across time, is the same for all individuals. It is more likely that individuals differ in their rates of growth across time. Not everyone changes at the same rate. Furthermore, the above model implies a compound symmetry assumption for the variances and covariances of the repeated measures. That is, both the variances and covariances across time are assumed to be the same, namely,

$$\begin{aligned} V(y_{ij}) &= \sigma_\upsilon^2 + \sigma^2 \\ C(y_{ij}, y_{ij'}) &= \sigma_\upsilon^2, \quad \text{where } j \neq j'. \end{aligned} \tag{5}$$

This assumption is usually untenable for most longitudinal data. In general, measurements at points

Figure 12.1 Random-Intercepts HLM



close in time tend to be more highly correlated than measurements further separated in time. Also, in many studies, subjects are more similar at baseline and grow at different rates across time. Thus, it is natural to expect that variability will increase over time.

For these reasons, a more realistic HLM allows both the intercept and time trend to vary by individuals. For this, the Level 1 model is as before in (3), but the Level 2 model is augmented as

$$b_{0i} = \beta_0 + \upsilon_{0i},$$
$$b_{1i} = \beta_1 + \upsilon_{1i}. \tag{6}$$

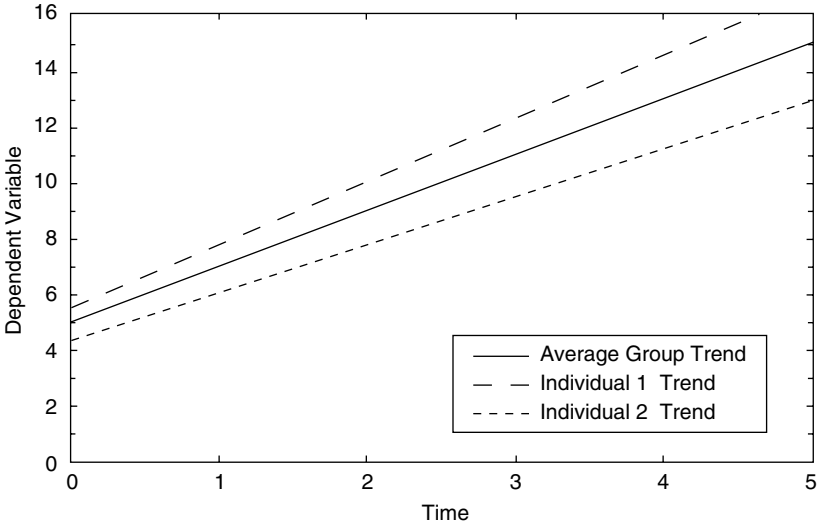
In this model, β_0 is the overall population intercept, β_1 is the overall population slope, υ_{0i} is the intercept deviation for subject i, and υ_{1i} is the slope deviation for subject i. As before, ε_{ij} is an independent error term distributed normally with mean 0 and variance σ^2. The assumption regarding the independence of the errors is one of conditional independence; that is, they are independent conditional on υ_{0i} and υ_{1i}. With two random individual-specific effects, the population distribution of intercept and slope deviations is assumed to be bivariate normal $N(0, \Sigma_\upsilon)$, with the random-effects variance-covariance matrix given by

$$\Sigma_\upsilon = \begin{bmatrix} \sigma^2_{\upsilon_0} & \sigma_{\upsilon_0\upsilon_1} \\ \sigma_{\upsilon_0\upsilon_1} & \sigma^2_{\upsilon_1} \end{bmatrix}.$$

This model can be thought of as a personal trend or change model because it represents the measurements of y as a function of time, both at the individual (υ_{0i} and υ_{1i}) and population (β_0 and β_1) levels. The intercept parameters indicate the starting point, and the slope parameters indicate the degree of change over time. The population intercept and slope parameters represent the overall (population) trend, whereas the individual parameters express how subjects deviate from the population trend. Figure 12.2 represents this model graphically.

Again, the figure represents the population trend with the solid line and the trends from two individuals, who now deviate both in terms of the intercept and slope. Because the slope varies for individuals, this model allows the possibility that some individuals do not change across time, whereas others can exhibit dramatic change. The population trend is the average across the individuals, and the variance terms indicate how much heterogeneity there is in the population. Specifically, the variance term $\sigma^2_{\upsilon_0}$ indicates how much spread there is around the population intercept, and $\sigma^2_{\upsilon_1}$ represents the spread in slopes. To the degree that each individual's deviation from the population trend is only due to random error, these variance terms will approach zero. Alternatively, as each individual's deviation from the population trend is nonrandom but characterized by the individual trend parameters υ_{0i} and υ_{1i} as being nonzero, these variance terms will increase from zero. In addition, the covariance term, $\sigma_{\upsilon_0\upsilon_1}$, represents the

Figure 12.2 Random-Intercept and Slopes HLM



degree to which the individual intercept and slope parameters covary. For example, a positive covariance term would suggest that individuals with higher initial values have greater positive slopes, whereas a negative covariance would suggest the opposite.

The coding of the time variable t has implications for the interpretation of the model parameters. For example, in growth models, t sometimes starts with the value zero for baseline and is incremented according to the measurement timeline (e.g., 1, 2, 3, 4 for, say, four monthly follow-ups). In this formulation, the intercept parameters (β_0, υ_{0i}, and $\sigma^2_{\upsilon_0}$) then characterize aspects of the baseline time point. Alternatively, t can be expressed in centered form, where the average of time is subtracted from each time value (e.g., -2, -1, 0, 1, 2). In this case, the meaning of the intercept parameters changes to reflect aspects about the midpoint of time and not the baseline time point. As yet another coding choice, sometimes substantive interest focuses on the end of the measurement timeline. Here, time could be coded as -4, -3, -2, -1, and 0 (in this example with five time points), so that the intercept parameters reflect aspects of the final time point. The choice of which representation to use often depends on ease of interpretation and the hypotheses of interest.

The occasions range from $j = 1$ to n_i in the model specification, with each person being measured on n_i time points. Because n carries the i subscript, each subject may vary in terms of the number of measured occasions. Furthermore, there are no restrictions on the number of observations per individual; subjects who are missing at a given time point are not excluded from the analysis. Also, because the time variable t carries the i subscript, subjects can be measured on different occasions. The underlying assumption of the model is that the data that are available for a given individual are representative of how that individual deviates from the population trend across the timeframe of the study.

Regarding missing data, as Laird (1988) points out, HLMs for longitudinal data using maximum likelihood estimation provide valid statistical tests in the presence of ignorable nonresponse. By *ignorable nonresponse,* it is meant that the probability of nonresponse is dependent on observed covariates *and* previous values of the dependent variable from the subjects with missing data. The notion here is that if subject attrition is related to previous performance, in addition to other observable subject characteristics, then the model provides valid statistical inferences for the model parameters. Because many instances of missing data are related to previous performance or other subject characteristics, HLMs provide a powerful method for dealing with longitudinal data sets in the presence of missing data.

12.2.1. Matrix Formulation

A more compact representation of the model is afforded using matrices and vectors. This formulation is particularly useful in model programming and helps

to summarize statistical aspects of the model. For this, the HLM for the $n_i \times 1$ response vector $\mathbf{y}$ for individual i can be written as

$$\underset{n_i\times 1}{\mathbf{y}_i} = \underset{n_i\times p}{\mathbf{X}_i}\ \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n_i\times r}{\mathbf{Z}_i}\ \underset{r\times 1}{\boldsymbol{\upsilon}_i} + \underset{n_i\times 1}{\boldsymbol{\varepsilon}_i} \tag{7}$$

with $i = 1 \ldots N$ individuals and $j = 1 \ldots n_i$ observations for individual i. Here, $\mathbf{y}_i$ is the $n_i \times 1$ dependent variable vector for individual i, $\mathbf{X}_i$ is the $n_i \times p$ covariate matrix for individual i, $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed regression parameters, $\mathbf{Z}_i$ is the $n_i \times r$ design matrix for the random effects, $\boldsymbol{\upsilon}_i$ is the $r \times 1$ vector of random individual effects, and $\boldsymbol{\varepsilon}_i$ is the $n_i \times 1$ residual vector.

For example, in the random intercepts and slopes HLM just considered, we would have

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \ldots \\ \ldots \\ y_{in_i} \end{bmatrix} \quad \text{and} \quad \mathbf{X}_i = \mathbf{Z}_i = \begin{bmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \ldots & \ldots \\ \ldots & \ldots \\ 1 & t_{in_i} \end{bmatrix}$$

for the data matrices and

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\upsilon}_i = \begin{bmatrix} \upsilon_{0i} \\ \upsilon_{1i} \end{bmatrix}$$

for the population and individual trend parameter vectors, respectively. The distributional assumptions about the random effects and residuals are

$$\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i}),$$
$$\boldsymbol{\upsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\upsilon).$$

As a result, it can be shown that the variance-covariance matrix of the repeated measures $\mathbf{y}$ is of the following form:

$$V(\mathbf{y}_i) = \mathbf{Z}_i \boldsymbol{\Sigma}_\upsilon \mathbf{Z}'_i + \sigma^2 \mathbf{I}_{n_i}. \tag{8}$$

For example, with $r = 2$, $n = 3$, and

$$\mathbf{Z}_i = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix},$$

the variance-covariance matrix equals $\sigma^2 \mathbf{I}_{n_i} +$

$$\begin{bmatrix} \sigma^2_{\upsilon_0} & \sigma^2_{\upsilon_0} + \sigma_{\upsilon_0\upsilon_1} & \sigma^2_{\upsilon_0} + 2\sigma_{\upsilon_0\upsilon_1} \\ \sigma^2_{\upsilon_0} + \sigma_{\upsilon_0\upsilon_1} & \sigma^2_{\upsilon_0} + 2\sigma_{\upsilon_0\upsilon_1} + \sigma^2_{\upsilon_1} & \sigma^2_{\upsilon_0} + 3\sigma_{\upsilon_0\upsilon_1} + 2\sigma^2_{\upsilon_1} \\ \sigma^2_{\upsilon_0} + 2\sigma_{\upsilon_0\upsilon_1} & \sigma^2_{\upsilon_0} + 3\sigma_{\upsilon_0\upsilon_1} + 2\sigma^2_{\upsilon_1} & \sigma^2_{\upsilon_0} + 4\sigma_{\upsilon_0\upsilon_1} + 4\sigma^2_{\upsilon_1} \end{bmatrix},$$

which allows the variances and covariances to change across time. For example, if both $\sigma_{\upsilon_0\upsilon_1}$ and $\sigma^2_{\upsilon_1}$ are positive, then clearly the variance increases across time. Diminishing variance across time is also possible if, for example, $-2\sigma_{\upsilon_0\upsilon_1} > \sigma^2_{\upsilon_1}$. Other patterns are possible depending on the values of these variance and covariance parameters.

Models with more than random intercepts and linear trends are also possible, as are models that allow autocorrelated errors; that is, $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 \boldsymbol{\Omega}_i)$. Here, $\boldsymbol{\Omega}$ might, for example, represent an autoregressive (AR) or moving average (MA) process for the residuals. Autocorrelated error regression models are common in econometrics. Their application within an HLM formulation is treated by Chi and Reinsel (1989) and Hedeker (1989) and extensively described in Verbeke and Molenberghs (2000). By including both random effects and autocorrelated errors, a wide range of variance-covariance structures for the repeated measures is possible. This flexibility is in sharp contrast to the traditional ANOVA models, which assume either a compound symmetry structure (univariate ANOVA) or a totally general structure (MANOVA). Typically, compound symmetry is too restrictive, and a general structure is not parsimonious. HLMs, alternatively, provide these two and everything in between and so allow efficient modeling of the variance-covariance structure of the repeated measures.

12.3. HLM EXAMPLE

To illustrate an HLM application, we will consider data from a psychiatric study described in Reisby et al. (1977). This study focused on the longitudinal relationship between imipramine (IMI) and desipramine (DMI) plasma levels and clinical response in 66 depressed inpatients. Imipramine is the prototypic drug in the series of compounds known as tricyclic antidepressants and is commonly prescribed for the treatment of major depression (Seiden & Dykstra, 1977). Because imipramine biotransforms into the active metabolite desmethylimipramine (or desipramine), measurement of desipramine was also done in this study. Major depression is often classified in terms of two types. The first type, nonendogenous or reactive depression, is associated with some tragic life event such as the death of a close friend or family member, whereas the second type, endogenous depression, is not a result of any specific event and appears to occur spontaneously. It is sometimes held that antidepressant medications are more effective for endogenous depression (Willner, 1985). In this sample, 29 patients were classified as

nonendogenous, and the remaining 37 patients were deemed to be endogenous.

The study design was as follows. Following a placebo period of 1 week, patients received 225-mg/day doses of imipramine for 4 weeks. In this study, subjects were rated with the Hamilton depression (HD) rating scale (Hamilton, 1960) twice during the baseline placebo week (at the start and end of this week), as well as at the end of each of the 4 treatment weeks of the study. Plasma level measurements of both IMI and its metabolite DMI were made at the end of each treatment week. The sex and age of each patient were recorded, and a diagnosis of endogenous or nonendogenous depression was made for each patient. Although the total number of subjects in this study was 66, the number of subjects with all measures at each of the weeks fluctuated: 61 at Week 0 (start of placebo week), 63 at Week 1 (end of placebo week), 65 at Week 2 (end of first drug treatment week), 65 at Week 3 (end of second drug treatment week), 63 at Week 4 (end of third drug treatment week), and 58 at Week 5 (end of fourth drug treatment week). Of the 66 subjects, only 46 had complete data at all time points. Thus, complete case analysis under repeated-measures MANOVA, for example, would discard approximately one third of the data set. HLM, alternatively, uses the data that are available from all 66 subjects.

12.3.1. Heterogeneous Growth Model

The first model fit to these data corresponds to the within-subjects model (3) and the between-subjects model (6). Here, time is treated using incremental values from 0 to 5. The results are presented in Table 12.1.

Focusing first on the estimated regression parameters, this model indicates that patients start, on average, with an HD score of 23.58 and change by -2.38 points each week. Lower scores on the HD reflect less depression, so patients are improving across time by about 2 points per week. The estimated HD score at Week 5 equals $23.58 - (5 \times 2.38) = 11.68$. In their report, Reisby et al. (1977) classified patients into three groups based on their final HD scores: Responders had scores below 8, partial responders were between 8 and 15, and nonresponders had final HD scores above 15. By this criterion, the average trend is in the partial response range at the final time point.

Both the intercept and slope are statistically significant ($p < .0001$) by the so-called "Wald test" (Wald, 1943), which uses the ratio of the maximum likelihood parameter estimate to its standard error to determine statistical significance. These test statistics (i.e., $z =$ ratio of the parameter estimate to its standard error) are compared to a standard normal frequency table to test the null hypothesis that the parameter equals 0. Alternatively, these z-statistics are sometimes squared, in which case the resulting test statistic is distributed as chi-square on 1 degree of freedom. In either case, the p-values are identical. The intercept being significant is not particularly meaningful; it just indicates that HD scores are different from zero at baseline. However, because the slope is significant, we can conclude that the rate of improvement is significantly different from zero in this study. On average, patients are improving across time.

Table 12.1 HLM Results for Level 1 Model (3) and Level 2 Model (6)

Parameter	*Estimate*	*SE*	z	$p <$
β_0	23.58	0.55	43.22	.0001
β_1	–2.38	0.21	–11.39	.0001
σ^2_{v0}	12.63	3.47		
σ^2_{v0v1}	–1.42	1.03		
σ^2_{v1}	2.08	0.50		
σ^2	12.22	1.11		

NOTE: $-2 \log L = 2219.04$.

For the variance and covariance terms, there are concerns with using the standard errors in constructing Wald test statistics, particularly when the population variance is thought to be near zero and the number of subjects is small (Bryk & Raudenbush, 1992). This is because variance parameters are bounded; they cannot be less than zero, and so using the standard normal for the sampling distribution is not reasonable. As a result, statistical significance is not indicated for the variance and covariance parameters in the tables. However, the magnitude of the estimates does reveal the degree of individual heterogeneity in both the intercepts and slopes. For example, although the average intercept in the population is estimated to be 23.58, the estimated population standard deviation for the intercept is $3.55 (= \sqrt{12.63})$. Similarly, the average population slope is -2.38, but the estimated population standard deviation for the slope equals 1.44, and so approximately 95% of subjects are expected to have slopes in the interval $-2.38 \pm (1.96 \times 1.44) = -5.20$ to .44. That the interval includes positive slopes reflects the fact that not all subjects improve across time. Thus, there is considerable heterogeneity in terms of patients' initial level of depression and in their change across time. Finally, the covariance between the intercept and linear trend is negative; expressed as a correlation, it

equals −.28, which is moderate in size. This suggests that patients who are initially more depressed (i.e., greater intercepts) improve at a greater rate (i.e., more pronounced negative slopes). An alternative explanation, though, is that of a floor effect due to the HD rating scale. Simply put, patients with less depressed initial scores have a more limited range of lower scores than those with higher initial scores.

An interesting question, at this point, is whether the between-subjects model in equation (6) is necessary over that in equation (4). In other words, is the assumption of compound symmetry rejected or not? Fitting the more restrictive compound symmetry model (not shown) yields $-2\log L = 2285.14$. Because these are nested models, they can be compared using a likelihood ratio test. For this, one compares the model deviance values (i.e., $-2\log L$) to a chi-square distribution, where the degrees of freedom equal the number of parameters set equal to zero in the more restrictive model. In the present case, $\chi^2_2 = 2285.14 - 2219.04 = 66.1$, $p < .0001$, for $\mathrm{H}_0 : \sigma_{\upsilon_0\upsilon_1} = \sigma^2_{\upsilon_1} = 0$. It should be noted that use of the likelihood ratio test for this purpose also suffers from the variance boundary problem mentioned above (Verbeke & Molenberghs, 2000). Based on simulation studies, it can be shown that the likelihood ratio test is too conservative (for testing null hypotheses about variance parameters)—namely, it does not reject the null hypothesis often enough. This would then lead to accepting a more restrictive variance-covariance structure than is correct. As noted by Berkhof and Snijders (2001), this bias can largely be corrected by dividing the p-value obtained from the likelihood ratio test (of variance terms) by 2. In the present case, it does not really matter, but this modification yields $p < .0001/2 = .00005$. Thus, there is clear evidence that the assumption of compound symmetry is rejected.

Using the estimated population intercept (β_0) and slope (β_1), we can estimate the average HD score across all time points. These are displayed in Table 12.2, along with the observed means and sample sizes at each time point.

As can be seen, there is close agreement between the observed and estimated means. Thus, the average change across time is very consistent with the posited linear change model. For a more quantitative assessment, the interested reader is referred to Kaplan and George (1998), who describe the use of econometric forecasting statistics to assess various forms of fit between observed and estimated means.

Similarly, we can address the fit of the observed variance-covariance matrix of the repeated measures,

Table 12.2 Observed and Estimated Means

	Week					
	0	*1*	*2*	*3*	*4*	*5*
Observed	23.44	21.84	18.31	16.42	13.62	11.95
Estimated	23.58	21.21	18.82	16.45	14.07	11.69
Sample size	61	63	65	65	63	58

which is given below. These are calculated based on the pairwise data for the covariances and the available data for each of the variances.

$$V(\mathbf{y}) = \begin{bmatrix} 20.55 & & & & & \\ 10.50 & 22.07 & & & & \\ 10.20 & 12.74 & 30.09 & & & \\ 9.69 & 12.43 & 25.96 & 41.15 & & \\ 7.17 & 10.10 & 25.56 & 36.54 & 48.59 & \\ 6.02 & 7.39 & 18.25 & 26.31 & 32.93 & 52.12 \end{bmatrix}$$

Based on the model estimates, we get

$$\hat{V}(\mathbf{y}) = \mathbf{Z}\hat{\mathbf{\Sigma}}_{\upsilon}\mathbf{Z}' + \hat{\sigma}^2\mathbf{I}$$
$$= \begin{bmatrix} 24.85 & & & & & \\ 11.21 & 24.08 & & & & \\ 9.79 & 12.52 & 27.48 & & & \\ 8.37 & 13.18 & 18.00 & 35.03 & & \\ 6.95 & 13.84 & 20.73 & 27.63 & 46.74 & \\ 5.53 & 14.50 & 23.47 & 32.44 & 41.41 & 62.60 \end{bmatrix},$$

where the design matrix of the random effects and the estimates of the random-effects variance-covariance matrix are given by

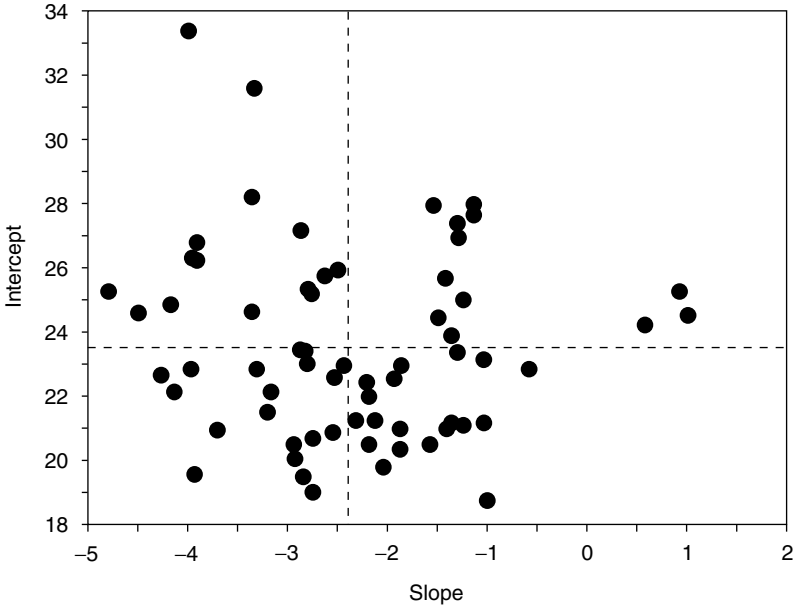
$$\mathbf{Z}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 \end{bmatrix},$$

$$\hat{\mathbf{\Sigma}}_{\upsilon} = \begin{bmatrix} 12.63 & -1.42 \\ -1.42 & 2.08 \end{bmatrix},$$

and $\hat{\sigma}^2 = 12.22$. Given that this variance-covariance matrix of 21 elements is represented by four parameter estimates, the fit is reasonably good. The model is clearly picking up on the increasing variance across time and the diminishing covariance away from the diagonal.

Finally, estimates of the individual random effects, $\hat{b}_{0i}$ and $\hat{b}_{1i}$, are often of interest. These are plotted in Figure 12.3. The dashed lines indicate the estimated population intercepts and slopes. Thus, $\hat{\upsilon}_{0i}$ is represented by the horizontal distance between a point and the horizontal line, whereas $\hat{\upsilon}_{1i}$ is represented by the vertical distance between a point and the vertical line.

Figure 12.3 Reisby Data: Estimated Random Effects



This scatter plot reveals the wide range of observed intercepts and slopes in this sample. In particular, some patients are very depressed initially but improve to a great degree (upper left-hand corner). Similarly, some patients show little or no improvement over time (toward the right side).

It is worth noting that the estimates of the individual random effects, presented in Figure 12.3, are empirical Bayes (EB) estimates, which reflect a compromise between an estimate based solely on an individual's data and an estimate for the population of interest. Thus, they are not equivalent to ordinary least squares (OLS) estimates, which would only rely on an individual's data. An important advantage of EB estimates relative to OLS estimates is that they are not as prone to the undue influence of outliers. This is especially true when an individual has few measurements by which to base these estimates on. Because of this, the EB estimates are said to be *shrunken to the mean*, where the mean of the random effects equals zero in the population. The degree of shrinkage depends on the number of measurements an individual has. Thus, if a subject has few measurements, then the EB estimate will be smaller (in absolute value) than the corresponding OLS estimate. Alternatively, if the subject has many measurements across time, then the EB and OLS estimates would be very similar. These EB estimates are readily available from most HLM software programs.

12.3.2. Effect of Diagnosis on Growth

At this point, it may be interesting to examine whether we can explain some of the heterogeneity in intercepts and slopes, depicted in Figure 12.3, in terms of particular subject characteristics. For this, we will augment the Level 2 model to include a covariate *DX*, which equals 0 if the patient's diagnosis is nonendogenous (NE) and 1 if the patient is endogenous (E). This variable enters the Level 2 model rather than the Level 1 model because it varies only with subjects (i) and not with time (j).

$$\begin{aligned} b_{0i} &= \beta_0 + \beta_2 DX_i + \upsilon_{0i}, \\ b_{1i} &= \beta_1 + \beta_3 DX_i + \upsilon_{1i}. \end{aligned} \tag{9}$$

Now, β_0 represents the average Week 0 HD level for NE patients, and β_1 is the average HD weekly improvement for NE patients. Similarly, β_2 represents the average Week 0 HD difference for E patients (relative to NE patients), and β_3 is the average difference in HD weekly improvement rates for E patients (relative to NE patients). Thus, β_3 represents the diagnosis-by-time interaction, indicating the degree to which the time trends vary by diagnostic group. In this augmented model, υ_{0i} is the individual's deviation from his or her diagnostic group intercept, and υ_{1i} is the individual's deviation from his or her diagnostic group slope. To the degree that the variable *DX* is

Table 12.3 HLM Results for Level 1 Model (3) and Level 2 Model (9)

Parameter	*Estimate*	*SE*	*z*	*p <*
NE intercept β_0	22.48	0.79	28.30	.0001
NE slope β_1	–2.37	0.31	–7.59	.0001
E intercept difference β_2	1.99	1.07	1.86	.063
E slope β_3 difference	–0.03	0.42	–0.06	.95
$\sigma^2_{\upsilon_0}$	11.64	3.53		
$\sigma_{\upsilon_0\upsilon_1}$	–1.40	1.00		
$\sigma^2_{\upsilon_1}$	2.08	0.50		
σ^2	12.22	1.11		

NOTE: –2 log $L = 2214.94$

Table 12.4 HLM Results for Level 1 Model (10) and Level 2 Model (11)

Parameter	*Estimate*	*SE*	*z*	*p <*
β_0	23.76	0.55	43.04	.0001
β_1	–2.63	0.48	–5.50	.0001
β_2	0.05	0.09	0.58	.56
$\sigma^2_{\upsilon_0}$	10.44	3.58		
$\sigma_{\upsilon_0\upsilon_1}$	–0.92	2.42		
$\sigma^2_{\upsilon_1}$	6.64	2.75		
$\sigma_{\upsilon_0\upsilon_2}$	–0.11	0.42		
$\sigma_{\upsilon_1\upsilon_2}$	–0.94	0.48		
$\sigma^2_{\upsilon_2}$	0.19	0.09		
σ^2	10.52	1.10		

NOTE: –2 log $L = 2207.64$.

useful in explaining intercept and slope variation, these individual deviations and their corresponding variances ($\sigma^2_{\upsilon_0}$ and $\sigma^2_{\upsilon_1}$) will be reduced. Results for this model are listed in Table 12.3.

A likelihood ratio test comparing this model to the previous one can be used to test the null hypothesis that the diagnosis-related effects (i.e., β_2 and β_3) are zero. This yields $\chi^2_2 = 2219.04 - 2214.94 = 4.1$, which is not statistically significant. Inspection of the estimates in Table 12.3 reveals a marginally significant difference in terms of their initial scores, with endogenous patients about 2 points higher and absolutely no difference in their trends across time. This is also borne out if one compares the variance estimates from Tables 12.2 and 12.3. Notice that the intercept variance has diminished slightly from to 12.63 to 11.64 as a result of the marginally significant intercept difference, whereas the slope variance is the same. Taken together, there is no real evidence that the two diagnostic groups differ in terms of their HD scores across time.

12.3.3. Curvilinear Growth Model

In many situations, it is too simplistic to assume that the change across time is linear. In the present example, for instance, it may be that the depression scores diminish across time in a curvilinear manner. A curvilinear trend would allow a leveling off of the improvement across time. This is clearly plausible for rating scale data, like the HD scores, where values below zero are impossible. Here, we will consider a curvilinear growth model by adding a quadratic term to the Level 1 model. More general polynomial growth models can also be obtained by adding cubic terms, quartic terms, and so on to the Level 1 model.

$$y_{ij} = b_{0i} + b_{1i}t_{ij} + b_{2i}t^2_{ij} + \varepsilon_{ij}. \tag{10}$$

Here, b_{0i} is the Week 0 HD level for patient i, b_{1i} is the weekly linear change in HD for patient i, and b_{2i} is the weekly quadratic change in HD for patient i. This model can also be written as

$$y_{ij} = b_{0i} + (b_{1i} + b_{2i}t_{ij})t_{ij} + \varepsilon_{ij}$$

to point out that the overall effect of time is $b_{1i} + b_{2i}t_{ij}$—namely, it is not constant but changes across time. The Level 2 between-subjects model is now

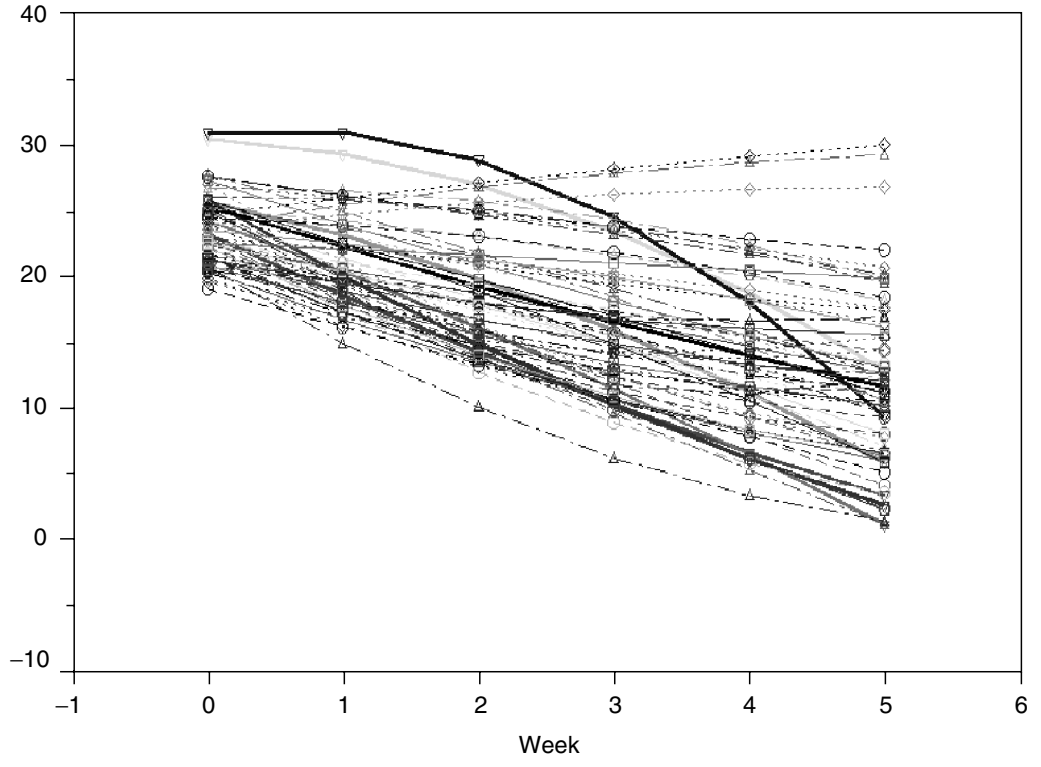
$$\begin{aligned} b_{0i} &= \beta_0 + \upsilon_{0i}, \\ b_{1i} &= \beta_1 + \upsilon_{1i}, \\ b_{2i} &= \beta_2 + \upsilon_{2i}, \end{aligned} \tag{11}$$

where β_0 is the average Week 0 HD level, β_1 is the average HD weekly linear change, and β_2 is the average HD weekly quadratic change. Similarly, υ_{0i} is the individual deviation from average intercept, υ_{1i} is the individual deviation from average linear change, and υ_{2i} is the individual deviation from average quadratic change. Thus, the model allows curvilinearity at both the population (β_2) and individual (υ_{2i}) levels.

Fitting this model yields the results given in Table 12.4.

Comparing this model to that of Table 12.1 (i.e., a model with $\beta_2 = \sigma^2_{\upsilon_2} = \sigma_{\upsilon_0\upsilon_2} = \sigma_{\upsilon_1\upsilon_2} = 0$) yields a deviance of 11.4, which is statistically significant on 4 degrees of freedom. This is interesting given that the Wald test for β_2 is clearly nonsignificant. In fact, comparing the above model to one with $\sigma^2_{\upsilon_2} = \sigma_{\upsilon_0\upsilon_2} = \sigma_{\upsilon_1\upsilon_2} = 0$ (not shown) yields a deviance

Figure 12.4 Reisby Data: Estimated Curvilinear Trends



of 11.0. Nearly all of the improvement in model fit is through the inclusion of the quadratic term as a random effect and not as a fixed effect. This suggests that although the trend across time is essentially linear at the population level, it is curvilinear at the individual level.

Figure 12.4 contains a plot of the individual trend estimates from this model. These are obtained by calculating $\hat{y}_{ij} = \hat{b}_{0i} + \hat{b}_{1i}t_{ij} + \hat{b}_{2i}t_{ij}^2$, for $t = 0, 1, \ldots, 5$, and then connecting the time point estimates for each individual.

The plot makes apparent the wide heterogeneity in trends across time, as well as the increasing variance in HD scores across time. Some individuals have accelerating downward trends, suggesting a delay in the drug effect. Alternatively, others have decelerating downward trends, which are consistent with a leveling off of the drug effect. Some individuals even have positive trends, indicating a worsening of their depressive symptoms across time. This is not too surprising given that antidepressants, such as imipramine, are known to be ineffective for some patients. The figure is also interesting in showing that many of the individual trend lines are approximately linear. Thus, the improvement that the curvilinear model provides in describing change across time is perhaps modest.

Finally, the fit of the observed variance-covariance matrix of the repeated measures is provided as follows:

$$\hat{V}(\mathbf{y}) = \mathbf{Z}\hat{\boldsymbol{\Sigma}}_{\upsilon}\mathbf{Z}' + \hat{\sigma}^2\mathbf{I}$$

$$= \begin{bmatrix} 20.96 & & & & & \\ 9.41 & 23.86 & & & & \\ 8.16 & 15.57 & 31.07 & & & \\ 6.68 & 16.08 & 23.11 & 38.31 & & \\ 4.98 & 14.88 & 23.26 & 30.12 & 45.98 & \\ 3.06 & 11.97 & 20.98 & 30.09 & 39.29 & 59.11 \end{bmatrix},$$

where

$$\mathbf{Z}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 1 & 4 & 9 & 16 & 25 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_{\upsilon} = \begin{bmatrix} 10.44 & -0.92 & -0.11 \\ -0.92 & 6.64 & -0.94 \\ -0.11 & -0.94 & 0.19 \end{bmatrix}.$$

By comparing this matrix with the observed variance-covariance matrix, presented earlier, we see that the estimated variances are close to the observed, and the model is clearly picking up the pattern of diminishing covariance away from the diagonal and at the earlier time points. Comparing this model to one with a totally general variance-covariance structure (not shown) yields a likelihood ratio $\chi^2_{14} = 14.9$, which is not statistically significant. Thus, this curvilinear model with seven variance-covariance parameters (σ^2 and six unique parameters in $\mathbf{\Sigma}_\upsilon$) provides a parsimonious fit of the variance-covariance matrix $V(\mathbf{y})$, which, being of dimension 6×6, has 21 unique elements. More details on methods for assessing and comparing model fit of the variance-covariance structure are described by Wolfinger (1993) and Grady and Helms (1995).

12.3.4. Orthogonal Polynomials

For trend models, it is often beneficial to represent the polynomials in orthogonal form (Bock, 1975). Mathematically, this avoids collinearity problems that can result from using multiples of t (t^2, t^3, etc.) as regressors. To see this, consider a curvilinear trend model with three time points. Then, $t = 0, 1$, and 2, whereas $t^2 = 0, 1$, and 4; these two variables are nearly perfectly correlated. To counter this, time is sometimes expressed in centered form—for example, $(t-\bar{t}) = -1, 0$, and 1 and $(t-\bar{t})^2 = 1, 0$, and 1. If there is the same number of observations at the three time points, this centering removes the correlation between the linear and quadratic trend components entirely. In the more usual situation of nonequal numbers of observations across time, this greatly diminishes the correlation between the polynomials. Another aspect of centering time is that the meaning of the model intercept changes. In the previous raw form of time, the intercept represented differences at the first time point (i.e., when time = 0). Alternatively, in centered form, the model intercept represents differences at the midpoint of time. For this reason, the intercept is often referred to as the constant or grand mean term in models using centered regressors.

An additional advantage of using orthogonal polynomials, over simply centering time, is that the polynomials are put on the same scale. Thus, their estimated coefficients can be compared in terms of their magnitude in the same way as standardized beta coefficients in ordinary regression analysis. For equal time intervals, tables of orthogonal polynomials can be found in Pearson and Hartley (1976), and Bock (1975) also describes how orthogonal polynomials can be obtained for unequal time intervals. For the current situation with six equally spaced time points, these are given as

$$\mathbf{X}' = \mathbf{Z}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -5 & -3 & -1 & 1 & 3 & 5 \\ 5 & -1 & -4 & -4 & -1 & 5 \end{bmatrix} \begin{matrix} /\sqrt{6} \\ /\sqrt{70}. \\ /\sqrt{84} \end{matrix}$$

Notice that these row vectors are independent of each other. Also, by dividing the values by the square root of the quantities on the right, which are simply the sum of squared values in a row, these polynomials have the same scale. Thus, these terms are simultaneously made independent of each other and standardized to the same (unit) scale. This holds exactly when the number of observations at each time point is equal and approximately so when they are unequal.

Fitting this orthogonal polynomial trend model yields the results given in Table 12.5.

Comparing the regression coefficients, as before, we see that only the constant and linear terms are significant. These terms also dominate in terms of magnitude; not only is the quadratic term nonsignificant, but it is also negligible. Thus, at the population average level, the trend is unquestionably linear. Turning to the variance estimates, we see that the estimated constant variance ($\hat{\sigma}^2_{\upsilon_0}$) is much larger than the estimated linear trend component ($\hat{\sigma}^2_{\upsilon_1}$), which is much larger than the estimated quadratic trend component ($\hat{\sigma}^2_{\upsilon_2}$). In terms of relative percentages, these three represent 71.7, 23.7, and 4.6, respectively, of the sum of the estimated individual variance terms. Thus, at the individual level, there is heterogeneity in terms of all three components but with diminishing return as the order of the polynomial increases. This analysis then quantifies what Figure 12.4 depicts.

Table 12.5 HLM Results for Orthogonal Polynomial Version of Level 1 Model (10) and Level 2 Model (11)

Parameter	*Estimate*	*SE*	z	*P* <
β_0	43.24	1.37	31.61	.0001
β_1	–9.94	0.86	–11.50	.0001
β_2	0.31	0.54	0.58	.56
$\sigma^2_{\upsilon_0}$	111.91	21.60		
$\sigma_{\upsilon_0\upsilon_1}$	37.99	10.92		
$\sigma^2_{\upsilon_1}$	37.04	8.90		
$\sigma_{\upsilon_0\upsilon_2}$	–10.14	6.19		
$\sigma_{\upsilon_1\upsilon_2}$	–0.82	3.50		
$\sigma^2_{\upsilon_2}$	7.23	3.50		
σ^2	10.52	1.10		

NOTE: $-2 \log L = 2207.64$.

Inspection of the covariance terms reveals a strong positive association between the constant and linear terms ($\hat{\sigma}^2_{\upsilon_1\upsilon_0} = 37.99$, expressed as a correlation = .59). This seems to be in contrast with the results for this term from the previous analysis in Table 12.4, in which there was a slight negative association between the intercept and linear terms ($\hat{\sigma}^2_{\upsilon_1\upsilon_0} = -.92$, expressed as a correlation = $-.11$). The reason for this apparent discrepancy is that in Table 12.4, the intercept represents the first time point, whereas the constant term in Table 12.5 represents the midpoint in time. Thus, an individual's linear trend is both negatively associated with his or her baseline depression level and positively associated with his or her mid-study depression level. Subjects with higher initial depression levels have slightly more negative linear slopes and, as a result, lower values at mid-study.

Finally, notice that the log-likelihood value is identical in Tables 12.5 and 12.6. Thus, the two solutions are equivalent; one is simply a reexpressed version of the other. Because of this, one can derive the results from Table 12.5 based on those from Table 12.6 and vice versa. Because the orthogonal polynomial representation greatly reduces any collinearity and scale differences in the regressors, it is computationally easier to obtain. For this reason, in cases where numerical difficulties are occurring with analyses using raw time values, investigators might consider using orthogonal polynomials instead.

12.3.5. Growth Model With Time-Varying Covariates

In this section, we examine the effects of the time-varying drug plasma levels IMI and DMI. Because an inspection of the data indicated that the magnitude of these measurements varied greatly between individuals (from 4 to 312 mg/L for IMI and from 0 to 740 mg/L for DMI), a log transformation is used for these covariates. This helps to ensure that the estimated regression coefficients are not unduly influenced by extreme values on these covariates. Also, these variables, ln IMI and ln DMI, are expressed in grand-mean centered form so that the model intercept represents HD scores for patients with average drug levels. To obtain the grand-mean centered versions of these variables, we subtract the variable's sample mean from each observation. For notational simplicity in the model equations, I_{ij} and D_{ij} will represent the grand-mean centered versions of ln IMI and ln DMI, respectively, in what follows. Also, whereas the previous models considered HD outcomes from Weeks 0 to 5, the models of this section only include HD outcome data from Weeks 2 to 5. This is because the drug plasma levels are not available at the first two time points of the study (i.e., Week 0, or baseline, and Week 1, or the end of the drug washout period). Although HLM does allow incomplete data across time, data must be complete within a given time point (in terms of both the dependent variable and covariates) for that time point to be included in the analysis. Thus, the analyses that follow are for the 4-week period following the drug washout period, with t_{ij} coded as 0, 1, 2, and 3 for these four respective time points. As a result, the intercept represents HD scores for Week 2 of the study (i.e., when $t_{ij} = 0$).

The first Level 1 model is given by

$$y_{ij} = b_{0i} + b_{1i}t_{ij} + b_{2i}I_{ij} + b_{3i}D_{ij} + \varepsilon_{ij}, \qquad (12)$$

where b_{0i} is the Week 2 HD level for patient i under average levels of both ln IMI and ln DMI, b_{1i} is the weekly change in HD for patient i, b_{2i} is the patient's change in HD due to ln IMI, and b_{3i} is the change in HD due to ln DMI. The between-subjects model is given as

$$\begin{aligned} b_{0i} &= \beta_0 + \upsilon_{0i}, \\ b_{1i} &= \beta_1 + \upsilon_{1i}, \\ b_{2i} &= \beta_2, \\ b_{3i} &= \beta_3, \end{aligned} \qquad (13)$$

where β_0 is the average Week 2 HD level for patients with average ln IMI and ln DMI values, β_1 is the average HD weekly change, β_2 is the average HD difference for a unit change in ln IMI, and β_3 is the average HD difference for a unit change in ln DMI. Also, υ_{0i} is the individual intercept deviation, and υ_{1i} is the individual slope deviation. Notice that the Level 2 model indicates that the drug effects could also be treated as random. This would be accomplished by adding υ_{2i} and υ_{3i} to the model and would allow individual variation in terms of the drug-level effect on HD scores. Given that antidepressants such as IMI and DMI are not effective for all individuals, it is plausible that the drug levels are more strongly related to changes in depression for some individuals, whereas for others they are less so. Similarly, one could add individual-level covariates (e.g., endogenous/nonendogenous group) into the models for b_{2i} and b_{3i} to examine whether the drug effects vary with individual-level covariates. Again, it is feasible that the drug effects on outcome are stronger for endogenous than nonendogenous patients. Although these possibilities will not be considered

Table 12.6 HLM Results for Level 1 Model (12) and Level 2 Model (13)

Parameter	*Estimate*	*SE*	*z*	*p <*
Intercept β_0	18.17	0.71	25.70	.0001
Time slope β_1	–2.03	0.28	–7.15	.0001
In IMI β_2	0.60	0.85	0.71	.48
In DMI β_3	–1.20	0.63	–1.90	.06
$\sigma^2_{v_0}$	24.83	5.79		
$\sigma_{v_0 v_1}$	–0.72	1.74		
$\sigma^2_{v_1}$	2.73	0.95		
σ^2	10.46	1.37		

NOTE: –2 log L = 1502.5.

Table 12.7 HLM Results for Level 1 Model (14) and Level 2 Model (13)

Parameter	*Estimate*	*SE*	*z*	*p <*
Intercept β_0	–5.18	0.66	–7.87	.0001
Slope β_1	–1.97	0.29	–6.90	.0001
In IMI β_2	0.63	0.82	0.77	*ns*
In DMI β_3	–1.97	0.60	–3.26	.0014
$\sigma^2_{v_0}$	20.50			
$\sigma_{v_0 v_1}$	0.84			
$\sigma^2_{v_1}$	2.78			
σ^2	10.53			

NOTE: –2 log L = 1498.8.

here, an example of an HLM allowing such individual variation in relationships is described by Hedeker, Flay, and Petraitis (1996).

Fitting the present model yields the results given in Table 12.6. It is interesting to note that neither of the drug levels seems to be significantly related to the depression scores across time. However, note that the model given in (12) specifies that a person's drug level is related to his or her depression score at that same time point. It might be more plausible to instead posit that a person's drug level is related to his or her *change* in depression score, or improvement, at that same time point. For this, the following alternative Level 1 model is considered:

$$(y_{ij} - y_{i0}) = b_{0i} + b_{1i} t_{ij} + b_{2i} I_{ij} + b_{3i} D_{ij} + \varepsilon_{ij}, \quad (14)$$

where y_{i0} is the individual's HD score at baseline (or at Week 1 for those few subjects with a missing baseline score). This yields the results presented in Table 12.7.

Interestingly, now the effect of DMI, the metabolite of IMI, is highly significant and negative. Thus, greater DMI values are associated with greater improvement (i.e., more negative HD change scores). However, the parent drug IMI is not significantly related to HD change scores; in fact, its coefficient is positive. It is important to remember that the model estimates the IMI effect, controlling for the DMI effect, and vice versa. These two drug levels are moderately correlated with each other (r = .18, .23, .22, and .18 for the four respective time points), and so the results above are not necessarily indicative of the marginal relationships of each drug with depression scores. Correlations of the drug plasma levels with the HD scores, both raw and expressed as change scores, are given in Table 12.8. These bear out the fact that the drug levels are much more associated with the HD change scores than the actual scores. These correlations also show the greater association between HD change scores and DMI, rather than IMI, drug levels.

Table 12.8 Correlation Between HD Scores and Plasma Levels (Natural Log Units)

Drug	*Week 2*	*Week 3*	*Week 4*	*Week 5*
HD total score				
IMI	–0.034	–0.038	–0.003	–0.189
DMI	–0.177	–0.075	–0.246	–0.293*
HD change from baseline				
IMI	–0.049	–0.106	–0.046	–0.240
DMI	–0.366*	–0.281*	–0.363*	–0.361*

NOTE: *$p < .05$.

12.3.5.1. Within- and Between-Subjects Effects for Time-Varying Covariates

When time-varying covariates are included in an HLM, as in the manner of the last analysis, an assumption is made that the between- and within-subjects effects of these variables are equal. To see this, express the time-varying covariates I_{ij} and D_{ij} as

$$I_{ij} = \bar{I}_i + (I_{ij} - \bar{I}_i),$$
$$D_{ij} = \bar{D}_i + (D_{ij} - \bar{D}_i),$$

where $\bar{I}_i$ and $\bar{D}_i$ are the means of these two time-varying covariates computed for each individual. Thus, the first term following the equality represents the individual's mean on the time-varying covariate (i.e., a between-subjects variable), and the second term represents the individual's deviation around his or her

mean (i.e., a within-subjects variable). Including both of these terms into the HLM yields

$$(y_{ij} - y_{i0}) = b_{0i} + b_{1i}t_{ij} + b_{2i}(I_{ij} - \bar{I}_i) + b_{3i}(D_{ij} - \bar{D}_i) + \varepsilon_{ij}, \quad (15)$$

and

$$\begin{aligned} b_{0i} &= \beta_0 + \beta_4\bar{I}_i + \beta_5\bar{D}_i + \upsilon_{0i}, \\ b_{1i} &= \beta_1 + \upsilon_{1i}, \\ b_{2i} &= \beta_2, \\ b_{3i} &= \beta_3, \end{aligned} \quad (16)$$

for the Level 1 and Level 2 models. Thus, the total effect of IMI, for example,

$$\beta_2(I_{ij} - \bar{I}_i) + \beta_4\bar{I}_i,$$

is partitioned into its within- and between-subjects effects (i.e., β_2 and β_4, respectively). The between-subjects part indicates the degree to which the individual's average drug level is related to his or her average depression level, averaging across time. In other words, it may be that subjects with consistently high drug levels have consistently low depression scores. Alternatively, the within-subjects component represents the degree to which variation in an individual's drug level is associated with a change in his or her depression scores (i.e., a within-subject change). Thus, it may be that a higher relative drug level for an individual is associated with a lower relative depression score for that individual at a particular time point. If these two are equal ($\beta_2 = \beta_4$), then the IMI effect is

$$\beta_2(I_{ij} - \bar{I}_i) + \beta_2\bar{I}_i = \beta_2 I_{ij},$$

which is exactly what was used in the last analysis. Thus, we implicitly assumed that the within- and between-subjects effects of these two drug levels were the same in the previous analysis. This assumption can be tested by comparing the model specified by (14) and (13) with the more general model of (15) and (16). Table 12.9 includes the results of this latter analysis.

Comparing the two models yields a likelihood-ratio statistic of $\chi^2_2 = 3.0$, which is not statistically significant. Thus, the assumption of homogeneity of the between- and within-subjects regressions cannot be rejected for these data. Inspecting the estimated coefficients for DMI supports this: -1.8 and -2.4 for the within- and between-subjects effects, respectively. Conversely, the estimates for IMI are very different and even of the opposite sign. However, neither is statistically significant, and the standard errors for these two IMI estimates are quite large. In conclusion, for these data, there is not sufficient evidence to reject the assumption of equality in the within- and between-subjects effects for these two drug levels.

Table 12.9 HLM Results for Level 1 Model (15) and Level 2 Model (16)

Parameter	*Estimate*	*SE*	*z*	*p <*
Intercept β_0	–5.09	0.66	–7.71	.0001
Slope β_1	–2.02	0.29	–6.94	.0001
Within In IMI β_2	2.44	1.46	1.68	.10
Within In DMI β_3	–1.80	1.00	–1.80	.075
Between In IMI β_4	–0.31	1.00	–0.31	*ns*
Between In DMI β_5	–2.37	0.80	–2.97	.004
$\sigma^2_{\upsilon_0}$	20.32			
$\sigma_{\upsilon_0\upsilon_1}$	0.50			
$\sigma^2_{\upsilon_1}$	2.83			
σ^2	10.38			

NOTE: $-2 \log L = 1495.8$.

12.3.5.2. Time Interactions With Time-Varying Covariates

In some cases, it can be of substantive interest to examine whether there are interactions between a time-varying covariate and time. For example, one might posit that the relationship between the time-varying covariate and the outcome either increases or decreases across time. This is clearly plausible in the present example because the effectiveness of antidepressants is not thought to be immediate but instead to develop over time (Reisby et al., 1977). Thus, it is of interest to examine the degree to which the effects of the time-varying drug plasma levels on the change in depression scores vary across time. To explore this possibility, we can augment the Level 1 model to include the time interactions, namely,

$$(y_{ij} - y_{i0}) = b_{0i} + b_{1i}t_{ij} + b_{2i}I_{ij} + b_{3i}D_{ij} + b_{4i}(I_{ij} \times t_{ij}) + b_{5i}(D_{ij} \times t_{ij}) + \varepsilon_{ij}, \quad (17)$$

with the accompanying Level 2 model,

$$\begin{aligned} b_{0i} &= \beta_0 + \upsilon_{0i}, \\ b_{1i} &= \beta_1 + \upsilon_{1i}, \\ b_{2i} &= \beta_2, \\ b_{3i} &= \beta_3, \\ b_{4i} &= \beta_4, \\ b_{5i} &= \beta_5. \end{aligned} \quad (18)$$

To correctly interpret the model parameters, one should remember that the drug levels have been grand-mean centered, that the week variable equals 0 for the second week of the study, and that interpretation of the "main effects" is altered when interactions are present (i.e., they represent the effect of the variable when the interacting variable equals 0). Thus, in this model, β_0 represents the average Week 2 HD change score for patients with average drug levels, β_1 is the average weekly change in HD change scores for patients with average drug levels, β_2 is the HD change score difference for a unit change of ln IMI at Week 2, and β_3 represents the HD change score difference per unit change of ln DMI at Week 2. One can think of β_2 as the regression slope corresponding to the plot of HD change scores versus ln IMI levels considering Week 2 data only (with the caveat that this regression slope is really a partial regression slope adjusting for the other drug level). Similar comments apply for interpreting β_3 in terms of ln DMI. Turning to the interactions β_4 and β_5, these indicate the per week change in the drug effects on the HD change scores. In terms of the plot analogy, these interactions correspond to the change in (partial) regression slopes associated with separate weekly plots of HD change scores versus drug levels as one goes across the weeks—in other words, how the slope for a given drug varies across time. Finally, υ_{0i} represents the individual intercept deviation, and υ_{1i} is the individual time-slope deviation. Table 12.10 lists the results of this analysis.

Comparing this model to the one without the drug-by-time interaction (i.e., from Table 12.7) yields a likelihood ratio statistic of $\chi^2_2 = 6.8$, which is statistically significant at the .05 level. Thus, there is evidence that the drug effects on depression do vary across time. Inspecting the estimates and their test statistics in Table 12.10 reveals that it is DMI, not IMI, that is interacting significantly with time. Specifically, DMI has an initial Week 2 effect that is significant ($p < .017$), indicating that higher levels of DMI are associated with greater improvement on the HD scale at this time point, and this beneficial effect of DMI gets more pronounced across time ($p < .01$). Concretely, the benefit of a one-unit change in ln DMI at Week 2 is a 1.5-point reduction on the HD change score, whereas by the last time point, it is a 4.5-point reduction (3×1.5).

At first glance, it might seem a bit unusual that the DMI-by-time interaction is so highly significant given the reported correlations in Table 12.8. To better understand this, consider the simple linear regression slopes that are obtained from regressing HD change scores on ln DMI values at each of the four time points separately: These are -2.081, -2.195, -3.370, and -3.3765, respectively. These regression slopes provide clearer evidence of the DMI-by-time interaction, as they increase (in absolute value) more dramatically across time than the analogous correlations in Table 12.8. Why do these two sets of descriptive statistics suggest different conclusions? Remembering that the correlation is essentially a scale-free representation of the slope (i.e., $r = \hat{\beta}\, s_x/s_y$), it is clear that the scales of the dependent and independent variables play a role here. Interestingly, the scale of these two go in opposite directions across time; the standard deviations of the HD change scores increase (5.38, 6.51, 7.35, and 7.88 across the four time points), whereas the standard deviations of the ln DMI values decrease (.95, .84, .79, and .76 across these same four time points). Thus, the metric for the slopes across time is very different (i.e., the ratio of standard deviations s_x/s_y equals .18, .13, .11, and .10, respectively), which explains why the simple slopes and correlations are not in such close agreement and why the significant DMI-by-time interaction of the HLM is a bit at odds with the apparent consistent pattern of the correlations across time. As this final HLM and the descriptive statistics make clear, it is the scale-dependent slope of DMI (i.e., how much change in depression is associated with a unit change in this blood level) that is increasing across time, not the scale-free association.

Table 12.10 HLM Results for Level 1 Model (17) and Level 2 Model (18)

Parameter	*Estimate*	*SE*	*z*	$p <$
Intercept β_0	–5.12	0.65	–7.82	.0001
Time slope β_1	–1.94	0.28	–7.04	.0001
In IMI β_2	0.40	0.87	0.46	*ns*
In DMI β_3	–1.51	0.62	–2.43	.017
In IMI by time β_4	0.16	0.41	0.39	*ns*
In DMI by time β_5	–0.90	0.34	–2.65	.01
$\sigma^2_{\upsilon_0}$	20.24			
$\sigma_{\upsilon_0\upsilon_1}$	0.99			
$\sigma^2_{\upsilon_1}$	2.50			
σ^2	10.35			

NOTE: $-2 \log L = 1492.0$.

12.4. Discussion

As demonstrated, HLM provide a useful way of analyzing longitudinal data. Specifically, HLM allows for the presence of missing data, irregularly spaced measurements across time, time-varying and invariant covariates, accommodation of individual-specific

deviations from the average time trend, and estimation of the population variance associated with these individual effects. In addition, methods and software exist for the analysis of continuous and categorical outcomes. Perhaps the most popular feature of HLM is its treatment of missing data. As has been illustrated, subjects are not assumed to be measured at the same number of time points. Because there are no restrictions on the number of observations per individual, subjects who are missing at a given interview wave are not excluded from the analysis. The assumption of the model is that the data that are available for a given subject are representative of that subject's deviation from the average trends across time (which are estimated based on the whole sample).

A slightly more sophisticated approach for handling missing data is to group subjects based on their available data pattern across time. For example, subjects might be classified as complete-data subjects or incomplete-data subjects. This between-subjects classification variable can then be included in the analysis to examine the degree to which these two types of subjects differ in terms of the outcome variable. Interactions can also be included to see if the treatment group-related effects vary by missing data pattern. This approach has been called *pattern-mixture modeling* by Little (1993, 1994, 1995). Hedeker and Gibbons (1997) illustrate the use of this approach as applied to psychiatric clinical trials data. Verbeke and Molenberghs (2000) further describe the pattern-mixture approach in much greater statistical detail, including how these models can be used to assess the sensitivity of the results to different assumptions about the missing data. Although not applied in this chapter, the pattern-mixture approach provides a further way of dealing with missing data in longitudinal studies.

Statistical software to perform HLM analysis has proliferated, especially for continuous outcomes: HLM 5 (Raudenbush, Bryk, Cheong, & Congdon, 2000), SAS PROC MIXED, MLwiN (Goldstein et al., 1998), and MIXREG (Hedeker & Gibbons, 1996b), to mention a few programs. For categorical data, software has become available for dichotomous (EGRET [CYTEL, 1999]) and ordinal or nominal outcomes (SAS PROC NLMIXED, HLM 5, MLwiN, and GLLAMM [Rabe-Hesketh, Pickles, & Skrondal, 2001]; MIXOR [Hedeker & Gibbons, 1996a]; MIXNO [Hedeker, 1999]). Of course, software for nominal and ordinal outcomes can be used to fit models for dichotomous outcomes. Review articles comparing some of these software programs include van der Leeden, Vrijburg, and de Leeuw (1996) and de Leeuw and Kreft (2001).

This chapter has focused on the modeling aspects of HLM without discussion of parameter estimation. In nearly all of the software programs for continuous outcomes, a combination of two complementary methods has generally been used: empirical Bayes (EB) methods for estimation of the individual effects (e.g., υ_{0i}) and maximum likelihood (ML) methods for estimation of variance and covariance parameters (e.g., σ^2, $\sigma^2_{\upsilon_0}$, $\sigma^2_{\upsilon_1}$, and $\sigma_{\upsilon_0\upsilon_1}$) and covariate effects ($\boldsymbol{\beta}$). Iterative solutions to estimate these two sets of parameters have been described using the EM algorithm (Bryk & Raudenbush, 1992; Laird & Ware, 1982) and the Fisher scoring algorithm (Bock, 1989a; Longford, 1987). Because these models are more complex than ordinary fixed-effects regression models, it is sometimes the case that the iterative procedure does not converge to a solution. If this occurs, it is often because the model is overly complex relative to the data being used to estimate it, and so model simplification is necessary. Although it is not always apparent why a particular model does not converge, building models in a sequential piecewise manner can help to isolate where troubles occur.

In the example, repeated observations were observed nested within individuals. In the terminology of multilevel analysis (Goldstein, 1995) and hierarchical linear models (Raudenbush & Bryk, 2002), this is termed a *two-level data structure,* with individuals representing Level 2 and the nested repeated observations Level 1. The models that we have presented are thus referred to as two-level models. Individuals themselves, though, are often observed clustered within some higher level unit, for example, a classroom, clinic, or worksite. Cross-sectional clustered data can also be considered as two-level data, with the clusters representing Level 2 and the clustered subjects Level 1. Analysis of cross-sectional clustered data using HLM is discussed by Hedeker, Gibbons, and Flay (1994) and Hedeker, McMahon, Jason, and Salina (1994). In some studies, subjects are clustered and also repeatedly measured, resulting in three levels of data: the cluster (Level 3), individual (Level 2), and repeated observation (Level 1). Analysis of three-level data is described in Goldstein (1995), Raudenbush and Bryk (2002), Longford (1993), and Gibbons and Hedeker (1997).

Because longitudinal designs are increasingly used in the social sciences, it is important that statistical methods are developed and used to extract the most out of these longitudinal data sets. HLM provides an attractive approach for addressing some key questions that emerge from longitudinal designs. It is hoped that this chapter has helped in increasing the understanding of

these methods and their potential for use in analyzing longitudinal outcomes.

REFERENCES

Agresti, A., & Natarajan, R. (2001). Modeling clustered ordered categorical data: A survey. *International Statistical Review, 69*, 345–371.

Albert, P. S. (1999). Longitudinal data analysis (repeated measures) in clinical trials. *Statistics in Medicine, 18*, 1707–1732.

Berkhof, J., & Snijders, T. A. B. (2001). Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics, 26*, 133–152.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.

Bock, R. D. (1983a). The discrete Bayesian. In H. Wainer & S. Messick (Eds.), *Modern advances in psychometric research* (pp. 103–115). Hillsdale, NJ: Lawrence Erlbaum.

Bock, R. D. (1983b). Within-subject experimentation in psychiatric research. In R. D. Gibbons & M. W. Dysken (Eds.), *Statistical and methodological advances in psychiatric research* (pp. 59–90). New York: Spectrum.

Bock, R. D. (1989a). Measurement of human variation: A two stage model. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 319–342). New York: Academic Press.

Bock, R. D. (Ed.). (1989b). *Multilevel analysis of educational data.* New York: Academic Press.

Brown, H., & Prescott, R. (1999). *Applied mixed models in medicine*. New York: John Wiley.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Burchinal, M. R., Bailey, D. B., & Snyder, P. (1994). Using growth curve analysis to evaluate child change in longitudinal investigations. *Journal of Early Intervention, 18*, 403–423.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. In D. Berliner (Ed.), *Review of research in education* (Vol. 8, pp. 158–233). Washington, DC: American Educational Research Association.

Campbell, S. K., & Hedeker, D. (2001). Validity of the test of infant motor performance for discriminating among infants with varying risks for poor motor outcome. *Journal of Pediatrics, 139*, 546–551.

Carroll, K. M., Rounsaville, B. J., Nich, C., Gordon, L. T., Wirtz, P. W., & Gawin, F. (1994). One-year follow-up of psychotherapy and pharmacotherapy for cocaine dependence. *Archives of General Psychiatry, 51*, 989–997.

Chi, E. M., & Reinsel, G. C. (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Society, 84*, 452–459.

Cnaan, A., Laird, N. M., & Slasor, P. (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine, 16*, 2349–2380.

Collins, L. M., & Sayer, A. G. (Eds.). (2001). *New methods for the analysis of change*. Washington, DC: American Psychological Association.

Curran, P. J., Stice, E., & Chassin, L. (1997). The relation between adolescent and peer alcohol use: A longitudinal random coefficients model. *Journal of Consulting and Clinical Psychology, 65*, 130–140.

CYTEL. (1999). *Egret for Windows*. Cambridge, MA: Author.

Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*. New York: Springer.

de Leeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics, 11*, 57–85.

de Leeuw, J., & Kreft, I. (2001). Software for multilevel analysis. In A. H. Leyland & H. Goldstein (Eds.), *Multilevel modelling of health statistics* (pp. 187–204). New York: John Wiley.

Delucchi, K., & Bostrom, A. (1999). Small sample longitudinal clinical trials with missing data: A comparison of methods. *Psychological Methods, 4*, 158–172.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance component models. *Journal of the American Statistical Society, 76*, 341–353.

Diggle, P., Liang, K.-Y., & Zeger, S. L. (1994). *Analysis of longitudinal data*. New York: Oxford University Press.

Elkin, I., Gibbons, R. D., Shea, M. T., Sotsky, S. M., Watkins, J. T., Pilkonis, P. A., & Hedeker, D. (1995). Initial severity and differential treatment outcome in the NIMH treatment of depression collaborative research program. *Journal of Consulting and Clinical Psychology, 63*, 841–847.

Everitt, B. S. (1998). Analysis of longitudinal data: Beyond MANOVA. *British Journal of Psychiatry, 172*, 7–10.

Fitzmaurice, G. M., Laird, N. M., & Rotnitzky, A. G. (1993). Regression models for discrete longitudinal responses. *Statistical Science, 8*, 284–309.

Gallagher, T. J., Cottler, L. B., Compton, W. M., & Spitznagel, E. (1997). Changes in HIV/AIDS risk behaviors in drug users in St. Louis: Applications of random regression models. *Journal of Drug Issues, 27*, 399–416.

Gibbons, R. D., & Hedeker, D. (1994). Application of random-effects probit regression models. *Journal of Consulting and Clinical Psychology, 62*, 285–296.

Gibbons, R. D., & Hedeker, D. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics, 53*, 1527–1537.

Gibbons, R. D., & Hedeker, D. (2000). Application of mixed-effects models in biostatistics. *Sankhya, Series B, 62*, 70–103.

Gibbons, R. D., Hedeker, D., Elkin, I., Waternaux, C. M., Kraemer, H. C., Greenhouse, J. B., et al. (1993). Some conceptual and statistical issues in analysis of longitudinal psychiatric data. *Archives of General Psychiatry, 50*, 739–750.

Gibbons, R. D., Hedeker, D., Waternaux, C. M., & Davis, J. M. (1988). Random regression models: A comprehensive approach to the analysis of longitudinal psychiatric data. *Psychopharmacology Bulletin, 24*, 438–443.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). New York: Halstead.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., et al. (1998). *A user's guide to MLwiN*. London: Institute of Education, University of London.

Grady, J. J., & Helms, R. W. (1995). Model selection techniques for the covariance matrix for incomplete longitudinal data. *Statistics in Medicine, 14*, 1397–1416.

Halikas, J. A., Crosby, R. D., Pearson, V. L., & Graves, N. M. (1997). A randomized double-blind study of carbamazepine

in the treatment of cocaine abuse. *Clinical Pharmacology and Therapeutics, 62*, 89–105.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology and Neurosurgical Psychiatry, 23*, 56–62.

Hand, D., & Crowder, M. (1996). *Practical longitudinal data analysis*. New York: Chapman & Hall.

Hedeker, D. (1989). *Random regression models with autocorrelated errors.* Unpublished doctoral dissertation, University of Chicago, Department of Psychology.

Hedeker, D. (1999). MIXNO: A computer program for mixed-effects nominal logistic regression. *Journal of Statistical Software, 4*(5), 1–92.

Hedeker, D., Flay, B. R., & Petraitis, J. (1996). Estimating individual differences of behavioral intentions: An application of random-effects modeling to the theory of reasoned action. *Journal of Consulting and Clinical Psychology, 64*, 109–120.

Hedeker, D., & Gibbons, R. D. (1996a). MIXOR: A computer program for mixed-effects ordinal probit and logistic regression analysis. *Computer Methods and Programs in Biomedicine, 49*, 157–176.

Hedeker, D., & Gibbons, R. D. (1996b). MIXREG: A computer program for mixed-effects regression analysis with autocorrelated errors. *Computer Methods and Programs in Biomedicine, 49*, 229–252.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods, 2*, 64–78.

Hedeker, D., Gibbons, R. D., & Flay, B. R. (1994). Random-effects regression models for clustered data: With an example from smoking prevention research. *Journal of Consulting and Clinical Psychology, 62*, 757–765.

Hedeker, D., McMahon, S. D., Jason, L. A., & Salina, D. (1994). Analysis of clustered data in community psychology: With an example from a worksite smoking cessation project. *American Journal of Community Psychology, 22*, 595–615.

Hedeker, D., & Mermelstein, R. J. (1996). Application of random-effects regression models in relapse research. *Addiction, 91*(Suppl.), S211–S229.

Hedeker, D., & Mermelstein, R. J. (2000). Analysis of longitudinal substance use outcomes using random-effects regression models. *Addiction, 95*(Suppl. 3), S381–S394.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.

Hui, S. L., & Berger, J. O. (1983). Empirical Bayes estimation of rates in longitudinal studies. *Journal of the American Statistical Association, 78*, 753–759.

Huttenlocher, J. E., Haight, W., Bryk, A. S., & Seltzer, M. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27*, 236–248.

Kaplan, D., & George, R. (1998). Evaluating latent growth models through ex post simulation. *Journal of Educational and Behavioral Statistics, 23*, 216–235.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1999). A comparison of recent approaches to the analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology, 52*, 63–78.

Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine, 7*, 305–315.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*, 963–974.

Lesaffre, E., Asefa, M., & Verbeke, G. (1999). Assessing the goodness-of-fit of the Laird and Ware model—an example: The Jimma infant survival differential longitudinal study. *Statistics in Medicine, 18*, 835–854.

Leyland, A. H., & Goldstein, H. (Eds.). (2001). *Multilevel modelling of health statistics*. New York: John Wiley.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association, 88*, 125–133.

Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika, 81*, 471–483.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association, 90*, 1112–1121.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika, 74*, 817–827.

Longford, N. T. (1993). *Random coefficient models*. New York: Oxford University Press.

Manor, O., & Kark, J. D. (1996). A comparative study of four methods for analysing repeated measures data. *Statistics in Medicine, 15*, 1143–1159.

Moskowitz, D. S., & Hershberger, S. L. (Eds.). (2002). *Modeling intraindividual variability with repeated measures data*. Mahwah, NJ: Lawrence Erlbaum.

Niaura, R., Spring, B., Borrelli, B., Hedeker, D., Goldstein, M. G., Keuthen, N., et al. (2002). Multicenter trial of fluoxetine as an adjunct to behavioral smoking cessation treatment. *Journal of Consulting and Clinical Psychology, 70*, 887–896.

Omar, R. Z., Wright, E. M., Turner, R. M., & Thompson, S. G. (1999). Analysing repeated measures data: A practical comparison of methods. *Statistics in Medicine, 18*, 1587–1603.

Pearson, E. S., & Hartley, H. O. (1976). *Biometrika tables for statisticians* (Vol. 1). London: Biometrika Trust.

Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M., & Fisher, M. R. (1996). A survey of methods for analyzing clustered binary response data. *International Statistical Review, 64*, 89–118.

Rabe-Hesketh, S., Pickles, A., & Skrondal, A. (2001). GLLAMM: A class of models and a Stata program. *Multilevel Modelling Newsletter, 13*, 17–23.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. (2000). *HLM 5: Hierarchical linear and nonlinear modeling*. Chicago: Scientific Software International.

Reisby, N., Gram, L. F., Bech, P., Nagy, A., Petersen, G. O., Ortmann, J., et al. (1977). Imipramine: Clinical effects and pharmacokinetic variability. *Psychopharmacology, 54*, 263–272.

Seiden, L. S., & Dykstra, L. A. (1977). *Psychopharmacology: A biochemical and behavioral approach*. New York: Van Nostrand Reinhold.

Serretti, A., Lattuada, E., Zanardi, R., Franchini, L., & Smeraldi, E. (2000). Patterns of symptom improvement during antidepressant treatment of delusional depression. *Psychiatry Research, 94*, 185–190.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis*. New York: Oxford University Press.

Strenio, J. F., Weisberg, H. I., & Bryk, A. S. (1983). Empirical Bayes estimation of individual growth curve parameters and their relationship to covariates. *Biometrics, 39*, 71–86.

Sullivan, L. M., Dukes, K. A., & Losina, E. (1999). An introduction to hierarchical linear modelling. *Statistics in Medicine, 18*, 855–888.

van der Leeden, R., Vrijburg, K., & de Leeuw, J. (1996). A review of two different approaches for the analysis of growth data using longitudinal mixed linear models. *Computational Statistics and Data Analysis, 21*, 583–605.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society, 54*, 426–482.

Willner, P. (1985). *Depression: A psychobiological synthesis.* New York: John Wiley.

Wolfinger, R. D. (1993). Covariance structure selection in general mixed models. *Communications in Statistics, Simulation and Computation, 22*, 1079–1106.

Zeger, S. L., & Liang, K.-Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine, 11*, 1825–1839.

Chapter 13

Multilevel Models for School Effectiveness Research

Russell W. Rumberger

Gregory J. Palardy

One of the major topics for social science research is the study of school effectiveness. Beginning with the first large-scale study of school effectiveness in 1966, known as the "Coleman report" (Coleman et al., 1966), literally hundreds of empirical studies have been conducted that have addressed two fundamental questions:

1. Do schools have measurable impacts on student achievement?
2. If so, what are the sources of those impacts?

Studies designed to answer these questions have employed different sources of data, different variables, and different analytic techniques. Both the results of those studies and the methods used to conduct them have been subject to considerable academic debate.

In general, there has been widespread agreement on the first question. Most researchers have concluded that schools indeed influence student achievement. Murnane's (1981) early review captured this consensus well:

> There are significant differences in the amount of learning taking place in different schools and in different classrooms within the same school, even among inner city schools, and even after taking into account the skills and backgrounds that children bring to school. (p. 20)

Another reviewer concluded more succinctly, "Teachers and schools differ dramatically in their effectiveness" (Hanushek, 1986, p. 1159). Despite this general level of agreement on the overall impact of schools, how much impact schools and teachers have is less clear, an issue we address later in this chapter.

It is the second question, however, that has generated the biggest debate. Coleman et al. began this debate with the publication of their report in 1966 by concluding that schools had relatively little impact on student achievement compared to the socioeconomic background of the students who attend them. Moreover, Coleman (1990) found that "the social composition of the student body is more highly related to achievement, independent of the student's own social background, than is any school factor" (p. 119). The publication of the Coleman report also marked the beginning of the methodological debate on how to estimate school effectiveness, a debate that has continued to this day. The Coleman study was criticized on a number of methodological grounds, including the lack of

AUTHORS' NOTE: We would like to acknowledge the helpful comments of David Kaplan and especially Michael Selzter.

controls for prior background and the regression techniques used to assess school effects (Mosteller & Moynihan, 1972).

Since the publication of the original Coleman report, there have been a number of other controversies on sources of school effectiveness and the methodological approaches to assess them. One debate has focused on whether school resources make a difference. In a major review of 187 studies that examined the effects of instructional expenditures on student achievement, Hanushek (1989) concludes, "There is no strong or systematic relationship between school expenditures and student performance" (p. 47). As noted earlier, Hanushek does acknowledge widespread differences in student achievement among schools but does not attribute these differences to the factors commonly associated with school expenditures—teacher experience, teacher education, and class size. A recent reanalysis of the same studies used by Hanushek, however, reaches a different conclusion: "Reanalysis with more powerful analytic methods suggests strong support for at least some positive effects of resource inputs and little support for the existence of negative effects" (Hedges, Laine, & Greenwald, 1994, p. 13).

Another debate has focused on the effectiveness of public versus private schools. Several empirical studies found that average achievement levels are higher in private schools, in general, and Catholic schools, in particular, than in public schools, even after accounting for differences in student characteristics and resources (Bryk, Lee, & Holland, 1993; Chubb & Moe, 1990; Coleman & Hoffer, 1987; Coleman, Hoffer, & Kilgore, 1982). Yet although some (Chubb & Moe, 1990) argue that all private schools are better than public and thus argue for private school choice as a means to improve education, other researchers have argued that Catholic schools, but not other private schools, are both more effective and more equitable than public schools (Bryk et al., 1993). Still other researchers find little or no Catholic school advantage (Alexander & Pallas, 1985; Gamoran, 1996; Willms, 1985). Moreover, it has been suggested that controlling for differences in demographic characteristics may still not adequately control for fundamental and important differences among students in the two sectors (Witte, 1992, p. 389).

Much of the debate about school effectiveness has centered on methodological issues. These issues concern such topics as data, variables, and statistical models used to estimate school effectiveness. Since the research and debate on school effectiveness began almost 50 years ago, new, more comprehensive sources of data and new, more sophisticated statistical models have been developed that have improved school effectiveness studies. In particular, the development of multilevel models and the computer software to estimate them have given researchers more and better approaches for investigating school effectiveness. This chapter reviews some of the major methodological issues surrounding school effectiveness research, with a particular emphasis on how multilevel models can be used to investigate a number of substantive issues concerning school effectiveness.[1]

We will illustrate these issues by conducting analyses of a large-scale national longitudinal study that has been the source of a lot recent research on school effectiveness, the National Education Longitudinal Study of 1988 (NELS). NELS is a national longitudinal study of a representative sample of 25,000 eighth graders begun in 1988. Base year data were collected from questionnaires administered to students, their parents and teachers, and the principal of their school. Follow-up data were collected in 1990, 1992, 1994, and, most recently, in 2000 on a subset of the original sample (Carroll, 1996). Students were also given a series of achievement tests in English, math, science, and history/social studies in the spring of 1988, 1990, and 1992, when most respondents were enrolled in Grades 8, 10, and 12, respectively. In this chapter, we will use a subsample of the NELS data for 14,199 students with valid questionnaires from the 1988, 1990, and 1992 survey years who attended 912 high schools in 1990.[2] The appendix provides descriptive information on the variables in the data set that were used to test the models in this chapter.

We begin this chapter by presenting a conceptual model of schooling that can be used to frame studies of school effectiveness. Next we discuss several issues regarding the selection of data and variables used to test multilevel models. Then we review various types and uses of multilevel models for estimating school effectiveness. Finally, we review techniques for identifying effective schools. For each topic, we will explain some of the important decisions that researchers must make in undertaking school effectiveness studies and how those decisions can influence the outcomes and conclusions of the study.

---

1. Many of the concepts and techniques we discuss can be used to study the effectiveness of other types of organizations, such as hospitals.

2. To generate accurate school-level composition measures, we restricted the sample to respondents who had a valid school ID in 1990, had valid test scores in 1988 and 1990, and attended a high school with at least five students.

13.1. A Conceptual Model of Schooling

To undertake quantitative research on school effectiveness, we should have a conceptual model of the schooling process. A conceptual model can be used to guide the initial design of the study, such as the selection of participants and the collection of data, as well as the selection of variables and the construction of statistical models. Although several different conceptual frameworks have been developed and used in school effectiveness research over the years (e.g., Rumberger & Thomas, 2000; Shavelson, McDonnell, Oakes, & Carey, 1987; Willms, 1992), all have portrayed schooling as a multilevel or nested phenomenon in which the activities at one level are influenced by those at a higher level (Barr & Dreeben, 1983; Willms, 1992). For example, student learning is influenced by experiences and activities of individual students, such as the amount and nature of the homework that they do. But student learning is also influenced by the amount and nature of the instruction that they receive within their teachers' classrooms, as well as by the qualities of the schools they attend, such as school climate and the nature of the courses that are provided. Ignoring or incorrectly specifying these multilevel influences can yield misleading conclusions about their effects on student learning (e.g., Summers & Wolfe, 1977).

In addition to its multilevel nature, the process of schooling can be divided into distinct components. One framework is based on the sociological view of schooling (Tagiuri, 1968; Willms, 1992), which identifies four major dimensions of schooling: ecology (physical and material resources), milieu (characteristics of students and staff), social system (patterns and rules of operating and interacting), and culture (norms, beliefs, values, and attitudes). Another framework is based on an economic model of schooling (e.g., Hanushek, 1986; Levin, 1994), which identifies three major components of schooling: the inputs of schooling—students, teachers, and other resources; the educational process itself, which describes how those inputs or resources are actually used in the educational process; and the outputs of schooling—student learning and achievement.[3]

An example of a conceptual framework based on the economic model is illustrated in Figure 13.1. The framework shows the educational process operating at the three levels of schooling—schools, classrooms, and students. It also identifies two major types of factors that influence the outcomes of schooling: (a) inputs to schools, which consist of structure (size, location), student characteristics, and resources (teachers and physical resources), and (b) school and classroom processes and practices. School inputs are largely "given" to a school and therefore are not alterable by the school itself (Hanushek, 1989). The second set of factors refers to practices and policies that the school does have control over and thus are of particular interest to school practitioners and policymakers in developing indicators of school effectiveness (Shavelson et al., 1987).

13.1.1. Dependent Variables

The framework suggests that school effectiveness research can focus on a number of different educational outcomes. The most common measure of school effectiveness is academic achievement, as reflected in student test scores, which is considered one of the most important outcomes of schooling. Although student academic achievement is affected by the background characteristics of students, research has clearly demonstrated that achievement outcomes are also affected by the characteristics of schools that students attend (Coleman et al., 1982; Gamoran, 1996; Lee & Bryk, 1989; Lee & Smith, 1993, 1995; Lee, Smith, & Croninger, 1997; Witte & Walsh, 1990).

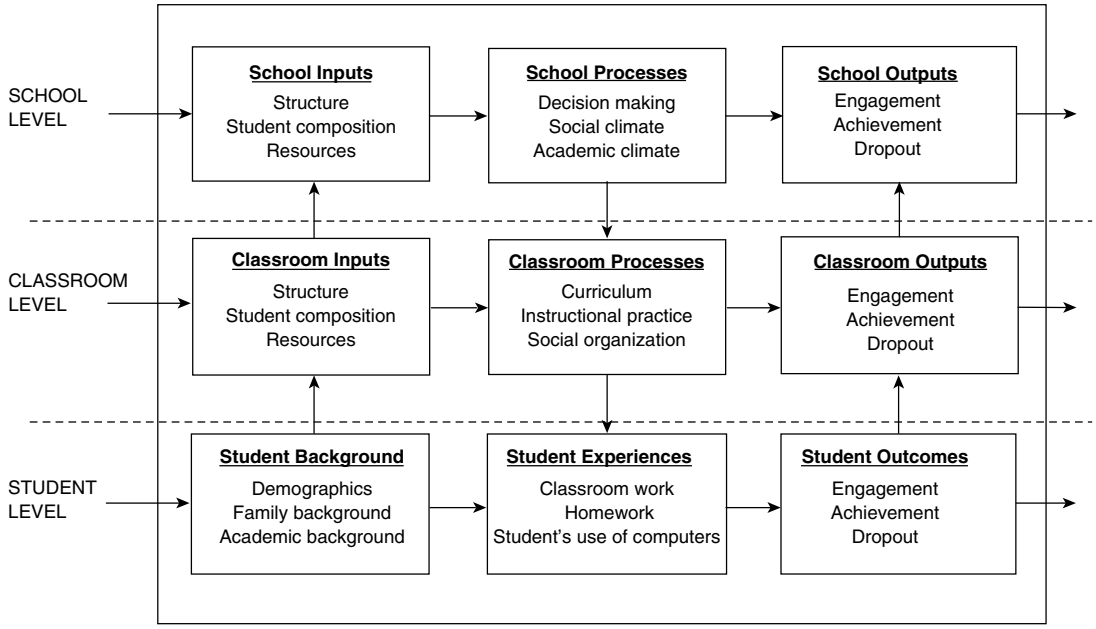
Other student outcomes have also been examined in studies of school effectiveness. One of these is school dropout, which studies have shown is also affected by the characteristics of schools that students attend (Bryk et al., 1993; Bryk & Thum, 1989; Coleman & Hoffer, 1987; McNeal, 1997; Rumberger, 1995; Rumberger & Thomas, 2000). Other studies have examined the impact of school characteristics on absenteeism (Bryk & Thum, 1989), engagement (Johnson, Crosnoe, & Elder, 2001), and social behavior (Lee & Smith, 1993). One reason for examining alternative student outcomes is that schools and school characteristics that are effective in improving student performance in one outcome may not be effective in improving student performance in another outcome (Rumberger & Palardy, 2003b).

13.1.2. Independent Variables

The conceptual framework suggests that several types of variables are valuable in constructing statistical models of school effectiveness. We provide a very brief review of some of these variables.

3. In his landmark study of school effectiveness, sociologist James Coleman employed an input-output model of the schooling process (see Coleman, 1990).

Figure 13.1 A Multilevel Conceptual Framework for Analyzing School Effectiveness



13.1.2.1. Student Characteristics

Research has demonstrated that a wide variety of individual student characteristics are related to student outcomes. These include demographic characteristics, such as ethnicity and gender; family characteristics, such as socioeconomic status and family structure; and academic background, such as prior achievement and retention. These characteristics have been shown to relate to such student outcomes as engagement, achievement (test scores), and dropout (Bryk & Thum, 1989; Chubb & Moe, 1990; Lee & Burkam, 2003; Lee & Smith, 1999; McNeal, 1997; Rumberger, 1995; Rumberger & Palardy, 2003b; Rumberger & Thomas, 2000).

Student characteristics influence student achievement not only at an individual level but also at an aggregate or social level. That is, the social composition of students in a school (sometimes referred to as *contextual effects*) can influence student achievement apart from the effects of student characteristics at an individual level (Coleman et al., 1966; Gamoran, 1992). Studies have found that the social composition of schools predicts school engagement, achievement, and dropout rates, even after controlling for the effects of individual background characteristics of students (Bryk & Thum, 1989; Chubb & Moe, 1990; Jencks & Mayer, 1990; Lee & Smith, 1999; McNeal, 1997; Rumberger, 1995; Rumberger & Thomas, 2000).

13.1.2.2. School Resources

School resources consist of both fiscal resources and the material resources that they can buy. As mentioned earlier, there is considerable debate in the research community about the extent to which school resources contribute to school effectiveness. But there is much less debate that material resources matter, particularly the number and quality of teachers. Yet the exact nature of teacher characteristics that contribute to school effectiveness, such as credentials and experience, is less clear (Goldhaber & Brewer, 1997). Beyond the quality of teachers, there is at least some evidence that the quantity of teachers—as measured by the pupil/teacher ratio—has a positive and significant effect on some student outcomes (McNeal, 1997; Rumberger & Palardy, 2003b; Rumberger & Thomas, 2000).

13.1.2.3. Structural Characteristics of Schools

Structural characteristics, such as school location (urban, suburban, rural), size, and type of control (public, private), also contribute to school performance. Although widespread achievement differences have been observed among schools based on structural characteristics, what remains unclear is whether structural characteristics themselves account for these differences or whether they are related to differences

in student characteristics and school resources often associated with the structural features of schools. As we pointed out earlier, this issue has been most widely debated with respect to one structural feature: the difference between public and private schools. More recently, there has been considerable interest in another structural feature of schools: school size (Lee & Smith, 1997).

13.1.2.4. School Processes

Despite all the attention and controversy surrounding the previous factors associated with school effectiveness, it is the area of school processes that many people believe holds the most promise for understanding and improving school performance. Although most *individual* schools, or at least most public schools, have little control over student characteristics, resources, and their structural features, they can and do have a fair amount of control over how they are organized and managed, the teaching practices they use, and the climate they create for student learning—features referred to as *school processes.* Some researchers have also referred to them as "Type B effects" because, when statistical adjustments are made for the effects of other factors, they provide a better and more appropriate basis for comparing the performance of schools (Raudenbush & Willms, 1995; Willms, 1992; Willms & Raudenbush, 1989). A number of school processes have been shown to affect student achievement, such as school restructuring and various policies and practices that affect the social and academic climate of schools (Bryk & Thum, 1989; Croninger & Lee, 2001; Gamoran, 1996; Lee & Smith, 1993, 1999; Lee et al., 1997; Phillips, 1997; Rumberger, 1995).

13.2. Data and Sample Selection

13.2.1. Data

Like all quantitative studies, school effectiveness research requires suitable data. The conceptual framework discussed earlier shows that student outcomes are influenced by a number of different factors operating at different levels within the educational system, including student factors, family factors, and school factors. Generally, insightful school effectiveness research requires data on all those factors. Moreover, as we discuss below, longitudinal models are useful for addressing certain research questions and required repeated measurements of student outcomes over time. For these reasons, the data requirements of multilevel school effectiveness models can be extensive.

Meeting these extensive data requirements necessitates considerable resources, which are not often available to small-scale studies. For this reason, the federal government has invested in the design and collection of several large-scale longitudinal studies that have been the basis for most school effectiveness studies conducted over the past 40 years or so. Early studies were based on national and some local (state) longitudinal surveys conducted on cohorts of high school students (e.g., see Alexander & Eckland, 1975; Hauser & Featherman, 1977; Jencks & Brown, 1975; Summers & Wolfe, 1977). The U.S. Department of Education conducted the 1972 National Longitudinal Study of the High School Class of 1972, the 1980 High School and Beyond study of 10th- and 12th-grade students, the 1988 National Education Longitudinal Study of 8th graders, and, most recently, the 1998 Early Childhood Longitudinal Study (ECLS) of the kindergarten class of 1998–1999 and the birth cohort of 2000, as well as the 2002 Educational Longitudinal Study of 10th graders.[4] All these survey programs involve large samples of students and schools along with student, parent, teacher, and school surveys as well as specially designed student assessments of academic achievement. One drawback of these studies is that they rarely have adequate classroom-level sample sizes, which makes investigations of teaching and classroom effects problematic. Until recently, all the federal education studies focused on middle and high school students, which has resulted in an inordinate proportion of the school effectiveness research in the past 20 years being directed at middle and high schools. With the availability of ECLS data, that focus seems to be shifting toward elementary schools.

13.2.2. Sample Selection

Once an appropriate set of data is selected, the next step in conducting a school effectiveness study is to select an appropriate sample. In addition to selecting a set of data and a sample based on the types of research questions that are to be addressed, two other issues are important to consider: missing data and sampling bias.

4. For further information, visit the National Center for Education Statistics Web site at http://nces.ed.gov/surveys/.

13.2.2.1. Missing Data

Missing data are a reality in social research and especially problematic in longitudinal analyses in which attrition tends to exacerbate the problem. In panel studies, attrition may occur when families move or students drop out between waves or students cannot be located for some other reason at the follow-up survey. Another situation is nonresponse on certain items. Deciding how to deal with missing values is a common dilemma. Perhaps the most widely used approach is to omit cases with missing data, although the general consensus is that deletion is only an appropriate course of action when data are missing completely at random (see Little & Rubin, 1987, for a detailed treatment of types of "missingness" and remedies). Deletion of cases in other situations can bias the sample and parameter estimates. For that reason, it is important to consider alternatives to deletion.

13.2.2.2. Sampling Bias

Sampling bias arises when some part of the target population is inadequately represented in the sample. This problem is often an outcome of deleting cases with missing data and, as mentioned above, can lead to distorted results.[5] Other times, researchers may choose to exclude some valid cases for one reason or another. For example, dropouts and mobile students may be excluded from a school effectiveness evaluation analysis because their achievement growth cannot be attributed to a single school. Whether cases have missing data or are being considered for removal for another reason, deletion is an option that should only be considered after establishing that those cases do not differ systematically from the rest. In general, the larger the percentage of cases being excluded, the greater the potential for selection bias. However, to be safe against sampling bias, cases with missing values should not be deleted but rather handled using an appropriate missing value routine.

As the title of this chapter suggests, school effectiveness research generally necessitates a multilevel model because students are nested in classrooms and schools. The previous discussion of selection bias focused on omission of student cases. Omissions at the student level can also bias the school-level sample. A simple example of this is the effect of deleting students with missing achievement data. If the omitted cases have lower achievement levels than the retained cases, mean achievement estimates at the school level will also be biased. Furthermore, omitting cases at the student level decreases the average number of students per school, which generally reduces the reliability of the fixed and random coefficients in the model.

5. The problem can also arise due to sampling techniques often used in collecting multilevel longitudinal studies, such as the large-scale federal studies mentioned earlier. Such studies typically provide sampling weights that researchers can use to produce accurate estimates of population parameters (e.g., see Carroll, 1996).

13.3. Using Multilevel Models to Address Research Questions

A wide range of multilevel models can and have been used to conduct school effectiveness research. The choice of models depends both on the questions the investigator wishes to answer and on the data available to answer them. Two key aspects of the data are relevant in selecting models: whether the data represent measures at a single point in time (cross-sectional) or multiple points in time (longitudinal) and whether the outcome measures are continuously distributed (e.g., standard test scores) or categorical (e.g., dropout rates). In this section, we review a number of different models. We group the models by the types of dependent or outcome variables used in the models and whether the data are cross-sectional or longitudinal:

- achievement (cross-sectional) models with continuous outcomes,
- achievement growth (longitudinal) models with continuous outcomes,
- models with categorical outcomes.

For each group of models, we pose a series of research questions and the models most suited to address them. Then we illustrate the procedures for using them with the sample NELS data.

13.3.1. Achievement Models

The most commonly used type of multilevel model for school effectiveness is one in which the dependent variable is student achievement at a single point in time. One reason for the popularity of these models is that they only require one round of data collection, which is both easier and less expensive than multiple rounds of data collection found in longitudinal studies. Moreover, even though there are some inherent limitations in these models, as we discuss below, they can still be used to address a wide range of research questions.

Student achievement models typically specify two distinct components or submodels: (a) models for

student-level outcomes within schools, known as *within-school models,* and (b) models for school-level outcomes, known as *between-school models,* in which the parameters from the within-school model serve as dependent variables in the between-school model. Because the within-school model may contain a number of parameters, each parameter produces its own between-school equation. In most applications, a series of models are estimated that begin with relatively simple models and then add parameters to develop more complete models. Each model is useful for addressing particular types of research questions, so school effectiveness studies typically employ a number of distinct models.

13.3.1.1. Do Schools Make a Difference?

This is the most fundamental research question in school effectiveness research that focuses on how much of the variation in student achievement can be attributed to the schools that students attend. Coleman was the first researcher to address this question, and he did it by partitioning the total variation in student achievement into two components: One component consisted of the variation in individual test scores around their respective school means, and the other component consisted of the variation in school means around the grand mean for the entire sample (Coleman, 1990, p. 76). Coleman found that schools only accounted for a small amount of the total variation in student test scores, ranging from 5% to 38% among different grade levels, ethnic groups, and regions of the country (Coleman, 1990, p. 77).

This research question can easily be addressed using a multilevel *unconditional* or *null model.* The first model has no predictor variables in either the within-school or between-school model and is known as a null or one-way ANOVA model:

$$\text{Level 1 model: } Y_{ij} = \beta_{0j} + r_{ij}, r_{ij} \sim N(0, \sigma^2).$$
$$\text{Level 2 model: } \beta_{0j} = \gamma_{00} + \mu_{0j}, \mu_{0j} \sim N(0, \tau_{00}).$$
$$\text{Combined model: } Y_{ij} = \gamma_{00} + \mu_{0j} + r_i.$$

In this case, the Level 1 model represents the achievement of student i in school j as a function of the average achievement in school j (β_{0j}) and a student-level error term (r_{ij}), and the Level 2 model represents the average achievement in school j as a function of the grand mean of all the school means (γ_{00}) and a school-level error term (μ_{0j}). In addition to providing an estimate of the one fixed effect, the grand mean for achievement (γ_{00}), the model also provides estimates for the student-level (σ^2) and at the school-level (τ_{00}) variance components, which can be used to determine how much of the total variance is accounted for by students and schools.

We can illustrate the usefulness of the null model with the NELS data using 10th-grade math test scores as the dependent variable. The estimated parameters from this model are shown in Table 13.1 (column 1).[6] The estimate for the grand mean of the mean math achievement ($\hat{\gamma}_{00}$) among the sample of 912 high schools is 50.85, which is very close to the actual mean for the students in the sample (see appendix). The estimated values for the two variance components can be used to partition the variance in student math scores between the student and school levels, as shown as follows:

Student-level variance ($\hat{\sigma}^2$) : 73.88

School-level variance ($\hat{\tau}_{00}$) : 24.12

Total variance: 98.00

Proportion of variance at school level : .25

The results show that 25% of the total variance is at the school level, which suggests that schools do indeed contribute to differences in student math scores. This result is within the range that Coleman et al. found in their 1966 study[7] and the range found in other recent studies of student achievement using similar models (e.g., Lee & Bryk, 1989; Rumberger & Willms, 1992). Once the total variance is decomposed into its student and school components, subsequent models can be constructed to explain each component, much the way single-level regression models are used to explain variance.

13.3.1.2. To What Degree Does Mean Achievement Vary Across Schools?

This is a related question that allows the researcher to determine the extent of the variation in average school achievement among schools. This question can also be addressed by using the parameter estimates from the unconditional model to calculate a 95% confidence interval, referred to as a range of plausible values, under the assumption that the school-level variance

6. Because of space considerations, we only provide estimates of fixed and random effects. Raudenbush and Bryk (2002) also suggest that researchers examine other statistics, including reliability.

7. Coleman (1990) provides a summary of the findings in Table 3.22.1 on page 77.

Table 13.1 Parameter Estimates for Alternative Multilevel Math Achievement Models

	Null Model (1)	*Means-as-Outcomes Model 1 (2)*	*Means-as-Outcomes Model 2 (3)*	*One-Way ANCOVA Model (4)*	*Random-Coefficient Model (5)*	*Intercepts- and Slopes-as-Outcomes Model (6)*
Fixed effects						
Model for school mean achievement (β_0)						
INTERCEPT (γ_{00})	50.85**	49.93**	50.85**	50.96**	50.84**	50.84**
	(0.18)	(0.17)	(0.12)	(0.12)	(0.18)	(0.11)
MEANSES (γ_{01})			8.11**			8.11**
			(0.25)			(0.25)
CATHOLIC (γ_{02})		3.22**	−0.21			−0.23
		(0.62)	(0.43)			(0.43)
PRIVATE (γ_{03})		9.35**	0.76			0.73
		(0.64)	(0.53)			(0.53)
Model for SES achievement slope (β_1)						
INTERCEPT (γ_{10})				4.95**	4.22**	4.51**
				(0.10)	(0.12)	(0.13)
MEANSES (γ_{11})						1.09**
						(0.30)
CATHOLIC (γ_{12})						−1.78**
						(0.55)
PRIVATE (γ_{13})						−3.55**
						(0.55)
Variance components						
Within school (Level 1) (σ^2)	73.88	73.91	73.95	66.55	65.88	65.97
Between school (Level 2)						
School means (τ_{00})	24.12**	17.33**	5.35**	9.00**	24.75**	5.93**
SES achievement slopes (τ_{11})					1.34**	0.82*
Proportion explained						
School means		.28	.77	.63		.75
SES achievement slopes						.29

NOTE: SES = socioeconomic status; PRIVATE = private schools; CATHOLIC = Catholic schools; MEANSES = mean socioeconomic status. $^*p < .05$; $^{**}p < .01$.

is normally distributed (Raudenbush & Bryk, 2002, p. 71):

$$\begin{aligned}\text{Range of plausible values} &= \hat{\gamma}_{00} \pm 1.96\,(\hat{\tau}_{00})^{1/2} \\ &= 50.85 \pm 1.96\,(24.12)^{1/2} \\ &= (41.23, 60.47).\end{aligned}$$

These results indicate a substantial range in average achievement among high schools, with average achievement 50% higher in the highest performing (97.5th percentile) compared to the lowest performing (2.5th percentile) high schools.

13.3.1.3. What School Inputs Account for Differences in School Outputs?

Another fundamental research question on school effectiveness concerns the relationship between school inputs and school outputs. Again, this is one of the main questions that Coleman et al. (1966) addressed in their landmark study (summarized in Coleman, 1990, p. 2), and it continues to have importance for policy initiatives designed to address disparities in school inputs.

This research question can be addressed using a second type of multilevel model, known as a *means-as-outcomes model.* This model attempts to explain school-level variance, but not student-level variance, by adding school-level predictors to the model, as shown in the following example in which we add two indicator or dummy variables for school sector:

$$\text{Level 1 model: } Y_{ij} = \beta_{0j} + r_{ij}.$$

$$\text{Level 2 model: } \beta_{0j} = \gamma_{00} + \gamma_{01}\text{CATHOLIC}_j + \gamma_{02}\text{PRIVATE}_j + u_{0j}.$$

In this example, there are three fixed effects: one for the mean math achievement in public high schools (γ_{00}), one for the mean achievement difference

between public and Catholic schools (γ_{01}), and one for the mean achievement difference between public and private, non-Catholic schools (γ_{02}). The results of this model (see Table 13.1, column 2) show that mean student math achievement is 49.93 in public schools and averages more than 3 points higher in Catholic schools and more than 9 points higher in private schools. Both predictor variables are statistically significant.[8]

With these two predictors in the model, the school-level variance (τ_{00}) is now a conditional variance or the variance that remains after controlling for the effects of school sector (CATHOLIC, PRIVATE). Consequently, it is generally smaller than the variance in the unconditional model. The difference in the two variance estimates can be used to determine how much of the unconditional variance is explained by the model containing these two predictors:

$$\begin{aligned}&\text{Proportion of variance explained}\\&= [\hat{\tau}_{00}(\text{Model 1}) - \hat{\tau}_{00}(\text{Model 2})]/\hat{\tau}_{00}(\text{Model 1})\\&= [24.12 - 17.33]/24.12\\&= .28.\end{aligned}$$

The results indicate that 28% of the total variance between schools in mean math achievement is accounted for by the two school sector variables.

Next we added a third predictor to the school-level model, mean socioeconomic status of students in each school (MEANSES_j):

$$\text{Level 2 model: } \beta_{0j} = \gamma_{00} + \gamma_{01}\text{MEANSES}_j + \gamma_{02}\text{CATHOLIC}_j + \gamma_{03}\text{PRIVATE}_j + u_{0j}.$$

In this example, there are four fixed effects: the mean math achievement in public high schools, where MEANSES is zero (γ_{00});[9] the effect of school mean socioeconomic status (SES) on mean math achievement (γ_{01}); the mean achievement difference between public and Catholic schools, holding constant school mean SES (γ_{02}); and the mean achievement difference between public and private, non-Catholic schools, holding constant school mean SES (γ_{03}). The results of this model (see Table 13.1, column 3) show that MEANSES has a large and statistically significant effect on mean math achievement ($\hat{\gamma}_{01} = 8.11$, $p < .01$)—a one standard deviation increase in MEANSES increases mean test scores by 4.22 ($8.11 \times .52$) points. After controlling for school mean SES, the coefficients for Catholic and private schools are no longer statistically significant. This example illustrates the importance of correctly specifying a model to yield valid and unbiased results. Although this issue applies to all statistical models, it is particularly important in multilevel models because the researcher must draw on a broader array of research literature pertaining to both individual and school determinants of student achievement to correctly specify models at each level of analysis.

This model explains 77% of the school-level variance. In other words, only three predictors explain the majority of the variability in average achievement among schools.[10]

13.3.1.4. What Difference Does the School a Child Goes to Make in the Child's Achievement?

This is another fundamental question that Coleman (1990, p. 2) addressed in his landmark study and one particularly important to parents. Parents are often interested in selecting a school that will improve their child's academic achievement. They are also aware that the average achievement varies widely among schools, in part because schools, state education agency Web sites, and newspapers often report such information. Yet, all the variance in student achievement at the school level cannot be attributed to the effects of schools. Some of that variance is due to the individual background characteristics of the students, which affect student outcomes no matter where they attend school.

This research question can be addressed using another type of multilevel model, known as a *one-way ANCOVA model.* One helpful technique to control for the effects of student background characteristics in this model is through "centering" student-level predictors around their grand or sample mean.

A simple illustration of this model is shown in the following model, in which a single student-level predictor, SES, is introduced and centered on the grand mean:

$$\begin{aligned}&\text{Level 1 model: } Y_{ij} = \beta_{0j} + \beta_{1j}(\text{SES}_{ij} - \overline{\text{SES}}_{..}) + r_{ij}.\\&\text{Level 2 model: } \beta_{0j} = \gamma_{00} + u_{0j}.\\&\qquad\qquad\qquad\quad \beta_{1j} = \gamma_{10}.\end{aligned}$$

8. Hypothesis testing for both fixed and random effects is explained in detail in Raudenbush and Bryk (2002, pp. 56–65). The p-values shown in Tables 13.1 and 13.2 are from single-parameter tests, which are based on t-tests for fixed effects and chi-square tests for the variance components.

9. This is extremely close to the sample mean of .01.

10. In fact, mean SES alone explains 77% of the variance, which is why Coleman concluded that the social composition of the school is the most important school input.

Grand-mean centering alters the meaning of the intercept term (β_{0j}). Instead of representing the actual mean achievement of students in each school, it now represents the expected achievement of a student whose background characteristics are equal to the grand mean of all students in the larger sample of students (Raudenbush & Bryk, 2002, p. 33). In other words, the school means are adjusted for differences in the background characteristics of the students attending them and now represent the expected achievement of an "average" student. In this example, there are two fixed effects: one for the school mean of the expected math achievement for students with mean SES (γ_{00}) and one for the predicted effect of student SES on math achievement (γ_{10}).[11] In addition, the equation for the student-level predictor is "fixed" at Level 2 in this model because no random school effect is specified, which assumes that the effect of student SES does not vary among schools (like a classical ANCOVA model)—an assumption that we test below. In this case, the student-level variance (σ^2) represents the residual variance of student achievement after controlling for student SES, and the school-level variance (τ_{00}) represents the variance among schools in adjusted school means.

The estimated parameters of this model (see Table 13.1, column 4) show that student SES is a powerful predictor of academic achievement ($\hat{\gamma}_{10} = 4.95$, $p < .01$). A one standard deviation increase in student SES implies a 4-point ($4.95 \times .81$) increase in student achievement. This single predictor, grand-mean centered, explains 63% of the school-level variance. In other words, almost two thirds of the observed variance in mean math achievement among schools can be explained by differences in the SES background of the students who attend them. The magnitude of this impact can also be illustrated by calculating the adjusted range of plausible values:

$$\begin{aligned}\text{Range of plausible values} &= \hat{\gamma}_{00} \pm 1.96(\hat{\tau}_{00})^{1/2} \\ &= 50.85 \pm 1.96(9.00)^{1/2} \\ &= (45.08, 56.84).\end{aligned}$$

These results indicate that for a student from an average SES background, his or her expected achievement would be about 26% higher in the highest performing compared to the worst-performing high school. Although such a difference is only about half of the range in the overall means shown earlier, it may still be considered meaningful.

13.3.1.5. Do the Effects of Student Background Characteristics Vary Among Schools?

In the preceding example, we assumed that the effects of the student-level predictors were the same across schools. In most cases, the investigator should test this assumption by first specifying them as random at the school level. If the variance of the random effect is not significantly different from zero, the researcher can "fix" the predictor by removing the random effect. If the variance is significantly different from zero, the researcher can then try to explain the variance by adding school-level predictors much the same way that school-level predictors are added to the intercept term.

This type of multilevel model is known as a *random-coefficient model.* To derive accurate estimates of all the variance parameters in this type of model, we must use a different form of centering known as *group-mean centering* (see Raudenbush & Bryk, 2002, pp. 143–149). In this case, the student-level predictors are centered at the mean for the students in their respective schools, and, by doing so, the intercept term (β_{0j}) represents the unadjusted mean achievement for the school (Raudenbush & Bryk, 2002, p. 33).[12]

To illustrate this model, we estimated a model similar to the one above, but SES was group-mean centered, and a random term was added to its Level 2 equation:

$$\text{Level 1 model: } Y_{ij} = \beta_{0j} + \beta_{1j}(\text{SES}_{ij} - \overline{\text{SES}}_{\cdot j}) + r_{ij}.$$

$$\text{Level 2 model: } \beta_{0j} = \gamma_{00} + u_{0j}.$$

$$\beta_{1j} = \gamma_{10} + u_{1j}.$$

In this example, there are two fixed effects—the grand mean of the mean math achievement among schools (γ_{00}) and the mean of the SES achievement slope among schools (γ_{10})—and three random effects: the residual variance of student achievement after controlling for student SES (σ^2), the variance in the average math achievement among schools (τ_{00}), and the variance in the SES achievement slopes among schools (τ_{11}). The results from this model (see Table 13.1, column 5) show similar parameter estimates for mean achievement and student SES compared to the previous ANCOVA model (column 4), but now the variance parameter for the intercept term is similar to that of the unconditional model (column 1), and there is a variance estimate for the SES equation,

11. In cases in which student characteristics affect educational outcomes at both the individual and school levels, as we discuss below, then the student-level predictors in this model produce biased estimators of the within-school effects of those characteristics (see Raudenbush & Bryk, 2002, pp. 135–139).

12. In addition, group-mean centering provides an unbiased estimator of the student-level effects (see Raudenbush & Bryk, 2002, pp. 135–139).

which in this case is statistically significant.[13] This suggests that the effects of SES on achievement, sometimes referred to as the SES achievement slope, vary among schools. The magnitude of this variation can be illustrated by calculating a range of plausible values:

$$\begin{aligned}\text{Range of plausible values} &= \hat{\gamma}_{10} \pm 1.96\,(\hat{\tau}_{11})^{1/2} \\ &= 4.22 \pm 1.96\,(1.34)^{1/2} \\ &= (1.95, 6.49).\end{aligned}$$

The results suggest that the effects of student SES on achievement are more than three times as great in some high schools as in other high schools, which suggests that some schools are more equitable in that they attenuate the effects of student background characteristics on achievement.

13.3.1.6. How Effective Are Different Kinds of Schools?

One of the most important policy questions concerns measuring school effectiveness. Policymakers are interested in identifying effective and ineffective schools to recognize the effective schools and intervene in the ineffective schools. But this is easier said than done. Schools should only be accountable for the factors that they have control over. In most cases, at least in the public sector, schools do not have control over the types of students who are enrolled in them (as well as other types of school inputs). As we demonstrated earlier, the background characteristics of students explain much of the variation in mean achievement among schools. In addition, student background characteristics can affect student outcomes at the school level, which are known as compositional or contextual effects (Gamoran, 1992). For example, the average SES of a school may have an effect on student achievement above and beyond the individual SES levels of students in that school. In other words, a student attending a school where the average SES of the student body is low may have lower achievement outcomes than a student from a similar background attending a school where the average SES of the student body is high. Data from the 2000 National Assessment of Educational Progress confirm this: Low-income students attending schools with less than 50% low-income students had higher scores in the fourth-grade math exam than middle-income students attending schools with more than 75% low-income students (U.S. Department of Education, 2003, p. 58).

School effectiveness may be judged not simply by determining which schools have higher average achievement, after controlling for certain inputs, but also by how successful they are in attenuating the relationship between student background characteristics and achievement, as we suggested earlier. Coleman (1990, p. 2) argued that there is another important question about school effectiveness: *How much do schools overcome the inequalities with which children come to school?* For example, some earlier studies found that not only did Catholic schools have higher achievement than public schools, even after controlling for differences in the average SES of students, but the relationship between student SES and achievement was lower, meaning that disparities between high and low SES students was lower (Byrk et al., 1993; Lee & Bryk, 1989). In other words, Catholic schools were found to be more equitable.

A type of multilevel model that can be used to assess both questions on school effectiveness is referred to as a *means- and slopes-as-outcomes model.* This model incorporates school-level predictors in both the intercept and random slopes equations. To generate accurate parameter estimates in these types of models, one must introduce a common set of school-level predictors in all the Level 2 equations (see Raudenbush & Bryk, 2002, p. 151). In addition, to disentangle the individual and compositional effects of student-level predictors, one should include school-level means of all the student-level predictors in the model (see Raudenbush & Bryk, 2002, p. 152).

An example of this model is the following:

$$\text{Level 1 model: } Y_{ij} = \beta_{0j} + \beta_{1j}(\text{SES}_{ij} - \overline{\text{SES}}_{\cdot j}) + r_{ij}.$$

$$\begin{aligned}\text{Level 2 model: } \beta_{0j} &= \gamma_{00} + \gamma_{01}\text{MEANSES}_j \\ &\quad + \gamma_{02}\,\text{CATHOLIC}_j + \gamma_{03}\,\text{PRIVATE}_j + u_{0j}.\end{aligned}$$

$$\begin{aligned}\beta_{1j} &= \gamma_{10} + \gamma_{11}\text{MEANSES}_j \\ &\quad + \gamma_{12}\text{CATHOLIC}_j + \gamma_{13}\text{PRIVATE}_j + u_{1j}.\end{aligned}$$

In this example, there are eight fixed effects and three random effects. The meaning of the student-level random effect and the effects for the model for school means (β_{0j}) are similar to those described earlier. In the model for the SES achievement slope (β_{1j}), there are now four fixed effects: the SES achievement slope in public high schools, where the school mean SES is zero (γ_{10}); the effect of school mean SES on the SES achievement slope (γ_{11}); the difference between public and Catholic schools in the SES achievement slope, holding constant school mean SES (γ_{12}); and the difference between public and private, non-Catholic schools on the SES achievement slope,

13. The SES achievement slope in this model is lower than in the ANCOVA model (4.22 vs. 4.95), which suggests that there are both student-level and school-level effects of SES, something we confirm in the next model.

holding constant school mean SES (γ_{13}). In this model, the variance (τ_{11}) now represents the residual variance in the SES achievement slopes after controlling for school sector and school SES.

The estimated parameters from this model (see Table 13.1, column 6) yield several important conclusions about differences in school effectiveness among public, private, and Catholic schools. First, unlike the earlier reported studies, the average achievement at private and Catholic schools is not significantly higher than the average achievement at public schools after controlling for the effects of school mean SES. Second, consistent with earlier studies, the effects of student SES on achievement are higher in high-SES schools than lower SES schools and lower in Catholic and private schools than in public schools. For example, the effect of student SES is 4.51 at public schools, with a school mean SES equal to zero; at a Catholic school, it is 2.73 (= 4.51 − 1.78), and at a private school, it is 0.96 (= 4.51 − 3.55). Third, the SES of students affects school achievement at both the individual and schools levels—that is, student SES has both individual and compositional or contextual effects on student achievement.[14]

13.3.2. Achievement Growth Models

Achievement models only examine the relationship between student outcomes and predictor variables at discrete points in time. A drawback of this approach is that it fails to account for the fact that an unknown proportion of the achievement that students demonstrate at a particular point in a school is due to learning that took place prior to their arrival at that school. Although this problem can be partially corrected by including measures of prior achievement in the model, using an outcome measure that isolates the student learning that occurred while students where actually attending that school is a far better choice.

Growth models are a special class of multilevel model in which repeated measurements are collected for each individual in the sample (Singer & Willett, 2003). Growth models are useful for understanding mean patterns of change as well as individual differences in those patterns. Growth models include two or more level of analyses. A growth trajectory is estimated for each individual at Level 1 of the multilevel model, and between-individual differences in the change pattern are estimated at Level 2.[15] A multilevel achievement growth model for schools will typically include three levels of analysis (e.g., Lee, Smith, & Croninger, 1997; Seltzer, Choi, & Thum, 2003). A special situation arises when there is a need to estimate teacher or classroom effects in addition to school effects. Typically, students will have been members of more than one classroom in a growth model, which means that they are not strictly nested within classrooms over time. In this scenario, a cross-classified random-effects model can be used to partition the variance in student learning into both classroom and school components (see Raudenbush & Bryk, 2002, chap. 12).

In this section, we discuss two different ways of specifying and estimating achievement growth models: one using the multilevel regression models similar to the ones we discussed above and the other using multilevel latent growth curves. As we did earlier, we discuss these models in relation to the types of research questions about school effectiveness they can be used to address.

13.3.2.1. Multilevel Growth Models

We begin with a Level 1 model for individual growth, where repeated, within-student measurements of achievement are modeled as a function time. The simplest model depicts a linear growth trajectory, although piecewise linear and polynomial terms can be added to examine nonlinear trends if there are sufficient observations (see Raudenbush & Bryk, 2002, chap. 6). A Level 1 linear growth model can be written as follows:

$$\text{Level 1 model: } Y_{tij} = \pi_{0ij} + \pi_{1ij} a_{tij} + e_{tij},$$
$$e_{tij} \sim N(0, \sigma^2),$$

where Y_{tij} represents the achievement outcome measure of student i in school j at time t; π_{0ij} and π_{1ij} represent, respectively, the initial status (when time equals zero) and rate of change for student i in school j; a_{tij} is a measure of time; and e_{tij} is a random error term. For the NELS data, we coded time 0, 0.5, and 1 for 1988, 1990, and 1992, respectively. Coding the time variable this way offers two advantages in

14. As Raudenbush and Bryk (2002) point out, there is more than one way to disentangle the individual and compositional effects of student background characteristics, with the choice of method depending on whether the analyst wishes to test for random slopes (pp. 139–149). In this example, the conditional individual effect of SES (i.e., expected within-school effects on achievement in public schools with MEANSES equal to zero) is 4.51, and the compositional effect of SES = 8.11 − 4.51 = 3.6.

15. One of the advantages of this approach is that individuals only have to have a single observation to be included in the analysis (Raudenbush & Bryk, 2002, p. 199).

interpreting the results. First, the intercept can be interpreted as an approximation of student achievement level upon entering high school since the first wave of testing was conducted in the spring of 1988, just before most students entered high school. Second, the slope represents achievement gains during the 4-year period of high school.[16]

13.3.2.1.1. Do schools make a difference in student learning? This question is similar to the one addressed earlier, except here we are interested in whether schools make a difference in student learning, not simply student achievement. This question can be addressed with a fully unconditional model with no predictors at Levels 2 and 3:

Level 2 model:

$$\pi_{0ij} = \beta_{00j} + r_{0ij}, r_{0ij} \sim N(0, \tau_{\pi 00})$$
$$\pi_{1ij} = \beta_{10j} + r_{1ij}, r_{1ij} \sim N(0, \tau_{\pi 11}).$$

Level 3 model:

$$\beta_{00j} = \gamma_{000} + u_{00j}, u_{00j} \sim N(0, \tau_{\beta 00}).$$
$$\beta_{10j} = \gamma_{100} + u_{10j}, u_{10j} \sim N(0, \tau_{\beta 11}).$$

Note that Level 2 here is equivalent to Level 1 in the multilevel cross-sectional model. In this model, there are two fixed effects: one for initial status or achievement (γ_{000}) and one for achievement growth (γ_{100}), with the latter being of primary focus. There are also five random effects, which can be used to partition the variance in both initial achievement and achievement growth into their within- and between-school components.

We can illustrate this technique with the NELS data and math test scores in Grades 8, 10, and 12 as the dependent variables. The estimated parameters from this model are shown in Table 13.2 (column 1). The results indicate that the average math score for students entering high school ($\hat{\gamma}_{000}$) is 45.87 points and that students increase their math scores ($\hat{\gamma}_{100}$) by an average of 8.76 points over 4 years. The estimated values for the variance components can be used to partition the variance in both initial status and learning between students and schools as we did earlier.[17] The results show that about one quarter of the total variance in both initial achievement and achievement growth occurs at the school level in this sample of data (see Table 13.3).

The proportion of variance in achievement growth at the school level is similar to the proportion we calculated earlier for 10th-grade achievement. In another study using this same data set, we found the proportion varied by subject area—ranging from a low of 20% in reading to a high of 60% in history (Rumberger & Palardy, 2003a). One study of elementary schools in Chicago found that almost 60% of the variance in achievement growth occurred at the school level (Raudenbush & Bryk, 2002, p. 239). In general, these studies suggest schools account for a sizable amount of variance in both student achievement and achievement growth.[18]

To illustrate the usefulness of the growth outcome and to draw comparisons between it and the achievement outcome, we estimate a series of achievement growth models to address the questions we posed above for the achievement models. The results are shown in Table 13.2. Because of space limitations, we will not discuss all of the results of these models, but instead we will point out where the results of these models yield different answers to the set of questions about school effectiveness.

For example, consider the following question: *Do the effects of student background characteristics vary among schools?* To address this question, we specify a random-coefficient model similar to the one estimated earlier, where student SES is group-mean centered:

Level 2 model:

$$\pi_{0ij} = \beta_{00j} + \beta_{01j}(\mathrm{SES}_{ij} - \overline{\mathrm{SES}}_{.j}) + r_{0ij}.$$
$$\pi_{1ij} = \beta_{10j} + \beta_{11j}(\mathrm{SES}_{ij} - \overline{\mathrm{SES}}_{.j}) + r_{1ij}.$$

Level 3 model:

$$\beta_{00j} = \gamma_{000} + u_{00j}.$$
$$\beta_{01j} = \gamma_{010} + u_{01j}.$$
$$\beta_{10j} = \gamma_{100} + u_{10j}.$$
$$\beta_{11j} = \gamma_{110} + u_{11j}.$$

The achievement model estimated earlier (see Table 13.1, column 5) found that the effect of student

16. An alternative scheme is 0, 2, 4, which also sets the intercept as achievement upon entering high school, but now the growth parameter is scaled so that it is interpreted as achievement gains per year.

17. The variance components can also be used to examine the correlation between initial status and growth at both the individual and school levels (see Raudenbush & Bryk, 2002, p. 240). In this example, the correlation at the student level is .34, and the correlation at the school level is .39, which suggests that students who begin high school with higher math achievement have higher achievement growth rates than lower achieving students.

18. Because of the size and heterogeneity of course offerings (tracking) found in high schools, schools may account for a great proportion of the variance at the elementary level compared to the secondary level.

Table 13.2 Parameter Estimates for Alternative Multilevel Math Achievement Growth Models

	Null Model (1)	*Means-as-Outcomes Model 1 (2)*	*Means-as-Outcomes Model 2 (3)*	*Random-Coefficient Model (4)*	*One-Way ANCOVA Model 1 (5)*	*One-Way ANCOVA Model 2 (6)*
Fixed effects						
Model for initial status (π_{0ij})						
Model for school mean of initial status (β_{00j})						
INTERCEPT (γ_{000})	45.87**	45.10**	45.82**	45.87**	45.97**	45.92**
	(0.16)	(0.15)	(0.10)	(0.16)	(0.11)	(0.10)
MEANSES (γ_{001})			7.17**			3.60**
			(0.22)			(0.25)
CATHOLIC (γ_{002})		2.18**	−0.86*			−0.87*
		(0.54)	(0.39)			(0.39)
PRIVATE (γ_{003})		8.21**	0.61			0.33
		(0.62)	(0.50)			(0.50)
Model for within-school relationship between SES and initial status (β_{01j})						
INTERCEPT (γ_{010})				3.58**	4.19**	3.59**
				(0.10)	(0.09)	(0.10)
Model for 4-year learning rate (π_{1ij})						
Model for school mean of 4-year learning rate (β_{10j})						
INTERCEPT (γ_{100})	8.76**	8.49**	8.66**	8.76**	8.79**	8.69**
	(0.08)	(0.08)	(0.08)	(0.18)	(0.07)	(0.11)
MEANSES (γ_{101})			1.65**			0.53**
			(0.16)			(0.25)
CATHOLIC (γ_{102})		1.84**	1.15**			1.15**
		(0.36)	(0.34)			(0.37)
PRIVATE (γ_{103})		2.15**	0.40			0.40
		(0.33)	(0.37)			(0.37)
Model for within-school relationship between SES and 4-year learning rate (β_{01j})						
INTERCEPT (γ_{110})				1.12**	1.28**	1.12**
				(0.08)	(0.07)	(0.08)
Variance components						
Within students (Level 1) (σ^2)	8.18	8.18	8.18	8.19	8.20	8.19
Within school (Level 2)						
Initial status ($\tau_{\pi 00}$)	49.81**	49.83**	49.86**	44.04**	44.62**	44.50**
Four-year learning rate ($\tau_{\pi 11}$)	13.05**	13.05**	13.04**	12.23**	12.47**	12.46**
Between school (Level 3)						
Initial status ($\tau_{\beta 00}$)	19.11**	13.98**	4.66**	19.59**	7.75**	5.03**
SES/initial status ($\tau_{\beta 01}$)				0.98*		
Four-year learning rate ($\tau_{\beta 10}$)	4.00**	2.91**	2.39**	3.43**	2.57**	2.42*
SES/4-year learning rate ($\tau_{\beta 11}$)				0.55		
Proportion school-level variance explained						
Initial status		.27	.76		.59	.74
Four-year learning rate		.27	.40		.36	.40

NOTE: SES = socioeconomic status; PRIVATE = private schools; CATHOLIC = Catholic schools; MEANSES = mean socioeconomic status.
*$p < .05$; **$p < .01$.

SES (group-mean centered) on math achievement ($\gamma_{10} = 4.22$, $p < .01$) varied significantly between schools ($\tau_{11} = 1.34$, $p < .01$). As a result, several predictors were added to the model, and it was found that the effect of SES on math achievement was lower in Catholic and private schools—that is, the distribution of achievement appeared to be more equitable in Catholics schools. In the achievement growth model (see Table 13.2, column 4), however, the effect of student SES (group-mean centered) on math

Table 13.3 Decomposing the Variance in a Linear Math Achievement Growth Model

	Initial Status	*Achievement Growth*
Student-level variance (%)	49.81	13.05
School-level variance (%)	19.11	4.00
Total student-level and school-level variance (%)	68.92	17.05
Proportion of variance at school level	.28	.24

achievement growth ($\gamma_{110} = 8.76, p < .01$) did not vary significantly between schools ($\hat{\tau}_{\beta 11} = .055$, $p \geq .05$).[19]

Consequently, the effect of student SES on initial status and achievement growth was fixed, and a one-way ANCOVA model (with SES was grand-mean centered) was estimated:[20]

Level 2 model:

$$\pi_{0ij} = \beta_{00j} + \beta_{01j}(\text{SES}_{ij} - \overline{\text{SES}}_{..}) + r_{0ij}.$$

$$\pi_{1ij} = \beta_{10j} + \beta_{11j}(\text{SES}_{ij} - \overline{\text{SES}}_{..}) + r_{1ij}.$$

Level 3 model:

$$\beta_{00j} = \gamma_{000} + u_{00j}.$$

$$\beta_{01j} = \gamma_{010}.$$

$$\beta_{10j} = \gamma_{100} + u_{10j}.$$

$$\beta_{11j} = \gamma_{110.}$$

The results (see Table 13.2, column 5) show that not only do differences in student SES explain a large proportion of variance between schools in initial achievement (.59), but these differences also explain a substantial proportion of the variance between schools in achievement growth (.36). Nonetheless, even after controlling for student SES, significant variation in student achievement growth remains. This model answers a similar yet more important question that the earlier model could not address: *What difference does the school a child goes to make in the child's learning (as opposed to achievement)?*

13.3.2.1.2. How effective are different kinds of schools? To address this question, we estimated a second ANCOVA model with the same set of predictors as in the earlier achievement model. Because student SES is grand-mean centered in this model, the model estimates the effects of the school-level predictors on the adjusted school mean—in this case, the expected achievement growth for a student with average SES. As a result, the coefficient for school SES provides a direct estimate of the contextual or compositional effect of student SES. In this example, the individual (γ_{110}) and contextual (γ_{101}) effects of student SES are both significant—a one standard deviation increase in student SES increases 4-year learning rates by .91(1.12 × .81) units, or about 4 months of learning over a 4-year period, and a one standard deviation increase in school SES increases learning rates by .28(.53 × .51), or about 1 month of learning over a 4-year period. After controlling for school SES, the results also show that learning rates in math are not significantly higher in private schools than in public schools ($\gamma_{103} = .40, p \geq .05$), but they are significantly higher in Catholic schools than in public schools ($\gamma_{102} = 1.15, p < .05$) : 8.69 in public schools versus 9.84 (8.69 + 1.15) in Catholic schools.

The preceding question on school effectiveness focused on differences between different kinds of schools. To more thoroughly address this question, an investigator needs to develop a more comprehensive model that more adequately controls for a variety of differences in student background characteristics (e.g., prior achievement, aspirations, school experiences) and a variety of other school inputs that schools typically have little control over (e.g., teachers, textbooks, facilities, location), as suggested by the earlier conceptual framework (see, e.g., Rumberger & Palardy, 2003a). Yet one additional important question remains to be addressed: *Why are some schools more effective than others?* If schools are to be improved, it is important not just to more accurately identify effective and ineffective schools but also to determine why some schools are more effective than others. By identifying the factors that contribute to school effectiveness, it may be possible to use the information to improve existing schools. Based on the framework presented earlier, this involves identifying the factors

19. This result is based on a single-parameter hypothesis in which $p = .05$, the threshold of statistical significance. Investigators one can also use a multiparameter test that tests for significant differences in the entire array of variances and covariances between two separate models (Bryk & Raudenbush, 2002, pp. 63–65). In this case, the results of the multiparameter test confirmed the results of the single-parameter test—that is, a model with a fixed SES/achievement growth term was not significantly different from a model with a random SES/achievement growth term. Similar models and procedures can be used to examine differences in achievement growth rates between students within the same school (see Seltzer, Choi, & Thum, 2003).

20. Because we focused on achievement growth, we fixed the effect of student SES on initial status, even though its variance was significantly different from zero. As in the case of achievement models that we discussed earlier, the grand-mean centered student-level predictors produce biased estimators of the within-school effects of student characteristics when those characteristics affect educational outcomes at both the individual and school levels. The estimates in Table 13.2, column 6 suggest that is the case in this example, especially for initial status.

that mediate the relationship between school inputs and school outcomes as well as explain variance in mean student achievement, which we refer to as process variables.

We illustrate this by estimating two additional ANCOVA models. First, we estimate a model that includes a single school-level predictor, MEANSES, because it represents a school input that has been shown to strongly affect math learning. The second model adds two school process variables that teachers and other school personnel have at least partial control over: MEANNAEP, which measures the mean number of college-prep courses (as designated by the National Assessment of Educational Progress [NAEP]) students complete during high school, and MEANHW, the mean amount of time students spend on homework per week during the 10th grade (see the appendix for means and standard deviations for these measures). The first model estimates the total compositional effects of student SES (without additional school-level predictors), and the other can be used to see if the two school process variables mediate the relationship between student composition and achievement growth as well as affect student learning.

Table 13.4 displays the estimates for the school-level predictors in standardized form so that the relative magnitude of effects of these factors can be compared. Results from the first model show that the compositional effect of student SES (MEANSES) is highly significant. Results from the second model show that adding the two process variables reduces the effects of MEANSES to the point that it actually has a negative impact on math learning ($\hat{\gamma}_{101} = -0.139$, $p < .05$). That is, the compositional effects of mean SES is reversed after controlling for the average number of college-prep courses that students take in the school and by the average amount of homework that students do—what some investigators have found other studies and have labeled *academic press* (Lee & Smith, 1999; Phillips, 1997). Moreover, MEANNAEP and MEANHW both have significant positive effects on the mean rate of math learning at schools ($\hat{\gamma}_{102} = 0.324$, $p < .01$; $\hat{\gamma}_{103} = 0.276$, $p < .01$). Notice that although we have concluded that MEANNAEP and MEANHW mediate the effect of MEANSES on mean math learning, we have not examined the exact nature of that relationship. Multilevel regression models are not suited for estimating this type of indirect effects. To address this question, we introduce another class of models: multilevel latent growth curves (MLGC), an extension of the latent growth curve (LGC) in the structural equation modeling (SEM) literature.

Table 13.4 Standardized Parameter Estimates of School-Level Predictors in Model for School Mean of 4-Year Math Learning Rate (β_{10j})

	Composition Model (1)	*Process Model (2)*
MEANSES (γ_{101})	0.229**	−0.139*
MEANNAEP (γ_{102})		0.324**
MEANHW (γ_{103})		0.276**

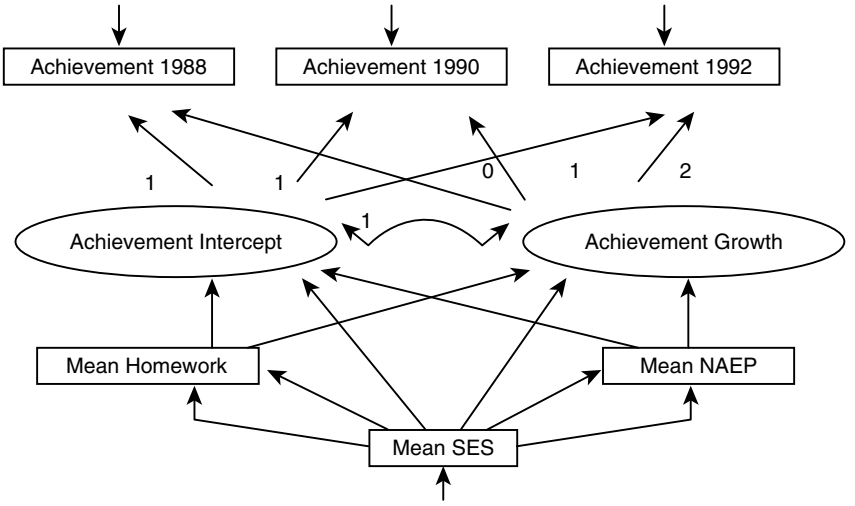
NOTE: MEANSES = mean socioeconomic status; MEANNAEP = mean number of NAEP (college-prep) courses students complete during high school; MEANHW = mean amount of time students spend on homework per week during the 10th grade. This model can be estimated as a three-level multilevel regression model or a two-level multilevel latent growth curve model. The models include student SES (grand-mean centered) in the student-level model (fixed) and a school intercept model with the same three school-level predictors, although those parameter estimates are not shown.
*$p < .05$; **$p < .01$.

13.3.2.2. Multilevel Latent Growth Curves

SEM is widely used among social scientists because of its flexibility for modeling covariance structure in both measurement and structural models. Multilevel SEM has evolved over the past few decades (Muthén, 1989, 1991) but has not received much attention from educational researchers until recently, although Kaplan and his associates have written on its usefulness to the study of school effects (Kaplan & Elliott, 1997; Kaplan & Kreisman, 2000). LGCs (see McArdle & Epstein, 1987; Meredith & Tisak, 1990) are a special class of SEMs designed for modeling between-person change in an outcome over time. LGCs are highly similar to regression-based individual growth trajectories in function, although these two methods evolved independently. Like other single-level SEMs, LGCs have limited applications in the study of school effects because they do not include a school level of analysis.

LGCs have only recently been formulated to analyze multilevel data (Muthén, 1997), resulting in a model that is especially appropriate for the study of school effects. Much like three-level hierarchical linear model (HLM) growth models, this method can accommodate individual growth trajectories, as well as within- and between-school analyses when at least three waves of longitudinal data are available for the student achievement outcome. The appeal of MLGC compared with the multilevel regression growth model is precisely the appeal of SEM over regression models—that is, additional flexibility in specifying covariance relationships, which can result in a more compelling model of school

Figure 13.2 School-Level Path Diagram of MLGC With Indirect Effects



effects. Additional modeling options include the estimation of latent variables from multiple observed variables, of measurement error on observed variables, of complex measurement error structures, and of group comparison models. One MLGC advantage that stands out in particular is the ability to estimate direct, indirect, and total effects between variables. Although MLGCs are in many ways ideally suited for the study of school effects, they have rarely been applied to this field of study (Palardy, 2003). In this section, we illustrate how this method can be used to estimate direct, indirect, and total effects.[21]

13.3.2.2.1. What are the magnitudes of the indirect effects of mean SES on mean math learning, flowing through mean NAEP and mean homework? Recall that this question evolved from our multilevel regression growth model in which we determined that two process variables, MEANNAEP and MEANHW, mediated the effects of MEANSES on mean student achievement growth in math. We now examine the magnitudes and significance levels of those indirect effects.

Figure 13.2 shows the path diagram for the school-level MLGC with the indirect effects of MEANSES flowing through MEANHW and MEANNAEP. Note that this model is highly similar to the "process" model for which results are displayed in Table 13.4. The same assumptions hold in this model. Here the math achievement intercept and growth factors are estimated by fixing path loadings to a linear arrangement with the intercept centered on Time 1 (1988), but the interpretation of these parameters, as well as their values and standard errors, is equivalent to the intercept and growth parameters of the multilevel regression growth model. Other than its multilevel nature, this model is like other LGCs. Table 13.5 shows the standardized coefficient estimates for the direct and indirect effects of MEANSES. The results show that MEANHW and MEANNAEP are significant mediators of MEANSES on mean achievement growth. Students attending higher SES schools took more college-prep courses (0.590, $p < .01$), which resulted in more learning (0.191, $p < .01$). Similarly, students attending higher SES schools did more homework (0.628, $p < .01$), which resulted in a greater learning rate (0.173, $p < .01$). The total effect of mean SES on student learning is the sum of its direct effect and indirect effects (0.225, $p < .01$). Note that the total effect of mean SES in Table 13.5 is equal to the direct effect of mean SES in the compositional model, with no other covariates shown in Table 13.4. Note that multilevel regression software can estimate some forms of indirect effects.[22]

21. MLGCs can be used to address some additional questions about school effectiveness, including whether parameter estimates vary among different samples of schools and alternative specifications for measurement models for independent and dependent variables. See Palardy (2003) for some examples.

22. For example, HLM software can estimate indirect effects that flow through variables with random effects (see Raudenbush & Bryk, 2002, pp. 356–360).

Table 13.5 Indirect Effects of Mean SES on Math Achievement Growth Mediated by Mean NAEP and Mean Homework

	Mediating Process Variable	*Effect on Mediator*	*Effect on Growth*
Mean SES (direct effect)			−0.139*
	NAEP courses	0.590**	0.191**
	Homework	0.628**	0.173**
Total effects			0.225**

NOTE: Standardized coefficients. Significance levels of indirect effects were computed using the Sobel method. SES = socioeconomic status; NAEP = National Assessment of Educational Progress.
*$p < .05$; **$p < .01$.

13.3.3. Categorical Outcome Models

Most school effectiveness studies have focused on student achievement and other student outcomes that can be estimated with linear models in which the random effects are normally distributed. But some student outcomes cannot be estimated with such models. In particular, student outcomes such as dropout rates are binary, taking on one value if the outcome is present and another value if the outcome is not (e.g., $Y = 1$ if the student is a dropout, $Y = 0$ otherwise). As a result, the random effect can also only take on two values and hence is not normally distributed. Other outcomes can involve several discrete conditions, such as attending a 4-year college, a 2-year college, or no college.[23]

Discrete outcomes require a different type of model from the standard multilevel or hierarchical linear models we have discussed up until this time. These models are known as hierarchical generalized linear models (HGLMs), or simply generalized linear models (see Raudenbush & Bryk, 2002, chap. 10). These methods can be used to estimate a wide range of models using multilevel data, including nonlinear models with random effects that are not normally distributed. In fact, hierarchical linear models simply represent a specific and simple type of generalized linear model.

Estimating generalized linear models requires several additional steps from those we have discussed so far. First, the researcher has to specify a Level 1 sampling model. In the linear case, the sampling model is simply a normal distribution with a mean, μ_{ij}, and a variance, σ.[24] Second, the researcher has to specify a link function that transforms the expected value, Φ_{ij}, into a predicted value that can be estimated with a linear model. In the linear case, this link function is simply the value 1 because no transformation is required. Finally, the researcher specifies a linear structural model to estimate the transformed expected value.

We can illustrate this process for the case of school dropouts. For binary student outcomes, such as dropout, the Level 1 sampling model is Bernoulli:

$$\text{Prob}(Y_{ij} = 1|\beta_j) = \Phi_{ij},$$

where Φ_{ij} represents the probability of student i in school j dropping out of school. The Level 1 link function is a log odds ratio:

$$\eta_{ij} = \log[\Phi_{ij}/(1 - \Phi_{ij})],$$

which has a range of $-u$ to $+u$ and takes on the value of 0 when the probability of an outcome equals .5 and the odds of success are even $[.5/(1 - .5) = 1]$. The log odds ratio can be converted to a probability through the following equation:

$$\Phi_{ij} = 1/[1 + \exp\{-\eta_{ij}\}].$$

The Level 1 structural model is similar to the previous Level 1 models. In the case of a null or unconditional model, it is simply

$$\eta_{ij} = \log[\Phi_{ij}/(1 - \Phi_{ij})] = \beta_{0j},$$

and the Level 2 model is exactly as in the linear case:

$$\beta_{0j} = \gamma_{00} + u_{0j}.$$

Conditional models can be constructed by adding Level 1 and Level 2 predictors. And as in the linear case, the analyst must also decide whether the Level 1 predictors should be centered and whether they should be fixed or random at Level 2.

We can illustrate the use of nonlinear models with the NELS data. We first estimated an unconditional model with no Level 1 or Level 2 predictors. The HLM program that we used actually produces two sets of estimates for the fixed effects. The first is a unit-specific estimate, which corresponds to the estimated log odds with a random effect of zero. The second is a population estimate, which provides a better estimate of the true population mean. The second estimate is needed because the nonlinear transformation of probability into log odds means that a symmetrical distribution of log odds results in an asymmetrical distribution of

23. For other examples, see Raudenbush and Bryk (2002, chap. 10).

24. To generate accurate school-level composition measures, we restricted the sample to respondents who had a valid school ID in 1990, had valid test scores in 1988 and 1990, and attended a high school with at least five students.

probabilities that is positively skewed and thus has a higher mean value than the mean of the log-odds distribution.[25] The difference in these two estimates for dropout rates is shown as follows:

Unit-specific estimated mean: 6.49%
Population average estimated mean: 6.94%
Sample mean: 6.81%

As the figures show, the population average dropout rate is higher than the unit-specific rate and closer to the sample mean. The two sets of estimates differ not only in the values they produce but also in their assumptions about the underlying distribution of random effects and in the type of questions they can be used to address. In general, unit-specific estimates are more useful for analyzing differences in the effects of Level 1 and Level 2 predictors across Level 2 units, whereas population-average estimates are more useful for estimating average probabilities for the population as a whole.

We next estimated the same model we did earlier with one student-level predictor, SES, and three school-level predictors: MEANSES, CATHOLIC, and PRVIATE. Both SES and MEANSES were centered on the grand mean, which affects the value and interpretation of the intercept term. The unit-specific estimated parameters are shown in Table 13.6. The parameter estimate for the student-level predictor is −.868. A student with average SES attending a typical public school with average MEANSES would have a predicted log-odds dropout rate of −2.843, corresponding to a predicted probability of $1/(1 + \exp\{2.843\}) = .055$. A student with an SES one unit higher than average attending a typical public school with average MEANSES would have a predicted log-odds dropout rate of $-2.843 - .868 = -3.711$, corresponding to a predicted probability of $1/(1 + \exp\{3.711\}) = .022$. The average SES of the school, MEANSES, would also affect the odds of dropping out, even after controlling for the individual effects of SES, something referred to as the contextual or compositional effect of SES (which we discuss below). A student with average SES attending a school with a MEANSES one unit higher than average (about two standard deviations, as shown in the appendix) would have a predicted log-odds dropout rate of $-2.843 - .295 = -3.138$, corresponding to a predicted probability of $1/(1 + \exp\{3.138\}) = .042$. Because of the nonlinear relationship between the log odds and probability, an additional one-unit increase in MEANSES (a 100% increase) would only lower the predicted probability to .031 (a 27% decrease).

Table 13.6 Estimated Parameters for Dropout Models

	Null Model	*School Model*
Fixed effects		
Model for school mean dropout rate (β_0)		
INTERCEPT (γ_{00})	−2.667**	−2.843**
MEANSES (γ_{01})		−0.295**
CATHOLIC (γ_{02})		−1.358**
PRIVATE (γ_{03})		−0.913**
Model for SES dropout slope (β_1)		
INTERCEPT (γ_{10})		−0.868**
Variance components		
Between school (τ_{00})	.455**	.207
Proportion school-level variance explained		.545
Reliability	.292	.154

NOTE: SES = socioeconomic status; PRIVATE = private schools; CATHOLIC = Catholic schools; MEANSES = mean socioeconomic status.
**$p < .01$.

13.4. Identifying Effective Schools

Although many school effectiveness studies attempt to identify school-level factors that predict student outcomes based on a sample of schools, some analysts are also interested in identifying individual schools that are particularly effective. That is, even after controlling for a given set of predictors, each school may have a mean student achievement that is above or below the mean predicted from the model. Schools whose mean achievement is above the level predicted by the model can be considered effective, whereas schools whose mean achievement is below the level predicted can be considered ineffective schools.

The unique contribution of each school to its effectiveness is captured in the school-level random effect or error term. Consider the following simple, two-level achievement model:

$$\text{Level 1 model: } Y_{ij} = \beta_{0j} + r_{ij}.$$
$$\text{Level 2 model: } \beta_{0j} = \gamma_{00} + \mu_{0j}.$$

The dependent variable in the Level 2 model, β_{0j}, which represents the average achievement of each school, is composed of a fixed effect, γ_{00}, and a random effect, μ_{0j}. Hierarchical analysis produces an empirical Bayes estimator for the random effect

25. The two estimates can be quite similar when the fixed effect is close to zero (which translates to a probability of .5) or when the random effect is close to zero. For a more complete discussion, see Raudenbush and Bryk (2002, pp. 297–304).

Table 13.7 Models and Estimated Random Effects From Two Dropout Models

	Unconditional Model	*Conditional Model*
Level 1 model	$\log[\Phi_{ij}/(1-\Phi_{ij})] = \beta_{0j}$	$\log[\Phi_{ij}/(1-\Phi_{ij})] = \beta_{0j}$
Level 2 model	$\beta_{0j} = \gamma_{00} + u_{0j}$	$\beta_{0j} = \gamma_{00} + \gamma_{01}\text{MEANSES}_j + u_{0j}$
Estimated random effect (u_{0j})		
Mean, standard deviation	0, .365	0, .179
Minimum, maximum	−.759, 1.466	−.409, .710

Table 13.8 Estimated Fixed and Random Effects for Two Schools

			Unconditional Model			*Conditional Model*		
Case	*n*	*MEANSES*	*Fixed Effect*	*Random Effect*	*Dropout Rate*	*Fixed Effect*	*Random Effect*	*Dropout Rate*
81	15	−.332	−2.667	−.326	4.78	−2.348	−.219	7.13
393	16	.785	−2.667	−.342	4.70	−3.812	−.066	2.03

NOTE: Fixed and random effects are log odds. Dropout rate = $1/\{1 + \exp[-(\text{Fixed effect} + \text{Random effect})]\}$.

that provides a better and more stable estimate of the unique school effect than other methods (e.g., ordinary least squares [OLS] estimates) by taking into account group membership and the within-school sample size (Raudenbush & Bryk, 2002, p. 154). More accurate estimates can be obtained by adding school-level variables to the Level 2 model, which provides conditional shrinkage estimates of the random effects (Raudenbush & Bryk, 2002, pp. 90–94). In achievement models, schools with positive random effects have higher than predicted achievement rates and should be considered effective, whereas in dropout models (as we illustrate below), schools with negative random effects have lower than predicted dropout rates and should be considered effective.

To illustrate how this technique can be used to identify effective schools, we can compare the empirical Bayes estimates for the Level 2 random effects from two simple dropout models, one unconditional and one conditional. The two models and descriptive statistics for the empirical Bayes estimates of the Level 2 random effects are shown in Table 13.7.

As the descriptive statistics show, the estimated random effects in the conditional model have a much narrower range and hence smaller standard deviation than the unconditional estimates.

Moreover, the conditional model provides a better way to identify effective schools. Consider the two schools shown in Table 13.8. Based on the unconditional model, both schools are equally effective—their unique or random log-odds dropout rate are both about one third of a logit less than the fixed or expected rate—and hence both schools have similar estimated dropout rates that are considerably smaller than the average dropout rate of 6.49 for the entire sample of schools. Estimates from the conditional model tell another story, however. School 393 has a much higher average SES than School 81, so its expected log-odds dropout rate is much higher. Yet the unique contribution to its dropout rate—that is, its random effect—is not very large, and hence the school is not particularly effective. In contrast, School 81 has a much higher expected dropout rate (i.e., lower log-odds dropout rate) because its average SES is much lower, yet its estimated dropout rate is actually lower than the expected rate. Hence, School 81 should be considered more effective than School 393, even though it has a higher dropout rate.

13.5. Summary and Future Directions

The need for useful and methodologically sound school effectiveness studies has never been greater. Fortunately, the development of large-scale, comprehensive, longitudinal studies of student development has coincided with the development of new and powerful statistical techniques for analyzing the data for these studies. The result has been the continued growth of more sophisticated and comprehensive school effectiveness studies.

Several important challenges remain, however. One is to develop even more comprehensive studies. Although earlier studies were particularly useful for identifying student, family, and school factors related

to student achievement over time, they were not particularly well suited for studying teacher and classroom effects. In part, this was due to the nature of the sampling frame that was used, in which relatively small samples of students were selected within schools. Future designs should sample intact classrooms and develop better measures of classroom practices to focus on teacher and classroom effects (Mullens & Gayler, 1999).

looseness1 Another challenge is to encourage researchers to develop and use more comprehensive conceptual frameworks for their studies of school effectiveness. For example, although economists routinely examine resource variables in their studies of school effectiveness, sociologists and educational researchers frequently do not. Conversely, economists often ignore important process variables in their models, such as school climate. To the extent that models are misspecified at any level of analysis, the resulting estimates can be biased and the conclusions faulty (Raudenbush & Byrk, 2002, chap. 9). A similar argument can be made regarding outcome measures: School effectiveness studies focus predominately on student achievement as measured by test scores, thereby ignoring outcomes, such as dropout or attrition, that can be influenced by different factors and could lead to different conclusions about effective schools (Rumberger & Palardy, 2003a).

The final challenge is to encourage better use of the growing advances in statistical modeling techniques in school effectiveness studies. Although statistical advances in multilevel and structural equation modeling have been quite rapid, these advances are slow to find their way into mainstream school effectiveness studies. Although there is always a lag between the initial development of new statistical techniques and their widespread use in the field, as the techniques become more sophisticated, that lag could increase. This may be particularly problematic for existing scholars who were most likely trained in earlier techniques and who will require a sort of in-service training to learn the new approaches. Fortunately, many professional associations, such as the American Educational Research Association and the American Sociological Association, sponsor such training sessions in conjunction with their national meetings each year.

Appendix: Variable Descriptive Statistics and Labels for NELS Data

Variable Name	*M*	*SD*	*Minimum*	*Maximum*	*Description and (NELS:88 Variables)*
Measurement variables (n = 39, 241)					
Math	50.31	10.28	23.34	80.67	Math IRT theta score (BY2XRTH, F12XRTH, F2XRTH)
Time	0.46	0.40	0.00	1.00	Time (0 = 8th grade; 0.5 = 10th grade; 1 = 12th grade)
Student variables (n = 14, 199)					
Math, Grade 10	51.11	9.88	24.87	72.90	Math IRT theta score (F12XRTH)
SES	0.04	0.81	−2.95	2.75	10th-grade SES composite (F1SES)
Transfer	0.06	0.24	0.00	1.00	Transferred schools between 10th and 12th grades (F2F1SCFG = 1)
Dropout	0.07	0.25	0.00	1.00	Dropped out of school (F2DOSTAT = 3, 4, 5)
School variables (n = 912)					
Mean SES	0.01	0.52	−1.33	1.54	Mean SES of students (F1SES)
Catholic	0.07	0.25	0.00	1.00	(G10CTRL1 = 2)
Private	0.08	0.27	0.00	1.00	(G10CTRL1 = 3–5)
Homework time	4.61	2.05	1.06	14.00	Mean number of hours spent on homework per week (F1S36A2)
NAEP composite	13.76	2.27	6.00	27.74	Number of NAEP units in math, science, English, and social science earned in high school (F2ra11_C + a12_C + geo_C, tri_C + pre_C + cal_C + bio_C + che_C + phy_C + soc_C + his_C)

NOTE: NELS = National Education Longitudinal Study; SES = socioeconomic status; IRT = item response theory; NAEP = National Assessment of Educational Progress.

REFERENCES

Alexander, K. L., & Eckland, B. K. (1975). Basic attainment processes. *Sociology of Education, 48*, 457–495.

Alexander, K. L., & Pallas, A. (1985). School sector and cognitive performance: When is a little a little? *Sociology of Education, 58*, 115–128.

Barr, R., & Dreeben, R. (1983). *How schools work.* Chicago: University of Chicago Press.

Bryk, A. S., Lee, V. E., & Holland, P. B. (1993). *Catholic schools and the common good.* Cambridge, MA: Harvard University Press.

Bryk, A. S., & Thum, Y. M. (1989). The effects of high school organization on dropping out: An exploratory investigation. *American Educational Research Journal, 26*, 353–383.

Carroll, D. (1996). *National Education Longitudinal Study (NELS:88/94): Methodology report.* Washington, DC: Government Printing Office.

Chubb, J. E., & Moe, T. M. (1990). *Politics, markets, and America's schools.* Washington, DC: Brookings Institution.

Coleman, J. S. (1990). *Equality and achievement in education.* San Francisco: Westview.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, F., Weinfeld, F., et al. (1966). *Equality of educational opportunity.* Washington, DC: Government Printing Office.

Coleman, J. S., & Hoffer, T. (1987). *Public and private high schools: The impact of communities.* New York: Basic Books.

Coleman, J. S., Hoffer, T., & Kilgore, S. B. (1982). *High school achievement: Public, Catholic, and private schools compared.* New York: Basic Books.

Croninger, R. G., & Lee, V. E. (2001). Social capital and dropping out of high school: Benefits to at-risk students of teachers' support and guidance. *Teachers College Record, 103*, 548–581.

Gamoran, A. (1992). Social factors in education. In M. C. Alkin (Ed.), *Encyclopedia of educational research* (pp. 1222–1229). New York: Macmillan.

Gamoran, A. (1996). Student achievement in public magnet, public comprehensive, and private city high schools. *Educational Evaluation and Policy Analysis, 18*, 1–18.

Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources, 33*, 505–523.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature, 24*, 1141–1177.

Hanushek, E. A. (1989). The impact of differential expenditures on school performance. *Educational Researcher, 18*, 45–62.

Hauser, R. M., & Featherman, D. L. (1977). *The process of stratification: Trends and analysis.* New York: Academic Press.

Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher, 23*, 5–14.

Jencks, C. S., & Brown, M. D. (1975). Effects of high schools on their students. *Harvard Educational Review, 45*, 273–324.

Jencks, C., & Mayer, S. E. (1990). The social consequences of growing up in a poor neighborhood. In L. Lynn Jr. & M. G. H. McGeary (Eds.), *Inner-city poverty in the United States* (pp. 111–186). Washington, DC: National Academy Press.

Johnson, M. K., Crosnoe, R., & Elder, G. H., Jr. (2001). Students' attachment and academic engagement: The role of race and ethnicity. *Sociology of Education, 74*, 318–340.

Kahlenberg, R. D. (2001). *All together now: Creating middle-class schools through public school choice.* Washington, DC: Brookings Institution.

Kaplan, D., & Elliott, P. R. (1997). A model-based approach to validating education indicators using multilevel structural equation modeling. *Journal of Educational and Behavioral Statistics, 22,* 323–347.

Kaplan, D., & Kreisman, M. B. (2000). On the validation of indicators of mathematics education using TIMSS: An application of multilevel covariance structure modeling. *International Journal of Educational Policy, Research, and Practice, 1,* 217–242.

Lee, V. E., & Burkam, D. T. (2003). Dropping out of high school: The role of school organization and structure. *American Educational Research Journal, 40,* 353–393.

Lee, V. E., & Bryk, A. S. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education, 62*, 172–192.

Lee, V. E., & Smith, J. B. (1993). Effects of school restructuring on the achievement and engagement of middle-grade students. *Sociology of Education, 66*, 164–187.

Lee, V. E., & Smith, J. B. (1995). Effects of high school restructuring and size on gains in achievement and engagement for early secondary school students. *Sociology of Education, 68*, 241–279.

Lee, V. E., & Smith, J. B. (1997). High school size: Which works best for whom? *Educational Evaluation and Policy Analysis, 19*, 205–227.

Lee, V. E., & Smith, J. B. (1999). Social support and achievement for young adolescents in Chicago: The role of school academic press. *American Educational Research Journal, 36*, 907–945.

Lee, V. E., Smith, J. B., & Croninger, R. G. (1997). How high school organization influences the equitable distribution of learning in mathematics and science. *Sociology of Education, 70*, 128–150.

Levin, H. M. (1994). Production functions in education. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education* (pp. 4059–4069). New York: Pergamon.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: John Wiley.

McArdle, J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58,* 110–133.

McNeal, R. B. (1997). High school dropouts: A closer examination of school effects. *Social Science Quarterly, 78,* 209–222.

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

Mosteller, F., & Moynihan, D. P. (Eds.). (1972). *On equality of educational opportunity.* New York: Random House.

Mullens, J. E., & Gayler, K. (1999). *Measuring classroom instructional processes: Using survey and case study field-test results to improve item construction* (Working Paper No. 1999–08). Washington, DC: National Center for Education Statistics.

Murnane, R. J. (1981). Interpreting the evidence on school effectiveness. *Teachers College Record, 83*, 19–35.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations: Presidential address to the Psychometric Society. *Psychometrika, 54,* 557-585.

Muthén, B. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28,* 338–354.

Muthén, B. (1997). Latent variable modeling with longitudinal and multilevel data. In A. Raftery (Ed.), *Sociological methodology* (pp. 453–480). Boston: Blackwell.

Palardy, G. J. (2003). *A comparison of hierarchical linear and multilevel structural equation growth models and their application in school effectiveness research.* Unpublished doctoral dissertation, University of California, Santa Barbara.

Phillips, M. (1997). What makes schools effective? A comparison of the relationships of communitarian climate and academic climate to mathematics achievement and attendance during middle school. *American Educational Research Journal, 34*, 633–662.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*, 307–335.

Rumberger, R. W. (1995). Dropping out of middle school: A multilevel analysis of students and schools. *American Educational Research Journal, 32*, 583–625.

Rumberger, R. W., & Palardy, G. J. (2003a). *Does segregation (still) matter? The impact of student composition on academic achievement in high school.* Revised paper originally presented at the annual meeting of the American Educational Research Association, April 10–14, 2001, Seattle, WA.

Rumberger, R. W., & Palardy, G. J. (2003b). *Test scores, dropout rates, and transfer rates as alternative measures of school performance.* Revised paper originally presented at the annual meeting of the American Educational Research Association, April 1–5, 2002, New Orleans, LA.

Rumberger, R. W., & Thomas, S. L. (2000). The distribution of dropout and turnover rates among urban and suburban high schools. *Sociology of Education, 73*, 39–67.

Rumberger, R. W., & Willms, J. D. (1992). The impact of racial and ethnic segregation on the achievement gap in California high schools. *Educational Evaluation and Policy Analysis, 14*, 377–396.

Seltzer, M., Choi, K., & Thum, Y. M. (2003). Examining relationships between where students start and how rapidly they progress: Using new developments in growth modeling to gain insight into the distribution of achievement within schools. *Educational Evaluation and Policy Analysis, 25*, 263–286.

Shavelson, R., McDonnell, L., Oakes, J., & Carey, N. (1987). *Indicator systems for monitoring mathematics and science education.* Santa Monica, CA: RAND.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford University Press.

Summers, A. A., & Wolfe, B. L. (1977). Do schools make a difference? *American Economic Review, 67*, 639–652.

Tagiuri, R. (1968). The concept of organizational climate. In R. Tagiuri & G. H. Litwin (Eds.), *Organizational climate: Exploration of a concept* (pp. 1–32). Boston: Harvard University, Division of Research, Graduate School of Business Administration.

U.S. Department of Education, National Center for Education Statistics. (2003). *The condition of education, 2003* (NCES 2003-67). Washington, DC: Government Printing Office.

Willms, J. D. (1985). Catholic-school effects on academic achievement: New evidence from the High School and Beyond follow-up study. *Sociology of Education, 59*, 98–114.

Willms, J. D. (1992). *Monitoring school performance: A guide for educators.* Washington, DC: Falmer.

Willms, J. D., & Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement, 26*, 209–232.

Witte, J. F. (1992). Private school versus public school achievement: Are there findings that should affect the educational choice debate? *Economics of Education Review, 11*, 371–394.

Witte, J. F., & Walsh, D. J. (1990). A systematic test of the effective schools models. *Educational Evaluation and Policy Analysis, 12*, 188–212.

Chapter 14

The Use of Hierarchical Models in Analyzing Data From Experiments and Quasi-Experiments Conducted in Field Settings

Michael Seltzer

14.1. Introduction

In studies of programs and interventions in a variety of fields (e.g., education, social welfare, epidemiology), individuals are typically nested within different sites or organizational units (e.g., schools, communities, clinics). Ignoring the nested structure of the data in such studies (e.g., using standard regression techniques to analyze student or client outcomes) can give rise to a host of problems. (Note that we often use the terms *site* and *organizational unit* interchangeably in this chapter.)

First, in such studies, individuals nested in different sites experience different implementations of programs. In addition, the background characteristics of study participants may vary appreciably from site to site. Factors such as these give rise to a certain degree of dependency or similarity among the observations nested within a site. Ignoring such dependencies (i.e., ignoring the intra-class correlational structure of multisite data) can result in standard errors for treatment effect estimates that are misleadingly small.

Moreover, when we ignore the nesting of individuals in different sites in our analyses, we run the risk

AUTHOR'S NOTE: This chapter is dedicated to Leigh Burstein, who made seminal contributions to the development of multilevel modeling techniques. I wish to thank Maryl Gearhart and Geoff Saxe for permission to use the data from their study, "Integrating Assessment With Instruction in Elementary Mathematics," which was supported by NSF grant MDR 9154512. I would also like to thank the University of Chicago School Mathematics Project for permission to use the data from the Transition Mathematics Field Study. The development and evaluation of the Transition Mathematics curriculum was supported by grants from the Amoco Foundation and Carnegie Foundation. I am grateful to Jin-Ok Kim for extremely valuable discussions regarding the issues addressed in this chapter and for her many thoughtful suggestions and comments. I also wish to thank Noreen Webb and Kilchan Choi for reading this chapter with care and for their many helpful comments.

of inadvertently concealing potentially substantial between-site heterogeneity in program effects. Such heterogeneity is not surprising when we consider that sites can vary considerably in terms of implementation, background characteristics of program participants, and numerous other factors that may dampen or magnify the effects of a program (see Campbell & Stanley, 1963, pp. 19–22; Cohen, Raudenbush, & Ball, 1999; Cronbach, 1975, 1982; McLaughlin, 1987; Patton, 1980). Failing to attend to differences in results across sites can, as will be seen, result in erroneous conclusions concerning the effects of programs and, in addition, missed opportunities to investigate how differences in implementation and other key aspects of program settings relate to differences in program effectiveness.

In this chapter, we show how hierarchical models (Kreft & de Leeuw, 1999; Goldstein, 2003; Longford, 1993; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) can be used to obtain more appropriate standard errors for estimates of treatment effects and other key parameters in multisite studies of programs and interventions. Furthermore, we show how these models can be used to study how differences in such factors as implementation and the background characteristics of program participants relate to differences in results across sites. Such analyses can potentially provide insight into questions of the following kind: Under what conditions does a program of interest appear to be successful, and for whom?

Drawing sound conclusions regarding the effects of programs in field settings can be extremely challenging. In this connection, we emphasize the importance of collecting data on implementation and the continual need to attend to possible confounding variables.

In the next section of this chapter, we discuss two general types of designs that are commonly encountered in multisite studies of programs and interventions. One type involves blocking and essentially gives rise to a series of "mini" experiments or "mini" quasi-experiments. An example would be a study in which both treatment and control conditions are implemented in each of a number of schools. The second type, which does not involve forming blocks, entails assigning organizational units to the different conditions that are being investigated in a study. An example would be a study in which schools are randomly assigned to treatment or control conditions, giving rise to a sample of treatment schools and a sample of control schools.

We then present analyses of the data from two multisite studies, which provide examples of these major design types. Both studies focus on innovative mathematics curricula and instruction. The analyses that we present form the heart of this chapter. In particular, they provide opportunities to discuss the logic of hierarchical models (HMs) and to illustrate their value in analyzing data from experiments and quasi-experiments in field settings.

In the final section of this chapter, we recap key points and discuss their implications for designing multisite studies and for analyzing multisite data. We also discuss some of the possibilities that arise when longitudinal data are collected, that is, when constructs of interest (e.g., key outcomes) are measured on a series of occasions during the course of a study.

14.2. Two General Types of Designs in Multisite Studies

The kinds of designs commonly encountered in multisite evaluation studies typically fall into two broad categories. Designs in the first category involve blocking and take on two basic forms, which we term *Forms A* and *B*. Form A designs involve implementing both treatment and comparison conditions in each of a series of sites (e.g., schools or communities). Consider, for example, the portion of Pinnell, Lyons, DeFord, Bryk, and Seltzer's (1994) study that focuses on the relative effectiveness of Reading Recovery versus more conventional remedial reading instruction: In each of 10 schools, first graders at risk for failure in reading were randomly assigned to Reading Recovery or to a more standard remedial program. Other examples include Raffe's (1991) study of a vocational educational initiative in Britain. Each of the sites (schools) in this study provided a comparison of individuals who participated in the initiative with a group of individuals who did not. Assignment to program and comparison conditions in this evaluation, in contrast to the Reading Recovery study, was not random.

Form B designs entail forming matched pairs of groups or organizational units (e.g., matched pairs of communities) and assigning one group within a pair to the program or intervention of interest and the other to the comparison condition. One example is the evaluation of a community-based intervention called COMMIT, which sought to promote smoking cessation among heavy smokers (Gail, Byar, Pechacek, & Corle, 1992). Within each of 11 carefully matched pairs of communities, 1 community was randomly assigned to COMMIT, and the other served as a comparison community. Another example is the study of DARE (Drug Abuse Resistance Education) conducted by Rosenbaum, Flewelling, Bailey, Ringwalt,

and Wilkinson (1994), which entailed forming 18 well-matched pairs of elementary schools; in the case of 12 pairs, 1 school was randomly assigned to DARE, whereas the other served as a comparison school, and in the case of 6 pairs, assignment was nonrandom.

In contrast to Form A designs, assignment to different conditions in the case of Form B designs does not occur at the level of the individual; rather, intact groups within each matched pair are assigned to different conditions. However, in either case, our samples consist of a series of blocks (e.g., schools in the case of the Reading Recovery evaluation and matched pairs in the case of the COMMIT study), in which both treatment and comparison conditions are implemented. Thus, each block can be viewed as a "mini" experiment (or "mini" quasi-experiment).

Rather than providing us with a series of experiments or quasi-experiments, the second general type of design provides us with a sample of organizational units in each condition (e.g., treatment, control) that is being investigated. One example is a study of a social influence-based drug abuse prevention program reported in Pentz et al. (1989) and Chou, Bentler, and Pentz (1998), in which 32 middle schools were randomly assigned to the program and 25 assigned to the control condition. A second example is a study of the effectiveness of two school-based violence prevention programs conducted by Flay and his colleagues. In this study, which involved 12 schools, there were two treatment conditions and a control condition; 4 schools were randomly assigned to each condition.

In the first general category of designs discussed above, blocks (e.g., schools in the Reading Recovery study) are viewed as a random factor crossed with treatment type, which is viewed as fixed. In the second category, organizational units (e.g., schools in Pentz et al.'s [1989] study) are viewed as a random factor nested within treatment type, which again is viewed as fixed. (See Raudenbush, 1993, and, for example, Kirk, 1982, for discussions of these designs.) Thus, the analysis of data arising from these designs requires the use of models containing both random and fixed effects (i.e., mixed models). As Raudenbush (1993) notes, in the simplest of cases, efficient estimates of the fixed effects and variance components in such models are available in closed form. Consider, for example, a design in which classrooms are nested within treatment type. If the number of students per classroom is identical across classrooms, if the number of classrooms per treatment type is identical, and if we do not need to adjust for various pretest measures, estimation can proceed in a straightforward manner (see, e.g., Kirk, 1982, chap. 10).

In field settings, however, our data will almost always be unbalanced. Moreover, there will almost always be a need to include covariates in our models to adjust for possible confounding variables or to obtain more precise estimates of parameters of interest. As will be seen, hierarchical modeling, with parameter estimation carried out via iterative techniques such as the EM algorithm, provides a viable way of proceeding in such realistically complex settings. Note that the above designs will also often contain a longitudinal component, thus giving rise to time-series observations nested within individuals. Furthermore, in some settings, we may need to explicitly represent the nesting of students in different classrooms and, in turn, the nesting of classrooms in different schools. Complex nested structures of this kind can be modeled readily using HMs.

We first present a series of analyses of the data from an evaluation of an innovative prealgebra curriculum (University of Chicago School Mathematics Project, 1986), and this is followed by a set of analyses of the data from a study of the effects of reform-minded mathematics instructional practices on upper elementary students' understanding of fractions (Gearhart et al., 1999; Saxe, Gearhart, & Seltzer, 1999). The former study provides an example of a matched-pair design, whereas the latter can be viewed as a design in which organizational units (i.e., classes) are nested within treatment type.

Note that these studies are by no means perfect from a methodological standpoint. Rather, they provide, we believe, examples of thoughtful efforts, given limited resources, to address important substantive questions in field settings and to tackle the kinds of methodological challenges that arise in such settings. Furthermore, they provide valuable opportunities to illustrate the kinds of questions that we can begin to address using HMs in analyses of multisite evaluation data.

14.3. Designs in Which Blocks Are Crossed With Treatment Type: Reanalyses of the Transition Mathematics Data

14.3.1. Background

Transition Mathematics (TM) is an innovative prealgebra curriculum that seeks to prepare students for greater success in algebra and geometry. A distinctive feature of TM is the importance placed on reading in learning mathematics. Appreciable amounts

of reading are included in each chapter of the TM text in efforts to clarify key concepts and to integrate material presented in previous chapters. A second notable feature of TM is its focus on real-world applications of mathematics.

A large-scale study of the effectiveness of TM was conducted during the 1985–1986 school year. The study's sample consisted of 20 carefully matched pairs of classrooms located within various school districts throughout the United States. Each pair of classrooms was matched on the basis of pretests administered at the start of the school year as well as on the basis of information supplied by district math coordinators and teachers. Within each pair, the students in one class were taught by a teacher who used the TM text, whereas the students in the other class were taught by a teacher who used the materials already in place at that particular school. An alternative design possibility would have been to have the same teacher teach both the TM and comparison classes at a given site. However, a concern was that a teacher might consciously or unconsciously draw on elements of the TM curriculum in teaching the comparison class and vice versa. Therefore, it was decided that different teachers would teach the classes within a pair. Note that all teachers who participated in the study volunteered to do so and tended to have substantial teaching experience. The decision as to which teacher at a site would use TM and which would use the materials already in place was based on random assignment in the case of 10 sites; logistical reasons precluded this in the case of the 10 other pairs. Later in this section, we will show how HMs can be used to assess whether differences in certain key facets of design are systematically related to program outcomes. As will be seen, the effects of TM seem to be similar at sites in which teachers were assigned randomly and at sites where random assignment was not possible.

Thus, the 20 well-matched pairs of classes that provide the basis of this study can in essence be viewed as 20 studies (i.e., mini-experiments or mini-quasi-experiments) of the effects of TM. For ease of exposition, we refer to each matched pair as a site.

In addition to pretests administered at the study's outset, a battery of posttests was administered at the end of the 1985–1986 school year. Note also that information on program implementation was obtained through classroom observations, diaries kept by a sample of teachers, and questionnaires completed by all participating teachers. In the analyses that follow, the outcome that we focus on is geometry readiness, which was measured by a student's total score on a 19-item test. The analyses that we present represent an extension of those presented in Seltzer (1994).

14.3.2. Ignoring the Nested Structure of the Data: A Conventional Ordinary Least Squares Analysis

We first conduct an analysis that ignores the nesting of students in different sites. To estimate the expected difference in geometry readiness scores between students who work with TM materials versus those who do not, we fit the following regression model to the $N = 572$ student-level cases in our data set:

$$Y_i = \beta_0 + \beta_1 TRT_i + \beta_2 PRE_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2), \tag{1}$$

where Y_i is the geometry readiness score for student i, TRT_i is an indicator variable that takes on a value of 1 if student i is a TM student (0 otherwise), and PRE_i represents the score for student i on a general mathematics pretest. Note that geometry readiness scores range from 1 to 19, out of a possible score of 19, and general math pretest scores range from 5 to 38, out of a possible score of 40. The parameter of primary interest in this model is β_1, which represents the expected difference in geometry readiness scores between TM and comparison group students, holding constant pretest scores. Note that the ε_i are errors assumed independent and normally distributed with mean 0 and variance σ^2. As will be seen below, the assumption of independent errors is problematic.

Fitting the above model to the data using ordinary least squares (OLS), we obtain an estimate of the effect of TM of 1.08 points ($SE = 0.26$; $t = 4.15$). Thus, these results suggest that students who use TM materials outperform students who do not by approximately 1 point on average.

Given that TM was implemented in 20 sites throughout the United States, a valuable way of proceeding at this stage would be to reanalyze the data site by site. Thus, we fit the model specified in equation (1) to each site's data. As can be seen in Table 14.1, the OLS estimates of the site TM effects vary substantially across sites, ranging from approximately -2 points to values exceeding 4.6 points. Furthermore, the 95% intervals displayed in Table 14.1 suggest positive TM effects for 7 sites and a negative effect for 1 site (Site 11). The resulting 90% intervals also suggest a negative effect for Site 8 and a positive effect for Site 12. We also see that there are a number of sites whose point estimates are extremely close to 0 and

Table 14.1 Site-by-Site Analyses: Ordinary Least Squares (OLS) Estimates of Site Transition Mathematics (TM) Effects

Site (j)[a]	*Size (n_j)*	*TM Effect ($\hat{\beta}_{1j}$) [SE($\hat{\beta}_{1j}$)]*	*95% CI of TM Effect*	*90% CI of TM Effect*	*Implementation of Reading (0 = low, 1 = high)*	*Random Assignment of Teachers (0 = no, 1 = yes)*	*Site Pretest Mean*
1	31	−0.25 [0.78]	[−1.85, 1.35]	[−1.58, 1.08]	0	0	23.55
2	27	2.69 [0.86]	[0.91, 4.47]	[1.21, 4.17]	1	1	16.82
3	34	0.44 [0.77]	[−1.13, 2.01]	[−0.86, 1.74]	0	0	11.79
4	44	0.10 [0.75]	[−1.41, 1.61]	[−1.16, 1.36]	0	0	19.14
5	17	0.33 [1.20]	[−2.25, 2.91]	[−1.79, 2.45]	0	0	16.41
6	35	0.78 [0.94]	[−1.14, 2.70]	[−0.82, 2.38]	1	0	21.94
7	37	1.40 [0.66]	[0.05, 2.75]	[0.28, 2.52]	1	1	28.00
8	23	−1.68 [0.85]	[−3.45, 0.09]	[−3.14, −0.22]	1	0	17.39
9	42	4.67 [0.78]	[3.10, 6.24]	[3.36, 5.98]	1	0	14.69
10	17	4.64 [1.50]	[1.43, 7.85]	[2.00, 7.28]	1	1	15.24
11	28	−2.15 [0.93]	[−4.07, −0.23]	[−3.74, −0.56]	0	1	14.50
12	31	1.68 [0.93]	[−0.22, 3.58]	[0.10, 3.26]	1	0	25.13
13	31	0.73 [1.30]	[−1.93, 3.39]	[−1.48, 2.94]	0	1	23.32
14	25	3.33 [1.18]	[0.89, 5.77]	[1.31, 5.35]	1	1	22.44
15	23	−0.25 [0.85]	[−2.02, 1.52]	[−1.71, 1.21]	0	1	21.70
16	33	−1.74 [1.10]	[−3.99, 0.51]	[−3.61, 0.13]	0	1	20.06
17	33	1.07 [0.92]	[−0.81, 2.95]	[−0.49, 2.63]	0	1	20.27
18	27	0.77 [1.16]	[−1.63, 3.17]	[−1.22, 2.76]	1	1	17.63
19	17	2.61 [1.07]	[0.31, 4.91]	[0.72, 4.50]	0	0	16.06
20	17	4.64 [1.82]	[0.75, 8.53]	[1.44, 7.84]	1	0	17.59

NOTE: CI = confidence interval.
a. The subscript j provides a way of referencing each of the sites in the sample.

whose 90% and 95% intervals comfortably include a value of 0. Finally, the estimated effect of TM based on the analysis, ignoring the nesting of students within sites (i.e., 1.08), lies outside the 95% intervals for 5 sites (8, 9, 10, 11, and 16).

Given that teachers may vary substantially in their use of instructional materials and in other critical aspects of practice, and given the appreciable differences across sites in the TM study in various student compositional characteristics, the results in Table 14.1 are not very surprising. A problem, however, is that the results based on the initial analysis mask this heterogeneity. Such an analysis gives stakeholders the misleading impression that the effects of TM are uniform across sites. Moreover, the results from such an analysis do not prompt one to ask whether and, if so, why TM may be more successful in some sites than others.

We now show how HMs provide a means of reflecting the location or nesting of program participants in different sites and enable us to study the variability in program effects across sites.

14.3.3. Assessing the Variability in TM Effects Across Sites

The models that we present are often referred to as two-stage or two-level HMs (see Mason, Wong, & Entwistle, 1983; Raudenbush & Bryk, 2002). Each of the HMs that we employ, as will be seen, consists of two models: a Level 1 or within-site model and a Level 2 or between-site model.

We now pose the following within-site model:

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j}(TRT_{ij} - \overline{TRT}_{.j}) \\ &\quad + \beta_{2j}(PRE_{ij} - \overline{PRE}_{.j}) + \varepsilon_{ij} \\ \varepsilon_{ij} &\sim N(0, \sigma^2), \end{aligned} \tag{2}$$

where Y_{ij} is the geometry readiness score for student i in site j; because there are 20 sites in our sample, our subscript or index for sites takes on values from 1 to 20 (i.e., $j = 1 \ldots 20$). TRT_{ij} is a treatment indicator variable that takes on a value of 1 if student i in site j is a member of the TM class (0 otherwise), and PRE_{ij} is the pretest score for student i in site j. The parameter of primary interest in this equation is β_{1j}, which represents the expected TM/comparison class contrast for site j, holding constant pretest performance; β_{2j} is the pretest/posttest slope for site j, holding constant *TRT*. Note that TRT_{ij} and PRE_{ij} are centered around their site means. This is termed *group-mean centering* (see Raudenbush & Bryk, 2002, chaps. 2, 5). By virtue of this centering, β_{0j} represents the mean geometry score for site j. The ε_{ij} are errors assumed independent and normally distributed with mean 0 and variance σ^2.

A defining characteristic of HMs is that Level 1 parameters—for example, site means (β_{0j}), TM effects (β_{1j}), and pretest/posttest slopes (β_{2j})—can be viewed as varying across sites. To represent this in the form of a model, we treat Level 1 parameters as outcomes in a between-site model. We now pose a relatively simple between-site model in which Level 1 parameters (e.g., site TM effects [β_{1j}]) are viewed as varying around corresponding grand means (e.g., a mean TM effect). Thus, we have

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + U_{0j} \qquad U_{0j} \sim N(0, \tau_{00}), \\ \beta_{1j} &= \gamma_{10} + U_{1j} \qquad U_{1j} \sim N(0, \tau_{11}), \\ \beta_{2j} &= \gamma_{20} + U_{2j} \qquad U_{2j} \sim N(0, \tau_{22}), \end{aligned} \tag{3}$$

where γ_{00}, γ_{10}, and γ_{20} represent, respectively, the grand mean for geometry readiness, an overall average TM effect, and an average pretest/posttest slope. The residuals in the above model are termed *random effects*. Thus, U_{0j} captures the deviation of the mean readiness score for site $j(\beta_{0j})$ from γ_{00}, U_{1j} captures the deviation of the TM effect for site $j(\beta_{1j})$ from γ_{10}, and U_{2j} captures the deviation of the pretest/posttest slope for site $j(\beta_{2j})$ from γ_{20}. The random effects are assumed normally distributed as in equation (3). Thus, τ_{00} represents the variation in site means around the grand mean, τ_{11} represents the variation in site TM effects around the average TM effect, and τ_{22} represents the variation in site pretest/posttest slopes around the average slope. Note that part of the variation among the OLS estimates of β_{1j} ($\hat{\beta}_{1j}$) shown in Table 14.1 is likely attributable to estimation error, as well as to underlying between-site differences in the effectiveness of TM. It is the latter source of variation that is captured by the variance parameter τ_{11}. Note also that the random effects in the Level 2 model are assumed to covary: $\mathrm{Cov}(U_{0j}, U_{1j}) = \tau_{01}$, $\mathrm{Cov}(U_{0j}, U_{2j}) = \tau_{02}$, and $\mathrm{Cov}(U_{1j}, U_{2j}) = \tau_{12}$.

We refer to the HM defined by equations (2) and (3) as Model 1. We now fit Model 1 to the data using the HLM5 program (Raudenbush, Bryk, Cheong, & Congdon, 2000). Note that the HLM program uses the EM algorithm and Fisher scoring to obtain maximum likelihood estimates of the Level 1 and Level 2 variance components (i.e., σ^2, τ_{00}, τ_{11}, τ_{22}, and the Level 2 covariances) and then uses these estimates in computing generalized least squares (GLS) estimates of the fixed effects in the model (i.e., γ_{00}, γ_{10}, and γ_{20}) (see Raudenbush & Bryk, 2002, chap. 3 for details). The two parameters of primary interest in Model 1 are γ_{10} (i.e., the average TM effect) and τ_{11} (i.e., the variance component capturing the extent to which the effects of TM vary across sites). As can be seen in Table 14.2, the resulting estimate of the average TM effect is 1.16 and is more than twice its standard error. Although this estimate is extremely close to the estimated TM effect obtained in the single-level OLS analysis (1.08), note that the standard error that we obtain in the HLM analysis is nearly twice as large (0.45 vs. 0.26). Before explaining why this is so, it will first be helpful to focus on the results for τ_{11}. We see that the resulting point estimate is 2.96. In addition, a chi-square test of the hypothesis that $\tau_{11} = 0$ (i.e., a test of homogeneity) results in a test statistic that is highly significant. (See Raudenbush & Bryk, 2002, pp. 63–65, for a discussion of such tests.)

Table 14.2 Treating the Effects of Transition Mathematics (TM) as Varying Across Sites: Hierarchical Models 1 and 2

	Model 1		*Model 2*	
Fixed Effects	*Estimate [SE] (95% CI)*	*t-Ratio*	*Estimate [SE] (95% CI)*	*t-Ratio*
Grand mean (γ_{00})	9.10 [.64] (7.76, 10.44)	14.28**	9.11 [.64] (7.77, 10.44)	14.30**
Overall TM effect (γ_{10})	1.16 [.45] (0.22, 2.10)	2.60*	1.14 [.44] (0.22, 2.06)	2.59*
Average within-site pretest/posttest slope (γ_{20})	0.27 [.03] (0.21, 0.34)	10.69**	0.29 [.02] (0.25, 0.32)	13.94**
Variance Components	*Estimate*	χ^2 *(df)*	*Estimate*	χ^2 *(df)*
Between site				
Variance in site mean readiness (τ_{00})	7.87	708.44** (19)	7.86	695.77** (19)
Variance in site TM effects (τ_{11})	2.96	76.29** (19)	2.84	74.62** (19)
Variance in pretest/posttest slopes (τ_{22})	0.01	26.12 (19)	—	—
Within site				
Residual variance (σ^2)	6.56		6.68	

NOTE: CI = confidence interval.
$^*p < .05$; $^{**}p < .001$.

To grasp the practical significance of this result, it is helpful to consider that the above between-site model essentially constitutes a model for the population of sites similar to those in our study. More specifically, site TM effects, for the population of sites of interest, are conceived as being normally distributed around a mean effect (γ_{10}) with variance τ_{11}. Of course, there is some uncertainty attached to our estimates of γ_{10} and τ_{11}, but a "best guess" based on the above results is that site TM effects are normally distributed with a mean of 1.16 and variance of 2.96. Thus, the effect of TM for sites located near the mean of the distribution is a little over 1 point. However, the effect of TM for a site that is two standard deviations above the mean would, based on this analysis, be equal to $1.16 + 2(\sqrt{2.96}) = 4.60$. In contrast, the effect of TM for sites located two standard deviations below the mean would be equal to $1.16 + 2(\sqrt{2.96}) = -2.28$. Thus, the point estimate of γ_{10}, coupled with the point estimate of τ_{11}, points to substantial variation in the effects of TM across sites.[1]

As noted above, the standard error of the estimate of the average TM effect obtained in the HM analysis is nearly twice as large as the standard error obtained in the single-level OLS analysis. The reason for this is that in the single-level analysis, it is assumed that the 572 student-level observations contained in the study's sample, conditional on the predictors included in the model (i.e., *TRT*, *PRE*), are independent. As such, the observations from the 20 sites are simply pooled in estimating the effects of TM. If this assumption were true, however, one implication would be that there is no between-site variance in the effectiveness of TM (i.e., $\tau_{11} = 0$). The above results of the HLM analysis clearly indicate that how well TM students perform relative to comparison group children will depend to some extent on site membership. This is likely due to a variety of reasons. For example, the TM students located in a particular site will experience a particular implementation of TM, the students in a particular site will likely differ from students in other sites in terms of prior educational experiences, and the like. Such dependencies or clustering in the data are taken into account in HM analyses. Intuitively, given the large amount of between-site variance in TM effects, it is clear that the precision with which we are able to estimate the effects of TM will not depend solely on the number of students in our sample. Rather, this will also depend on the number of sites in our sample (J). As such, the standard error for the point estimate of γ_{10} consists of a part involving the estimate of the between-site variance in TM effects ($\hat{\tau}_{11}$) and a part involving

1. Note that various diagnostics one can compute to assess the plausibility of normality assumptions at Level 2 (e.g., plots of Mahalanobis distances; see Raudenbush & Bryk, 2002, chap. 9) suggest that the assumption of normality is reasonable in the case of this particular analysis. Other model-checking procedures are discussed in other parts of this chapter.

the estimate of within-site error variance ($\hat{\sigma}_2$). As the number of sites in a sample increases, the magnitude of the part involving ($\hat{\tau}_{11}$) diminishes. Note further that if $\hat{\tau}_{11}$ is close to 0, the number of sites becomes immaterial, and the resulting standard error would be driven essentially by the total number of students in TM classrooms and the total number in comparison classrooms.[2]

In Table 14.2, we also see that the point estimate of the grand mean for geometry readiness is approximately 9 points. The resulting estimate for τ_{00}, however, points to substantial variation in site geometry readiness means around the grand mean. To see this, note, for example, that the geometry readiness mean for a site that is two standard deviations above the grand mean would, based on the results for Model 1, be equal to $9.10 + 2(\sqrt{7.89}) = 14.72$.

In addition, a test of the hypothesis that the variance in site pretest/posttest slopes (τ_{22}) is 0 provides some grounds for retaining the null hypothesis. In the interests of parsimony, we now refit our HM with τ_{22} constrained to a value of 0; that is, we remove U_{2j} from our between-site model. In doing so, we are essentially viewing site pretest/posttest slopes as being homogeneous (or parallel). Thus, our Level 2 model is as follows:

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + U_{0j} \qquad U_{0j} \sim N(0, \tau_{00}),\\
\beta_{1j} &= \gamma_{10} + U_{1j} \qquad U_{1j} \sim N(0, \tau_{11}),\\
\beta_{2j} &= \gamma_{20}.
\end{aligned} \tag{4}$$

As can be seen in Table 14.2, the results based on the model defined by equations (2) and (4) (termed Model 2) are extremely similar to those based on our first HM analysis.

We now illustrate the use of HMs in exploring potentially important sources of variability in the effects of TM: differences across sites in implementation, in the characteristics of program participants, and in design.

2. To take into account the uncertainty that stems from substituting point estimates of the variance components into the standard errors for the fixed effects, the HLM program employs critical values based on the family of t-distributions in conducting hypothesis tests regarding fixed effects. Thus, for example, in a test of the hypothesis that the overall TM effect (γ_{10}) is equal to 0, the HLM program employs critical values based on a t-distribution with $J - 1 = 19$ degrees of freedom. Note that when J is small, critical values based on the z-distribution will give rise to rejection rates that are too high and 95% intervals whose levels of coverage are less than nominal. Provided that one's data are not too unbalanced, basing critical values on the family of t-distributions will tend to provide appropriate rejection rates and levels of coverage in small-sample settings. See Raudenbush and Bryk (2002, chap. 9) for further details.

14.3.4. Testing Program Assumptions: The Role of Reading in TM

The analyses above reveal substantial variability in the effectiveness of TM across sites. We now illustrate the use of HMs in helping to identify those aspects of a program that may be critical to its success.

The developers of TM view daily discussion of the reading passages in the TM text as a key element of the program. As such, information regarding the usage of reading in the text was obtained through a teacher questionnaire administered at the end of the school year. As can be seen in Table 14.1, the responses of the TM teachers fall into two categories: those who indicated that they discussed the reading in the text on a daily basis, which we term *high implementation* ($IMPLRDG_j = 1$), and those who indicated that reading was discussed frequently but was not part of the daily routine, which we term *low implementation* ($IMPLRDG_j = 0$).

As can be seen from the data presented in Table 14.1, TM effect estimates tend to be higher in those sites in which the reading passages in the text are discussed on a daily basis. We now examine this more formally by including $IMPLRDG_j$ as a predictor in our between-site model for site TM effects (β_{1j}):

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + \gamma_{01}(\overline{PRE}_{.j} - \overline{PRE}) + U_{0j}\\
&\quad U_{0j} \sim N(0, \tau_{00}),\\
\beta_{1j} &= \gamma_{10} + \gamma_{11} IMPLRDG_j + U_{1j}\\
&\quad U_{1j} \sim N(0, \tau_{11}),\\
\beta_{2j} &= \gamma_{20}.
\end{aligned} \tag{5}$$

Given the coding scheme employed for $IMPLRDG_j$, γ_{10} is the expected effect of TM at low-implementation sites, and γ_{11} represents the expected increment in the effectiveness of TM when the level of implementation at a site is high. Analogous to a regression model, U_{1j} is a residual capturing the deviation of β_{1j} from an expected value based on $IMPLRDG_j$. As such, τ_{11} now represents the remaining variance in site TM effects after taking into account $IMPLRDG_j$.

As can be seen, we have also modeled differences in site geometry readiness means as a function of site pretest means (i.e., $\overline{PRE}_{.j}$). Thus, γ_{01} captures the *between-site* relationship between pretest scores and geometry readiness (i.e., the expected change in site geometry readiness means when site pretest means increase one unit). Conversely, γ_{20} captures the *within-site* relationship between pretest scores and geometry readiness (i.e., the expected difference in geometry scores for two students in the same site whose pretest scores differ by one unit). Note that $\overline{PRE}$ represents the

Table 14.3 Modeling Site Transition Mathematics (TM) Effects as a Function of Implementation and Other Site Characteristics: Hierarchical Models 3, 4, and 5

	Model 3		*Model 4*		*Model 5*	
Fixed Effects	*Estimate [SE] (95% CI)*	*t-Ratio*	*Estimate [SE] (95% CI)*	*t-Ratio*	*Estimate [SE] (95% CI)*	*t-Ratio*
Model for site mean readiness						
Grand mean (γ_{00})	9.10 [.28] (8.51, 9.69)	32.41**	9.10 [.28] (8.51, 9.69)	32.41**	9.10 [.28] (8.51, 9.69)	32.42**
Between-site pretest/posttest slope (γ_{01})	0.62 [.07] (0.47, 0.77)	8.93**	0.62 [.07] (0.48, 0.77)	8.96**	0.62 [.07] (0.47, 0.77)	8.94**
Model for site TM effects						
Expected TM effect at low-implementation sites (γ_{10})	0.12 [.53] (−0.99, 1.23)	0.22	0.09 [.54] (−1.05, 1.23)	0.16	0.13 [.55] (−1.04, 1.29)	0.23
Expected increase in effects of TM at high-implementation sites (γ_{11})	2.03 [.76] (0.43, 3.62)	2.68*	2.11 [.78] (0.46, 3.76)	2.72*	2.02 [.78] (0.38, 3.67)	2.59*
Relationship between site pretest means and the effects of TM (γ_{12})	—	—	−0.07 [.10] (−0.28, 0.14)	−0.78	—	—
Expected difference in effects of TM between RA and non-RA sites (γ_{13})	—	—	— (−1.86, 1.43)	—	−0.22 [.78]	−0.28
Model for within-site pretest/posttest slopes						
Average within-site slope (γ_{20})	0.29 [.02] (0.25, 0.33)	13.99**	0.29 [.02] (0.25, 0.33)	14.00**	0.29 [.02] (0.25, 0.33)	13.98**
Variance Components	*Estimate*	*χ^2 (df)*	*Estimate*	*χ^2 (df)*	*Estimate*	*χ^2 (df)*
Between site						
Variance in site mean readiness (τ_{00})	1.33	115.56** (18)	1.33	115.59** (18)	1.33	115.56** (18)
Variance in site TM effects (τ_{11})	1.86	51.56** (18)	1.93	49.58** (17)	2.02	51.42** (17)
Within site						
Residual variance (σ^2)	6.69		6.69		6.69	

NOTE: The resulting estimate for $\hat{\tau}_{11}$ based on Models 4 and 5 is slightly larger than the estimate based on Model 3. This can occur when one adds predictors to a Level 2 equation that are unrelated to the Level 1 parameter that is being modeled (e.g., β_{1j}). See Raudenbush and Bryk (2002) for details. CI = confidence interval.
*$p < .05$; **$p < .001$.

mean of the $\overline{PRE}_{.j}$ values for the 20 sites. By virtue of centering $\overline{PRE}_{.j}$ around $\overline{PRE}$, γ_{00} retains its meaning as the grand mean for geometry readiness.

We refer to the HM defined by equations (2) and (5) as Model 3. As can be seen in Table 14.3, the resulting estimate for the expected effect of TM at low-implementation sites is approximately a tenth of point, and the corresponding t-ratio is extremely small. Thus, when reading in the TM text is not discussed on a daily basis, the results suggest that, on average, TM and more conventional curricula are equally effective with respect to student performance in the domain of geometry readiness. However, the point estimate for γ_{11} is approximately 2 points and more than twice its standard error. This suggests that when reading is discussed on a daily basis, the expected effect of TM is over 2 points: $0.12 + 2.03 = 2.15$.

Note that the estimate of τ_{11} that we obtain when implementation of reading is included in the analysis is substantially smaller than the estimate we obtain based on Model 2 (i.e., 1.86 vs. 2.84). Thus, $IMPLRDG_j$

accounts for approximately 35% of the variability in site TM effects.

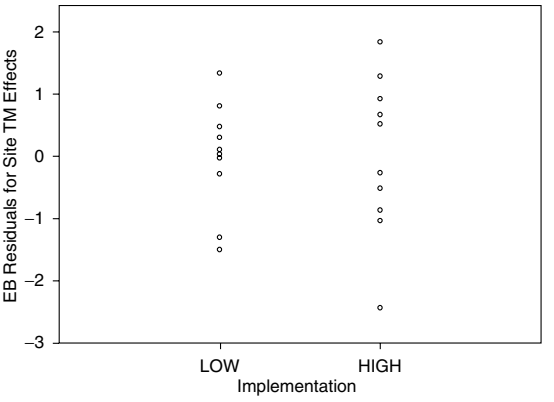
14.3.5. Taking a Closer Look at the Results Concerning Implementation: Examining Residuals

In Table 14.1, we see that the OLS estimate of the TM effect for one of the high-implementation sites (Site 8) is negative (-1.68) and substantially smaller than the TM effect estimate for any of the other high-implementation sites. Thus, Site 8 appears to be an outlier. More formally, to help identify outlying sites, we can construct plots based on OLS or empirical Bayes (EB) residuals for each site. An OLS residual would be computed by taking the OLS estimate of the site TM effect for site j (i.e., $\hat{\beta}_{1j}$) and subtracting a fitted value based on the *IMPLRDG* value for site j (i.e., $FV_{1j} = 0.12 + 2.03\,IMPLRDG_j$): $\hat{U}_{1j} = (\hat{\beta}_{1j} - FV_{1j})$. Computing EB residuals (i.e., U^*_{1j}) entails shrinking OLS residuals toward a value of 0. In essence, for sites in which the precision of $\hat{\beta}_{1j}$ is relatively high, the degree of shrinkage will be minimal. However, for sites in which the precision is low, the degree of shrinkage toward 0 will be substantial. Thus, EB residuals are, in a sense, adjusted for estimation error connected with the $\hat{\beta}_{1j}$ (see Raudenbush & Bryk, 2002, pp. 45–51).

As can be seen in the plot of EB residuals versus fitted values (see Figure 14.1), Site 8 stands clearly apart from the other high-implementation sites. One possible explanation centers on the difficulties encountered by many of the TM students at this site in reading the text; for many of these students, English was their second language. Although we cannot say with certainty that this explanation is the correct one, it is consistent with the notion that reading plays a key role in the TM curriculum.

To what extent is Site 8 affecting our results? When we set aside Site 8 and reestimate Model 3, we obtain, as might be expected, a larger estimate for the coefficient for $IMPLRDG_j$ (i.e., 2.29 [$SE = 0.65$]). However, from a practical standpoint, the conclusions we might reach concerning the effects of TM given a high level of implementation are quite similar.[3]

Figure 14.1 EB Residuals Versus Implementation of Reading



14.3.6. Considering Possible Confounding Variables

An issue of fundamental importance that we must consider in our analysis concerning implementation is that the 20 TM teachers in this study were not randomly assigned to different levels of implementation. Rather, for reasons that are unclear, 10 of the TM teachers discussed the reading in the TM text with their students on a daily basis, and 10 did not. Put differently, the TM teachers self-selected into different levels of implementation. Thus, in terms of trying to assess whether TM is in fact more effective when reading in the text is discussed on a daily basis, we are clearly in a quasi-experimental setting. As such, we need to consider whether there are other factors at work that account for the results that we obtained for $IMPLRDG_j$ (i.e., factors that are associated with $IMPLRDG_j$ *and* with TM/comparison class contrasts). That is, we need to attend to possible confounding variables. As will be seen, we take up this issue in several places below.

14.3.7. Who Benefits From the Program?

A key question that often arises when an innovative curriculum is developed is the following: Will students from a broad range of backgrounds benefit from the newly developed curriculum, or will the curriculum primarily be successful at sites that serve students who tend to have higher levels of prior achievement or are from more advantaged backgrounds? In the TM study,

3. Note that by employing recently developed estimation tools termed *Markov chain Monte Carlo* (MCMC) methods, it is possible to refit HMs under heavy-tailed distributional assumptions. That is, rather than assume normality at Levels 1 and 2, one can specify t-distributional assumptions with small degrees of freedom (e.g., 4) at each level. This has the effect of downweighting possible outliers, thereby yielding robust results for parameters of interest (see, e.g., Seltzer, Novak, Choi, & Lim, 2002). One can readily fit HMs under t-distributional assumptions using the software program WinBUGS (Spiegelhalter, Thomas, Best, & Gilks, 2000), which is freely available via the Web.

there are substantial differences across sites in terms of prior levels of student achievement. This provides an opportunity to explore whether the effects of TM are related to differences in prior levels of achievement. Thus, we now include site mean pretest scores ($\overline{PRE}_{.j}$) as a predictor in the Level 2 model for site TM effects:

$$\begin{aligned}\beta_{1j} &= \gamma_{10} + \gamma_{11} IMPLRDG_j \\ &\quad + \gamma_{12}(\overline{PRE}_{.j} - \overline{PRE}) + U_{1j} \\ U_{1j} &\sim N(0, \tau_{11}). \end{aligned} \tag{6}$$

As in a multiple regression analysis, γ_{12} represents the expected increment in β_{1j} when $\overline{PRE}_{.j}$ increases one unit, holding constant $IMPLRDG_j$, and, similarly, γ_{11} now represents the expected increment in β_{1j} when reading is discussed on a daily basis, holding constant $\overline{PRE}_{.j}$. By virtue of centering $\overline{PRE}_{.j}$ around the grand mean of the site mean pretest scores, γ_{10} represents the expected effect of TM in a low-implementation site whose pretest mean is equal to the grand mean.

In Table 14.3, under the heading Model 4, we see that the resulting point estimate of γ_{12} is -0.07 ($SE = 0.10$). Thus, there appears to be no evidence of a systematic relationship between site pretest means and the effectiveness of TM.

Given this result, it is not surprising that the estimate for the fixed effect connected with $IMPLRDG_j$ is extremely similar to the estimate obtained in the previous analysis. Furthermore, note that even if there were, for example, a positive relationship between site pretest means and the effectiveness of TM, this would still have little impact on the estimated coefficient for $IMPLRDG_j$. This is due to the fact that there is no evidence of association between site pretest means and level of implementation. For example, regressing $IMPLRDG_j$ on $\overline{PRE}_{.j}$ in a logistic regression analysis, we obtain an estimate of the coefficient for $\overline{PRE}_{.j}$ of 0.06 ($SE = 0.11$). Thus, selection into different levels of implementation on the part of TM teachers does not appear to be related to differences in site mean pretest performance.

14.3.8. Do Differences in Design Relate to Differences in the Magnitude of TM Effects Across Sites?

As noted above, the assignment of teachers to TM was random in the case of 10 sites. In the case of those sites in which assignment was not random, one potential concern is that district math coordinators or school principals, in selecting teachers to teach TM, may have selected those individuals whom they viewed as being highly skilled and successful math teachers. If this were the case, then estimates of the effects of TM would tend to be biased in these sites; specifically, they would be larger, on average, than the TM effect estimates for those sites in which assignment was random. To explore this possibility, we expand our Level 2 model for site TM effects as follows:

$$\begin{aligned}\beta_{1j} &= \gamma_{10} + \gamma_{11} IMPLRDG_j + \gamma_{13} RA_j + U_{1j} \\ U_{1j} &\sim N(0, \tau_{11}), \end{aligned} \tag{7}$$

where RA_j takes on a value of 1 if the assignment of teachers at site j is random (0 otherwise); γ_{13} represents the expected decrement (or increment) in the effects of TM when assignment is random, holding constant $IMPLRDG_j$; γ_{11} now captures how differences in implementation relate to differences in site TM effects, holding constant the type of assignment; and finally, γ_{10} represents the expected TM effect in sites in which implementation is low and the assignment of teachers is nonrandom.

In Table 14.3, under the heading Model 5, we see that the resulting point estimate for γ_{13} is less than a quarter of a point, and that the corresponding t-ratio is extremely small. Thus, the effects of TM in sites in which assignment was not random appear to be similar to the effects of TM in sites in which assignment was random.

Not surprisingly, given this result, we also see that the results concerning the relationship between implementation and the effectiveness of TM are extremely similar to those based on Model 3. Note that even if there were evidence of larger TM effects in sites in which assignment was not random, the results concerning implementation of reading would remain essentially unchanged. This is due to the fact that type of assignment and level of implementation are unassociated: Among the 10 sites in which assignment is nonrandom, implementation is high in 5 sites and low in 5, and the pattern is the same among the 10 sites in which assignment is random. Thus, selection into different levels of implementation is unrelated to the type of assignment employed at a site.

Note more generally that variables capturing other aspects of the design and conduct of a study that are of potential concern to investigators could be employed as predictors in Level 2 models.

14.3.9. Taking Other Factors Into Consideration

In addition to the variables considered above, we also explored a number of other site characteristics

that could conceivably relate to differences in the magnitude of TM effects (e.g., differences across sites in the text employed in comparison classes) and found no systematic relationships. Furthermore, in addition to the factors considered above (site pretest performance, type of assignment), we also examined a number of other factors that we thought could conceivably account for the results that we obtained regarding implementation of reading (e.g., how far in the text each TM instructor had gotten by the end of the school year), and no confounding variables emerged. Our search was by no means exhaustive. For example, 10 TM teachers were randomly selected to keep diaries. Combing through these diaries might suggest further factors to explore.

Note that the estimates that we obtained of the overall effects of TM reflect the fact that implementation was less than ideal in numerous sites. On one hand, such estimates may be of interest to policymakers because they reflect the various challenges and difficulties experienced by practitioners in the field (see Shadish, Cook, & Campbell's [2002, pp. 319–320] discussion of intent-to-treat analyses). But on the other hand, if we are interested in the effects of TM when it is implemented with high fidelity, such estimates are misleading.

As can be seen in Table 14.3, even after including $IMPLRDG_j$ in our analyses, substantial between-site variability in the effects of TM remains. In general, an appreciable amount of between-site variability is likely to be connected with key differences across sites in implementation, in the characteristics of program participants, and in certain aspects of design. However, at least some of the variability will likely be due to unique site characteristics (e.g., unexpected events, unusually strong administrative support).

14.4. Designs in Which Organizational Units Are Nested Within Treatment Type: Reanalyses of the Integrative Mathematics Assessment Data

14.4.1. Background

In contrast to more traditional mathematics instruction, which is characterized by drill and the memorization and application of algorithms, influential documents such as the National Council of Teachers of Mathematics (NCTM) Standards (1989) call for instructional practices that involve eliciting and building on students' thinking and that provide students with opportunities to engage with mathematical concepts in solving problems. Carrying out instruction in this way, however, can be extremely challenging. In this connection, Saxe et al. (1999) and Gearhart et al. (1999) conducted a study that entailed developing and implementing two programs intended to help teachers develop the skills necessary for teaching mathematics for upper elementary students in ways consistent with the NCTM standards, and comparing the mathematics learning of students taught by teachers participating in these programs.

Both programs focused primarily on instruction in the domain of fractions and centered on teachers' use of a textbook titled *Seeing With Fractions.* One program, called Integrated Mathematics Assessment (IMA), focused on helping teachers (a) develop more sophisticated understandings of fractions and related topics, (b) gain insight into the ways in which students' understandings of fractions change over time, and (c) learn instructional practices that entail, for example, eliciting and building on students' understandings of fractions. The aim of the second program, called Collegial Support (SUPPORT), "was to provide teachers opportunities to reflect on their practices with a community of practitioners engaged in similar efforts" (Gearhart et al., 1999, p. 291). In this model, participating teachers develop an agenda of topics that they wish to discuss at each session.

A total of 16 upper elementary teachers in the greater Los Angeles area participated in IMA and SUPPORT. All teachers had prior experience using the *Seeing With Fractions* text and volunteered to participate in the study. Nine teachers were assigned to IMA and 7 to SUPPORT via a random assignment procedure described in Gearhart et al. (1999).

Prior to the start of the school year, the IMA teachers participated in a 5-day summer institute. During the school year, the IMA teachers attended 13 evening meetings, which were held approximately every 2 weeks. The SUPPORT teachers attended 2 full-day meetings and 7 evening meetings held monthly.

The students of each teacher in the study were administered a series of pretests at the start of the school year and a series of posttests immediately after the last module of the *Seeing With Fractions* text was completed. In addition, a measure of student English-language proficiency was obtained at the start of the school year. In this section of our chapter, we present a series of HM analyses that focuses on the

performance of students in IMA and SUPPORT classrooms on a posttest that measures problem-solving skills in the fractions domain. The items on this test cannot be solved through the mere application of computational algorithms.

Although the IMA teachers and SUPPORT teachers in our sample were, on average, very similar in terms of experience and training, students in the IMA classes tended to be somewhat more advantaged than students in SUPPORT classes with respect to several intake characteristics (see below). This is connected to the fact that although six of the nine IMA teachers were located in six different schools, three were located in the same school. The latter three teachers were randomly assigned as a group to IMA. A concern was that assigning, say, two of these teachers to IMA and one to SUPPORT might result in the diffusion of information and ideas from the IMA teachers to the SUPPORT teacher during the school year. This particular school, however, served relatively advantaged students. As will be seen, we adjust for differences in various key intake characteristics in our analyses and conduct a number of sensitivity analyses, including refitting key models with these three teachers set aside.

An important feature of this study is that in-depth information was obtained via classroom observations of each teacher's instructional practices over the course of fractions instruction. Data from these observations were used to construct scales that capture the extent to which a teacher's instructional practices are aligned with various reform-minded principles. Of particular interest is a scale capturing the extent to which a teacher provides opportunities for engagement with mathematical (i.e., fractions) concepts in discussions of problem solving in ways that build on students' thinking (Gearhart et al., 1999, p. 303). This measure, which we term *ALIGN,* will be used as a Level 2 predictor in two of the key analyses presented below. Studying how differences in this aspect of practice relate to differences in student outcomes figures prominently in Saxe et al.'s (1999) and Gearhart et al.'s (1999) work.

14.4.2. Specifying a Within-Class Model: The Use of Grand-Mean Centering

The HMs that we employ in our analyses are two-level models, each of which consists of a within-class model (Level 1) and a between-class model (Level 2). We now pose the following within-class model for the $J = 16$ classrooms in our sample ($j = 1 \ldots 16$):

$$\begin{aligned} Y_{ij} = \beta_{0j} &+ \beta_{1j}(PREPS_{ij} - \overline{PREPS}_{..}) \\ &+ \beta_{2j}(INCIP_{ij} - \overline{INCIP}_{..}) \\ &+ \beta_{3j}(ELP_{ij} - \overline{ELP}_{..}) + \varepsilon_{ij} \\ &\varepsilon_{ij} \sim N(0, \sigma^2). \end{aligned} \tag{8}$$

Y_{ij} and $PREPS_{ij}$ are, respectively, the problem-solving posttest and pretest scores for student i in classroom j. Note that the maximum possible score on both tests is 13 and that the tests are composed of very similar though not identical items. $INCIP_{ij}$ is an indicator variable that takes on a value of 1 if student i in classroom j demonstrated an incipient understanding of fractions based on a special pretest (0 otherwise; see Saxe et al., 1999, for details). ELP_{ij} is an indicator variable that takes on a value of 1 if student i in classroom j is categorized as being fluent in English (0 otherwise), and β_{1j}, β_{2j}, and β_{3j} are slopes capturing the relationships between the predictors *PREPS*, *INCIP*, and *ELP* and problem-solving posttest scores. The ε_{ij} are errors assumed independent and normally distributed with mean 0 and variance σ^2.

In contrast to the Level 1 model that we employed in our analyses of the TM data (equation (2)), the predictors in equation (8) have been centered around their grand means. The type of centering that we choose has important implications for the interpretation of the intercept term (β_{0j}) in the model. When we employ group-mean centering, β_{0j} represents the (unadjusted) mean outcome score for group j. When we employ grand-mean centering, β_{0j}, analogous to ANCOVA models, represents an adjusted mean outcome score for group j. Thus, for example, if *PREPS* scores in a particular class are, on average, lower than the grand-mean pretest score, and if pretest and posttest scores are positively related, then the expected outcome score for that class (β_{0j}) will be adjusted upwards. For classes with pretest scores that are, on average, higher than the grand mean, the expected outcome score for these classes will be adjusted downwards. Thus, grand-mean centering at Level 1 provides a way of controlling for differences among classes in their student intake characteristics (see Raudenbush & Bryk, 2002, chaps. 2, 5).

Note that class mean *PREPS* scores ($\overline{PREPS}_{.j}$) are, on average, higher for IMA classes than for SUPPORT classes (3.45 vs. 2.02; $t = 2.72$, $p = .02$), where 3.45 represents the average of the $\overline{PREPS}_{.j}$ values for the nine IMA classes, and 2.02 represents the average of

the $\overline{PREPS}_{.j}$ values for the seven SUPPORT classes. Class mean *ELP* and *INCIP* scores (i.e., the proportions of students in a class who, respectively, are English-language proficient and who demonstrated incipient understanding of fractions) are also somewhat higher, on average, for IMA classes than for SUPPORT classes (*ELP* : 0.94 vs. 0.81, $t = 1.93$, $p = .07$; *INCIP* : 0.77 vs. 0.55, $t = 1.88$, $p = .08$). When we set aside the three IMA teachers located in the school serving relatively advantaged students, the differences for *PREPS*, *ELP*, and *INCIP* are, respectively, 2.68 versus 2.02 ($t = 1.87$, $p = .09$), 0.91 versus 0.81 ($t = 1.20$, $p = .26$), and 0.68 versus 0.55 ($t = 1.01$, $p = .34$).

14.4.3. Assessing Contextual Effects

In HMs for nested designs, such as the ones employed in Gearhart et al.'s (1999) and Saxe et al.'s (1999) studies, adjusted means are the Level 1 parameters of primary interest. In a Level 2 model, we then model differences in adjusted means as a function of various key characteristics of the organizational units (e.g., classes) in our sample.

Before comparing outcomes for students in IMA and SUPPORT classes, we first pose a between-class model of the following form:

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + \gamma_{01}\,(\overline{ELP}_{.j} - \overline{ELP}) + U_{0j} \\
U_{0j} &\sim N(0, \tau_{00}), \\
\beta_{1j} &= \gamma_{10}, \\
\beta_{2j} &= \gamma_{20}, \\
\beta_{3j} &= \gamma_{30}. \qquad (9)
\end{aligned}$$

In the above model, γ_{10}, γ_{20}, and γ_{30} represent, respectively, the average within-class pretest/posttest slope, the average within-class *INCIP*/posttest slope, and the average within-class *ELP*/posttest slope. In initial models that we fit to the data, tests of homogeneity pointed to little variation in Level 1 slopes. Thus, as can be seen in equation (9), we did not specify random effects in the equations for β_{1j}, β_{2j}, and β_{3j}. In addition to including ELP_{ij} as a predictor in our within-class model, we have included the proportion of English-language proficient students in a class ($\overline{ELP}_{.j}$) as a predictor of β_{0j}. Note that γ_{30} represents the expected difference in problem-solving posttest scores between two students—one of whom is proficient in English and the other who is not—who are in the same class and who have identical pretest scores and levels of incipient knowledge of fractions. Raudenbush and Bryk (2002) term γ_{30} a person-level effect. In contrast, γ_{01} in equation (9) is a contextual effect connected with the extent to which students in a class are proficient in English. To grasp the meaning of γ_{01}, consider two students with identical levels of English proficiency, as well as identical pretest values and levels of incipient understanding. Suppose, however, that one student is in a class in which virtually all students are English-language proficient (Class A), whereas the other is in a class in which approximately half of the students are English-language proficient (Class B). By virtue of these differences in classroom composition, one can imagine that pacing might be faster in Class A and/or that coverage of the material might be more thorough in Class A. As a result, the student in Class A might learn appreciably more than the student in Class B, even though the two students are identical in terms of their *ELP*, *PREPS*, and *INCIP* values. Such differences in outcomes, termed a *contextual effect*, would be captured by γ_{01}.

The equation for β_{0j}, in contrast to the other Level 2 equations, contains a random effect (U_{0j}). The random effects are assumed normally distributed with variance τ_{00}, which represents the variance in adjusted means that remains after taking into account the effects of $\overline{ELP}_{.j}$.

We now fit the model defined by equations (8) and (9), termed Model 1 in Table 14.4, to the data. One noteworthy finding is that whereas the person-level effect of English-language proficiency (γ_{30}) is negligible, the contextual effect (γ_{01}) appears to be substantial. Consider two students who are identical in terms of their level of English-language proficiency, problem-solving pretest score, and level of incipient knowledge. If one student is in a class in which all children are proficient ($\overline{ELP}_{.j} = 1.00$) and the other is in a class in which six tenths of the students are proficient ($\overline{ELP}_{.j} = 0.60$), the expected difference in posttest scores between two such students would be $4.02 \times (1 - 0.60) = 1.61$. (Note that the minimum and maximum values for $\overline{ELP}_{.j}$ in our sample are 0.60 and 1.00, respectively.)

Turning to the results for the other fixed effects in the model, we see that the estimate of the grand-mean posttest score (γ_{00}) is 5.85 points and that the point estimates for the average within-class *PREPS* and *INCIP* slopes (γ_{10}, γ_{20}) are positive and statistically significant. Finally, the results for τ_{00} indicate that the remaining variability in adjusted class mean posttest performance is appreciable.

Before moving to the next section, note that when the 3 IMA teachers in the school serving relatively advantaged students are set aside, the resulting point estimate for the contextual effect of *ELP* is slightly larger than the result based on the entire sample of

Table 14.4 IMA Study Analyses: Models 1, 2, 3, and 4

	Model 1		*Model 2*		*Model 3*		*Model 4*	
Fixed Effects	*Estimate [SE] (95% CI)*	*t-Ratio*	*Estimate [SE] (95% CI)*	*t-Ratio*	*Estimate [SE] (95% CI)*	*t-Ratio*	*Estimate [SE] (95% CI)*	*t-Ratio*
Model for adjusted class means								
Grand mean (γ_{00})	5.85 [.24] (5.33, 6.36)	23.95**	5.84 [.23] (5.34, 6.34)	25.26**	5.84 [.20] (5.41, 6.28)	29.17**	5.85 [.20] (5.42, 6.28)	29.13**
Class ELP (γ_{01})	4.02 [1.84] (0.08, 7.97)	2.18*	2.69 [1.94] (−1.51, 6.88)	1.38	3.51 [1.75] (−.30, 7.32)	2.01	4.30 [1.56] (0.93, 7.67)	2.76*
IMA/SUPPORT contrast (γ_{02})	—	—	0.86 [.53] (−.28, 2.01)	1.63	0.49 [.49] (−.58, 1.56)	1.00	—	—
Alignment (γ_{03})	—	—	—	—	0.80 [.35] (0.04, 1.56)	2.29*	0.92 [.33] (0.21, 1.63)	2.78*
Average within-class pretest/posttest slope (γ_{10})	0.59 [.07] (0.45, 0.72)	7.96**	0.57 [.07] (0.44, 0.71)	7.77**	0.58 [.07] (0.44, 0.72)	7.91**	0.59 [.07] (0.45, 0.73)	8.11**
Average within-class incipient/posttest slope (γ_{20})	1.13 [.35] (0.44, 1.82)	3.20**	1.12 [.35] (0.43, 1.80)	3.17**	1.17 [.35] (0.48, 1.86)	3.34**	1.18 [.35] (0.49, 1.87)	3.39**
Average within-class ELP/posttest slope (γ_{30})	0.02 [.48] (−0.92, 0.96)	0.04	0.03 [.48] (−0.92, 0.97)	0.06	0.02 [.48] (−0.93, 0.96)	0.04	0.01 [.48] (−0.93, 0.95)	0.02
Variance Components	*Estimate*	*χ^2 (df)*	*Estimate*	*χ^2 (df)*	*Estimate*	*χ^2 (df)*	*Estimate*	*χ^2 (df)*
Between class								
Variance in adjusted means (τ_{00})	0.64	42.90** (14)	0.54	35.65** (13)	0.33	24.28* (12)	0.33	26.44* (13)
Within class								
Residual variance (σ^2)	7.24		7.23		7.24		7.23	

NOTE: IMA = Integrated Mathematics Assessment; SUPPORT = Collegial Support; ELP = English-language proficiency.
*$p < .05$; **$p < .001$.

16 teachers—that is, 4.34 ($SE = 2.09, t = 2.08$, $p = .06$) versus 4.02.

14.4.4. Comparing the Performance of Students in IMA and SUPPORT Classes

We now add a predictor to the Level 2 equation for β_{0j} that denotes whether class j was taught by a teacher who participated in IMA ($IMA_j = 1$) or by a teacher who participated in SUPPORT ($IMA_j = 0$):

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\overline{ELP}_{.j} - \overline{ELP}) + \gamma_{02}(IMA_j - \overline{IMA}) + U_{0j}$$

$$U_{0j} \sim N(0, \tau_{00}). \quad (10)$$

In this model, γ_{02} represents the expected difference in posttest performance between students taught by IMA teachers and students taught by SUPPORT teachers, holding constant the student-level predictors in our model and the proportion of English-language proficient students in a class. That is, for students with similar values for the student-level predictors and who are in classes with similar proportions of English-language proficient students, the expected difference in posttest performance between those students in classes taught by IMA teachers and those taught by SUPPORT teachers is γ_{02}.

As can be seen in Table 14.4 under the heading Model 2, the point estimate for the IMA/SUPPORT contrast is positive and slightly under a point. However, the corresponding t-ratio is substantially less than 2. We also see that the resulting point estimate of the coefficient for $\overline{ELP}_j$ is appreciably smaller than the estimate obtained in the previous analysis (2.68 vs. 4.02) and that the corresponding t-ratio is substantially smaller than 2. A plot of residuals versus fitted values did not reveal any unusual cases.

Note that if we conduct an OLS analysis in which we simply regress student posttest problem-solving scores on the set of student and class characteristics

in the above model, we obtain an estimate of the IMA/SUPPORT contrast that is very similar to the estimate reported in Table 14.4 (i.e., 0.89). However, the resulting standard error is substantially smaller (0.34 vs. 0.53), and the corresponding t-ratio and p-value ($t = 2.62$, $p = .01$) point strongly toward a positive contrast. But, as in the case of the OLS student-level analysis of the TM data, such an analysis ignores the dependencies among student observations nested within Level 2 units and hence can result in standard errors for parameters of interest that are misleadingly small. As explained earlier, such dependencies or clustering in the data are reflected in the standard errors obtained in HM analyses. Note that, for similar reasons, the OLS analysis also results in a substantially smaller standard error for the estimate of the contextual effect of English-language proficiency and, in connection with this, a t-ratio of 2.07 and a p-value of .04.

Before proceeding to the next section, note that when we delete the three IMA teachers in the school serving relatively advantaged children, we obtain slightly larger estimates for the contextual effect for English-language proficiency (3.11; $SE = 2.04$, $t = 1.52$, $p = .16$) and for the IMA/SUPPORT contrast (1.02; $SE = 0.57$, $t = 1.79$, $p = .10$). Thus, we obtain similar results whether we include or exclude these three teachers.

14.4.5. Studying the Relationship Between Teacher Practice and Problem-Solving Outcomes

Compared with the estimate of τ_{00} based on Model 1, adding the IMA indicator variable to the model results in a reduction in variance of approximately 15% (i.e., 0.54 vs. 0.64). This would seem to signal that the information contained in one predictor is not simply duplicating the information contained in the other.

The predictor IMA_j simply denotes participation in the IMA or SUPPORT programs. Is there a more proximal factor (i.e., a factor more directly connected with instruction) underlying the positive estimate obtained for the IMA/SUPPORT contrast?

At this juncture, we employ the teacher practice measure described above (i.e., *ALIGN*) as a Level 2 predictor. Note that a value of 2 on this scale corresponds to a very high level of implementation, and a value of -2 corresponds to an extremely low level of implementation. The *ALIGN* values for the teachers in our sample vary considerably, ranging from -0.75 to 1.62. The mean *ALIGN* value for IMA teachers is somewhat higher than the mean value for SUPPORT teachers (0.57 vs. 0.24), although the difference is modest in magnitude and not statistically significant ($t = 1.10$, $p = .30$).

We now expand the between-class equation for β_{0j} as follows:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}\,(\overline{ELP}_{.j} - \overline{ELP}) + \gamma_{02}(IMA_j - \overline{IMA}) \\ &\quad + \gamma_{03}(ALIGN_j - \overline{ALIGN}) + U_{0j} \\ U_{0j} &\sim N(0, \tau_{00}). \end{aligned} \tag{11}$$

As can be seen in Table 14.4 under the heading Model 3, adding *ALIGN* to the model results in a substantial reduction in the estimate of the IMA/SUPPORT contrast (0.49 [$t = 1.00$] vs. 0.86). Moreover, the results signal an appreciable positive relationship between *ALIGN* and posttest performance. Holding constant all other predictors in the model, when we consider two teachers with *ALIGN* values that are 2 points apart (e.g., 1.50 vs. -0.50), the expected difference in student posttest performance is 1.60 points (i.e., 2×0.80). Furthermore, we see that the point estimate for τ_{00} drops markedly from a value of 0.54 in the previous analysis to a value of 0.33, which represents a decrease of nearly 40%. We also see that adding *ALIGN* to the model results in an appreciable increase in the point estimate for the contextual effect of English-language proficiency (3.51; $t = 2.01$).

Note that the number of Level 2 units in a sample (J) sets limits on the number of predictors that we can include in a Level 2 equation at one time. Analogous to conducting regression analyses in small-sample settings, when $J = 16$, one should be very cautious about including more than two predictors in a Level 2 equation at one time. Of course, much care is called for when employing one or two predictors. One mitigating factor in the case of Model 3 is that *ALIGN* and $\overline{ELP}_{.j}$ are essentially uncorrelated ($r = -0.09$). However, to help gauge the sturdiness of these results, we obtained robust estimates of the fixed effects in Model 3 using an estimation approach outlined in an article by Seltzer, Novak, Choi, and Lim (2002).[3] These estimates were virtually identical to those reported in Table 14.4. In addition, we sequentially deleted a number of possible leverage points (e.g., the class with the largest $ALIGN_j$ value; the class with the smallest $\overline{ELP}_{.j}$ value). These "leave-one-out" analyses also yielded results consistent with those in Table 14.4.

Omitting IMA_j and retaining $ALIGN_j$ in our Level 2 model, we find that the point estimate for τ_{00} remains

virtually unchanged (i.e., 0.33; see the results for Model 4 in Table 14.4). In addition, we see that the point estimates of the coefficients for $ALIGN_j$ and $\overline{ELP}_{.j}$ increase somewhat. Note also that setting aside the three IMA teachers in the school serving relatively advantaged children and refitting Model 4, we obtain extremely similar results for *ALIGN* (0.98; $SE = 0.43$, $t = 2.26$, $p = .05$) and for the contextual effect of *ELP* (4.26; $SE = 1.81$, $t = 2.36$, $p = .04$).

14.4.6. Attending to Possible Confounding Variables in Drawing Conclusions Regarding Alignment

Although the above results suggest that differences in alignment with reform-minded instructional practices underlie differences in student problem-solving performance, it is important that we attend to other variables that may be associated with *ALIGN* and with problem-solving outcomes. We saw above that the correlation between *ALIGN* and the proportion of English-language proficient students in a class is very low. Correlations between *ALIGN* and other class compositional characteristics (e.g., class mean problem-solving pretest scores) are weak as well. Thus, the extent to which teacher practice is aligned with reform-minded principles appears not to depend on the compositional characteristics of classes.

In terms of teacher experience and training, although the correlation between *ALIGN* and years of teaching experience is close to 0 ($r = 0.02$), we do see a moderately large correlation between *ALIGN* and the amount of relevant prior professional development and training (*PDT*) that a teacher received ($r = 0.52$). However, adding *PDT* to Model 4, we obtain results for *ALIGN* (1.07; $SE = 0.39$, $t = 2.78$, $p = .02$) and for the contextual effect of *ELP* (3.85; $SE = 1.67$, $t = 2.31$, $p = .04$) that are very similar to those reported in Table 14.4. Furthermore, the results for *PDT* suggest that holding constant *ALIGN* and $\overline{ELP}_{.j}$, *PDT* is not systematically related to problem-solving outcomes (-0.31; $SE = 0.38$, $t = -0.81$, $p = .44$).[4]

Thus, as in the case of the TM study, we see that the collection of implementation data greatly increased the value of the IMA study. Although various conceptual examinations of mathematics learning point to the potential value of reform-minded instructional practices, Saxe et al. (1999) note that few studies empirically examine the relationship between student learning and these practices. As we saw above, the degree to which instruction was aligned with reform-minded practices emerged as a key factor with respect to student problem-solving outcomes.[5]

14.5. Recap, Implications, and New Directions

14.5.1. Obtaining More Appropriate Standard Errors

The above examples helped illustrate the use of HMs in obtaining more appropriate standard errors for fixed effects of interest. We saw that conducting OLS analyses that ignore the nested structure of multisite data can result in misleadingly small standard errors. Specifically, in the case of the IMA example, recall that the standard error for the estimate of the IMA/SUPPORT contrast was 40% smaller than the standard error based on an HM analysis and that the corresponding t-ratios based on the OLS and HM analyses were 2.62 and 1.63, respectively.

4. The procedure by which teachers were assigned to IMA and SUPPORT involved a priori matching based on years of teaching experience (*EXPER*) and *PDT* (see Gearhart et al., 1999). Background information concerning, for example, the demographic characteristics of students in a given teacher's school entered into this procedure as well. (Because the IMA and SUPPORT programs began during the summer, it was not possible to employ class compositional characteristics such as prior mean achievement scores in the matching process.) The goal of the matching procedure in the IMA study was not to construct a series of matched pairs as in the case of the TM study; that is, the IMA study was not conceived as a series of experiments. Rather, given the relatively small number of teachers in the study, the purpose of the matching procedure was to try to ensure that the resulting samples of IMA and SUPPORT teachers were, on average, comparable in terms of *EXPER* and *PDT*, which they were. Setting aside the three IMA teachers in the school serving relatively advantaged students, the students of the IMA and SUPPORT teachers were also fairly comparable in terms of language proficiency and various baseline skills.

More generally, Murray (1998) notes that simple random assignment is an unreliable means of achieving baseline comparability among conditions when the sample of organizational units in a study is small. In such situations, he recommends the use of matching or stratification to help increase comparability. If it is not possible to obtain very close matches, as in the case of the IMA study, a sensible strategy is to use matching to try to achieve overall comparability but to ignore the matching in the analysis phase.

Note, finally, that we fit a series of HMs to the IMA data employing the Level 1 model depicted in equation (8) and various Level 2 models involving *EXPER* and *PDT*. Neither of these variables appeared to be systematically related to adjusted class mean outcome scores (β_{0j}). In some respects, this is not too surprising because all but 3 of 16 teachers in the sample had 12 or more years of teaching experience; furthermore, although teachers varied in terms of the amount of relevant prior training they had received, all of the teachers had received at least some prior training.

5. As we saw above, the *ALIGN* values for IMA teachers were, on average, somewhat higher than the *ALIGN* values for SUPPORT teachers. Note that the IMA program may help teachers develop instructional practices other than those captured by the *ALIGN* measure, which help promote student learning (see Gearhart et al., 1999, p. 308).

14.5.2. The Importance of Collecting and Using Data on Implementation

In multisite studies of programs and interventions, there will very likely be variability in implementation across sites. This was clearly the case in both the TM and IMA studies. Due to a variety of difficulties and challenges that practitioners may encounter in the field and, in connection with this, differences in the adaptation of programs to local settings, implementation may be as intended in some sites but partial or poor in others.

In this regard, estimates of the overall, average effectiveness of a program (e.g., estimates of the average effect of TM in Table 14.2) may be of interest to policymakers because such estimates will reflect a variety of difficulties that may have arisen in the field; they will reflect the fact that implementation, on the whole, may have been less than ideal (see Shadish et al., 2002). It is easy to see that information regarding the extent to which implementation was good or poor would be essential for making sense of such estimates.

But clearly, if implementation is poor in some sites and good in others, such estimates are problematic if the goal is to draw inferences concerning the effects of a fully implemented version of the program of interest. However, when data on implementation have been collected, we can, for example, begin to explore differences in the effectiveness of a program when it is implemented with high fidelity and when it is not. Furthermore, as we saw in the above examples, we can begin to test some of the assumptions and ideas that inform the development of a program and focus on factors that may be more proximal with respect to outcomes of interest.

Although analyses involving the use of implementation data provide opportunities to learn more about the conditions under which a program might be particularly effective, an important theme that ran throughout the above examples is that we must be aware that differences in implementation may be associated with other factors (e.g., differences in staff experience, differences in the background characteristics of program participants) that are associated with outcomes of interest. That is, we must attend to possible confounding variables. Thus, not only is it essential to collect data on implementation, but it is also important to collect data on factors that, on the basis of relevant theory and previous research, are likely to be related to differences in implementation and program outcomes.

An implication of this is the need for multisite studies involving larger numbers of sites (J). Analogous to multiple regression models, increases in J enable us to specify Level 2 models containing larger sets of predictors (i.e., measures of implementation, site compositional characteristics, and other Level 2 characteristics). This becomes crucial in situations in which we have identified a number of Level 2 covariates that we must adjust for in efforts to draw sound conclusions regarding, for example, interactions between a particular aspect of implementation and program effectiveness. In this connection, note that we frequently encounter multisite studies in which $J < 20$. This is understandable, given the expense of such studies. But the constraints this places on the number of Level 2 covariates that we can adjust for in our analyses can hamper our ability to draw sound inferences. Note also that with larger J comes increases in the precision with which we can estimate fixed effects of interest.

14.5.3. Interactions Between Site Compositional Characteristics and Program Effectiveness

We also saw how HMs can be used to investigate whether the effects of a program might depend on (i.e., interact with) the compositional characteristics of the individuals at a site. But there are things we must be mindful of in such investigations. Suppose, for example, we find that site mean prior achievement in reading is positively related to the effectiveness of an innovative fourth-grade reading curriculum. It then becomes important to consider what factor or factors may be driving this relationship. Had many of the students in those sites with high mean prior achievement acquired certain skills by the end of Grade 3 that helped them reap the full benefits of the innovative curriculum? Was more time spent on reading instruction in those sites with high mean prior achievement? Was more of the innovative curriculum covered in such sites? We could attempt to address such questions by including measures of these factors, if available, as predictors in subsequent HM analyses. Again, we see the importance of collecting detailed information on implementation and other potentially relevant factors.[6]

6. When we model site treatment effects (e.g., β_{1j}) as a function of site characteristics, we are in effect specifying cross-level interactions. For example, in the case of the TM study, student outcomes are modeled as a function of treatment group membership (TRT_{ij}) in a Level 1 model, and the corresponding regression coefficient (β_{1j}) represents the effect of TM at site j (see equation (2)). Now consider the Level 2 model specified in equation (6), in which site TM effects are modeled as a function of $IMPLRDG_j$ and $\overline{PRE}_{.j}$. Note that replacing β_{1j} in equation (2) with the terms on the right-hand side of equation (6) would give rise to the product

14.5.4. Assessing the Adequacy of Models

Checking the adequacy of models is an essential part of any data analysis project. This becomes vital in the context of multisite studies because the samples in such studies often contain relatively small numbers of sites (J). As we saw above, it is important to examine plots of Level 2 residuals, search for possible outliers, conduct "leave-one-out" analyses, and the like. Seltzer et al. (2002) also point out that Level 1 outliers (e.g., a student who has an outcome score that is extremely high vis-à-vis the other students in her class) can affect the estimation of parameters of interest in multisite studies, and they provide an estimation strategy that downweights such cases. Although outliers are often viewed as nuisances that can adversely affect one's results, it is also important to note that examining field data pertaining to outlying individuals and sites can potentially yield insights concerning the conditions under which a program may be unusually successful and for whom.

14.5.5. Some Comparisons of Designs Employing Blocking and Designs in Which Sites Are Nested Within Treatment Type

Those who have some familiarity with research synthesis will immediately recognize similarities between HMs for meta-analysis and HMs for the analysis of data from multisite evaluation studies that employ blocking (see Chapter 15, this volume). In such studies, each block can be viewed as a mini-study of the effectiveness of a particular program or intervention. Thus, the application of HMs in such settings essentially provides a means of synthesizing results from a series of mini-studies. This has a great deal of conceptual appeal. We can, for example, assess the degree of heterogeneity in treatment effects across sites ("studies") and explore how differences in various site characteristics relate to differences in treatment effects.

In terms of the precision with which we can estimate key fixed effects, blocking can be extremely beneficial (see, e.g., Browne & Liao, 1999). The reason for this is that the between-block variation in outcomes (e.g., τ_{00} in equations (3) through (7)), which can be quite substantial, is not a key component of the standard errors for fixed effects of interest (e.g., the overall effect of TM in equations (3) and (4); the coefficient for *IMPLRDG* in equations (5), (6), and (7)). Rather, as noted in the analyses of the TM data, a key factor is the variance component connected with the between-block variance in treatment effects (e.g., τ_{11}). Note that in the case of matched-pair designs, the degree of improvement in precision will depend crucially on how strongly the matching factors are related to the outcomes of interest (see, e.g., Murray, 1998, pp. 72–74). Note also that Shadish et al. (2002) discuss problems that can arise in employing matching in quasi-experimental settings and offer advice for obtaining better matches.

In contrast, in the case of designs that do not involve blocking, the variance component connected with between-site (e.g., between-class) variability in posttest scores (τ_{00}) appears in the numerator of standard errors for estimates of key fixed effects (e.g., the IMA/SUPPORT contrast, the coefficient for *ALIGN*; see equations (14) and (11)). But as Raudenbush (1997) notes, incorporating Level 1 and Level 2 covariates that are strongly related to outcomes of interest can greatly increase the power of such designs.

Another consideration that arises in connection with designs for multisite studies revolves around the issue of contamination. For example, when both treatment and comparison conditions are implemented within each of a series of schools, there may be a concern that comparison class teachers will adopt certain instructional techniques that are being employed by treatment class teachers. In such cases, researchers would likely opt instead for a design in which schools do not serve as blocks. This was a definite consideration in the case of the IMA study. In addition, for various administrative and organizational reasons, it may be difficult to implement and staff two (or more) programs within the same school, and this may apply in some instances to community-based studies as well.

For detailed discussions and advice regarding an array of important issues that arise in designing and implementing multisite studies, see Browne and Liao (1999), Donner and Klar (2000), Murray (1998), Raudenbush (1997), Raudenbush and Liu (2000), Shadish (2002), and Shadish et al. (2002).

---

terms $TRT_{ij} \times IMPLRDG_j$ and $TRT_{ij} \times \overline{PRE}_{.j}$, and the coefficients of these terms would be γ_{11} and γ_{12}, respectively.

In contrast, in the case of designs in which organizational units are nested within treatment type, interactions between treatment type and a Level 2 predictor of interest would be specified through the inclusion of a product term at Level 2. Consider the IMA study. To test whether the effects of program type (IMA_j) interact with the proportion of English-language proficient students in a class ($\overline{ELP}_{.j}$), we would model adjusted class posttest means (β_{0j}) as a function of IMA_j, $\overline{ELP}_{.j}$, and the product term $IMA_j \times \overline{ELP}_{.j}$.

14.5.6. Looking Longitudinally: Attending to the Effects of Programs Over Time

Expanding the above designs by collecting longitudinal data on study participants opens up an array of modeling opportunities that enriches the kinds of questions we can address. For illustrative purposes, consider a setting in which individual growth is essentially linear over time. In the case of designs that involve blocking, each block would provide us with a contrast of the rates of change between individuals in treatment and control conditions. Analogous to the analyses of the TM data, we could model differences across sites in growth rate contrasts as a function of implementation and other site characteristics. In the case of designs in which organizational units are nested within treatment type, each organizational unit would provide us with a mean (or adjusted mean) growth rate. We could then contrast the growth rates for organizational units assigned to the program of interest with the growth rates for organizational units assigned to the comparison condition. Analyses of the data arising from these kinds of longitudinal multisite designs can be accomplished readily via the use of three-level HMs (see Raudenbush & Bryk, 2002, chap. 8).

We encourage, whenever feasible, the use of designs in which data are collected at several time points prior to the start of the treatment phase. This allows us to examine whether treatment and comparison group members differ, for example, in their rates of change prior to the start of the intervention, which is a serious potential threat to internal validity in settings where random assignment is not employed (see, e.g., Bryk & Weisberg, 1977; Raudenbush, 2001). In addition, recent advances in growth modeling would make it possible to compare growth for treatment and comparison group members during the treatment phase of a study, controlling for possible differences in, for example, status at the end of the pretreatment phase *and* growth rates in the pretreatment phase (see, e.g., Muthén & Curran, 1997; Raudenbush & Bryk, 2002, chap. 11; Seltzer, Choi, & Thum, 2003).

When the collection of time-series data also entails collecting data during a follow-up phase, it is then possible to consider how well individuals fare once a program has come to an end. For example, do rates of change tend to slow down, remain constant, or speed up? What are the factors that appear to promote sustained progress? Questions of this kind could be addressed by employing piecewise models for individual growth in three-level HMs (for discussions of piecewise models, see, e.g., Raudenbush & Bryk, 2002, chap. 6; Seltzer, Frank, & Bryk, 1994; Singer & Willett, 2003).

Note that for some individuals, growth rates might change (e.g., decline) immediately after the treatment phase, but for others, we might not see a decline in growth rates until a certain amount of time has elapsed. The point in time at which growth rates begin to change is termed a *change point.* An important extension of piecewise modeling presented by Thum and Bhattacharya (2001) treats change points as potentially varying across individuals.

14.5.7. Studying Sequences of Treatments

Recent work by Raudenbush, Hong, and Rowan (in press) on the application of HMs to study the effects of sequences of instructional treatments on student learning deserves special mention. Specifically, Raudenbush and his colleagues attempted to assess the causal effects during Grades 4 and 5 of mathematics instruction that emphasizes relatively high-level content and involves appreciable amounts of class time, which they term *intensive* mathematics instruction. Thus, for example, some students may experience *intensive* instruction in Grades 4 and 5, whereas others may experience *nonintensive* instruction in both grades. Still others may experience different forms of instruction in Grades 4 and 5 (e.g., *nonintensive* in Grade 4 and *intensive* in Grade 5). Of particular interest in studies of sequences of treatments are possible interactions between the types of treatment received at different points in time. Thus, for example, the type of mathematics instruction received in Grade 4 may magnify or dampen the effects of the type of instruction received in Grade 5.

Raudenbush et al. (2002) note that implementing and maintaining random assignment in studies of sequences of treatments can pose serious problems, including ethical ones. As such, many studies of sequences of treatments tend to be quasi-experiments. Thus, not only does treatment group membership change over time in such studies, but the treatment one receives at one point in a sequence may depend on the types of treatments one received at earlier points in time, how well one fared, various baseline measures, and the like. Clearly, taking into account possible confounding variables in such studies is crucial. To this end, Robins and his colleagues have developed an extremely valuable strategy (i.e., "inverse probability-of-treatment weighting") for properly taking into account possible confounding variables in settings in which the goal is to estimate the causal

effects of sequences of treatments (Robins, 2000; Robins, Hernan, & Brumback, 2002).

In their analyses, Raudenbush et al. (2002) show how this strategy can be adapted to the kinds of complex multilevel modeling settings that arise in studies of sequences of instructional treatments. Studies of sequences of treatments give rise to longitudinal data (e.g., time-series observations nested within students). However, the data analyzed by Raudenbush et al. also have a cross-classified structure. That is, the students in a given school who are taught by a particular teacher in Grade 4 are then split into different classrooms in Grade 5. Furthermore, the students and teachers in this sample are nested within different schools. The complex longitudinal and multilevel character of these data is explicitly represented in the HMs posed by Raudenbush et al.

Note, finally, that Raudenbush et al.'s (2002) work is explicitly grounded in Rubin's framework for causal inference (see, e.g., Holland, 1986; Rosenbaum & Rubin, 1983; Rubin, 1974, 1978). In sum, we feel that Raudenbush et al.'s work on studying sequences of treatments is going to generate considerable interest in education and related fields.

We hope that this chapter has helped convey the value of HMs in analyzing data from experiments and quasi-experiments in field settings. In particular, we have tried to highlight some of the possibilities that arise when rich data on implementation have been collected. We believe that hierarchical modeling, coupled with the use of implementation data, holds great promise for gaining insight into the effects of programs in an array of fields, including education, social welfare, the behavioral sciences, and epidemiology.

REFERENCES

Browne, C. H., & Liao, J. (1999). Principles for designing randomized preventive trials in mental health: An emerging developmental epidemiology paradigm. *American Journal of Community Psychology, 27*(5), 673–710.

Bryk, A., & Weisberg, H. (1977). Use of the nonequivalent control group design when subjects are growing. *Psychological Bulletin, 84*(5), 950–962.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Chou, C.-P., Bentler, P. M., & Pentz, M. A. (1998). Comparisons of two statistical approaches to study growth curves: The multilevel model and the latent curve analysis. *Structural Equation Modeling, 5*(3), 247–266.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2002). Resources, instruction, and research. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 80–119). Washington, DC: Brookings Institution.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*, 116–127.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London: Arnold.

Flay, B. R., Graumlich, S., Segawa, E., Burns, J. L., Holliday, M. Y., & Aban Aya Investigators. *Effects of two prevention programs on high-risk behaviors among African-American youth: A randomized trial*. Manuscript under review.

Gail, M. H., Byar, D. P., Pechacek, T. F., & Corle, D. K. (1992). Aspects of statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT). *Controlled Clinical Trials, 13*, 6–21.

Gearhart, M., Saxe, G. B., Seltzer, M. H., Schlackman, J., Ching C. C., Nasir, N., et al. (1999). Opportunities to learn fractions in elementary mathematics classrooms. *Journal for Research in Mathematics Education, 30*(3), 286–315.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Edward Arnold.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–970.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Pacific Grove, CA: Brooks/Cole.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage.

Longford, N. (1993). *Random coefficient models*. Oxford, UK: Clarendon.

Mason, W. M., Wong, G. M., & Entwistle, B. (1983). Contextual analysis through the multilevel linear models. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 72–103). San Francisco: Jossey-Bass.

McLaughlin, M. (1987). Implementation realities and evaluation design. In W. Shadish & C. Reichardt (Eds.), *Evaluation studies review annual* (pp. 73–97). Newbury Park, CA: Sage.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.

Muthén, B., & Curran, P. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power. *Psychological Methods, 2*, 371–402.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

Patton, M. Q. (1980). *Qualitative evaluation methods*. Beverly Hills, CA: Sage.

Pentz, M. A., Dwyer, J. H., MacKinnon, D. P., Flay, B. R., Hansen, W. B., Wang, E. Y. I., et al. (1989). A multicommunity trial for primary prevention of adolescent drug use. *Journal of the American Medical Association, 261*, 3259–3266.

Pinnell, G., Lyons, C., DeFord, D., Bryk, A. & Seltzer, M. (1994). Studying the effectiveness of early intervention approaches for first grade children having difficulty in reading. *Reading Research Quarterly, 39*, 8–39.

Raffe, D. (1991). Assessing the impact of a decentralised initiative: The British Technical and Vocational Education Initiative. In S. Raudenbush & D. Willms (Eds.), *Schools,*

classrooms and pupils: International studies of schooling from a multilevel perspective (pp. 149–166). San Diego: Academic Press.

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

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*(2), 173–185.

Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology, 52*, 501–525.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y., & Congdon, R. T. (2000). *HLM 5: Hierarchical linear and nonlinear modeling*. Chicago: Scientific Software International.

Raudenbush, S. W., Hong, G., & Rowan, B. (in press). Studying the causal effects of instruction with application to primary-school mathematics. In J. M. Ross, G. W. Bohrnstedt, & F. C. Hemphill (Eds.), *Instructional and performance consequences of high poverty schooling*. Washington, DC: National Council for Educational Statistics.

Raudenbush, S., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199–213.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In E. M. Halloran & D. Berry (Eds.), *Statistical models in epidemiology, the environment, and clinical trials* (pp. 95–134). New York: Springer.

Robins, J. M., Hernan, M., & Brumback, B. (2002). Marginal structural models and causal inference in epidemiology. *Epidemiology, 11*(5), 550–560.

Rosenbaum, D. P., Flewelling, R. L., Bailey, S. L., Ringwalt, C. L., & Wilkinson, D. L. (1994). Cops in the classroom: A longitudinal evaluation of drug abuse resistance education (DARE). *Journal of Research in Crime and Delinquency, 31*(1), 3–31.

Rosenbaum, P. R., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics, 6*, 34–58.

Saxe, G. B., Gearhart, M., & Seltzer, M. (1999). Relations between classroom practices and student learning in the domain of fractions. *Cognition and Instruction, 17*, 1–24.

Seltzer, M. (1994). Studying variation in program success: A multilevel modeling approach. *Evaluation Review, 18*(3), 342–361.

Seltzer, M., Choi, K., & Thum, Y. (2003). Examining relationships between where students start and how rapidly they progress: Using new developments in growth modeling to gain insight into the distribution of achievement within schools. *Educational Evaluation and Policy Analysis, 25*(3), 263–286.

Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The metric matters: The sensitivity of conclusions concerning growth in student achievement to choice of metric. *Educational Evaluation and Policy Analysis, 16*(1), 41–49.

Seltzer, M. H., Novak, J., Choi, K., & Lim, N. (2002). Sensitivity analysis for hierarchical models employing *t* level-1 assumptions. *Journal of Educational and Behavioral Statistics, 27*(2), 181–222.

Shadish, W. R. (2002). Revisiting field experimentation: Field notes for the future. *Psychological Methods, 7*(1), 3–18.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Spiegelhalter, D., Thomas, A., Best, N., & Gilks, W. (2000). *WinBUGS, Version 1.3 user manual*. MRC Biostatistics Unit, Cambridge University.

Thum, Y. M., & Bhattacharya, S. K. (2001). Detecting a change in school performance: A Bayesian analysis for a multilevel join point problem. *Journal of Educational and Behavioral Statistics, 26*(4), 443–468.

University of Chicago School Mathematics Project. (1986). *Transition mathematics field study* (Evaluation Report 85/86-TM-2). Chicago: University of Chicago, Department of Education.

Chapter 15

META-ANALYSIS

SPYROS KONSTANTOPOULOS

LARRY V. HEDGES

The growth of the social science research enterprise has led to a large body of related research studies. The sheer volume of research related to many topics of scientific or policy interest poses a problem of how to organize and summarize these findings to identify and exploit what is known and focus research on promising areas (see Garvey & Griffith, 1971). This problem is not unique to the social sciences. It has arisen in fields as diverse as physics, chemistry, experimental biology, medicine, and public health. In each of these fields, as in the social sciences, accumulation of quantitative research evidence has led to the development of systematic methods for the quantitative synthesis of research (see Cooper & Hedges, 1994). Although the term *meta-analysis* was coined to describe these methods in the social sciences (Glass, 1976), the methods used in other fields are remarkably similar to those in the social sciences (Cooper & Hedges, 1994; Hedges, 1987).

Meta-analysis refers to an analysis of the results of several studies for the purposes of drawing general conclusions. Meta-analysis involves describing the results of each study via a numerical index of effect size (such as a correlation coefficient, a standardized mean difference, or an odds ratio) and then combining these estimates across studies to obtain a summary. The specific analytic techniques involved will depend on the question the meta-analytic summary is intended to address. Sometimes, the question of interest concerns the typical or average study result. For example, in studies that measure the effect of some treatment or intervention, the average effect of the treatment is often of interest (see, e.g., Smith & Glass, 1977). In other cases, the degree of variation in results across studies will be of primary interest. For example, meta-analysis is often used to study the generalizability of employment test validities across situations (see, e.g., Schmidt & Hunter, 1977). In yet other cases, the primary interest is in the factors that are related to study results. For example, meta-analysis is often used to identify the contexts in which a treatment or intervention is most successful or has the largest effect (see, e.g., Cooper, 1989b).

The term *meta-analysis* is sometimes used to connote the entire process of quantitative research synthesis. More recently, it has begun to be used specifically for the statistical component of research synthesis. This chapter deals exclusively with that narrower usage of the term to describe statistical methods only. However, it is crucial to understand that in research synthesis, as in any research, statistical methods are only one part of the enterprise. Statistical methods cannot remedy the problem of data that are of poor quality. Excellent treatments of the nonstatistical aspects of research synthesis are available in Cooper (1989b), Cooper and Hedges (1994), and Lipsey and Wilson (2001).

15.1. Effect Sizes

Effect sizes are quantitative indexes that are used to summarize the results of a study in meta-analysis. That is, effect sizes reflect the magnitude of the association between variables of interest in each study. There are many different effect sizes, and the effect size used in a meta-analysis should be chosen so that it represents the results of a study in a way that is easily interpretable and comparable across studies. In a sense, effect sizes should put the results of all studies "on a common scale" so that they can be readily interpreted, compared, and combined.

It is important to distinguish the effect size estimate in a study from the effect size parameter (the true effect size) in that study. In principle, the effect size estimate will vary somewhat from sample to sample that might be obtained in a particular study. The effect size parameter is, in principle, fixed. One might think of the effect size parameter as the estimate that would be obtained if the study had a very large (essentially infinite) sample, so that the sampling variation is negligible.

The choice of an effect size index will depend on the design of the studies, the way in which the outcome is measured, and the statistical analysis used in each study. Most of the effect size indexes used in the social sciences will fall into one of three families of effect sizes: the standardized mean difference family, the odds ratio family, and the correlation coefficient family.

15.1.1. The Standardized Mean Difference

In many studies of the effects of a treatment or intervention that measure the outcome on a continuous scale, a natural effect size is the standardized mean difference. The standardized mean difference is the difference between the mean outcome in the treatment group and the mean outcome in the control group divided by the within-group standard deviation. That is, the standardized mean difference is

$$d = \frac{\bar{Y}^T - \bar{Y}^C}{S},$$

where $\bar{Y}^T$ is the sample mean of the outcome in the treatment group, $\bar{Y}^C$ is the sample mean of the outcome in the control group, and S is the within-group standard deviation of the outcome. The corresponding standardized mean difference parameter is

$$\delta = \frac{\mu^T - \mu^C}{\sigma},$$

where μ^T is the population mean in the treatment group, μ^C is the population mean outcome in the control group, and σ is the population within-group standard deviation of the outcome. This effect size is easy to interpret because it is just the treatment effect in standard deviation units. It can also be interpreted as having the same meaning across studies (see Hedges & Olkin, 1985).

The sampling uncertainty of the standardized mean difference is characterized by its variance, which is

$$v = \frac{n^T + n^C}{n^T n^C} + \frac{d^2}{2(n^T + n^C)},$$

where n^T and n^C are the treatment and control group sample sizes, respectively. Note that this variance can be computed from a single observation of the effect size if the sample sizes of the two groups within a study are known. Because the standardized mean difference is approximately normally distributed, the square root of the variance (the standard error) can be used to compute confidence intervals for the true effect size or effect size parameter δ. Specifically, a 95% confidence interval for the effect size is given by

$$d - 2\sqrt{v} \leq \delta \leq d + 2\sqrt{v}.$$

Several variations of the standardized mean difference are also sometimes used as effect sizes (see Rosenthal, 1994).

15.1.2. The Log Odds Ratio

In many studies of the effects of a treatment or intervention that measure the outcome on a dichotomous scale, a natural effect size is the log odds ratio. The log odds ratio is just the log of the ratio of the odds of a particular one of the two outcomes (the target outcome) in the treatment group to the odds of that particular outcome in the control group. That is, the log odds ratio is

$$\begin{aligned}\log(\text{OR}) &= \log\left(\frac{p^T/(1-p^T)}{p^C/(1-p^C)}\right) \\ &= \log\left(\frac{p^T(1-p^C)}{p^C(1-p^T)}\right),\end{aligned}$$

where p^T and p^C are the proportions of the treatment and control groups, respectively, that have the target outcome. The corresponding odds ratio parameter is

$$\omega = \log\left(\frac{\pi^T/(1-\pi^T)}{\pi^C/(1-\pi^C)}\right) = \log\left(\frac{\pi^T(1-\pi^C)}{\pi^C(1-\pi^T)}\right),$$

where π^T and π^C are the population proportions in the treatment and control groups, respectively, that have the target outcome. The log odds ratio is widely used in the analysis of data that have dichotomous outcomes and is readily interpretable by researchers who frequently encounter these kinds of data. It also has the same meaning across studies, so it is suitable for combining (see Fleiss, 1994).

The sampling uncertainty of the log odds ratio is characterized by its variance, which is

$$v = \frac{1}{n^T p^T} + \frac{1}{n^T(1-p^T)} + \frac{1}{n^C p^C} + \frac{1}{n^C(1-p^C)},$$

where n^T and n^C are the treatment and control group sample sizes, respectively. As in the case of the standardized mean difference, the log odds ratio is approximately normally distributed, and the square root of the variance (the standard error) can be used to compute confidence intervals for the true effect size or effect size parameter ω. Specifically, a 95% confidence interval for the effect size is given by

$$d - 2\sqrt{v} \leq \omega \leq d + 2\sqrt{v}.$$

There are several other indexes in the odds ratio family, including the *risk ratio* (the ratio of the proportion having the target outcome in the treatment group to that in the control group, or p^T/p^C) and the *risk difference* (the difference between the proportion having a particular one of the two outcomes in the treatment group and that in the control group, or $p^T - p^C$). For a discussion of effect size measures for studies with dichotomous outcomes, including the odds ratio family of effect sizes, see Fleiss (1994).

15.1.3. The Correlation Coefficient

In many studies of the relation between two continuous variables, the correlation coefficient is a natural measure of effect size. Often, this correlation is transformed via the Fisher z-transform

$$z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$$

in carrying out statistical analyses. The corresponding correlation parameter is ρ, the population correlation, and the parameter that corresponds to the estimate z is ξ, the z-transform of ρ. The sampling uncertainty of the z-transformed correlation is characterized by its variance

$$v = \frac{1}{n-3},$$

where n is the sample size of the study, and it is used in the same way as are the variances of the standardized mean difference and log odds ratio to obtain confidence intervals.

The statistical methods for meta-analysis are quite similar, regardless of the effect size measure used. Therefore, in the rest of this chapter, we do not describe statistical methods that are specific to a particular effect size index but describe them in terms of a generic effect size measure T_i. We assume that the T_i are normally distributed about the corresponding θ_i with known variance v_i. That is, we assume that

$$T_i - N(\theta_i, v_i), i = 1, \ldots, k.$$

This assumption is very nearly true for effect sizes such as the Fisher z-transformed correlation coefficient and standardized mean differences. However, for effect sizes such as the untransformed correlation coefficient, or the log odds ratio, the results are not exact but remain true as large sample approximations. For a discussion of effect size measures for studies with continuous outcomes, see Rosenthal (1994), and for a treatment of effect size measures for studies with categorical outcomes, see Fleiss (1994).

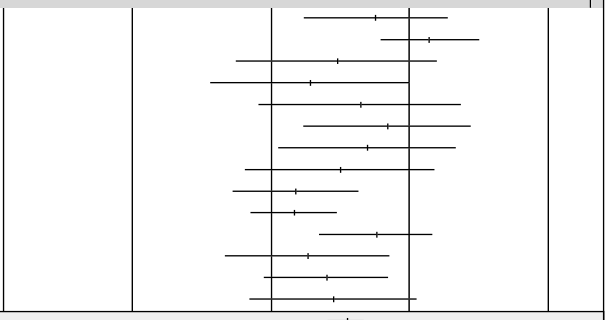
15.1.4. Example

Gender differences in field articulation ability (sometimes called visual-analytic spatial ability) were studied by Hyde (1981). She reported standardized mean differences from 14 studies that examined gender differences in spatial ability tasks that call for the joint application of visual and analytic processes (see Maccoby & Jacklin, 1974). The results of these 14 studies are shown in Figure 15.1, in which each study is depicted as an effect size estimate (a standardized mean difference) and a 95% confidence interval reflecting the sampling uncertainty of that estimate. These 95% confidence intervals are computed as the effect size estimate plus or minus two times the square root of the sampling variance of the effect size.

The figure raises several important issues that might be explored in the meta-analysis. First, the effect size estimates from the studies are not identical. This is to be expected because the estimates are based on data from samples, and random variations due to sampling should introduce fluctuations into the estimates. The confidence interval about each estimate suggests how large these fluctuations due to sampling might be. If all of the studies are estimating the same treatment effect, it is reasonable that combining estimates across studies (e.g., taking an average) will reduce the overall

Figure 15.1 Results of 14 Studies That Examined Gender Differences in Spatial Ability Tasks

Model	Study name	Statistics for each study: Std diff in means	Lower limit	Upper limit
	1.000	0.760	0.238	1.282
	2,000	1.150	0.794	1.506
	3,000	0.480	–0.245	1.205
	4,000	0.290	–0.430	1.010
	5,000	0.650	–0.803	1.383
	6,000	0.840	0.236	1.444
	7,000	0.700	0.062	1.338
	8,000	0.500	–0.182	1.182
	9,000	0.180	–0.271	0.631
	10,000	0.170	–0.140	0.480
	11,000	0.770	0.359	1.181
	12,000	0.270	–0.324	0.864
	13,000	0.400	–0.047	0.847
	14,000	0.450	–0.154	1.054
Fixed		0.547	0.414	0.680



SOURCE: Comprehensive Meta Analysis (www.Meta-Analysis.com).

sampling uncertainty by evening out the study-to-study sampling fluctuations.

Second, the amount of sampling uncertainty is not identical in every study, as reflected in the differing lengths of the confidence intervals. Therefore, it seems reasonable that, if an average effect size is to be computed across studies, it would be desirable to give more weight in that average to studies that have more precise estimates (smaller variances) than those with less precise estimates. How, exactly, should this be done?

Third, when we examine the confidence intervals, there is considerable overlap, but the effect size estimates of some studies are outside of the confidence intervals of other studies. This raises the question of whether the effect sizes of these studies might differ by more than would be expected due to sampling variation alone. To put it another way, is it reasonable to assume that all of the studies are estimating the same underlying effect size and differ in their estimates by sampling variation alone?

Fourth, there seems to be a trend over time in these data, with the studies that are conducted in earlier years tending to have larger effect sizes. How do we determine whether this trend is statistically reliable or is just an artifact of sampling variation?

Fifth, two of the studies, conducted in 1955 and 1959, appear to have somewhat larger effect sizes than the others. Are the effects of these studies really different from the others, and if these studies are omitted, is the pattern of the other studies more consistent?

In the sections that follow, we will introduce methods for exploring these questions as a paradigm for similar explorations that are sensible in meta-analyses generally.

15.2. Estimating the Average Effect Across Studies

Consider now the first question that was raised above—namely, combining the effect size estimates across studies to estimate the average effect size. Let θ_i be the (unobserved) effect size parameter (the true effect size) in the ith study, let T_i be the corresponding observed effect size estimate from the ith study, and let v_i be its variance. Thus, the data from a set of k studies are the effect size estimates $T_1, \ldots, T_k$, and their variances $v_1, \ldots, v_k$.

A natural way to describe the data is via a two-level hierarchical model, with one model for the data at the study level and another model for the between-study variation in effects. At the first (within-study) level, the effect size estimate T_i is just the effect size parameter plus a sampling error ε_i. That is,

$$T_i = \theta_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, v_i).$$

The parameter θ is the mean effect size parameter for all of the studies. It has the interpretation that θ is the mean of the distribution from which the study-specific effect size parameters $(\theta_1, \theta_2, \ldots, \theta_k)$ were sampled. Note that this is not conceptually the same as the mean of $\theta_1, \theta_2, \ldots, \theta_k$, the effect size parameters of the k studies that were observed.

At the second (between-study) level, the effect size parameters are determined by a mean effect size β_0 plus a study-specific random effect η_i. That is,

$$\theta_i = \beta_0 + \eta_i, \qquad \eta_i \sim N(0, \tau^2).$$

In this model, the η_i represent differences between the effect size parameters from study to study. The parameter τ^2, often called the between-studies variance component, describes the amount of variation across studies in the random effects (the η_is) and therefore effect parameters (the θ_is).

This model is identical in general form to the hierarchical linear model often used in the primary analysis of social science data. It has two features that are different from that model, however. First, in the usual model, the Level 1 variance is identical across Level 1 units. In the meta-analytic model, the Level 1 variances (the v_is) are different for each of the Level 1 units (in this case, studies). That is, each study has a *different* sampling error variance at Level 1. Second, in the usual model, the Level 1 variance is unknown and must be estimated from the data. In the meta-analytic model, the Level 1 variances, although differing across studies, are known.

The two-level model described above can be written as a one-level model as follows:

$$T_i = \beta_0 + \eta_i + \varepsilon_i = \beta_0 + \xi_i,$$

where ξ_i is a composite error defined by $\xi_i = \eta_i + \varepsilon_i$. Writing this as a one-level model, we see that each effect size is an estimate of β_0, with a variance that depends on both v_i and τ^2. In models such as this, it is necessary to distinguish between the variance of T_i, assuming a fixed θ_i and the variance of T_i incorporating the variance of the θ_i as well. The former is the *conditional sampling variance* of T_i (denoted by v_i), and the latter is the *unconditional sampling variance* of T_i (denoted by v_i^*). Because the sampling error ε_i and the random effect η_i are assumed to be independent and the variance of η_i is $\hat{\tau}^2$, it follows that the unconditional sampling variance of T_i is $v_i^* = v_i + \hat{\tau}^2$.

The least squares (and maximum likelihood) estimate of β_0 under the model is

$$\hat{\beta}_0^* = \frac{\sum_{i=1}^k w_i^* T_i}{\sum_{i=1}^k w_i^*}, \tag{1}$$

where $w_i^* = 1/(v_i + \hat{\tau}^2) = 1/v_i^*$, and $\hat{\tau}^2$ is the between-studies variance component estimate. Note that this estimator corresponds to a weighted mean of the T_i, giving more weight to the studies whose estimates have smaller unconditional variance (are more precise) when pooling.

The sampling variance $v_\bullet^*$ of $\hat{\beta}_0^*$ is simply the reciprocal of the sum of the weights,

$$v_\bullet^* = \left(\sum_{i=1}^k w_i^* \right)^{-1},$$

and the standard error $SE(\hat{\beta}_0^*)$ of $\hat{\beta}_0^*$ is just the square root of $v_\bullet^*$. Under this model, β_0^* is normally distributed, so a $100(1-\alpha)\%$ confidence interval for β_0 is given by

$$\hat{\beta}_0^* - t_{\alpha/2}\sqrt{v_\bullet^*} \le \beta_0 \le \hat{\beta}_0^* + t_{\alpha/2}\sqrt{v_\bullet^*},$$

where t_α is the $100\alpha\%$ point of the t-distribution with $(k-1)$ degrees of freedom. Similarly, a two-sided test of the hypothesis that $\beta_0 = 0$ at significance level α uses the test statistic $Z = \hat{\beta}_0^*/\sqrt{v_\bullet^*}$ and rejects if $|Z|$ exceeds $t_{\alpha/2}$.

To use the estimate of the average effect size given above and the tests and confidence intervals associated with it, one needs to know the between-studies variance component $\hat{\tau}^2$. Usually, this has to be estimated from the data. In any particular set of effect sizes, it may not be clear whether the variation in the observed effect size estimates is large enough to provide persuasive evidence that $\tau^2 > 0$. In the next section, we pursue the problem of testing whether $\tau^2 = 0$ and estimating a precise value of τ^2 to use in the estimation of β_0.

15.3. Testing Whether the Between-Studies Variance Component $\tau^2 = 0$

It seems reasonable that the greater the variation in the observed effect size estimates, the stronger the evidence that $\tau^2 > 0$. A simple test (the likelihood ratio test) of the hypothesis that $\tau^2 = 0$ uses the weighted sum of squares about the weighted mean that would be obtained if $\tau^2 = 0$. Specifically, it uses the statistic

$$Q = \sum_{i=1}^k (T_i - \hat{\beta}_0)^2 / v_i,$$

where $\hat{\beta}_0$ is the estimate of β_0 that would be obtained from equation (1) if $\tau^2 = 0$. The statistic Q has the chi-squared distribution with $(k-1)$ degrees of freedom if $\tau^2 = 0$. Therefore, a test of the null hypothesis that $\tau^2 = 0$ at significance level α rejects the hypothesis if Q exceeds the $100(1-\alpha)\%$ point of the chi-square distribution with $(k-1)$ degrees of freedom.

This (or any other statistical hypothesis test) should not be interpreted too literally. The test is not very

powerful if the number of studies is small or if the conditional variances (the v_i) are large (see Hedges & Pigott, 2001). Consequently, even if the test does not reject the hypothesis that $\tau^2 = 0$, the actual variation in effects across studies may be consistent with a substantial range on nonzero values of τ^2, some of them rather large. This suggests that it is important to consider estimation of τ^2 and use these estimates in constructing estimates of the mean using (1).

Note, however, that even when the estimate of τ^2 is equal to zero, there may still be many nonzero values of τ^2 that are quite compatible with the effect size data (see Raudenbush & Bryk, 1985). This has led many investigators to consider the use of Bayesian estimators that compute the average of the entire meta-analysis over a range of plausible values of τ^2 (see Hedges, 1998).

15.4. Estimating the Between-Studies Variance Component τ^2

Estimation of τ^2 can be accomplished without making assumptions about the distribution of the random effects or under various assumptions about the distribution of the random effects using other methods such as maximum likelihood estimation. Maximum likelihood estimation is more efficient if the distributional assumptions about the study-specific random effects are correct, but these assumptions are often difficult to justify theoretically and verify empirically. Thus, distribution-free estimates of the between-studies variance component are often attractive.

A simple, distribution-free estimate of τ^2 is given by

$$\hat{\tau}^2 = \begin{cases} \frac{Q-(k-1)}{a} & \text{if } Q \geq (k-1) \\ 0 & \text{if } Q < (k-1) \end{cases},$$

where a is given by

$$a = \sum_{j=1}^{k} w_i - \frac{\sum_{j=1}^{k} w_i^2}{\sum_{j=1}^{k} w_i}, \tag{2}$$

and $w_i = 1/v_i$. Estimates of τ^2 are set to 0 when $Q - (k-1)$ yields a negative value because τ^2, by definition, cannot be negative.

If the within-study sampling error variances $v_1, \ldots, v_k$ used to construct the weights w_i are known exactly and the estimate is *not* set to 0 when $Q - (k-1) < 0$, then the w_i are constants and the estimate is unbiased, a result that does not depend on assumptions about the distribution of the random effects (or the conditional distribution of the effect sizes themselves). Inaccuracies in the estimation of the v_i (and hence the w_i) may lead to biases, although they are usually not substantial. The truncation of the estimate at zero is a more serious source of bias, although it improves the accuracy (reduces its mean squared error about the true τ^2) of estimates of τ^2. This bias can be substantial when k is small but decreases rapidly when k becomes larger (see Hedges & Vevea, 1998). The relative bias of $\hat{\tau}^2$ can be well over 50% for $k = 3$ and $\hat{\tau}^2 = v/3$. This result underscores the fact that estimates of τ^2 computed from only a few studies should be treated with caution. For $k > 20$, the biases are much smaller, and relative biases are only a few percent (see Hedges & Vevea, 1998).

Bias is not the only concern in the estimation of τ^2. When the number of studies is small, $\hat{\tau}^2$ has a great deal of sampling uncertainty. Moreover, the sampling distribution of $\hat{\tau}^2$ is quite skewed (it is a distribution that is a constant times a chi-squared distribution). Although the standard error of $\hat{\tau}^2$ is known, it serves only to give a broad characterization of the uncertainty of $\hat{\tau}^2$ (see Hedges & Pigott, 2001). In particular, intervals of plus or minus 2 standard errors would be very poor approximations to 95% confidence intervals for τ^2 unless the number of studies was very large.

15.4.1. Example

Returning to our example of the studies of gender differences in field articulation ability, the data reported by Hyde (1981) are presented in Table 15.1. The effect size estimates in column 2 are standardized mean differences. All estimates are positive and indicate that, on average, males are performing higher than females in field articulation. The variances of the estimates are in column 3. Finally, the year that the study was conducted is in column 4.

First, we turn to the question of whether the effect sizes have more sampling variation than would be expected from the size of their conditional variances. Computing the test statistic Q, we obtain $Q = 24.103$, which is slightly larger than 22.36, which is the $100(1 - 0.05) = 95\%$ point of the chi-square distribution with $14 - 1 = 13$ degrees of freedom. Actually, a Q value of 24.103 would occur only about 3% of the time if $\tau^2 = 0$. Thus, there is some evidence that the

Table 15.1 Field Articulation Data From Hyde (1981)

ID	*ES*	*Var*	*Year*
1	0.76	0.071	1955
2	1.15	0.033	1959
3	0.48	0.137	1967
4	0.29	0.135	1967
5	0.65	0.140	1967
6	0.84	0.095	1967
7	0.70	0.106	1967
8	0.50	0.121	1967
9	0.18	0.053	1967
10	0.17	0.025	1968
11	0.77	0.044	1970
12	0.27	0.092	1970
13	0.40	0.052	1971
14	0.45	0.095	1972

NOTE: ID = study ID; ES = effect size estimate; Var = variance; Year = year of study.

variation in effects across studies is not simply due to chance sampling variation.

Hence, we investigate how much variation there might be across studies, and we compute the estimate of τ^2 using the distribution-free method given above. We obtain the estimate

$$\hat{\tau}^2 = \frac{24.103 - (14 - 1)}{195.384} = 0.057.$$

Comparing this value with the average of the conditional variances, we see that $\hat{\tau}^2$ is about 65% of the average sampling error variance. Thus, it cannot be considered negligible.

Now we compute the weighted mean of the effect size estimates, incorporating the variance component estimate $\hat{\tau}^2$ into the weights. This yields an estimate of

$$\beta_0^* = 58.488/106.498 = 0.549,$$

with a variance of

$$v_\cdot^* = 1/106.498 = 0.0094.$$

The 95% confidence interval for β_0 is given by

$$0.339 = 0.549 - 2160.\sqrt{0.0094} \leq \beta_0$$
$$\leq 0.542 - 2.160\sqrt{0.00974} = 0.758.$$

This confidence interval does not include 0, so the data are incompatible with the hypothesis that $\beta_0 = 0$.

15.5. Fixed-Effects Analysis

Two somewhat different statistical models have been developed for inference about effect size data from a collection of studies, called the random-effects and fixed-effects models, respectively (see, e.g., Hedges & Vevea, 1998). Random-effects models, which we discussed above, treat the effect size parameters as if they were a random sample from a population of effect parameters and estimate hyperparameters (usually just the mean and variance) describing this population of effect parameters (see, e.g., DerSimonian & Laird, 1986; Hedges, 1983a, 1983b; Schmidt & Hunter, 1977). The use of the term *random effects* for these models in meta-analysis is somewhat inconsistent with the use of the term elsewhere in statistics. It would be more consistent to call these models *mixed models* because the parameter structure of the models is identical to those of the general linear mixed model (and their important application in social sciences, hierarchical linear models).

Although the idea that studies (and their corresponding effect size parameters) are a sample from a population is conceptually appealing, studies are seldom a probability sample from any well-defined population. Consequently, the universe (to which generalizations of the random-effects model apply) is often unclear. Moreover, some scholars object to the idea of generalizing to a universe of studies that have not been observed. Why, they argue, should studies that have not been done influence inferences about the studies that have been done? This argument is part of a more general debate about conditionality in inference, which dates back at least to debates in the 1920s between Fisher and Yates about the proper analysis of 2×2 tables (e.g., the Fisher exact test vs. the Pearson chi-square tests) (see Camilli, 1990). This, like many debates about the foundations of inference, is unlikely to ever be resolved definitively.

Fixed-effects models treat the effect size parameters as fixed but unknown constants to be estimated and usually (but not necessarily) are used in conjunction with assumptions about the homogeneity of effect parameters (see, e.g., Hedges, 1982a; Rosenthal & Rubin, 1982). That is, fixed-effects models carry out estimation and testing as if $\tau^2 = 0$. The logic of fixed-effects models is that inferences are not about any putative population of studies but about the particular studies that happen to have been observed.

For example, if the fixed-effects approach had been applied to the example above, the weights would have been computed with $\tau^2 = 0$, so that the weight for each

study would have been $w_i = 1/v_i$. In this case, the weights assigned to studies would have been considerably more unequal across studies, and the mean would have been estimated as $\hat{\beta}_0 = 118.486/216.684 = 0.547$. Comparing this with the random-effects estimate of $\hat{\beta}_0^* = 0.549$, we see that the mean effects are very similar. This is often (but not necessarily) the case. However, the variance of the mean computed in the fixed-effects analysis is $v_\cdot = 1/216.684 = 0.0046$. This is much smaller (nearly 50%) than 0.0094, the variance computed for the mean effect in the random-effects analysis above. The variance of the fixed-effects estimate is always smaller than or equal to that of the random-effects estimate of the mean, and often it is much smaller. The reason is that between-studies variation in effects is included as a source of uncertainty in computing the variance of the mean in random-effects models but is not included as a source of uncertainty of the mean in fixed-effects models.

15.6. Modeling the Association Between Differences Among Studies and Effect Size

One of the fundamental issues facing meta-analysts is how to model the association between characteristics of studies and their effect sizes. In our example, we noted the fact that studies conducted earlier appeared to find gender differences that were larger. Does this mean that gender differences are getting smaller over time, or is this just an artifact of sampling fluctuation in effect sizes across studies? One might observe many differences among studies that appear to be associated with differences in effect sizes. One variety of treatment might produce bigger effects than others, more intensive treatments might produce bigger effects, and a treatment might be more effective in some contexts than others. To examine any of these questions, one must investigate the relation between study characteristics (study-level covariates) and effect size. The most obvious statistical procedures for doing so are mixed models for meta-analysis.

The mixed models considered in this chapter are closely related to the general mixed linear model, which has been studied extensively in applied statistics (e.g., Hartley & Rao, 1967; Harville, 1977). Mixed models have been applied to meta-analysis since very early in their use in the social sciences. For example, Schmidt and Hunter (1977) used the mixed-model idea (albeit not the term) in their models of validity generalization. Other early applications of mixed models to meta-analysis include Hedges (1983b), DerSimonian and Laird (1986), Raudenbush and Bryk (1985), and Hedges and Olkin (1985).

We first describe the general model and notations for mixed-model meta-analysis and point out the connection between this model and classical hierarchical linear models used in the social sciences. Then we consider distribution-free analyses of these models using software for weighted least squares methods. Finally, we show how to carry out mixed-model meta-analyses using software for hierarchical linear models such as SAS PROC MIXED, HLM, and MLwin.

15.6.1. Models and Notation

Suppose as before that the effect size parameters are $\theta_1, \theta_2, \ldots, \theta_k$, and we have k independent effect size estimates $T_1, \ldots, T_k$, with sampling variances $v_1, \ldots, v_k$. We assume, as before, that each T_i is normally distributed about θ_i. Thus, the Level 1 (within-study) model is as before:

$$T_i = \theta_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, v_i).$$

Suppose now that there are p known predictor variables for the fixed effects by $X_1, \ldots, X_p$, and assume that they are related to the effect sizes via a linear model. In this case, the Level 2 model for the ith effect size parameter becomes

$$\theta_i = \beta_0 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_p x_{ip} + \eta_i,$$
$$\eta_i \sim N(0, \tau^2),$$

where $x_{i1}, \ldots, x_{ip}$ are the values of the predictor variables $X_1, \ldots, X_p$ for the ith study (i.e., x_{ij} is the value of predictor variable X_j for study i), and η_i is a study-specific random effect with zero expectation and variance τ^2.

We can also rewrite the two-level model in a single equation as a model for the T_i as follows:

$$\begin{aligned} T_i &= \beta_0 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_p x_{ip} + \eta_i + \varepsilon_i \\ &= \beta_0 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_p x_{ip} + \xi_i, \end{aligned} \tag{3}$$

where $\xi_i = \eta_i + \varepsilon_i$ is a composite residual incorporating both study-specific random effect and sampling error. Because we assume that η_i and ε_i are independent, it follows that the variance of ξ_i is $\tau^2 + v_i$. Consequently, if τ^2 were known, we could estimate the regression coefficients via weighted least squares (which would also yield the maximum likelihood estimates of the β_is). When τ^2 is not known, there are four approaches to the estimation of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$.

One is to estimate τ^2 from the data and use the estimate in place of τ^2 to obtain weighted least squares estimates of $\boldsymbol{\beta}$. The second is to jointly estimate $\boldsymbol{\beta}$ and τ^2 via unrestricted maximum likelihood. The third is to define τ^2 to be 0 (which is effectively what is done in fixed-effects analyses). The fourth approach is to compute an estimate of $\boldsymbol{\beta}$ for each of a range of plausible values of τ^2 and compute a weighed average of those results, giving a weight to each according to its plausibility (prior probability), which is effectively what is done in Bayesian approaches to meta-analysis. We discuss each of these approaches below.

15.6.2. Analysis Using Software for Classical Hierarchical Linear Model Analysis

Hierarchical linear models are widely used in the social sciences (see, e.g., Goldstein, 1987; Longford, 1987; Raudenbush & Bryk, 2002). There has been considerable progress in developing software to estimate and test the statistical significance of parameters in hierarchical linear models.

There are two important differences in the hierarchical linear models (or general mixed models) usually studied and the model used in meta-analysis (equation (2)). The first is that in meta-analysis models, such as in equation (2), the variances of the sampling errors $v_1, \ldots, v_k$ are *not* identical across studies. That is, the assumption that $v_1 = \cdots = v_k$ is unrealistic. The sampling error variances usually depend on various aspects of study design (particularly sample size), which cannot be expected to be constant across studies. The second is that although the sampling error variances in meta-analysis are different for each study, they are generally assumed to be known.

Therefore, the model used in meta-analysis can be considered a special case of the general hierarchical linear model in which the Level 1 variances are unequal but known. Consequently, software for the analysis of hierarchical linear models can be used for mixed-model meta-analysis if it permits (as do to the programs SAS PROC MIXED, HLM, and MLwin) the specification of first-level variances that are unequal but known.

15.6.2.1. Mixed-Model Meta-Analysis Using SAS PROC MIXED

SAS PROC MIXED is a general-purpose computer program that can be used for fitting mixed models in meta-analysis (see Singer, 1998). The upper panel of Table 15.2 gives SAS input file for an analysis of the data on gender differences in field articulation ability from Hyde (1981). The first 20 lines of the upper panel of the table are the commands for the SAS data step, which name the data set (in this case, "genderdiff") and list the variable names and labels. The lines following the statement "datalines;" are the data consisting of an id number, the effect size estimate, the variance of the effect size estimate, and the year in which the study was conducted minus 1900 (the study-level covariate).

The last seven lines of the upper panel of Table 15.2 specify the analysis. The first line invokes PROC MIXED for the data set genderdiff. The command "cl" tells SAS to compute 95% confidence intervals for the variance component. The second line ("class = id") tells SAS that the aggregate units are defined by the variable "id," meaning that the Level 1 model is specific to each value of id (i.e., each study). The third line ("model effsize = year/solution = ddfm = bw notest") specifies that the Level 2 model (the model for the fixed effects) is to predict effect size from year and indicates that the "between/within" method of computing the denominator degrees of freedom will be used in tests for the fixed effects, which is generally advisable (see Littell, Milliken, Stroup, & Wolfinger, 1996). The third line also specifies first (via "random = intercept") that the intercept of the Level 1 model (i.e., 2_i) is a random effect and (via "/sub = id") that the Level 2 units are defined by the variable "id." The third line ("repeated/group = id") also specifies the structure of the Level 1 (within-study) error covariance matrix, and "/group = id" indicates that the covariance matrix of ε has a block-diagonal structure with one block for each value of the variable "id"; that is, there is a separate error variance for each study. The statement on lines 4 and 5 specifies the initial values of the 15 variances in this model ($\tau^2, v_1, v_2, \ldots, v_{14}$). Note that the values of the last 14 variances, the sampling error variances $v_1, v_2, \ldots, v_{14}$, are fixed, but τ^2 has to be estimated in the analysis. A good practice (used here) is to use half of the average of the sampling error variances as the initial value of τ^2. The part of the statement on line 5 ("eqcons = 2 to 15") specifies that variances 2 to 15 are to be fixed at their initial values throughout the analysis. Line 7 runs the analysis.

The lower panel of Table 15.2 gives the output file for the analysis specified in the upper panel of the table. It begins with the covariances of all the random effects. Because our model specifies only variances, the estimates are all variances, beginning with the variance of the intercept in the Level 2 model: τ^2,

Table 15.2 SAS Input and Output Files for the Field Articulation Data From Hyde (1981)

Input File

```
data genderdiff;   input label                id effsize variance year; id =’ID of the
                                              study’effsize =’effect size estimate’ variance
                                              =’variances of effect size estimates’ year
                                              =’year the study swas published’; datalines;
 1                 0.76          0.071        55
 2                 1.15          0.033        59
 3                 0.48          0.137        67
 4                 0.29          0.135        67
 5                 0.65          0.140        67
 6                 0.84          0.095        67
 7                 0.70          0.106        67
 8                 0.50          0.121        67
 9                 0.18          0.053        67
10                 0.17          0.025        68
11                 0.77          0.044        70
12                 0.27          0.092        70
13                 0.40          0.052        71
14                 0.45          0.095        72
;
 proc mixed data = genderdiff cl;
class id;
model effsize = year/solution ddfm = bw notest; random int/sub = id ; repeated/group =
id; parms (0.043) (0.071) (0.033) (0.137) (0.135) (0.140) (0.095) (0.106) (0.121)(0.053)
(0.025) (0.044) (0.092) (0.052) (0.095)/eqcons = 2 to 15;

run;
```

Output File

Covariance Parameter Estimates

Covariance Parameter	*Subject*	*Group*	*Estimate*	*Alpha*	*Lower*	*Upper*
Intercept	id		0.03138	0.05	0.008314	1.4730
Residual		id 1	0.0710			
Residual		id 2	0.0330			
Residual		id 3	0.1370			
Residual		id 4	0.1350			
Residual		id 5	0.1400			
Residual		id 6	0.0950			
Residual		id 7	0.1060			
Residual		id 8	0.1210			
Residual		id 9	0.0530			
Residual		id 10	0.0250			
Residual		id 11	0.0440			
Residual		id 12	0.0920			
Residual		id 13	0.0520			
Residual		id 14	0.0950			

Solution for Fixed Effects

Effect	*Estimate*	*Standard Error*	*df*	*t-Value*	*Pr > |t|*
Intercept	3.1333	1.2243	12	2.56	0.0250
Year	–0.03887	0.01837	12	–2.12	0.0560

showing that the estimate is 0.031. Notice that the 95% confidence interval of the variance component τ^2 does not include zero, suggesting that there is significant variation across studies or that effect size estimates vary considerably from study to study. This justifies the use of random-effects models, in which the effect

size parameter is a random variable and has its own distribution. The next 14 lines repeat the study-specific sampling error variances $v_1, v_2, \ldots, v_{14}$ as they were given as initial values and then fixed. The last two rows of the table give the estimates of the Level 2 parameters β_1, β_2 (the fixed effects), their standard errors, the test statistic t, and the associated p-value.

15.6.2.2. Sensitivity Analysis

A close inspection of the data suggests that the estimates of the two studies conducted in the 1950s are somewhat larger than the remaining estimates of the studies conducted in more recent years. Thus, we decided to conduct sensitivity analysis in which each of these estimates individually as well as both estimates simultaneously are omitted from our sample. First, we ran our mixed-models analysis, omitting the estimate of the 1955 study. Our results indicated that the year-of-study coefficient was negative (as the coefficient reported in Table 15.2) and significant at the .05 level. The between-study variance component estimate was comparable to the estimate reported in Table 15.2 and significantly different from zero. In our second analysis, we omitted the estimate of the 1959 study. The year-of-study coefficient was still negative but much smaller in magnitude and did not reach statistical significance. The between-study variance component estimate was statistically significant but nearly one half as large as the previous estimates. Finally, we ran a mixed-models analysis, omitting both estimates of the 1950s studies. In this specification, the year-of-study coefficient was practically zero, and the between-study variance component estimate was comparable to the estimate in our second specification. Overall, these results indicated that our coefficient and variance component estimates in our specification, in which all effect sizes were included in the analysis, are sensitive to omission of the 1950s study estimates, especially the 1959 estimate.

15.6.3. Estimation Using Weighted Least Squares

If τ^2 were known, we could estimate the regression coefficients via weighted least squares (which would also yield the maximum likelihood estimates of the β_is). In this section, we discuss here how to estimate $\boldsymbol{\beta}$ if an estimate of τ^2 is available. Actually obtaining the estimate of τ^2 is discussed in the appendix. The description of the weighted least squares estimation is facilitated by describing the model in matrix notation. The $k \times p$ matrix $\mathbf{X}$,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ . & . & \ldots & . \\ . & . & \ldots & . \\ x_{k1} & x_{k2} & \ldots & x_{kp} \end{bmatrix},$$

is called the *design matrix,* which is assumed to have no linearly dependent columns; that is, $\mathbf{X}$ has rank p. It is often convenient to define $x_{11} = x_{21} = \cdots = x_{k1} = 1$, so that the first regression coefficient becomes an intercept term, as in ordinary regression.

We denote the k-dimensional vectors of population and sample effect sizes by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ and $\mathbf{T} = (T_1, \ldots, T_k)'$, respectively. The model for the observations $\mathbf{T}$ as a one-level model can be written as

$$\mathbf{T} = \boldsymbol{\theta} + \boldsymbol{\varepsilon}, = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\varepsilon}, = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (4)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is the p-dimensional vector of regression coefficients, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)'$ is the k-dimensional vector of random effects, and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)'$ is a k-dimensional vector of residuals of $\mathbf{T}$ about $\mathbf{X}\boldsymbol{\beta}$. The covariance matrix of $\boldsymbol{\xi}$ is a diagonal matrix where the ith diagonal element is $v_i + \hat{\tau}^2$.

If the residual variance component τ^2 were known, we could use the method of generalized least squares to obtain an estimate of $\boldsymbol{\beta}$. Although we do not know the residual variance component τ^2, we can compute an estimate of τ^2 and use this estimate to obtain a generalized least squares estimate of $\boldsymbol{\beta}$. The unconditional covariance matrix of the estimates is a $k \times k$ diagonal matrix $\mathbf{V}^*$ be defined by

$$\mathbf{V}^* = \text{Diag}(v_1 + \hat{\tau}^2, v_2 + \hat{\tau}^2, \ldots, v_k + \hat{\tau}^2).$$

The generalized least squares estimator $\hat{\boldsymbol{\beta}}^*$ under the model (4) using the estimated covariance matrix $\hat{\mathbf{V}}^*$ is given by

$$\hat{\boldsymbol{\beta}}^* = [\mathbf{X}'(\mathbf{V}^*)^{-1}\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{V}^*)^{-1}\mathbf{T},$$

which is normally distributed, with mean $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Sigma}^*$ given by

$$\boldsymbol{\Sigma}^* = [\mathbf{X}'(\mathbf{V}^*)^{-1}\mathbf{X}]^{-1}.$$

Note that the estimate of the between-study variance component $\hat{\tau}^2$ is incorporated as a constant term in the computation of the fixed effects (or regression coefficients) and their dispersion via the variance-covariance matrix of the effect size estimates.

15.6.4. Tests and Confidence Intervals for Individual Regression Coefficients

The distribution of $\hat{\boldsymbol{\beta}}^*$ can be used to obtain tests of significance or confidence intervals for components of $\boldsymbol{\beta}$. If σ_{jj}^* is the jth diagonal element of $\boldsymbol{\Sigma}^*$, and $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_0^*, \hat{\beta}_1^*, \ldots, \hat{\beta}_p^*)'$, then the approximate $100(1-\alpha)\%$ confidence interval for β_j, $1 \leq j \leq p$, is given by

$$\hat{\beta}_j - C_{\alpha/2}\sigma_{jj} \leq \beta_j \leq \hat{\beta}_j + C_{\alpha/2}\sigma_{jj},$$

where $C_{\alpha/2}$ is the $100(1-\alpha)$ percentile of the standard normal distribution (e.g., for $\alpha = 0.05$, $C_{0.05} = 1.64$; for $\alpha = 0.025$, $C_{0.025} = 1.96$).

Approximate tests of the hypothesis that β_j equals some predefined value c_0 (typically 0), that is, a test of the hypothesis

$$H_0 : \beta_j = c_0,$$

uses the statistic

$$t^* = (\hat{\beta}_1^* - c_0)/(\sigma_{jj}^*)^{1/2}.$$

The one-tailed test rejects H_0 at significance level α when the $t^* > C_\alpha$, where C_α is the $100(1-\alpha)$ percentile of Student's t-distribution with $k-p$ degrees of freedom, and the two-tailed test rejects at level α if $|t^*| > C_{\alpha/2}$. The usual theory for the normal distribution can be applied if simultaneous confidence intervals are desired.

15.6.5. Tests for Blocks of Regression Coefficients

As in the fixed-effects model, we sometimes want to test whether a subset $\beta_1, \ldots, \beta_m$ of the regression coefficients is simultaneously zero, that is,

$$H_0 : \beta_1 = \cdots = \beta_m = 0.$$

This test arises, for example, in stepwise analyses when it is desired to determine whether a set of m of the p predictor variables ($m \neq p$) is related for effect size after controlling for the effects of the other predictor variables. For example, suppose one is interested in testing the importance of a conceptual variable such as research design, which is coded as a set of predictors. Specifically, such a variable can be coded as multiple dummies for randomized experiments, matched samples, nonequivalent comparison group samples, and other quasi-experimental designs, but it is treated as one conceptual variable, and its importance is tested simultaneously. To test this hypothesis, compute $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_0^*, \hat{\beta}_1^*, \ldots, \hat{\beta}_m^*, \hat{\beta}_{m+1}^*, \ldots, \hat{\beta}_p^*)'$ and the statistic

$$Q^* = (\hat{\beta}_1^*, \ldots, \hat{\beta}_m^*)(\boldsymbol{\Sigma}_{\mathbf{11}}^*)^{-1}(\hat{\beta}_1^*, \ldots, \hat{\beta}_m^*)', \qquad (5)$$

where $\boldsymbol{\Sigma}_{11}^*$ is the upper $m \times m$ submatrix of

$$\boldsymbol{\Sigma}^* = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{11}}^* & \boldsymbol{\Sigma}_{\mathbf{12}}^* \\ \boldsymbol{\Sigma}_{\mathbf{21}}^* & \boldsymbol{\Sigma}_{\mathbf{22}}^* \end{pmatrix}.$$

The test that $\beta_1 = \cdots = \beta_m = 0$ at the $100\alpha\%$ significance level consists of rejecting the null hypothesis if Q^* exceeds the $100(1-\alpha)$ percentage point of the chi-square distribution with m degrees of freedom.

If $m = p$, then the procedure above yields a test that all the β_j are simultaneously zero; that is, $\boldsymbol{\beta} = \mathbf{0}$. In this case, the test statistic Q^* given in (5) becomes the weighted sum of squares due to regression

$$Q_R^* = \hat{\boldsymbol{\beta}}' \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\beta}}.$$

The test that $\boldsymbol{\beta} = \mathbf{0}$ is simply a test of whether the weighted sum of squares due to the regression is larger than would be expected if $\boldsymbol{\beta} = \mathbf{0}$, and the test consists of rejecting the hypothesis that $\boldsymbol{\beta} = \mathbf{0}$ if Q_R^* exceeds the $100(1-\alpha)$ percentage point of a chi-square with p degrees of freedom.

15.6.6. Testing the Significance of the Residual Variance Component

It is sometimes useful to test the statistical significance of the residual variance component τ^2 in addition to estimating it. The test statistic used is

$$Q_E = \mathbf{T}'[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{T}, \qquad (6)$$

where $\mathbf{V} = \text{Diag}(v_1, \ldots, v_k)$. If the null hypothesis

$$H_0 : \tau^2 = 0$$

is true, then the weighted residual sum of squares Q_E given in (6) has a chi-square distribution with $k-p$ degrees of freedom (where p is the total number of predictors, including the intercept). Therefore, the test of H_0 at level α is to reject if Q_E exceeds the $100(1-\alpha)\%$ point of the chi-square distribution with $(k-p)$ degrees of freedom.

15.6.7. Example

We now return to the data from the 14 studies of gender differences in field articulation ability analyzed via SAS PROC MIXED. The hypothesis that gender

differences were changing over time was investigated using a linear model to predict effect size x from the year in which the study was conducted (minus 1900). The design matrix $\mathbf{X}$ is

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 55 & 59 & 67 & 67 & 67 & 67 & 67 & 67 & 68 & 70 & 70 & 71 & 72 \end{pmatrix}'$$

and the data vector

$$\begin{aligned} \mathbf{T} = (&0.76, 1.15, 0.48, 0.29, 0.65, \\ &0.84, 0.70, 0.50, 0.18, 0.17, \\ &0.77, 0.27, 0.40, 0.45)'. \end{aligned}$$

Using the method given in the appendix, we compute the constant c as $c = 174.537$. Therefore, $\hat{\tau}^2 = (15.11 - 12)/174.537 = 0.018$. Note that the standard error of $\hat{\tau}^2$ following the appendix is $SE(\hat{\tau}^2) = 0.0339$, which is very large compared with $\hat{\tau}^2$, suggesting that the data have relatively little information about τ^2. Using this value of $\hat{\tau}^2$, the covariance matrix $\hat{\mathbf{V}}^*$ becomes

$$\begin{aligned} \hat{\mathbf{V}}^* = \text{Diag}(&0.089, 0.051, 0.155, 0.153, 0.158, \\ &0.113, 0.124, 0.139, 0.071, 0.043, 0.062, \\ &0.110, 0.070, 0.113). \end{aligned}$$

Using SAS PROC REG with weight matrix $\mathbf{V}^*$ as described above, we obtain the estimated regression coefficients $\hat{\beta}_0^* = 3.215$ for the intercept term and $\hat{\beta}_1^* = -0.040$ for the effect of year. The covariance matrix of $\boldsymbol{\beta}^*$ is

$$\boldsymbol{\Sigma}^* = \begin{pmatrix} 1.25963 & -0.01887 \\ -0.01887 & 0.00028 \end{pmatrix},$$

which was obtained by requesting the inverse of the weighted $\mathbf{X}'\mathbf{X}$ matrix. The standard error of $\hat{\beta}_1^*$, the regression coefficient for year, is $\sqrt{0.00028} = 0.0167$, and a 95% confidence interval for β_1^* is given by $-0.0765 = -0.040 - 2.179(0.0167) \leq \beta_1^* \leq -0.040 + 2.179(0.0167) = -0.0035$.

Because the confidence interval does not contain zero, we reject the hypothesis that $\beta_1 = 0$.

Alternatively, we could have computed

$$z(\hat{\beta}_1^*) = \hat{\beta}_1^*/SE(\hat{\beta}_1^*) = -0.040/0.0167 = -2.395,$$

which is to be compared with the critical value of 2.179, so that the test leads to rejection of the hypothesis that $\beta_1^* = 0$ at the $\alpha = 0.05$ significance level.

The test statistic Q_R for testing that the slope and intercept are simultaneously zero has the value $Q_R = 54.14$, which exceeds 5.99, the 95th percentage point of the chi-square distribution with 2 degrees of freedom. Hence, we also reject the hypothesis that $\beta_0^* = \beta_1^* = 0$ at the $\alpha = 0.05$ significance level.

Note that the estimated regression coefficients and their standard errors are close to those estimated using SAS PROC MIXED.

15.6.8. Fixed-Effects Models

Another alternative is to carry out the analysis assuming that $\tau^2 = 0$. This is conceptually equivalent to assuming that any deviations from the linear model at Level 2 are either nonexistent (the linear model fits perfectly at Level 2) or not random (e.g., caused by fixed characteristics of studies that are not in the model). The fixed-effects analysis can be carried out by using the methods described above with $\tau^2 = 0$ or by using a weighted regression program and specifying that the weights to be used for the ith study are $w_i = 1/v_i$.

The weighted least squares analysis gives the regression coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$. The standard errors for the $\hat{\beta}_j$ printed by the program are incorrect by a factor of $\sqrt{MS_E}$, where MS_E is the error or residual mean square for the analysis of variance for the regression. If $S(\hat{\beta}_j)$ is the standard error of $\hat{\beta}_j$ printed by the weighted regression program, then the correct standard error $SE(\hat{\beta}_j)$ of $\hat{\beta}_j$ (the square root of the jth diagonal element of $(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}$) is most easily computed from the results given on the computed printout by

$$SE(\hat{\beta}_j) = S(\hat{\beta}_j)/\sqrt{MS_E}.$$

Alternatively, the diagonal elements of the inverse of the sum of squares and cross-products matrix $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$ also provide the correct sampling variances for $\hat{\beta}_0, \ldots, \hat{\beta}_p$.

The F-tests in the analysis of variance for the regression should be ignored, but the (weighted) sum of squares about the regression is the chi-square statistic Q_E for testing whether the residual variance component is 0, and the (weighted) sum of squares due to the regression gives the chi-square statistic Q_R for testing that all components $\boldsymbol{\beta}$ are simultaneously zero (or a related statistic for testing that all components of $\boldsymbol{\beta}$ except the intercept are zero if the program fits an intercept). Therefore, all the statistics necessary to compute the fixed-effects analysis can be computed with a single run of a weighted least squares program.

The test statistic (equation (5)) for the simultaneous test for blocks of regression coefficients can be computed from the matrices directly. Alternatively, it can

be computed from the output of the weighted stepwise regression as

$$Q_{\text{CHANGE}} = mF_{\text{CHANGE}}\, MS_E,$$

where F_{CHANGE} is the value of the F-test for the significance of the addition of the block of m predictor variables, and MS_E is the weighted error or residual mean square from the analysis of variance for the regression.

15.6.8.1. Example: Gender Differences in Field Articulation

We now return to the data from 14 studies of gender differences in field articulation ability presented by Hyde (1981). We fit a linear model to predict effect size θ from the year in which the study was conducted. As before, the regression model is linear with a constant or intercept term and a predictor, which is the year (minus 1900). The design matrix and the data vector are as before. The covariance matrix is

$$\begin{aligned}\mathbf{V} = \text{Diag}(&0.071, 0.033, 0.137, 0.135,\\ &0.140, 0.095, 0.106, 0.121, 0.053,\\ &0.025, 0.044, 0.092, 0.052, 0.095).\end{aligned}$$

Using SAS PROC REG and specifying the weight for the ith study as $w_i = 1/v_i$ as described above, we obtain the estimated regression coefficients $\hat{\beta}_0 = 3.422$ for the intercept term and $\hat{\beta}_1 = -0.043$ for the effect of the year. The covariance matrix of $\boldsymbol{\beta}$ is

$$\Sigma = \begin{pmatrix} 0.92387 & -0.01385 \\ -0.01385 & 0.00021 \end{pmatrix},$$

a result obtained by requesting the inverse of the weighted $\mathbf{X}'\mathbf{X}$ matrix.

Consequently, the fixed-effects standard error of $\hat{\beta}_1$, the regression coefficient for year, is $\sqrt{0.00021} = 0.0145$, and a 95% confidence interval for β_1 is given by $-0.043 \pm 2.179(0.0145)$ or $-0.0749 \leq \beta_1 \leq -0.0117$.

Because the confidence interval does not contain zero, we reject the hypothesis that $\beta_1 = 0$.

Alternatively, we could have computed

$$z(\hat{\beta}_1) = \hat{\beta}_1 / SE(\hat{\beta}_1) = -0.043/0.0145 = -2.986,$$

which is to be compared with the critical value of 2.179, so that the test leads to rejection of the hypothesis that $\beta_1 = 0$ at the $\alpha = 0.05$ significance level.

The coefficients in the fixed-effects model are comparable to the coefficients in the random-effects model. The standard errors of the regression coefficients, however, are larger in the random-effects model, as expected. In the random-effects model, $\hat{\tau}^2$ is included in the variance-covariance matrix of the effect size estimates (as a constant), and, therefore, the diagonal elements of $\mathbf{V}$ are somewhat larger than in the fixed-effects model, where $\hat{\tau}^2$ is zero. For example, in the random-effects model, the standard error of the year-of-study coefficient is 15% larger than in the fixed-effects model. It is noteworthy that year of study explained approximately 45% of the random variation across studies, suggesting that approximately one half of the between-study variance is associated with the year in which the study was conducted.

15.6.9. Bayesian Approaches

The third approach to carrying out the analysis is not to use *any* single value of τ^2 by using Bayesian methods. The problem with carrying out random-effects analyses by substituting any fixed value of τ^2 is that information about τ^2 comes from variation between studies, and when the number of studies is small, any estimate of the variance component must be rather uncertain. Therefore, an analysis (such as the conventional random-effects analysis) that treats an estimated value of the variance component as if it were known with certainty is problematic.

Bayesian analyses address this problem by recognizing that there is a family of random-effects analyses, one for each value of the variance component. The Bayesian analyses can be seen as essentially averaging over the family of results, assigning each one the weight that is appropriate given the posterior probability of each variance component value conditional on the observed data. Some approaches to Bayesian inference for meta-analysis do this directly (see, e.g., DuMouchel & Harris, 1983; Hedges, 1998; Rubin, 1981). They compute the summary statistics of the posterior distribution (such as the mean and variance) conditionally given τ^2, then average (integrate) these, weighting by the posterior distribution of τ^2 given the data. Although these approaches are transparent and provide a direct approach to obtaining estimates of parameters of the posterior distribution, they require numerical integrations that can be challenging.

An alternative is the use of Markov chain Monte Carlo methods, which provide the posterior distribution directly without difficult numerical integrations. These methods provide not only the mean and variance of the posterior distribution but also the entire posterior distribution, making it possible to compute

many descriptive statistics of that distribution. Another major advantage of these methods is that they permit models in which the study-specific random effects are not normally distributed but have heavier tailed distributions, such as a Student's t or gamma distribution (see Seltzer, 1993; Seltzer, Wong, & Bryk, 1996; Smith, Spiegelhalter, & Thomas, 1995).

15.7. Conclusion

This chapter presented several models for meta-analysis, situating these methods within the context of hierarchical linear models. Three different approaches to estimation were discussed, and the use of both random-effects (mixed) and fixed-effects models was illustrated with an example of data from 14 studies of gender differences in field articulation. Traditional statistical software packages such as SAS, SPSS, or Splus can be easily used to conduct weighted least squares analyses in meta-analysis. In addition, more specialized software packages, such as HLM, MLwin, and the mixed-models procedure in SAS, can carry out mixed-models analyses for meta-analytic data with nested structure.

The mixed-effects models presented here can be extended to three or more levels of hierarchy, capturing random variation at higher levels. For example, there is often reason to adopt a three-level structure in meta-analysis, in which studies themselves are clustered. For example, when the same investigators (or laboratories) carry out multiple studies, a three-level meta-analysis can model the variation between investigators (or laboratories) at the third level, as well as between studies within the same investigator (or laboratory) at the second level.

APPENDIX

A DISTRIBUTION-FREE ESTIMATE OF THE RESIDUAL VARIANCE COMPONENT

The distribution-free method of estimation involves computing an estimate of the residual variance component and then computing a weighted least squares analysis conditional on this variance component estimate. Whereas the estimates and their standard errors are "distribution free" in the sense that they do not depend on the form of the distribution of the random effects, the tests and confidence statements associated with these methods are only strictly true if the random effects are normally distributed.

ESTIMATION OF THE RESIDUAL VARIANCE COMPONENT τ^2

The first step in the analysis is the estimation of the residual variance component. Different estimators might be used, but the usual estimator is based on the statistic used to test the significance of the residual variance component, the inverse conditional variance-weighted residual sum of squares. It is the natural generalization of the estimate of the between-studies variance component, given, for example, by DerSimonian and Laird (1986).

The usual estimator of the residual variance component is given by

$$\hat{\tau}^2 = (Q_E - k + p)/c,$$

where Q_E is the test statistic used to test whether the residual variance component is zero (the residual sum of squares from the weighted regression using weights $w_i = 1/v_i$ for each study), and c is a constant given by

$$c = \text{tr}(\mathbf{V}^{-1}) - \text{tr}[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-2}\mathbf{X}],$$

where $\mathbf{V} = \text{diag}(v_1, \ldots, v_k)$ is a $k \times k$ diagonal matrix of conditional variances, and $\text{tr}(\mathbf{A})$ is the trace of the matrix $\mathbf{A}$.

The Standard Error of $\hat{\tau}^2$

When the random effects are normally distributed, the standard error of $\hat{\tau}^2$ is given by

$$[SE(\hat{\tau}^2)]^2 = 2\{\text{tr}(\mathbf{V}^{-2}\mathbf{V}^{*2}) - 2\text{tr}(\mathbf{M}\mathbf{V}^{-1}\mathbf{V}^{*2}) + \text{tr}(\mathbf{M}\mathbf{V}^{*}\mathbf{M}\mathbf{V}^{*})\}/c^2,$$

where the $k \times k$ matrix $\mathbf{M}$ is given by

$$\mathbf{M} = \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1},$$

$\mathbf{V}^*$ is a $k \times k$ symmetric matrix given by

$$\mathbf{V}^* = \mathbf{V} + \tau^2\mathbf{I},$$

and $\mathbf{I}$ is a $k \times k$ identity matrix. However, it is important to remember that the distribution of the residual variance component estimate is not close to normal unless $(k - p)$ is large. Consequently, probability statements based on $SE(\hat{\tau}^2)$ and the assumption of normality should not be used unless $(k - p)$ is large.

REFERENCES

Camilli, G. (1990). The test of homogeneity for 2 × 2 contingency tables: A review of and some personal opinions on the controversy. *Psychological Bulletin, 108*, 135–145.

Cooper, H. (1989a). *Homework.* New York: Longman.

Cooper, H. (1989b). *Integrating research* (2nd ed.). Newbury Park, CA: Sage.

Cooper, H. M., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*, 177–188.

DuMouchel, W. H., & Harris, J. E. (1983). Bayes method for combining the results of cancer studies in humans and other species. *Journal of the American Statistical Association, 78*, 293–315.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245–260). New York: Russell Sage Foundation.

Garvey, W., & Griffith, B. (1971). Scientific communication: Its role in the conduct of research and creation of knowledge. *American Psychologist, 26*, 349–361.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3–8.

Goldstein, H. (1987). *Multi level models in educational and social research.* Oxford, UK: Oxford University Press.

Hartley, H. O., & Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika, 54*, 93–108.

Harville, D. A. (1977). Maximum likelihood approaches to variance components estimation and to related problems. *Journal of the American Statistical Association, 72*, 320–340.

Hedges, L. V. (1982a). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92*, 490–499.

Hedges, L. V. (1982b). Fitting categorical models to effect sizes from a series of experiments. *Journal of Educational Statistics, 7*, 119–137.

Hedges, L. V. (1983a). Combining independent estimators in research synthesis. *British Journal of Mathematical and Statistical Psychology, 36*, 123–131.

Hedges, L. V. (1983b). A random effects model for effect sizes. *Psychological Bulletin, 93*, 388–395.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist, 42*, 443–455.

Hedges, L. V. (1998). Bayesian approaches to meta-analysis. In B. Everitt & G. Dunn (Eds.), *Recent advances in the statistical analysis of medical data* (pp. 251–275). London: Edward.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical test in meta-analysis. *Psychological Methods, 6,* 203–217.

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random effects models in meta analysis. *Psychological Methods, 3*, 486–504.

Hyde, J. S. (1981). How large are cognitive gender differences: A meta-analysis using omega and d. *American Psychologist, 36*, 892–901.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models.* Cary, NC: SAS Institute, Inc.

Longford, N. (1987). *Random coefficient models.* Oxford, UK: Oxford University Press.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences.* Stanford, CA: Stanford University Press.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10*, 75–98.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin, 92,* 500–504.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics, 6*, 377–401.

Schmidt, F. L., & Hunter, J. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529–540.

Seltzer, M. (1993). Sensitivity analysis for fixed effects in the hierarchical model: A Gibbs sampling approach. *Journal of Educational Statistics, 18*, 207–235.

Seltzer, M. H., Wong, W. H., & Bryk, A. S. (1996). Bayesian analysis in applications of hierarchical models: Issues and methods. *Journal of Educational and Behavioral Statistics, 21*, 131–167.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel growth models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24*, 323–355.

Smith, M., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*, 752–760.

Smith, T. C., Spiegelhalter, D. J., & Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine, 14*, 2685–2699.

Section V

Models for Latent Variables

Chapter 16

Determining the Number of Factors in Exploratory and Confirmatory Factor Analysis

Rick H. Hoyle

Jamieson L. Duvall

Researchers in the social sciences are interested in the unobserved, or latent, variables that are causes or consequences of the behaviors they observe. Latent variables such as attitudes, feelings, and motives are valuable because, in the context of a well-reasoned theory, they have the potential to explain a wide array of behavioral processes using a relatively small number of constructs. Moreover, they bring richness and detail to theoretical accounts of behaviors that are not adequately described by observable influences.

These latent variables, or factors, typically are inferred from patterns of association among sets of observed variables believed to be caused, at least in part, by one or more factors.[1] The patterns of association are conveyed in matrices of covariances or correlations, and to the extent that the associations among the observed variables are near zero when the influence of the factors is taken into account, the factors provide a parsimonious account of reliable variance in the observed variables without significant loss of information.

The primary statistical tool for drawing such inferences is factor analysis. The aim of factor analysis is to describe the associations among a potentially large number of observed variables, or indicators, using a relatively small number of factors. It is assumed that the indicators are fallible in their representation of the underlying factor. That is, a portion of the variability in observed scores on each indicator is shared with other indicators of the factor, whereas a portion is, with reference to the factor, unique to the indicator. In theory, the uniqueness component can further be decomposed into variability attributable to other factors, specificity, and variability attributable to random fluctuation,

1. The material presented in this chapter is not relevant for factor models in which the latent variable is presumed to be caused by the indicators. In such models, the indicators are referred to as *cause indicators* (Bollen & Lennox, 1991) or *formative indicators* (Cohen, Cohen, Teresi, Marchi, & Velez, 1990).

AUTHORS' NOTE: Correspondence should be addressed to Rick H. Hoyle, Department of Psychology: SHS, Box 90085, Duke University, Durham, NC 27708–0085. Send electronic mail to rhoyle@duke.edu.

error. In practice, this decomposition is possible only to the extent that the data matrix includes other indicators manifesting the same specificity, or covariates that are associated with the specific component (i.e., specific effects) (Bentler, 1990b).

A fundamental concern in applications of factor analysis is the determination of how many factors are necessary to adequately account for the commonality among the indicators in a set. Despite the fact that this concern received considerable attention from quantitative methodologists in the 1950s and 1960s (e.g., Cattell, 1966; Guttman, 1954; Horn, 1965), a half century later, it remains a topic of considerable debate (e.g., Bollen, 2000; Hayduk & Glaser, 2000; Herting & Costner, 2000; Mulaik & Millsap, 2000) and, judging from published applications of factor analysis, still is poorly understood by social and behavioral scientists (Fabrigar, Wegener, MacCallum, & Strahan, 1999).

In this chapter, we review five procedures for determining the number of factors in a factor analysis. Three apply exclusively to exploratory factor analysis, one applies either to exploratory or confirmatory factor analysis, and one applies exclusively to confirmatory factor analysis. We demonstrate that the two procedures used most frequently by substantive researchers frequently lead to incorrect inferences regarding the number of factors underlying a set of indicators. We illustrate the application of, and compare results from, the three remaining procedures in a factor analysis of a widely used self-report measure known to be factorially complex.

16.1. Exploratory and Confirmatory Factor Models

In exploratory factor analysis (i.e., principal axis, common factors), each of the p observed variables, X_i, is modeled as a linear combination of k factors, $\xi_1, \xi_2, \ldots, \xi_k$, and a uniqueness component, δ_i. The model can be expressed as a series of measurement equations:

$$X_1 = \sum_{j=1}^{k} \lambda_{1j}\xi_j + \delta_1$$

$$X_2 = \sum_{j=1}^{k} \lambda_{2j}\xi_j + \delta_2$$

$$\ldots$$

$$X_m = \sum_{j=1}^{k} \lambda_{mj}\xi_j + \delta_m.$$

The λs are, in effect, regression coefficients that index the degree to which variance in the indicator shared with the other indicators (i.e., its communality) is explained by each of the k factors.[2]

A significant drawback to exploratory factor analysis is the indeterminacy of the estimates of the λs and δs. That is, the λs and δs (as well as the variances of and covariances between the factors) are not uniquely determined by the observed data. In theory, there exist an infinite number of solutions to the measurement equations that would be equally consistent with the observed data. Indeed, it is possible to derive equally valid solutions that imply factors that are scarcely correlated (Steiger, 1996). This mathematical property of solutions generated by exploratory factor models stems from the fact that, as typically specified, such models require estimation of more unknowns, or free parameters, than there are observed data.[3] This is a concern about degrees of freedom that attends all forms of statistical inference, and it gives rise to the inferential problem that such models cannot be disconfirmed.

The solution to both the indeterminacy and inability-to-disconfirm problems is to impose constraints on the measurement equations by fixing a subset of the λs to a specific value. Often, this value is zero, which represents the hypothesis that the ξ does not contribute to the common variance in the variable. For instance, expanding the measurement equations presented earlier for the case of six indicators ($p = 6$), we might impose the following constraints:

$$X_1 = \lambda_{11}\xi_1 + 0\xi_2 + 0\xi_3 + 0\xi_4 + 0\xi_5 + 0\xi_6 + \delta_1,$$

$$X_2 = \lambda_{21}\xi_1 + 0\xi_2 + 0\xi_3 + 0\xi_4 + 0\xi_5 + 0\xi_6 + \delta_2,$$

$$X_3 = \lambda_{31}\xi_1 + 0\xi_2 + 0\xi_3 + 0\xi_4 + 0\xi_5 + 0\xi_6 + \delta_3,$$

$$X_4 = 0\xi_1 + \lambda_{42}\xi_2 + 0\xi_3 + 0\xi_4 + 0\xi_5 + 0\xi_6 + \delta_4,$$

$$X_5 = 0\xi_1 + \lambda_{52}\xi_2 + 0\xi_3 + 0\xi_4 + 0\xi_5 + 0\xi_6 + \delta_5,$$

$$X_6 = 0\xi_1 + \lambda_{62}\xi_2 + 0\xi_3 + 0\xi_4 + 0\xi_5 + 0\xi_6 + \delta_6.$$

The particular set of constraints imposed on these equations yields a two-factor model ($k = 2$) with a simple structure pattern of loadings. By fixing all

2. The term *exploratory factor analysis,* as used in this chapter, does not encompass principal components analysis.

3. The factor indeterminacy problem can be characterized in other ways and is substantially more complex than described here. A thoroughgoing presentation of the problem is beyond the scope of the chapter. Readers interested in greater detail would benefit from reading the historical review chapter by Steiger and Schönemann (1978). Readers interested in current thinking about the factor indeterminacy problem, from both mathematical and philosophical perspectives, would benefit from reading a series of articles published in Volume 31 of *Multivariate Behavioral Research.*

of the λs for ξ_3, ξ_4, ξ_5, and ξ_6 to zero, we have posited that two factors are sufficient to explain the commonality among the indicators. Moreover, by fixing the λs to zero for ξ_2 in the first three equations and ξ_1 in the last three equations, we have posited a model with no factorially complex indicators. Importantly, we have reduced the number of unknowns in this set of equations from 42 (36 λs and 6 λs) to 12 (6 λs and 6 δs).[4] Each zero in the measurement equations is a disconfirmable hypothesis, and the collection of zeroes and free parameters constitutes a disconfirmable model of the underlying causes of X_1 to X_6.

It is important to appreciate that we have ventured two hypotheses by imposing this set of constraints. The first, and most fundamental, is that $k = 2$. The second is that, given that $k = 2$, three indicators are exclusively caused by ξ_1, and three are exclusively caused by ξ_2. It is this first hypothesis, that k equals a specified number, that is the concern of the remainder of this chapter. For, as should be apparent from this example, if that hypothesis is refuted, then any hypothesis regarding the pattern of loadings on the factors is premature. In short, a fundamental decision that attends any application of factor analysis is how many factors to retain, in the case of exploratory factor analysis, or specify, in the case of confirmatory factor analysis.

16.2. Strategies for Determining the Number of Factors

Traditionally, the question of how many factors underlie a set of indicators has been viewed as an issue in applications of exploratory factor analysis but not a concern in applications of confirmatory factor analysis. For instance, if, upon evaluating the fit of a confirmatory factor model, a researcher determines that there is an unacceptable degree of misspecification (i.e., the model does not fit), the typical focus of respecification is on either the zero constraints imposed on the factor loadings (the λs) or the zero constraints imposed on the covariances between error terms (the off-diagonal $\Theta\delta$s). By allowing indicators to load on more than one factor or selected error terms to covary, researchers often can obtain acceptable statistical fit of the model to the data. By limiting the search for misspecification to these options, researchers fail to consider the possibility that their model is misspecified in a more fundamental way: It posits too many or too few factors. When faced with a misspecified confirmatory factor model, the researcher must address the same question that attends all applications of exploratory factor analysis: How many factors are required to adequately explain variance shared by the indicators?

In the remainder of this section, we review five strategies for addressing the number-of-factors question in applications of factor analysis. The first two—the Kaiser-Guttman (K-G) rule and the scree plot—apply exclusively to exploratory factor analysis, involve evaluating the latent roots of the correlation matrix, and are the overwhelming favorites of researchers in the social sciences. We argue that these strategies are problematic and should not be used. The other three—parallel analysis, maximum likelihood exploratory factor analysis, and the unrestricted factor model—are less subjective, statistically sound, and more accurate. Although parallel analysis is relevant only for applications of exploratory factor analysis, we demonstrate that maximum likelihood estimation of common factors and the unrestricted confirmatory factor model are equivalent approaches to determining the number of factors in applications of exploratory and confirmatory factor analysis, respectively.

16.2.1. K-G Rule

In a seminal paper on the determination of how many factors are necessary to account for the commonality among a set of indicators, Guttman (1954), concerned that investigators frequently retained too few factors, proposed three rules for determining the minimum number of factors to retain. Ironically, the rule that Guttman found least satisfactory, "latent root greater than or equal to unity," became the rule of choice for applied researchers. The popularity of this rule can be traced to Kaiser (1960). Kaiser found that the rule favored by Guttman frequently indicated more than half as many factors as variables, whereas the latent-root-greater-than-one rule routinely indicated from one sixth to one third as many factors as variables. Kaiser further argued that, when the latent root falls below 1.0, the internal consistency of the factor scores

4. Our discussion here does not require that we consider constraints that could be put on Θ_δ, which, in addition to the δs, includes the covariances between them, and Φ, which is the variance-covariance matrix for the factors. For instance, the zero constraints on all λs for ξ_3, ξ_4, ξ_5, and ξ_6 imply zero constraints on the variances for those factors and their covariances with each other and with ξ_1 and ξ_2. We could further constrain Φ by fixing the covariance between ξ_1 and ξ_2 to zero, which would yield a model with two orthogonal factors.

approaches zero.[5] Finally, on a more practical level, Kaiser argued that, based on his experience "of having carried out more factor analytic calculations of the theoretical sort on electronic computers than anyone walking the earth" (p. 141), the latent-root-greater-than-one rule "led to a number of factors corresponding almost invariably ... to the number of factors which practicing psychologists were able to interpret" (p. 145).

The results of subsequent studies of the K-G rule were not so favorable. On theoretical grounds, the rule is problematic because, although Guttman (1954) demonstrated its validity for determining the number of factors in a population matrix, its application is virtually always to a matrix based on data from a sample. Because of errors of measurement that abound in such data, factors emerge that are not substantively meaningful (Cattell, 1966). As such, it is possible that, for a matrix estimated from sample data, a relatively large number of factors with latent roots greater than one would be detected but that some portion of these factors would not be substantively meaningful. Indeed, simulation studies indicate that the K-G rule, when applied to data from a sample, virtually always leads to overextraction (e.g., Browne, 1968; Lee & Comrey, 1979; Yoemans & Golder, 1982; Zwick & Velicer, 1986).

Despite the now well-documented poor performance of the K-G rule in practice, it still is widely used. In large measure, this persistent adherence to a faulty strategy seems to be attributable to simplicity and availability. Indeed, as early as the mid-1960s, Cattell (1966) noted that the K-G rule had become widely accepted "because of its ease rather than its rationale" (p. 261). In an influential review of rules for determining the number of components to retain in principal components analysis, Zwick and Velicer (1986) concluded that the K-G remains popular because it continues to receive favorable coverage in general statistics textbooks and is invoked automatically by most general-purpose statistical software packages. Reflecting on their finding that the K-G rule often grossly overestimates the number of components to be retained, Zwick and Velicer stated,

> This pattern of explicit endorsement by textbook authors and implicit endorsement by computer packages, contrasted with empirical findings that the procedure is very likely to provide a grossly wrong answer, seems to guarantee that a large number of incorrect findings will continue to be reported. (p. 439)

This pessimistic projection has been borne out. Beginning 5 years after the publication of Zwick and Velicer's (1986) findings, Fabrigar et al. (1999) reviewed standard practice as evidenced in articles published in two major psychology journals. They found that, among those articles that specified the method by which they determined the number of factors, 28% use the K-G rule as the lone strategy, despite the fact that "we know of no study of this rule that shows it to work well" (p. 278). In short, application of the K-G rule is not a defensible strategy for determining the number of factors in applications of exploratory factor analysis.

16.2.2. Scree Plot

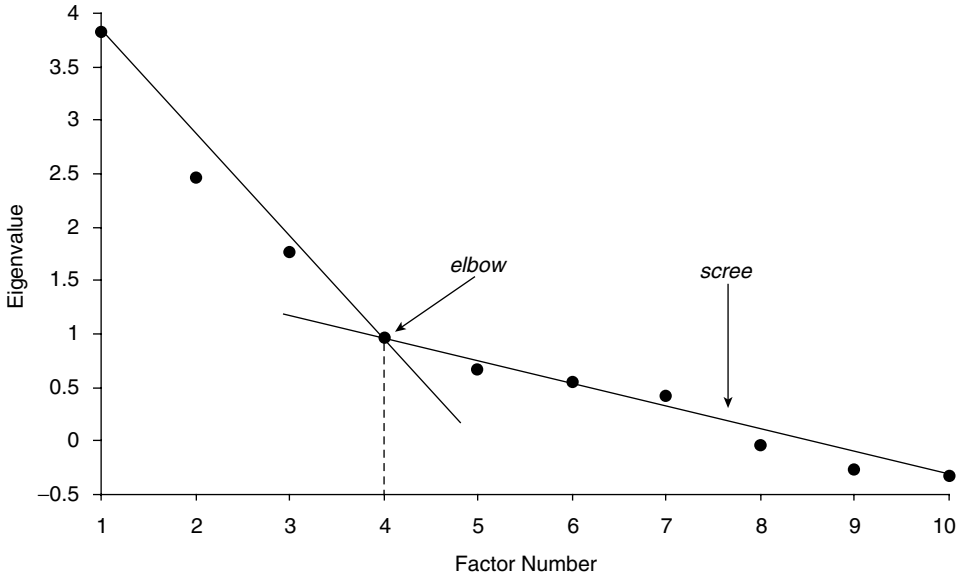
Another common approach to determining the number of factors to retain in exploratory factor analysis is the scree test (Cattell, 1966).[6] Like the K-G rule, the scree test involves evaluating the latent roots, or eigenvalues, of the observed correlation matrix. Cattell (1966) described the goal of factor analysis as detecting "non-trivial common variance" (p. 245). That is, although the fallible nature of observed data will give rise to many trivial sources of common variance, it is the goal of the factor analyst to distinguish these "rubble factors" from those factors that are prominent in the data and substantively meaningful. Cattell adopted the analogy of a rockslide for the purpose of distinguishing important and trivial common factors. He noted that, when eigenvalues are plotted sequentially, the resulting pattern drops precipitously before leveling off. Following the analogy, this relatively level portion of the plot corresponds to the scree, "the straight line of rubble and boulders which forms at the pitch of sliding stability at the foot of a mountain" (p. 249). A scree plot of hypothetical eigenvalues from a factor analysis of 10 variables is shown in Figure 16.1.

Cattell (1966) noted that, based on a number of studies in which the true number of factors was known, "this scree *invariably* began at the *k*th latent root

5. Cattell (1966) took issue with Kaiser's (1960) argument regarding the internal consistency of factor scores for latent roots less than 1. Specifically, he argued that, whereas Kaiser's point applied to all variables, one typically is concerned only with those variables having salient loadings on the factor. He also took issue with the implication by Kaiser that applied researchers were interested in using factor analysis for the purpose of measuring concepts, arguing instead that the focus is more on the interpretation of loadings for the purpose of defining theoretical concepts that are not directly measured.

6. Fabrigar, Wegener, MacCallum, and Strahan (1999) found that the scree test was used as the lone criterion for determining number of factors in 26% of the published articles included in their review.

Figure 16.1 Scree Plot From Hypothetical Factor Analysis of 10 Variables



when k was the true number of factors" (p. 249). As shown in Figure 16.1, the point at which the scree begins is referred to as the "elbow." Following Cattell's logic, the elbow falls at a point on the horizontal axis corresponding to the likely number of substantive factors ($k = 4$ in this example). Cattell confessed that "even a test as simple as this requires the acquisition of *some* art in administering it" (p. 256). It is this subjective quality of the scree test that raises concerns about its usefulness. Inter-rater reliability for judgments of the number of factors to retain based on the scree test using data typical of social science research is questionable. Following training using instructions provided by Cattell and Vogelman (1977), Zwick and Velicer (1986) observed inter-rater reliability estimates from .61 to 1.00 (see also Crawford & Koopman, 1979) and correlations between judgments by the trained raters and experts from .60 to .90. The mean correlation of .80 indicates that, on average, well-trained raters often will reach a conclusion regarding how many factors to retain that differs from the conclusion of experts.

Concerns stemming from the less-than-perfect reliability of the scree test when used by individuals who are not expert factor analysts plague any evaluation of the validity of the scree test. If trained raters' judgments are combined, yielding a judgment that is more reliable than any given rater's judgment, a "best-case" estimate of accuracy can be obtained. Under these conditions, Zwick and Velicer (1986) found that the scree test typically resulted in the retention of too many factors, although the magnitude of overextraction was considerably less than for the K-G rule. Under conditions of high factor saturation (i.e., salient loadings of .80), the scree test pointed to the correct number of factors about 70% of the time; however, accuracy dropped to 40% under conditions of saturation more typical of research in the social sciences. In short, the subjective nature of the scree test, coupled with its tendency to suggest too many factors in typical research conditions, argues persuasively against its use as the sole criterion for determining how many factors to extract in applications of exploratory factor analysis.

16.2.3. Parallel Analysis

A more systematic evaluation of the eigenvalues of an observed correlation matrix is parallel analysis, originally proposed by Horn (1965). Parallel analysis is based on differences between observed eigenvalues and expected eigenvalues from matrices of random data. The logic of parallel analysis is as follows: In population data, the eigenvalues of a matrix of uncorrelated variables would be equal to 1.0. In data from a sample, a matrix with the same properties would, as

a result of sampling error and bias due to estimation, produce some eigenvalues that exceed 1.0 and others that fall below 1.0. Indeed, if multiple samples were drawn from the same population, it would be possible to construct an empirical sampling distribution around the expected eigenvalues were there no common factors in the population (Glorfeld, 1995). One could then compare each observed eigenvalue to the distribution of its corresponding eigenvalue from random data to determine whether that eigenvalue differed significantly (i.e., exceeded the 95th percentile estimate) from the eigenvalue that would be expected were there no common variance to explain. The number of eigenvalues that significantly exceed the values expected from random data is interpreted as the number of factors or components to retain.

An aspect of parallel analysis that has received considerable attention is the method by which the values and distributions of the eigenvalues from random data are produced. One alternative is to generate multiple data sets of random-normal deviates that comprise the same number of variables and assume the same sample size as the observed data set. Using this strategy, Humphreys and Montanelli (1975) found parallel analysis to be accurate (and superior to maximum likelihood) across a range of analytic situations typical of research in the social sciences. Easily tailored computer programs that run under widely available statistical software (e.g., SAS and SPSS) are now available (e.g., O'Connor, 2000). These programs extend the strategy used by Humphreys and Montanelli to allow for the generation of random deviates whose distribution corresponds to the distribution of the observed variables (cf. Glorfeld, 1995).

The allure of parallel analysis, coupled with applied researchers' general unfamiliarity with data simulation at the time of Humphreys and Montanelli's (1975) work, led to the development of more straightforward methods of generating eigenvalues from random data. Most of this activity centered on the specification of regression equations for estimating eigenvalues of matrices with unities on the diagonal (e.g., Allan & Hubbard, 1986; Lautenschlager, Lance, & Flaherty, 1989; Montanelli, 1975; cf. Lautenschlager, 1989). Although computer programs and tabled values for selected sample sizes and numbers of variables are available (e.g., Kaufman & Dunlap, 2000; Longman, Cota, Holden, & Fekken, 1989a, 1989b), and equations are available for matrices with squared multiple correlations on the diagonal (Montanelli & Humphreys, 1976), the regression approach is less precise than the simulation approach. Moreover, the latter is now quite feasible for social scientists facile with SAS or SPSS and generally familiar with the activity of simulating data.

Early evaluations of parallel analysis revealed a slight tendency (about 5%) toward overextraction under certain conditions when, as recommended by Horn (1965), the random-data eigenvalues against which the observed eigenvalues were compared were simple means of a set of random-data eigenvalues at each serial position (Harshman & Reddon, 1983). This bias vanishes when the criterion is the 95th percentile estimate of the random-data eigenvalues at each serial position, a logic based on the standard hypothesis-testing criterion (Glorfeld, 1995).[7] Despite the accuracy of parallel analysis for determining the number of factors to extract under conditions typical of social science research, its use remains relatively rare. Fabrigar et al. (1999) found that, of the 129 articles that reported using factor analysis in two prominent psychology journals published from 1991 through 1995, only 1 reported the use of parallel analysis to determine how many factors to extract.

16.2.4. Maximum Likelihood Exploratory Factor Analysis

There are a variety of procedures for fitting data to a common factor model. These procedures, which produce estimates of the λs and δs in the measurement equations, vary in the assumptions they make regarding the observed data and the information they provide regarding the adequacy of a particular model for explaining the associations among indicators. Although the maximum likelihood procedure requires a relatively strong set of assumptions regarding the distribution of indicators and errors (Hu, Bentler, & Kano, 1992), it has the benefit of providing a test statistic for evaluating the tenability of particular models given a set of data. Because a fundamental aspect of the plausibility of a model is the number of factors it specifies, the test statistics that can be generated when maximum likelihood estimation is used provide an alternative means of determining the number of factors

7. Turner (1998) showed that, under certain conditions, use of the 95th percentile as a criterion for the random-data eigenvalues could lead to underextraction. Specifically, when the observed data were generated by a multilevel design or when all items are saturated by a single common factor, eigenvalues after the first eigenvalue for the observed data will be underestimated and likely fall below the 95th percentile estimate for the corresponding random-data eigenvalue. The circumstances that produce this underextraction by parallel analysis are rare in practice and can be overcome by building known features of the data structure into the parallel analysis. At present, there are no computer programs available for this sophisticated, sequential application of parallel analysis.

that underlie the pattern of associations among a set of indicators.

Whereas principal factor procedures focus on obtaining estimates that minimize residuals (or, alternatively, maximize variance accounted for), the maximum likelihood procedure focuses on obtaining the set of estimates of the free parameters in the model (principally the λs and δs) that has the highest probability of corresponding to the population values of the parameters as the size of the sample approaches the size of the population (Gorsuch, 1974). That is, the goal of estimation is to maximize the likelihood of the parameters given the data. When an iterative search yields the set of parameter estimates that achieve this goal, the estimation is said to have converged. The question of whether this set of parameter estimates offers an account of the data that is no worse than the generally uninteresting model—which is, in effect, the observed correlation matrix itself—can be evaluated using a test statistic that is, in theory, distributed as a chi-square. Because the goal is to obtain a set of parameter estimates in a specified model that fully accounts for the observed data, the goal of hypothesis testing is to fail to reject the null hypothesis of no difference between the observed correlation matrix and the set of correlations implied by the model.

When maximum likelihood estimation is used in the context of exploratory factor analysis, the statistical test is, in fact, a test of the adequacy of the number of factors specified. There are two slightly different approaches to deriving the test statistic; both are products involving an adjusted sample size and the minimized value of the fitting function, computed as

$$F_{ML} = \log^* \Gamma(\mathbf{\Theta})^* + \text{tr}(\mathbf{S}\Gamma^{-1}(\Theta)) - \log^* \mathbf{S}^* - p,$$

where $\Gamma(\mathbf{\Theta})$ is the implied covariance matrix of the parameters, $\mathbf{S}$ is the observed covariance matrix, and p is the number of indicators. The most widely used multiplier is $N-1$, where N is the sample size (Bollen, 1989; Browne, 1982). An alternative involves a correction attributable to Bartlett (1937) and uses the following as a multiplier:

$$(N - 1) - (2p + 4k + 5)/6,$$

where p is the number of indicators, and k is the number of factors (Lawley & Maxwell, 1971). Although Bartlett corrections, as a class, are useful in many contexts, the correction is rarely used in the factor analysis context and is not recommended here.

In the typical application of maximum likelihood factor analysis, a sequential set of analyses is run, which begins with the extraction of a single factor and continues by adding one factor at a time until the obtained chi-square is nonsignificant. A common problem with maximum likelihood estimation in the exploratory factor analysis context are Heywood cases—solutions in which the communality estimate associated with one or more factors approaches or exceeds 1.0, resulting in an associated uniqueness estimate that approaches zero or, counterintuitively, takes on a negative value. Heywood cases often result from overfitting (i.e., too many factors) but can result from underfitting as well.

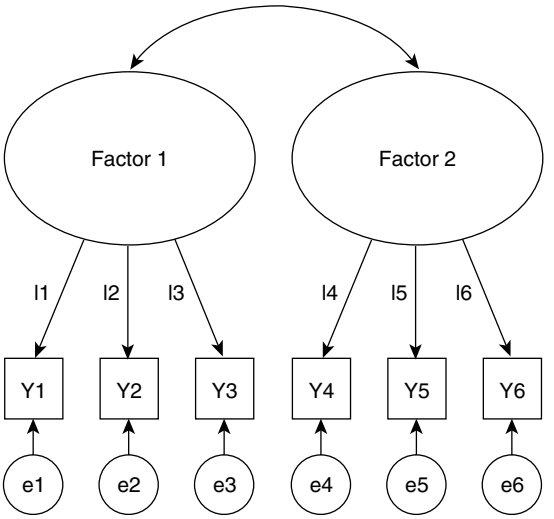
Concern over Heywood cases, the need for substantially greater computing power than required of principal axis procedures, and the validity of the chi-square test under conditions typical of social science research argue against unqualified endorsement of maximum likelihood exploratory factor analysis and the accompanying fit statistics as a routine strategy for determining the number of common factors (e.g., Jackson & Chan, 1980); however, this approach is the only one that offers a focused statistical test of the plausibility of a specific number of factors. Fortunately, computer power is rarely an issue in the current era, and a variety of criteria for evaluating model fit have been developed in the context of confirmatory factor analysis that can be used for maximum likelihood exploratory factor analysis as well (e.g., Browne & Cudeck, 1993). These are described and illustrated later in the chapter. Also, increasing interest in confirmatory factor analysis, as well as the accompanying familiarity with parameter estimation and evaluation, has resulted in social scientists who are better equipped than ever to detect Heywood cases and adjust model specifications to correct or avoid them.

16.2.5. Unrestricted Factor Model

Maximum likelihood is used to greater advantage as the typical estimation procedure in confirmatory factor analysis (Hoyle, 2000). The extension of maximum likelihood estimation from exploratory, or unrestricted, factor analysis to confirmatory, or restricted, factor analysis can be attributed to Jöreskog (1969). As typically applied, confirmatory factor models specify a particular number of factors and a pattern of loadings such that each indicator loads on one and only one factor. This simple-structure specification is illustrated in path diagram form in Figure 16.2.

As noted earlier, when omnibus fit indices indicate misspecification in a model such as the one shown in Figure 16.2, it is not clear whether the source of misspecification is the pattern of loadings or the number

Figure 16.2 Simple Structure Specification Typical of Applications of Confirmatory Factor Analysis



of factors. As such, the evaluation of model fit is an evaluation of a diffuse hypothesis and, therefore, ambiguous in meaning (Rosnow & Rosenthal, 1988). This problem is exacerbated by the typical strategy of implementing a specification search while holding the number of factors constant. Such a search would primarily be limited to cross-loadings and correlations between uniquenesses. A preferable strategy would be to disentangle the two major aspects of specification—the number of factors and the pattern of loadings on those factors—and evaluate each in turn. A strategy for accomplishing this disentanglement is the unrestricted factor model (Jöreskog, 1977), which is equivalent to the exploratory factor model estimated by maximum likelihood, with the important exception that, once the number of factors has been determined in the unrestricted factor model, one can readily move to a focus on the pattern of loadings on those factors.

The unrestricted factor model was described by Jöreskog (1979) and has been incorporated by Mulaik and Millsap (2000) into a comprehensive approach to evaluating the fit of structural equation models. Despite the appeal of the model as a bridge between exploratory and confirmatory approaches to factor analysis, it is relatively unknown to researchers in the social sciences (for the three published applications of which we are aware, see Browne & Cudeck, 1993; Hox et al., 1999; Tepper & Hoyle, 1996). Although specification of the model is unusual, it is not difficult. Estimation of the highly parameterized model can be challenging; however, we later illustrate a strategy that virtually always results in a proper solution.

The unrestricted factor model is specified as follows:

1. For each of the k factors, one of the p indicators is specified to load only on that factor; the remaining loadings for these marker variables are set to zero. The remaining $p - k$ indicators are left free to load on every factor.
2. The variances of the factors are fixed to unity, and the covariances between factors are estimated from the data.

Alternatively, the variances of the factors can be estimated from the data and the metric of the factors established by fixing a loading on each of the k factors to unity (Mulaik & Millsap, 2000). For a correctly specified model, k^2 parameters are fixed and degrees of freedom equal $[p(p+1)/2] - [(pk - k^2) + k(k+1)/2]$. When specified as described, the model will be identified and the solution unique (Howe, 1955).

Successful estimation of the unrestricted factor model requires certain knowledge about the indicators and their relation to the factors. Specifically, the marker variable for each factor should be the highest loading indicator on that factor, and the starting values for the free parameters, particularly the loadings, should be reasonably close to the final estimates. This requirement of considerable a priori knowledge about the model has been a target of criticism (e.g., Hayduk & Glaser, 2000) because the most straightforward means of acquiring the necessary knowledge is to submit the data to an exploratory factor analysis in which the number of factors is extracted that will be specified in the unrestricted model. And given that a maximum likelihood exploratory factor analysis implicitly is specified exactly as described above, it would not seem necessary to estimate the model using confirmatory factor analysis.

The benefit of estimating the model using confirmatory factor analysis is apparent when the model is considered in the larger context of a framework for evaluating structural equation models. It is widely acknowledged that, within such models, a distinction can (and perhaps should) be drawn between the measurement model, which concerns the association between indicators and latent variables, and the structural model, which concerns the association among latent variables (Anderson & Gerbing, 1988; Herting & Costner, 2000). The fit of the measurement model sets an upper bound for the fit of the full model, as the

latter involves adding restrictions to the former (i.e., they are nested models). As such, it is essential that the measurement portion of a structural equation model be properly specified if the full model is to receive support or, in the event of poor fit of the full model, the source of misspecification correctly diagnosed.

Mulaik and Millsap (2000) advocate expanding the two-step model testing sequence to four steps. The second and third steps in their nested sequence correspond to the standard steps just described, and the fourth step involves moving from the fit of the full model to hypothesis tests about specific free parameters within the model. Their first step is the unrestricted model described here, and its inclusion is based on the same logic that gave rise to the distinction between measurement and structural models. Specifically, if, within the measurement model, the wrong number of factors is specified, then the model is sure not to fit when restrictions to the loadings are imposed, and an evaluation of the pattern of loadings is premature. As such, it is necessary to pin down the correct number of factors before moving to a full-blown evaluation of the measurement model.[8]

For our purposes, the unrestricted model need not be viewed as part of Mulaik and Millsap's (2000) four-step nested sequence because we are interested in the measurement model as an end in itself. So we are, in effect, arguing for a two-step approach to fitting common factor models within confirmatory factor analysis when there is not an unequivocal basis for specifying a particular number of factors and pattern of loadings. The first step involves evaluating the fit of a series of unrestricted models that specify plausible numbers of factors. In this regard, the strategy does not differ from the standard implementation of maximum likelihood exploratory factor analysis. The second step involves imposing restrictions on the loadings using either a theoretical or empirical rationale. The fit of a given model at the first step establishes a ceiling for the fit of restricted models and, therefore, should be near perfect. Models at the second step are nested in their unrestricted counterpart and can be evaluated using both chi-square differences and absolute fit indices, such as the comparative fit index (Bentler, 1990a) or root mean square error of approximation (Steiger, 1990).

8. Although the unrestricted factor model is an appropriate test of the number of factors for the overwhelming majority of applications in the social sciences, there clearly are factor models for which it is not appropriate (Bollen, 2000; Hayduk & Glaser, 2000). Such examples include simplex models, models with correlated errors, and models in which subfactors are expected.

16.2.6. Summary

We have described and offered commentary on the performance of five strategies for addressing the number-of-factors question in applications of factor analysis. Although the Kaiser-Guttman rule and the scree plot are the most popular among researchers in the social sciences and are easily implemented using standard statistical software, empirical evaluations using variables for which the number of common factors is known indicate that they rarely result in a correct inference regarding how many factors to retain and interpret. Like the K-G rule and the scree plot, parallel analysis focuses on the eigenvalues of the correlation matrix, but it does so in a formal manner using statistical criteria. Empirical evaluations of parallel analysis indicate excellent performance under conditions typical of social science research, particularly when Horn's (1965) original implementation of the strategy is adjusted by using the 95th rather than the 50th percentile random-data eigenvalues at each serial position as criteria. Maximum likelihood exploratory factor analysis and the unrestricted factor model are equivalent strategies that provide a statistical test of whether a specific number of factors is sufficient given the observed data. The unrestricted factor model is estimated within the confirmatory factor analysis context and has the benefit of allowing restrictions to the pattern of loadings once the number of factors has been established.

In the remainder of the chapter, we illustrate parallel analysis, maximum likelihood exploratory factor analysis, and the unrestricted factor model using responses to items on a self-report measure designed to be multidimensional. These data are particularly appealing for this purpose because, although the measure was written to reflect three dimensions, there is ample evidence to suggest that it reflects four or five dimensions.

16.3. Example: Dimensionality of Self-Consciousness

The Self-Consciousness Scale (Fenigstein, Scheier, & Buss, 1975) is a widely used self-report measure of the tendency to direct attention toward the self. The 23-item measure comprises three subscales that reflect different manifestations of self-attention. The private self-consciousness items tap the tendency to focus on internal thoughts and feelings. The public self-consciousness items capture the tendency to focus on

oneself as a social object—that is, to see oneself from the imagined perspective of others. The social anxiety subscale measures the tendency to experience anxiety in the presence of others as a result of heightened public self-consciousness.

The numerous factor analyses of responses to these items have produced conflicting results regarding the correspondence of the structure evident in responses with the hypothesized tripartite structure. Fenigstein et al. (1975), as part of their original presentation of the measure, interpreted the pattern of loadings on three orthogonally rotated principal components. Scheier and Carver (1985) replicated the pattern of loadings using principal factors extraction followed by an orthogonal rotation. Burnkrant and Page (1984) used confirmatory factor analysis to evaluate the dimensionality of each subscale in isolation of the others before fitting a model to the entire set of items. They obtained good support for unidimensional models of public self-consciousness and social anxiety but clear evidence that the private self-consciousness items reflect two factors. They concluded that four, not three, factors underlie responses to the Self-Consciousness Scale. A drawback to their analysis and conclusions is that they advocated and fit models that excluded 5 of the 23 items; their recommendation that the five items be eliminated from the scale has not been accepted by researchers who use the measure. Mittal and Balasubramanian (1987) used manual iteration of communalities from maximum likelihood estimation to evaluate the dimensionality of the three originally prescribed subscales. They replicated the finding that the private self-consciousness items break into two factors and found that the public self-consciousness items break into two factors as well. A unidimensional model of the social anxiety items was tenable only after dropping two items.

A surprising feature of these analyses, as well as other factor analyses, of responses to the Self-Consciousness Scale (e.g., Britt, 1992; Piliavin & Charng, 1988) is that none of them employed any of the strategies for determining the number of factors reviewed in this chapter (including the widely used K-G rule and scree plot). Each began with the original three-factor structure as the assumed model and searched for ways to adjust that model to better account for observed data. None of the analyses that used maximum likelihood estimation obtained values of fit statistics that would support incorporating their measurement model into a model with structural paths. In each case, values of chi-square were very large relative to degrees of freedom, and in no case did the value of the standardized index used to evaluate fit reach the typical minimum criterion of .90.

As noted in the previous section, when evaluated in the context of a full structural equation model, a measurement model must fit exceptionally well because restrictions placed on the model to evaluate directional associations between factors will certainly lead to a decrement in fit. In other words, the fit of the measurement model sets the ceiling for fit of the full model of which it is part. Moreover, exceptional fit of a measurement model is predicated on specification of the correct number of factors. Hence, if the Self-Consciousness Scale is to ever be included in a structural equation model with latent variables, it is essential that the correct number of factors and the correct pattern of loadings on those factors be established.

16.3.1. Parallel Analysis

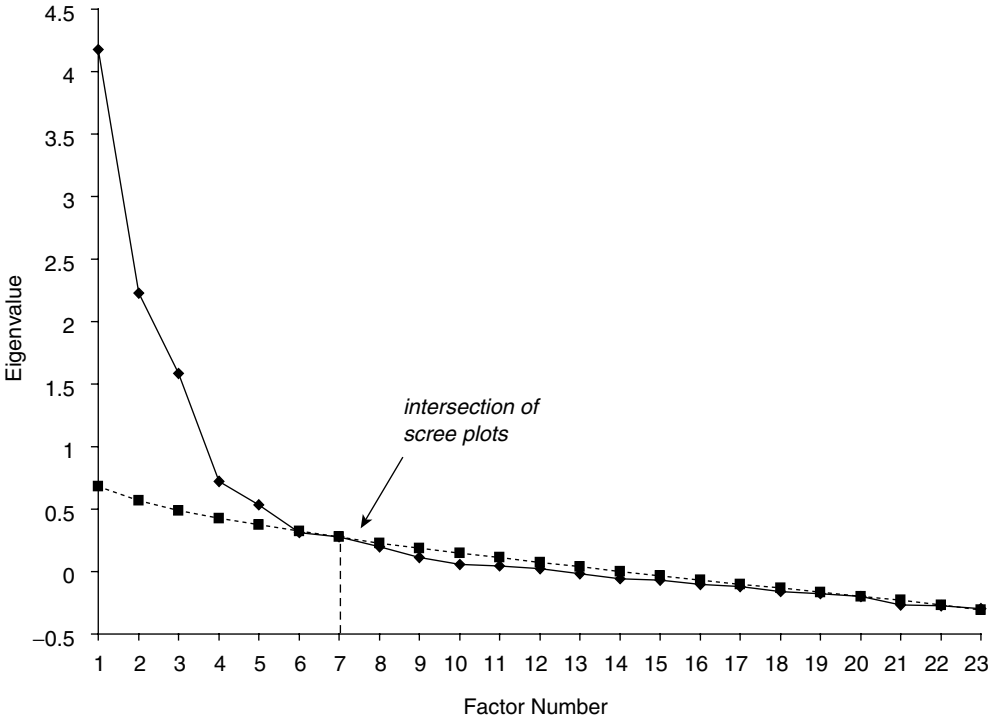
As detailed earlier, parallel analysis involves comparing eigenvalues from a factor analysis of a set of observed data with eigenvalues from a factor analysis of sets of random data comprising the same number of variables and observations as the observed data set. Only factors whose eigenvalue exceeds, at some predetermined probability, the corresponding random-data eigenvalue are retained for interpretation. Although relatively straightforward regression-based procedures for estimating random-data eigenvalues have been developed, these are not precise for most applications. Moreover, the more flexible and precise simulation approach to generating random-data eigenvalues is not difficult to implement on contemporary desktop computers using programs that run under widely accessible statistical software.

For the current analysis, we used O'Connor's (2000) SAS program for simulating data sets using the raw observed data as input.[9] We generated 500 random permutations of the raw data, an approach that preserves the distributional properties of the original data in the random data sets. We used as a criterion the 95th percentile estimate of the random-data eigenvalue for each factor. In other words, we retained all factors whose eigenvalue exceeded the corresponding random-data eigenvalue to a degree unlikely to occur by chance.

The two sets of eigenvalues are overlaid in the scree plot shown in Figure 16.3. As indicated in the figure, the two lines intersect between the seventh and eighth

9. The program we used, as well as related programs that run under SAS, SPSS, or MATLAB, are available for download from the Web at http://flash.lakeheadu.ca/~boconn02/nfactors.html. The raw observed data can be obtained from Rick Hoyle (rhoyle@duke.edu).

Figure 16.3 Plot of Eigenvalues in Serial Position for Observed Data (Solid Line) and 95th Percentile Eigenvalues From Distribution of 500 Random Permutations of the Raw Data (Broken Line)



eigenvalues, indicating that, beyond seven factors, the eigenvalues of the observed correlation matrix are not significantly greater than the values that would be expected were there no common factors at all. Hence, from the results of parallel analysis, we conclude that the correct number of factors is seven.

16.3.2. Maximum Likelihood Strategies

As we have established, maximum likelihood exploratory factor analysis and the unrestricted factor model are one and the same. As such, we present the statistical results that would be generated by either strategy, followed by additional analyses involving the pattern of loadings using the unrestricted factor model as a starting point.

Whether one uses exploratory factor analysis or formally specifies the unrestricted model in a confirmatory factor analysis, a separate analysis must be run for each number of factors to be considered. As parsimony is desirable in factor models, it is typical to begin by estimating a model with a single factor and, if necessary, continue sequentially until a particular number of factors can be justified statistically. Statistical justification comes in two forms. The traditional approach to hypothesis testing in the maximum likelihood context is a test statistic that, when certain assumptions are met, is distributed as a chi-square. If this test is used to justify the decision to retain a specific number of factors, the criterion is a p-value greater than .05, signifying a failure to reject the null hypothesis that the covariance matrix implied by the model is equivalent to the observed covariance matrix. An alternative approach is to consult one or more of a growing number of alternative indexes of fit. We advocate Bentler's (1990a) comparative fit index (CFI) and Steiger's (1990) root mean square error of approximation (RMSEA).[10] CFI indexes the proportionate

10. These fit indices are provided by all major software programs for estimating structural equation models (e.g., EQS, LISREL, AMOS). If the models are estimated using maximum likelihood exploratory factor analysis, the fit indices we reported, as well as several others, can be obtained using a SAS macro written by Steven Gregorich. The macro, which requires as input the number of variables, degrees of freedom, and chi-square for each model of interest, is available for download from the Web at http://mywebpage.netscape.com/segregorich/model_selection.html.

Table 16.1 Fit Statistics for Models Positing Zero to Eight Factors Underlying Responses to the Self-Consciousness Scale

k	df	F_{ML}	χ^2	p	*CFI*	*RMSEA*	$RMSEA_{.05}$	$RMSEA_{.95}$
0	253	6.868	2238.85	.00001	.000	.155	.150	.161
1	230	4.041	1317.24	.00001	.453	.120	.114	.127
2	208	2.595	845.91	.00001	.679	.097	.090	.104
3	187	1.392	453.78	.00001	.866	.066	.058	.074
4	167	1.038	338.41	.00001	.914	.056	.047	.065
5	148	0.807	263.07	.00001	.942	.049	.039	.025
6	130	0.657	211.54	.00001	*.959*	.044	.033	.054
7	113	0.508	165.68	.00092	*.973*	.038	.025	*.050*
8	97	0.365	119.10	*.06341*	*.989*	.026	.000	*.041*

NOTE: $N = 327$; k = number of factors; F_{ML} = maximum likelihood fitting function; CFI = comparative fit index; RMSEA = root mean square error of approximation (with 90% confidence limits). Consistent with common use in confirmatory factor analysis, chi-square values do not reflect Bartlett's correction. Italicized values indicate statistical support for the model they accompany.

improvement in fit of a specified model over a model that specifies no commonality among the indicators—the null, or independence, model. Following Mulaik and Millsap's (2000) recommendation, a value of .95 or greater would provide justification for a particular measurement model. RMSEA indexes the discrepancy between the observed covariance matrix and the covariance matrix implied by the model per degree of freedom. A value of zero indicates no discrepancy and, therefore, a perfect fit of the model to the data. It is now commonplace to put a 90% confidence interval around the point estimate of RMSEA. Although .08 typically is acceptable as the maximum value of the upper limit, the goal of a superior fit of the unrestricted factor model suggests that .05 is a more appropriate maximum for the upper limit of the confidence interval (Browne & Cudeck, 1993).

For the confirmatory factor analyses, models were specified by using output from an exploratory factor analysis to determine the marker variable (i.e., highest loading indicator) for each factor. As noted earlier, for these variables, only the loading on the designated factor was estimated; the remaining loadings on the factor were fixed at zero. All other loadings were free to be estimated. In addition, the variances of the factors were fixed at unity, and the covariances between factors were free to be estimated. Loadings and interfactor correlations from exploratory factor analyses were used as starting values.

Results from maximum likelihood estimation of a series of unrestricted models of the Self-Consciousness Scale are summarized in Table 16.1. The first model, for which $k = 0$, serves as the comparison for computation of CFI. It also can be viewed as a test for commonality among the indicators, and the highly undesirable values of the various indices of fit clearly indicate that at least one common factor underlies the covariance matrix. As indicated by the italicized values in the table, the results point to seven, perhaps eight, factors. Although CFI exceeds .95 for the six-factor solution, neither the chi-square test nor RMSEA support it. The RMSEA falls at the criterion, and CFI well exceeds the criterion for the seven-factor solution; however, the chi-square remains highly significant. A model with eight factors yields statistical support on all criteria.

The parallel analysis and two widely used fit indices support the interpretation of seven factors. The traditional chi-square test suggests the need for an additional factor. The chi-square test requires stringent, perhaps unrealistic, assumptions about the data and the model, and it is a somewhat unrealistic test, for it is a test of whether the model holds exactly in the population (Browne, 1984). Nonetheless, we examined parameter estimates for the eight-factor solution rather than dismiss the chi-square test out of hand.

Upon initial estimation, both the seven- and eight-factor solutions produced Heywood cases—apparent negative uniquenesses for Items 8 and 14. When the values for the suspect uniquenesses were permitted to dip below zero during iteration, both models yielded proper solutions. The final estimate for the Item 8 uniqueness in both solutions was negative but nonsignificant ($p > .60$). The final estimate in both solutions for the Item 14 uniqueness was greater than, but not significantly different from, zero ($p > .50$). Thus, we were able to obtain proper solutions for both the seven- and eight-factor models by allowing estimates of uniqueness to dip below zero during the iterative estimation process (Chen, Bollen, Paxton, Curran, & Kirby, 2001).

Before attempting to interpret the solutions, we used the multivariate Wald test to identify loadings and inter-factor correlations that could be constrained to

zero without significant loss in fit relative to the gain in degrees of freedom. For the seven-factor model, 8 of 21 inter-factor covariances were nonsignificant and, therefore, fixed at zero. Sixty-two loadings were nonsignificant and constrained to zero, and one loading that had been fixed in the original specification was freed. This resulted in a net gain of 69 degrees of freedom moving from the unrestricted model to a restricted model. The fit of this model exceeded our criteria for CFI and RMSEA, but the chi-square was significant: $\chi^2(182, N = 327) = 235.13$, $p = .004$, CFI = .973, RMSEA = .030(.017, .040). Importantly, the many zero constraints placed on parameters that had been free in the unrestricted model did not result in a decline in fit, $\Delta\chi^2(69, N = 327) = 69.45$, $p = .46$.

Moving to the parameter estimates, the solution indicated considerable factorial complexity in the item set. Of the 23 items, only 6 loaded on a single factor. Most items loaded significantly on two or three factors, although it is important to note that loadings as low as .15 were statistically significant. If standard saliency criteria are used (e.g., .30 or .40), factorial complexity diminishes, although the pattern still does not manifest simple structure. Consistent with documented attempts at factoring the measure, the private and public self-consciousness item sets each form two factors. The social anxiety items form a relatively coherent factor, although a subset of those items coalesces with a subset of the public self-consciousness factor to form a separate factor. Finally, Items 3 and 9, which typically fail to load in analyses that extract three or four factors, are the strongest indicators of a seventh factor.

Our suspicion that the eight-factor solution would constitute an overextraction was supported when we followed the same strategy as for the seven-factor solution in moving from the unrestricted to a restricted model. One factor (the fifth factor extracted in the maximum likelihood exploratory factor analysis) was represented by a single indicator; therefore, the restricted eight-factor model was underidentified and could not be estimated. This result underscores our reservations about using the chi-square test in either exploratory or confirmatory factor analysis for determining the number of factors to interpret. Unlike indices such as CFI and RMSEA, the chi-square test, even when assumptions regarding the data and model ensure that the statistic is in fact distributed as a chi-square, is a test of exact fit. This unrealistic hypothesis test, if taken seriously, is very likely to lead to overextraction, as illustrated by our eight-factor unrestricted model, the only model that was supported by the chi-square test.

16.4. Summary and Conclusions

We described and illustrated three formal approaches to determining the number of factors that underlie a set of variables. Parallel analysis focuses on the eigenvalues of the correlation matrix, a focus familiar to researchers accustomed to invoking the Kaiser-Guttman rule or using the scree test. Parallel analysis requires additional effort because the researcher must generate a corresponding set of random-data eigenvalues; however, readily available computer programs render this activity rather straightforward. More important, unlike the K-G rule and the scree test, for which there is scant empirical support, parallel analysis is virtually always accurate under conditions typical of social science research.

Maximum likelihood exploratory factor analysis and the unrestricted factor model, as specified in confirmatory factor analysis, represent two implementations of the same statistical model. In each case, a formal statistical test can be undertaken of the adequacy of particular numbers of factors to account for the associations in an observed correlation matrix. The default test in both models is a chi-square test; however, our recommendation is that researchers eschew this unrealistically stringent hypothesis test and, instead, take advantage of recently developed fit indexes such as the CFI and RMSEA. Because factor models estimated using confirmatory factor analysis often are incorporated into structural models including directional associations among factors, we argue that such models, particularly at the stage when the number-of-factors question is being considered, be evaluated against fit criteria that are more stringent than is typical of tests of structural models (e.g., CFI $>$ .95; upper limit of RMSEA confidence interval $<$ .05).

A by-product of our analysis is the illustration, in extreme form, that confirmatory factor models need not be specified as simple structure models. Indeed, simple structure specification, such as the model illustrated in Figure 16.2, is unlikely to hold for items typical of measures in the social sciences. For this reason, it seems unreasonable for researchers to rigidly adhere to this standard specification, leaving only covariances among uniqueness terms open to specification searching in misspecified models. And, as we have asserted, the consideration of cross-loadings (or any pattern of loadings) is reasonable only after the correct number of factors has been determined.

Factor analysis is a vital statistical tool for social scientists. When implemented and interpreted correctly, factor analysis provides a means of

operationally defining variables that cannot be measured directly. At the core of factor analysis is the question of how many factors underlie a given set of indicators. In some instances, this question is rendered moot by a detailed theoretical model or careful construction of indicators with a clear factor structure in mind. More often, however, the indicators predate the research of which they are part and are not closely tied to a detailed theoretical model. In such cases, the strategies described and illustrated in this chapter provide feasible and defensible means to determine the correct number of factors.

REFERENCES

Allan, S. J., & Hubbard, R. (1986). Regression equations for the latent roots of random data correlation matrices with unities on the diagonal. *Multivariate Behavioral Research, 21,* 393–398.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411–423.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London, Series A, 160,* 268–282.

Bentler, P. M. (1990a). Comparative fit indexes in structural models. *Psychological Bulletin, 88,* 588–606.

Bentler, P. M. (1990b). Latent variable structural models for separating specific from general effects. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), *Research methodology: Strengthening causal interpretations of nonexperimental data* (pp. 61–83). Rockville, MD: Department of Health and Human Services.

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

Bollen, K. A. (2000). Modeling strategies: In search of the holy grail. *Structural Equation Modeling, 7,* 74–81.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110,* 305–314.

Britt, T. W. (1992). The Self-Consciousness Scale: On the stability of the three-factor structure. *Personality and Social Psychology Bulletin, 18,* 748–755.

Browne, M. W. (1968). A comparison of factor analytic techniques. *Psychometrika, 33,* 267–334.

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in multivariate analysis* (pp. 72–141). Cambridge, UK: Cambridge University Press.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62–83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 132–162). Thousand Oaks, CA: Sage.

Burnkrant, R. E., & Page, T. J., Jr. (1984). A modification of the Fenigstein, Scheier, and Buss Self-Consciousness Scales. *Journal of Personality Assessment, 48,* 629–637.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 245–276.

Cattell, R. B., & Vogelman, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research, 12,* 289–325.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research, 29,* 468–508.

Cohen, P., Cohen, J., Teresi, J., Marchi, M., & Velez, C. N. (1990). Problems in the measurement of latent variables in structural equation causal models. *Applied Psychological Measurement, 14,* 183–196.

Crawford, C. G., & Koopman, P. (1979). Note: Inter-rater reliability of scree test and mean square ratio test of number of factors. *Perceptual and Motor Skills, 49,* 223–226.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4,* 272–299.

Fenigstein, A., Scheier, M. F., & Buss, A. H. (1975). Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology, 43,* 522–527.

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement, 55,* 377–393.

Gorsuch, R. L. (1974). *Factor analysis.* Philadelphia: Saunders.

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika, 19,* 149–162.

Harshman, R. A., & Reddon, J. R. (1983, May). *Determining the number of factors by comparing real with random data: A serious flaw and some possible corrections.* Paper presented at the annual meeting of the Classification Society of North America, Philadelphia.

Hayduk, L. A., & Glaser, D. N. (2000). Jiving the four-step, waltzing around factor analysis, and other serious fun. *Structural Equation Modeling, 7,* 1–35.

Herting, J. R., & Costner, H. L. (2000). Another perspective on "the proper number of factors" and the appropriate number of steps. *Structural Equation Modeling, 7,* 92–110.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179–185.

Howe, W. G. (1955). *Some contributions to factor analysis* (Rep. No. ORNL-1919). Oak Ridge, TN: Oak Ridge National Laboratory.

Hox, J., Auerbach, J., Erol, N., Fonseca, A. C., Mellenbergh, G. J., Nøvik, T. S., et al. (1999). Syndrome dimensions of the Child Behavior Checklist and the Teacher Report Form: A critical empirical evaluation. *Journal of Child Psychology and Psychiatry, 40,* 1095–1116.

Hoyle, R. H. (2000). Confirmatory factor analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 465–497). New York: Academic Press.

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112,* 351–362.

Humphreys, L. G., & Montanelli, R. G., Jr. (1975). An examination of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research, 10,* 193–206.

Jackson, D. N., & Chan, D. W. (1980). Maximum likelihood estimation in common factor analysis: A cautionary note. *Psychological Bulletin, 88,* 502–508.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183–202.

Jöreskog, K. G. (1979). Author's addendum. In J. Magidson (Ed.), *Advances in factor analysis and structural equation models* (pp. 40–43). Cambridge, MA: Abt.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20,* 141–151.

Kaufman, J. D., & Dunlap, W. P. (2000). Determining the number of factors to retain: Windows-based FORTRAN-IMSL program for parallel analysis. *Behavior Research Methods, Instruments, & Computers, 32,* 389–395.

Lautenschlager, G. J. (1989). A comparison of alternatives to conducting Monte Carlo analyses for determining parallel analysis criteria. *Multivariate Behavioral Research, 24,* 365–395.

Lautenschlager, G. J., Lance, C. E., & Flaherty, V. L. (1989). Parallel analysis criteria: Revised equations for estimating the latent roots of random correlation matrices. *Educational and Psychological Measurement, 49,* 339–345.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (2nd ed.). London: Butterworths.

Lee, H. B., & Comrey, A. L. (1979). Distortions in a commonly used factor analytic procedure. *Multivariate Behavioral Research, 14,* 301–321.

Longman, R. S., Cota, A. A., Holden, R. R., & Fekken, G. C. (1989a). PAM: A double-precision FORTRAN routine for the parallel analysis method in principal components analysis. *Behavior Research Methods, Instruments, & Computers, 21,* 477–480.

Longman, R. S., Cota, A. A., Holden, R. R., & Fekken, G. C. (1989b). A regression equation for the parallel analysis criterion in principal components analysis: Mean and 95th percentile eigenvalues. *Multivariate Behavioral Research, 24,* 59–69.

Mittal, B., & Balasubramanian, S. K. (1987). Testing the dimensionality of the Self-Consciousness Scales. *Journal of Personality Assessment, 51,* 53–68.

Montanelli, R. G., Jr. (1975). A computer program to generate sample correlation and covariance matrices. *Educational and Psychological Measurement, 35,* 195–197.

Montanelli, R. G., Jr., & Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A Monte Carlo study. *Psychometrika, 41,* 341–348.

Mulaik, S. A., & Millsap, R. E. (2000). Doing the four-step right. *Structural Equation Modeling, 7,* 36–73.

O'Connor, B. P. (2000). SPSS, SAS, and MATLAB programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers, 32,* 396–402.

Piliavin, J. A., & Charng, H.-W. (1988). What *is* the factorial structure of the private and public self-consciousness scales? *Personality and Social Psychology Bulletin, 14,* 587–595.

Rosnow, R. L., & Rosenthal, R. (1988). Focused tests of significance and effect size estimation in counseling psychology. *Journal of Counseling Psychology, 38,* 203–208.

Scheier, M. G., & Carver, C. S. (1985). The Self-Consciousness Scale: A revised version for use with general populations. *Journal of Applied Social Psychology, 15,* 687–699.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173–180.

Steiger, J. H. (1996). Dispelling some myths about factor indeterminacy. *Multivariate Behavioral Research, 31,* 539–550.

Steiger, J. H., & Schönemann, P. H. (1978). A history of factor indeterminacy. In S. Shye (Ed.), *Theory construction and data analysis* (pp. 136–178). San Francisco: Jossey-Bass.

Tepper, K., & Hoyle, R. H. (1996). Latent variable models of need for uniqueness. *Multivariate Behavioral Research, 31,* 467–494.

Turner, N. E. (1998). The effect of common variance and structure pattern on random data eigenvalues: Implications for the accuracy of parallel analysis. *Educational and Psychological Measurement, 58,* 541–568.

Yoemans, K. A., & Golder, P. A. (1982). The Guttman-Kaiser criterion as a predictor of the number of common factors. *The Statistician, 31,* 221–229.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432–442.

Chapter 17

Experimental, Quasi-Experimental, and Nonexperimental Design and Analysis With Latent Variables

Gregory R. Hancock

In the physical sciences, common variables such as temperature, pressure, mass, and volume, when considered in sufficient quantities, tend to be measured with relatively little error. In the social sciences, however, variables often contain fairly large amounts of measurement error. Educational policy researchers, for example, might want to know about teachers' feelings of burnout but only have measures of absenteeism and job satisfaction. Family studies researchers could be interested in maternal warmth for new mothers but only have mothers' responses to a few rating scale items regarding their interactions with their newborn infants. Health care researchers might want to know about AIDS patients' sense of hopelessness while in group therapy but only have measures of patients' medication and treatment compliance. Such is the nature of the social sciences—constructs of interest such as burnout, maternal warmth, or hopelessness are generally latent, so our analyses must rely on error-laden measured variables as surrogates.

Experimental design and analysis involving variables that are directly measured is well established. Analysis of variance (ANOVA) in its many forms constitutes the basis for inference regarding population means on some measured dependent variable as a function of one or more independent grouping variables (e.g., treatment group, sex). Such analyses strive to facilitate inferences about the construct underlying the measured dependent variable, but the sensitivity of those analyses for detecting construct-level relations is at the mercy of the reliability of the construct's operationalization in the measured variable employed. In short, unreliability of measured variables presents a signal-to-noise ratio problem within experimental design and analysis. Within the social sciences, detection of subtle but important population differences in

AUTHOR'S NOTE: Portions of this chapter appeared previously in Hancock (1997) and are used here with permission of the American Counseling Association.

an experiment, quasi-experiment, or nonexperiment might be thwarted by the imprecision arising from the inherent vagaries of human behavior, possibly leading to proclamations of no apparent population differences only because those that truly existed were masked by the outcome measure's unreliability.

This attenuation in sensitivity to detect specific population mean differences was described by Cohen (1988, p. 536). Drawing from Cleary and Linn (1969), he noted that for a measured variable, the standardized effect size ES equals $(ES^*)(\rho_{YY})^{1/2}$, where ES^* is the error-free (i.e., latent) standardized effect size measure (d for two groups or f for J groups) (Cohen, 1988), and ρ_{YY} is the reliability of a single measured variable Y. The direct implication is that any latent variability among population construct means is attenuated by the reliability of the measured indicator selected. Borrowing from classical test theory (see, e.g., Crocker & Algina, 1986), the ith score on a measured variable Y in the jth population may be expressed as $Y_{ij} = T_{ij} + E_{ij}$, where T is the hypothetical true score, and E is a random fluctuation from that true value that would be expected to increase in absolute magnitude when a less reliable Y variable is chosen to operationalize the construct of interest. Within the jth population, this implies (under standard assumed conditions) that $\mu_{Y_j} = \mu_{T_j}$ and $\sigma^2_{Y_j} = \sigma^2_{T_j} + \sigma^2_{E_j}$, with the latter facilitating the common definition of *reliability* as the proportion of score variance explained by the true underlying construct: $\rho_{YY} = \sigma^2_T/\sigma^2_Y$. Note that the term $\sigma^2_{Y_j}$ represents the assumed homogeneous intrapopulation (*within-groups*) variability, which appears below as $\sigma^2_{Y_{\text{within}}}$ in the expectations for the numerator and denominator for the one-way, between-subjects, fixed-effects ANOVA F-test statistic. For the equal-n case, these are

$$E[MS_{\text{between}}] = n\left[\frac{\sum(\mu_{Y_j} - \mu_{Y.})^2}{J-1}\right] + \sigma^2_{Y_{\text{within}}} \quad (1)$$

and

$$E[MS_{\text{within}}] = \sigma^2_{Y_{\text{within}}}. \quad (2)$$

Substituting the true and error score information leads to

$$E[MS_{\text{between}}] = n\left[\frac{\sum(\mu_{T_j} - \mu_{T.})^2}{J-1}\right] + \sigma^2_{T_{\text{within}}} + \sigma^2_{E_{\text{within}}} \quad (3)$$

and

$$E[MS_{\text{within}}] = \sigma^2_{T_{\text{within}}} + \sigma^2_{E_{\text{within}}}. \quad (4)$$

In equation (3), for $E[MS_{\text{between}}]$, the first term represents the signal, whereas the last two variance terms represent noise. The $\sigma^2_{T_{\text{within}}}$ term might be considered *true noise,* the natural variability of subjects on the underlying continuum of interest. The $\sigma^2_{E_{\text{within}}}$ term, on the other hand, is *measurement error noise,* induced by the unreliability of the dependent variable. Thus, when forming the observed F-ratio as $MS_{\text{between}}/MS_{\text{within}}$, the measurement error noise dampens the numerator and denominator, potentially masking the variability among true population means (the signal) contained in the first term of the $E[MS_{\text{between}}]$ expression in equation (3).

Two strategies to combat the dampening problem arising from $\sigma^2_{E_{\text{within}}}$ are, put simply, to boost the signal and to reduce the noise. With regard to the signal, the group differences would somehow need to be inflated (implying some change in the nature of the independent variable and, hence, the research question) and/or the sample size must be increased (often a costly alternative). On the other hand, somehow addressing the noise of unreliability would attempt to facilitate as closely as possible a test of population differences on the construct itself. For example, if the dependent variable is a summated scale instrument, perhaps it could be lengthened through the addition of quality items, thereby enhancing the variable's (i.e., total score's) representation of the construct. Alternatively, if the reliability of the instrument is known, one could build some form of correction for attenuation into the numerator and denominator of the F-ratio. Unfortunately, the reliability estimate is itself a statistic with a sampling distribution (see, e.g., Hakstian & Whalen, 1976), rendering any such corrections somewhat tenuous. Most promising, and the subject of the current chapter, are two approaches deriving from structural equation modeling (SEM). Multiple-indicator, multiple-cause (MIMIC) modeling (Jöreskog & Goldberger, 1975; Muthén, 1989) and structured means modeling (SMM) (Sörbom, 1974) employ multiple measured variables, rather than one alone, in a sort of "triangulation" in which their patterns of covariances (and means) are used to infer population differences on the underlying construct believed to have motivated those observed patterns. Although not fitting the F-ratio paradigm precisely, these methods effectively facilitate hypothesis testing of population means directly at the construct level, rather than at the level of the fallibly measured

proxies for those constructs, thereby attempting to pare away the measurement error noise from the testing process.

Given that the SEM strategies addressed in this chapter invoke multiple measures, it is critical to note that such methods differ fundamentally from another multivariate group comparison strategy, multivariate analysis of variance (MANOVA). This difference is rooted first and foremost in the nature of the assumed *variable system* (Bollen & Lennox, 1991; Cohen, Cohen, Teresi, Marchi, & Velez, 1990; Cole, Maxwell, Arvey, & Salas, 1993). In an *emergent* variable system, the variables are believed to have a causal bearing on the underlying trait of interest. As an example, one could imagine an unmeasured entity representing stress that is related to such variables as relationship with parents, relationship with spouse, and demands of the workplace. In this case, it is more reasonable to posit that changes in the variables lead to changes in stress, rather than changes in overall stress leading to changes in these individual variables. Because the measured variables are theoretically the causal agents, stress emerges as a linear composite of those observed variables upon which it is dependent. In such an emergent variable system, then, it is meaningful to talk about population differences in terms of a linear composite formed by the variable system. For this reason, population comparisons in an emergent variable system are best addressed using MANOVA, in which population differences are assessed using composites that maximally differentiate the groups in multivariate space.

In a *latent* variable system, on the other hand, the construct (or "factor" or "latent variable") is believed to have a causal influence on the observed variables, thereby necessitating the existence of covariance among those variables. As an example, consider measured variables that are Likert scale responses to the following questionnaire items: "I am comfortable with my child marrying a person of another race," "I believe schools should have students of all racial backgrounds," and "I would be comfortable if a family of a different race moved in next door to me." Here the construct is believed to have a causal bearing on the variables: Changes in a person's attitude would be expected to result in changes in responses to each of these questionnaire items. As a result, substantial covariance among these items should arise because their responses all derive from attitudes regarding race; in fact, the items might serve, at least in part, as measured indicators of an underlying construct of racial attitude. For answering research questions regarding population differences on such a construct, MANOVA methods are often used. However, they are less powerful than the latent variable SEM methods presented in this chapter, as demonstrated empirically by Hancock, Lawrence, and Nevitt (2000) across a broad spectrum of conditions and also as derived analytically by Kano (2001). Furthermore, and most important, MANOVA methods are fundamentally inconsistent with the nature of the variable system at hand, which may lead researchers to an inaccurate assessment of population differences on the construct of interest, if not to missing those differences altogether (Cole et al., 1993).

To sum, the SEM methods presented in this chapter help to address research questions about population differences in a latent variable system. Literally, one uses measured variable evidence to query whether populations' latent construct means differ. Do subjects randomly assigned to either group counseling sessions or individual counseling sessions differ in their overall perception of counselors? Do males and females differ in terms of mathematics self-efficacy? Do random samples of husbands and wives differ in the amount of spousal trust? MIMIC modeling and SMM techniques help to address such questions. In this chapter, an introduction to both methods is presented, using the two-group comparison scenario as a framework within which to describe their conceptual representations, unique underlying assumptions, and relative merits and limitations, as well as to point toward methodological extensions beyond the two-group case explicitly presented here.

17.1. Preliminary Information

From a pedagogical standpoint, it would be impossible to convey an understanding of the MIMIC modeling and SMM techniques without assuming some experience with the underlying principles and implementation of SEM (e.g., unstandardized path-tracing rules, parameter estimation techniques, etc.). The reader unfamiliar with such topics is referred to several excellent texts (e.g., Bollen, 1989; Hayduk, 1987; Kaplan, 2000; Kline, 1998; Loehlin, 1998; Mueller, 1996; Schumacker & Lomax, 1996). To assist the reader in the current chapter, we present an overview of the notation used next, as well as a brief reminder of some helpful algebraic relations.

17.1.1. Notation

Fairly traditional SEM notation is used throughout most of this chapter. Specifically, an independent

(exogenous) construct is denoted as ξ; the measured variables serving as indicators of this construct are each denoted as X. The path coefficient representing the impact of the construct ξ on each X (i.e., the unstandardized loading) is labeled with a λ_X. The variability in each X not explained by the construct ξ is a residual denoted as δ. As for dependent (endogenous) constructs, these are labeled as η; indicators of these dependent constructs are denoted as Y. The unstandardized path coefficient representing the impact of the construct η on each Y is labeled with a λ_Y, whereas residuals in each Y variable are denoted as ε. In this chapter, endogenous constructs (η) are either dependent on a group code ("dummy") variable or on an exogenous latent covariate. The group code variable is denoted as X, with its unstandardized path to the construct η being labeled with a γ. That portion of the construct η that is not explained by the group code variable X (and any latent covariates) is residual; a construct residual is labeled ζ. Finally, as will be introduced later, some structural equations will require intercept terms; these intercepts are designated by τ for measured variables and κ for latent variables.

Exceptions to traditional notation used in this chapter are as follows. The population mean, variance, and covariance are represented by $M(\_)$, $V(\_)$, and $C(\_,\_)$, respectively. As for estimates of those population parameters derived through the course of SEM, these are denoted by a circumflex (ˆ) atop the corresponding parameter symbol. An estimated unstandardized X loading, for example, would be designated as $\hat{\lambda}_X$; the estimated population mean, variance, and covariance would be denoted $\hat{M}(\_)$, $\hat{V}(\_)$, and $\hat{C}(\_,\_)$, respectively.

17.1.2. Useful Algebraic Relations

When conducting SEM methods, the researcher is actually attempting to solve a system of algebraic relations to get estimates of theoretically meaningful unknowns. Those relations to be solved involve the decomposition of population variances and covariances (and, in this chapter, means) into their component pieces, as implied by the theoretical structural model. The reader should have some familiarity with this process already. As a reminder, consider one possible example involving three independent variables (X_1, X_2, and X_3) and three dependent variables (Y_1, Y_2, and Y_3).[1] These dependent variables are theoretically related to the independent variables by the following familiar system of regression-type equations:

$$Y_1 = \tau_1 + \gamma_{11}X_1 + \zeta_1, \tag{5}$$

$$Y_2 = \tau_2 + \gamma_{21}X_1 + \zeta_2, \tag{6}$$

$$Y_3 = \tau_3 + \gamma_{32}X_2 + \gamma_{33}X_3 + \zeta_3. \tag{7}$$

In these equations, τ represents an intercept, γ represents a slope, and ζ represents an error ("residual" or "disturbance") variable. These can also be depicted in matrix form as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} \gamma_{11} & 0 & 0 \\ \gamma_{21} & 0 & 0 \\ 0 & \gamma_{32} & \gamma_{33} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \end{bmatrix} \tag{8}$$

or, symbolically, as

$$\mathbf{y} = \boldsymbol{\tau} + \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta}. \tag{9}$$

Assuming that the errors in $\boldsymbol{\zeta}$ do not covary with each other or with any independent variables in $\mathbf{x}$, unstandardized path-tracing rules (or the algebra of expected values) yield the following expected relations for the population:

$$M(Y_1) = \tau_1 + \gamma_{11}M(X_1), \tag{10}$$

$$V(Y_1) = \gamma_{11}^2 V(X_1) + V(\zeta_1), \tag{11}$$

$$C(Y_1, Y_2) = \gamma_{11}\gamma_{21}V(X_1), \tag{12}$$

$$V(Y_3) = \gamma_{32}^2 V(X_2) + \gamma_{33}^2 V(X_3) + 2\gamma_{32}\gamma_{33}C(X_2, X_3) + V(\zeta_3). \tag{13}$$

Certainly, more relations are implied as well; in fact, for the six measured X and Y variables, a total of 6 mean decompositions, 6 variance decompositions, and 15 covariance decompositions are possible. The individual decomposition relations in equations (10) through (13) were chosen because they illustrate many of the types of mean, variance, and covariance decompositions necessary for understanding the MIMIC modeling and SMM analyses presented in this chapter.

Finally, noteworthy in the above model-implied relations is that intercept terms do not appear in decompositions for variances or covariances. Whereas intercepts are commonly used in multiple regression to help capture the relation of the predictors' means to the mean of the criterion variable, in most SEM analyses, the intercepts are omitted from the structural equations because they are irrelevant to SEM's focus on variances and covariances. However, as will be shown later, these intercepts will need to be introduced into our structural equations when performing SMM.

1. Note that this use of X and Y notation differs from that described in the previous section and from that used in the MIMIC modeling and SMM portions of the chapter to follow.

17.2. MIMIC Modeling

17.2.1. Development

As the reader may recall, t-tests (and ANOVA in general) may be conducted in a regression model through the use of dummy or other group code variables. Dummy variables assume values of 0 or 1, respectively, indicating the absence or presence of some condition and are included as predictors in the linear model. This allows inferences to be made regarding population differences on a particular dependent variable. A similar idea may be used with a dependent construct, allowing questions to be answered about potential population differences on the construct of interest. This is the basis of the MIMIC modeling method of testing latent mean differences.

To elaborate, assume a researcher wishes to use two samples to make an inference about whether a difference exists between two population means on a latent construct η_1. In other words, η_1 may be dependent on the population to which a subject belongs. Given the context of a latent variable system, the construct η_1 is defined by the covariation among its measured indicators. Assume for this example that there are three such indicators—Y_1, Y_2, and Y_3—where Y_1 serves as the construct's scale indicator by fixing its loading to a value of 1. Assuming the Y variables are expressed as deviation scores, this measurement model may be expressed in matrix form as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ \lambda_{Y21} \\ \lambda_{Y31} \end{bmatrix} \eta_1 + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} \tag{14}$$

or, symbolically, as

$$\mathbf{y} = \mathbf{\Lambda}_Y \eta_1 + \boldsymbol{\varepsilon}. \tag{15}$$

Notice that the equations do not contain intercept terms. In fact, intercept terms could have been included in these equations; however, because MIMIC modeling (like most of SEM) focuses only on variances and covariances among measured variables, such intercepts are irrelevant to the necessary variance and covariance decompositions. Formally, the omission of intercept terms from structural equations implies zero values for those intercepts. Zero intercepts result when all variables have means of zero, such as when each score is expressed as a deviation from its variable mean. Thus, in MIMIC modeling, all variables are assumed to be scaled as deviation scores, thereby having zero means and intercepts. On the surface, this might seem to defeat the purpose of the MIMIC model—namely, the investigation of mean differences on a latent factor. However, because variables are scaled as deviations from their means in the combined sample rather than in each group separately, group differences on the measured variables and on the underlying construct are preserved.

As for the information about population membership, a dummy variable X_1 is constructed whose values represent the presence (e.g., $X_1 = 1$) and absence (e.g., $X_1 = 0$) of one of two conditions for each subject. To capture the potential relation between this dummy variable and the construct η_1, the data from both groups are combined into a single sample and the dependent construct is regressed on the dummy variable within a single structural model. The regression coefficient expressing the impact of the dummy variable X_1 on the construct η_1 is denoted by γ_{11}; this parameter is essential to the MIMIC analysis because a statistically significant impact of X_1 on η_1 allows the inference that populations differ in average amount of the underlying construct. That part of η_1 not explained by the dummy variable is captured by the disturbance term ζ_1. This structural model may be expressed as

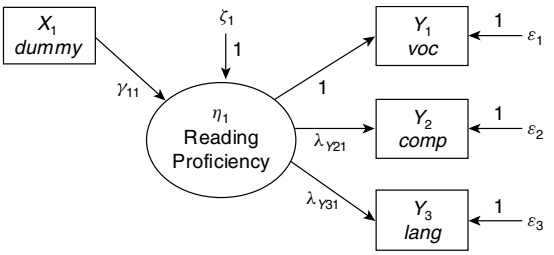
$$\eta_1 = \gamma_{11} X_1 + \zeta_1. \tag{16}$$

Notice again that no intercept term appears in this structural equation, reflecting the condition in which all variables, including the group code variable X_1, are treated as if they were expressed as deviation scores.

The model containing both measurement and structural portions is included in Figure 17.1 (variable labels may be ignored for now). Also shown are the model-implied variances and covariances among the four observed variables (i.e., including the dummy variable X_1), expressed as a function of the parameters to be estimated in the model and derivable from unstandardized path-tracing rules (or directly from the algebra of expected values). Consider, for example, the theoretical decomposition of the variance of Y_1, $V(Y_1) = V(\eta_1) + V(\varepsilon_1)$; however, because $V(\eta_1)$ may be further decomposed into $[\gamma_{11}^2 V(X_1) + V(\zeta_1)]$, the expression in brackets has been substituted into this relation and wherever $V(\eta_1)$ is part of a variance or covariance decomposition. In total, there are 10 equations and 8 unique parameters to be estimated: 2 loadings (λ_{Y21}, λ_{Y31}), 1 structural path (γ_{11}), 1 variable variance ($V(X_1)$), 1 disturbance variance ($V(\zeta_1)$), and 3 error variances ($V(\varepsilon_1)$, $V(\varepsilon_2)$, $V(\varepsilon_3)$). Overall, the model is overidentified with 2 ($10 - 8 = 2$) degrees of freedom.

Estimates of the model's eight parameters are made so as to reproduce the 10 model-implied variances and covariances in a 4×4 model-implied

Figure 17.1 Model and Implied Relations for the Multiple-Indicator, Multiple-Cause (MIMIC) Modeling Example



Model-implied relations

$V(Y_1) = [\gamma_{11}^2 V(X_1) + V(\zeta_1)] + V(\varepsilon_1)$
$V(Y_2) = \lambda_{Y21}^2[\gamma_{11}^2 V(X_1) + V(\zeta_1)] + V(\varepsilon_2)$
$V(Y_3) = \lambda_{Y31}^2[\gamma_{11}^2 V(X_1) + V(\zeta_1)] + V(\varepsilon_3)$
$V(X_1) = V(X_1)$
$C(Y_1,Y_2) = \lambda_{Y21}[\gamma_{11}^2 V(X_1) + V(\zeta_1)]$
$C(Y_1,Y_3) = \lambda_{Y31}[\gamma_{11}^2 V(X_1) + V(\zeta_1)]$
$C(Y_2,Y_3) = \lambda_{Y21}\lambda_{Y31}[\gamma_{11}^2 V(X_1) + V(\zeta_1)]$
$C(X_1,Y_1) = \gamma_{11} V(X_1)$
$C(X_1,Y_2) = \gamma_{11}\lambda_{Y21} V(X_1)$
$C(X_1,Y_3) = \gamma_{11}\lambda_{Y31} V(X_1)$

variance-covariance matrix $\hat{\boldsymbol{\Sigma}}$ as closely as possible to those 10 variance and covariance values observed in the $p = 4$ variables' data and contained in a 4×4 sample variance-covariance matrix **S.** In the context of maximum likelihood (ML) estimation specifically, parameter values are selected to imply (reproduce) a population variance-covariance matrix from which the observed data's sample variance-covariance matrix has the maximum likelihood of arising by random sampling. As derived elsewhere (e.g., Hayduk, 1987), this process is operationalized by choosing parameters so as to minimize the fit function

$$F_{\mathbf{ML}} = \ln|\hat{\boldsymbol{\Sigma}}| + \text{tr}(\mathbf{S}\hat{\boldsymbol{\Sigma}}^{-1}) - \ln|\mathbf{S}| - p. \qquad (17)$$

Assuming that reasonable data-model fit is achieved (i.e., that parameter estimates are found yielding a $\hat{\boldsymbol{\Sigma}}$ sufficiently close to **S**), the parameter of greatest interest is the path coefficient γ_{11} from the dummy variable X_1 to the construct η_1. This path represents the direct effect of the dummy variable on the construct, and its magnitude actually reflects the difference between the two population means on the construct η_1. To understand why, consider the structural equation relating the dummy variable X_1 to the following construct: $\eta_1 = \gamma_{11}X_1 + 1\zeta_1$. For the population coded $X_1 = 1$, this equation may be written as $\eta_1 = \gamma_{11} + 1\zeta_1$; for the population coded $X_1 = 0$, this equation simplifies to $\eta_1 = 1\zeta_1$. Because $M(\zeta_1) = 0$, the expected population mean for each group is ${}_1M(\eta_1) = \gamma_{11}$ and ${}_0M(\eta_1) = 0$ (where the prescript indicates the dummy code). Thus, $\gamma_{11} = {}_1M(\eta_1) - {}_0M(\eta_1)$, and its estimate $\hat{\gamma}_{11}$ represents the estimated difference between the two population means on the construct η_1.

If the dummy variable X_1 accounts for a statistically significant portion of the variability in the construct η_1, which is determined by a test of the magnitude of the parameter estimate $\hat{\gamma}_{11}$, then we would infer that a difference exists between the two population means on this construct. If, on the other hand, X_1 fails to account for a statistically significant portion of the variability in η_1, we must retain as tenable the null hypothesis stating ${}_1M(\eta_1) = {}_0M(\eta_1)$. The test of $\hat{\gamma}_{11}$ requires its associated standard error, designated here as $SE(\hat{\gamma}_{11})$, which is part of the typical SEM computer output. A simple z-test, where $z = \hat{\gamma}_{11}/SE(\hat{\gamma}_{11})$, tests whether the two population means appear to differ on the construct of interest.

Given statistical significance, one must be able to interpret which population has "more" of the construct in question and how much more. To do this, the researcher must first have a correct interpretation of the construct, which is done by examining the signs of the loading estimates ($\hat{\lambda}$ values) in light of what each measured variable represents. Once this is done, the sign of the value of $\hat{\gamma}_{11}$ implies which population has more of the construct—if positive, the population coded $X_1 = 1$ has more; if negative, the population coded $X_1 = 0$ has more. Finally, as detailed in Hancock (2001), the magnitude of this difference may be expressed as a standardized effect size estimate $\hat{d}$, where

$$\hat{d} = |\hat{\gamma}_{11}|/[\hat{V}(\zeta_1)]^{1/2}. \qquad (18)$$

Because $\hat{V}(\zeta_1)$ is effectively a pooled within-group factor variance, $\hat{d}$ may be interpreted as the estimated number of latent standard deviations separating the two population means on the latent continuum of interest. This estimated standardized effect size might also be used in post hoc power analysis and sample size determination (Hancock, 2001).

17.2.2. Two-Group MIMIC Example

To illustrate a MIMIC modeling approach to assessing latent mean differences, we present a hypothetical quasi-experimental example. Imagine two intact samples of 500 kindergartners, the first assigned to receive a traditional phonics-based curriculum for 2 years and the other to receive a whole-language

Table 17.1 Contrived Summary Statistics for Separate and Combined Samples

	Phonics							
	voc	*comp*	*lang*	*phon*	*alpha*	*print*	*SD*	*M*
voc	1.000						4.654	70.944
comp	0.770	1.000					3.943	69.606
lang	0.650	0.659	1.000				3.771	66.962
phon	0.223	0.260	0.181	1.000			1.343	3.274
alpha	0.381	0.413	0.313	0.511	1.000		1.325	3.616
print	0.325	0.349	0.277	0.419	0.640	1.000	1.329	3.300
	Whole Language							
	voc	*comp*	*lang*	*phon*	*alpha*	*print*	*SD*	*M*
voc	1.000						4.687	69.166
comp	0.747	1.000					3.901	68.046
lang	0.641	0.631	1.000				4.083	65.400
phon	0.199	0.238	0.174	1.000			1.300	2.774
alpha	0.289	0.315	0.237	0.507	1.000		1.276	2.794
print	0.267	0.311	0.253	0.439	0.618	1.000	1.238	2.564
	Combined							
	voc	*comp*	*lang*	*phon*	*alpha*	*print*	*dummy*	*SD*	*M*
voc	1.000							4.752	70.055
comp	0.768	1.000						3.997	68.826
lang	0.658	0.657	1.000					4.005	66.181
phon	0.239	0.277	0.207	1.000				1.342	3.024
alpha	0.371	0.400	0.315	0.533	1.000			1.363	3.205
print	0.332	0.365	0.303	0.456	0.660	1.000		1.335	2.932
dummy	0.187	0.195	0.195	0.186	0.302	0.276	1.000	0.500	0.500

curriculum for 2 years.[2] At the end of first grade, three reading assessments are made, focusing on vocabulary (*voc*), comprehension (*comp*), and language (*lang*). These are all believed to be observable manifestations of an underlying construct of reading proficiency. Summary statistics for these fabricated data are presented in Table 17.1 for the two samples separately, as well as in a combined sample with a dummy variable ($X_1 = 1$, phonics; $X_1 = 0$, whole language). Note that information on other variables appears in the table as well; this will be used later.

The model depicted previously in Figure 17.1 was fit to the covariance matrix (with the dummy variable) for the combined sample, whose correlations and standard deviations (*SD*) are shown in Table 17.1, using ML estimation in EQS 5.7b (Bentler, 1998). First and foremost, the fit of the model was excellent by any standards: $\chi^2(2, N = 1{,}000) = 2.128$, comparative fit index (CFI) = 1.000, standardized root mean square residual (SRMR) = .009, and root mean squared error of approximation (RMSEA) = .008, with a 90% confidence interval (CI) = (.000, .064). Second, key parameter estimates (all with $p < .05$) may be summarized as follows: $\hat{\lambda}_{Y21} = 0.841$, $\hat{\lambda}_{Y31} = 0.723$, $\hat{\gamma}_{11} = 1.872$, and $\hat{V}(\zeta_1) = 16.447$. Because $\hat{\gamma}_{11}$ is positive and statistically significant, we may *tentatively* infer that the phonics population ($X_1 = 1$) has a higher mean than the whole-language population ($X_1 = 0$) on the latent reading proficiency continuum. As for the magnitude of the effect, the estimated standardized effect size would be computed as $\hat{d} = |\hat{\gamma}_{11}|/[\hat{V}(\zeta_1)]^{1/2} = 1.872/(16.447)^{1/2} = .462$. This value implies that the phonics population mean sits almost one half of a standard deviation higher on the latent reading proficiency continuum than that of the whole-language population. By social science standards (e.g., Cohen, 1988), this could be considered a medium effect size.

2. Although such data would typically be multilevel in practice (e.g., students in common classrooms, etc.), assume for simplicity that the data here may be treated as simple random samples.

17.2.3. Extensions of the Basic MIMIC Model

First, the simple two-group scenario illustrated in this chapter is actually a special case of a more general class of MIMIC models (see Muthén, 1989). In MIMIC models in general, as the acronym's full name implies, a construct has several measured predictors causally impinging on it.[3] Therefore, just as the situation of a single dichotomous predictor allows us to make inferences about latent means for the case of two populations, using $J - 1$ group coded variables allows inferences about J populations in general. Furthermore, these J groups may represent populations varying on a single dimension, as in a one-way ANOVA, or along multiple dimensions, as in a factorial ANOVA. All the clever group coding schemes (e.g., dummy, contrast, effect) may be employed to target the population inferences of interest, just as they are when conducting ANOVA within a regression model (see, e.g., Pedhazur, 1997).

Second, when including the path from the group code predictor(s) directly to the construct, as shown in Figure 17.1 and analyzed in the MIMIC example for the two-group case, an assumption is made about the relation between the group code variable and the measured indicator variables. Specifically, this assumption is that the only relation between the group code predictor(s) and the indicator variables is the indirect one mediated by the construct of interest. Put more practically, this implies that the only reason for population differences in the measured indicator variables is the existence of population differences on the measured construct. Indeed, this is typically assumed and perhaps is the ideal case. However, one can imagine an indicator variable in which population differences are inflated or attenuated in addition to the construct-level differences (Muthén, 1989). Consider a researcher in a nonexperimental setting who wishes to assess population differences between Caucasian and Hispanic students on an English-language proficiency construct, whereby three English vocabulary tests are used as measured indicators of that construct. If one of the vocabulary tests consists of many words whose Latin roots are much more common in Spanish, then performance on that particular indicator variable will indicate not just the English-language proficiency construct but also something about cultural background. This illustrates what is known in the testing literature as *differential test functioning,* which is the more general case of *differential item functioning* (see, e.g., Holland & Wainer, 1993), and failing to accommodate this could lead to improper inference regarding population differences on the construct of interest. In the MIMIC model, this problem can be detected by model comparisons or modification indices (Lagrange multiplier tests), and it can be addressed by including an additional path directly from the ethnicity dummy predictor to that particular indicator variable; doing so will thereby preserve the integrity of the population inference at the construct level.

Third, MIMIC models also facilitate the accommodation of covariates, even when the covariates are latent constructs having their own indicator variables. Typically in analysis of covariance (ANCOVA) applications, a covariate has the potential to contain considerable measurement error. Covariate measurement error, whether specifically in ANCOVA settings (see Trochim, 2001) or in the more general set of mediator-variable models (see Hoyle & Kenny, 1999), can lead to inaccurate inference regarding mean differences and/or structural relations. However, when the desired covariate is operationalized as a construct from a latent variable system of measured covariates, the theoretically error-free latent covariate may be included in the model as an additional covarying predictor of the construct on which population differences are being investigated. For notational purposes, we may express the original group code predictor X_1 as a single-indicator factor ξ_1 (i.e., $X_1 = \xi_1$) and the new latent covariate as ξ_2 (with, say, indicators X_2 through X_4); ξ_1 and ξ_2 have covariance $C(\xi_1, \xi_2)$, often designated as φ_{21}. Thus, equation (16) may be extended as

$$\eta_1 = [\gamma_{11} \quad \gamma_{12}] \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \zeta_1 \tag{19}$$

or, symbolically, as

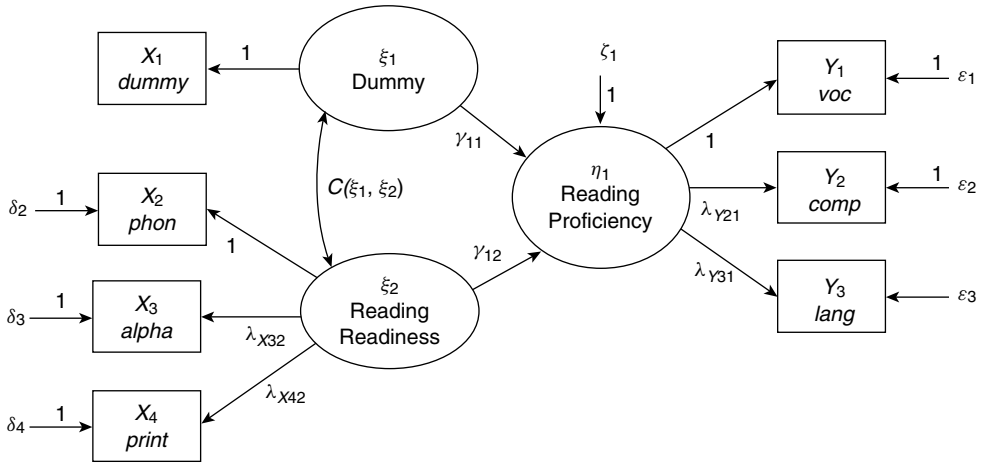
$$\eta_1 = \boldsymbol{\Gamma}\boldsymbol{\xi} + \zeta_1. \tag{20}$$

In such situations, as in ANCOVA in general, population differences above and beyond those resulting from the covariate are of interest. Thus, the key path in this model is still γ_{11}—that from the group code variable (designated herein as ξ_1) to the outcome construct η_1, which now represents a covariate-adjusted latent mean difference. Figure 17.2 depicts such a model with a three-indicator latent covariate, and a numerical example of this approach appears below.[4]

3. This also implies that, with the single dummy variable as the lone causal variable in the two-group case, the term *MIMIC* is technically a misnomer.

4. An interesting variation on ANCOVA within MIMIC models is useful with nonrandom selection into groups (e.g., treatment and control). Kaplan (1999) proposed using propensity scores (Rosenbaum & Rubin, 1983), which are conditional probabilities of selection into a target group (e.g.,

Figure 17.2 Model for the Multiple-Indicator, Multiple-Cause (MIMIC) Example With Latent Covariate



17.2.4. Two-Group MIMIC Example, With Latent Covariate

In the previous MIMIC example, the final inference was that the phonics population was higher in average latent reading proficiency than the whole-language population. This inference was noted to be made *tentatively,* however, given that the use of intact groups introduces a selection threat to the study's internal validity. To help compensate for the lack of experimental control, we may employ the statistical control of a latent covariate. Extending the previous MIMIC example, imagine that at the start of kindergarten (i.e., prior to receiving reading instruction), each teacher had used 6-point rating scales to form ratings for each child on phonemic awareness (*phon*), the alphabetic principle (*alpha*), and print concepts and conventions (*print*). These three measures are all believed to be observable manifestations of an underlying construct of reading readiness. Summary statistics for these fabricated data appear in Table 17.1 for the two samples separately, as well as in a combined sample, with a dummy variable ($X_1 = 1$, phonics; $X_1 = 0$, whole language).

The model depicted in Figure 17.2 was fit to the covariance matrix (with dummy variable) for the combined sample, whose correlations and standard deviations are shown in Table 17.1, using ML estimation in EQS 5.7b (Bentler, 1998). The fit of the model was excellent: $\chi^2(12, N = 1{,}000) = 10.507$, CFI $= 1.000$, SRMR $= .012$, and RMSEA $= .000$, with 90% CI $= (.000, .029)$. Also, all loading parameter estimates were statistically significant ($p < .05$) : $\hat{\lambda}_{Y21} = 0.853$, $\hat{\lambda}_{Y31} = 0.726$, $\hat{\lambda}_{X32} = 1.462$, and $\hat{\lambda}_{X42} = 1.252$. As for the structure, key parameters were as follows: $\hat{\gamma}_{11} = 0.464$ ($p = .087$), $\hat{\gamma}_{12} = 2.481 (p < .05)$, $\hat{C}(\xi_1, \xi_2) = 0.141$ ($p < .05$), and $\hat{V}(\zeta_1) = 12.674$. In these structural parameters, the two effects of using a covariate are apparent. First, there is a reduction in the endogenous construct's residual variance, from 16.447 in the previous example to 12.674 with the latent covariate. Second, the estimated population mean difference is adjusted, from a statistically significant 1.872 previously to a nonsignificant 0.464 with the latent covariate. This result implies that once the differences in the intact samples' reading readiness are taken into account, the null hypothesis that phonics and whole-language yield equivalent average reading proficiency appears tenable.

17.3. Structured Means Modeling (SMM)

17.3.1. Development

The relation between a MIMIC modeling strategy and using the SMM approach is similar to that between the regression and t-test approaches to assessing

treatment) derived from probit or logistic regression using covariates as predictors. Individuals are then grouped into propensity score strata, and separate MIMIC models are fit simultaneously to data from cases in each stratum. Under specific conditions regarding the original selection into groups and the invariance of the measurement model across populations (see Kaplan, 1999, for details), conclusions can be reached about the nature of group differences and its generalizability across strata.

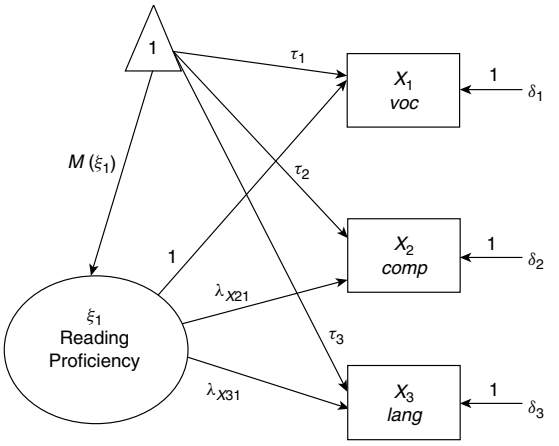
univariate differences between two populations. A simple regression approach, like MIMIC modeling, uses the data from both groups as part of a single sample. The criterion variable is regressed on the dummy predictor, allowing inferences to be made about potential differences on that criterion through the magnitude, sign, and statistical significance of the regression (path) coefficient. No variable means were directly required in this process, nor did the intercept commonly appearing as part of a regression equation have any bearing on the interpretation regarding group differences on the criterion. In MIMIC modeling—the SEM analog to a simple regression approach to assessing population differences—intercepts are generally ignored altogether (for reasons discussed previously). Similar to regression, population differences on the construct of interest are inferred directly from the variables' covariance structure; specifically, the path coefficient relating the predictor (dummy variable) and the criterion (construct) is used. Again, individual group means on the measured variables are not required.

In contrast, the SMM approach to assessing population differences, like a t-test, keeps the data from the two groups separate. This eliminates the need for a group code variable to differentiate because scores from different groups are not combined for analysis. Instead, also like a t-test, the means of the variables are used in the analyses. Thus, SMM will use equations involving means and, as was shown in the introductory section, the accompanying variable intercepts. These new equations constitute the mean structure, which will be estimated in addition to the covariance structure that is part of all SEM analyses. The simultaneous estimation of covariance and mean structures associated with the latent and observed variables will facilitate the ultimate goal of making inferences about population means on the construct of interest.

To understand how SMM operates, again assume there are only two groups to be compared in terms of their construct means. The construct of interest ξ_1 has three indicators—X_1, X_2, and X_3—and the scale of the construct is determined by fixing the X_1 loading to 1. This model, which is assumed to hold for both populations, is included in Figure 17.3. The structural equations for a single population include intercept terms and may be represented in matrix form as follows:

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} 1 \\ \lambda_{X21} \\ \lambda_{X31} \end{bmatrix} \xi_1 + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} \qquad (21)$$

Figure 17.3 Model and Implied Relations for Structured Means Modeling (SMM)



Model-implied relations

$V(X_1) = V(\xi_1) + V(\delta_1)$
$V(X_2) = \lambda_{X21}^2 V(\xi_1) + V(\delta_2)$
$V(X_3) = \lambda_{X31}^2 V(\xi_1) + V(\delta_3)$
$C(X_1, X_2) = \lambda_{X21} V(\xi_1)$
$C(X_1, X_3) = \lambda_{X31} V(\xi_1)$
$C(X_2, X_3) = \lambda_{X21}\lambda_{X31} V(\xi_1)$
$M(X_1) = \tau_1 + M(\xi_1)$
$M(X_2) = \tau_2 + \lambda_{X21} M(\xi_1)$
$M(X_3) = \tau_3 + \lambda_{X31} M(\xi_1)$

or, symbolically, as

$$\mathbf{x} = \boldsymbol{\tau} + \boldsymbol{\Lambda}_X \xi_1 + \boldsymbol{\delta}. \qquad (22)$$

Notice also that a unit constant may be inserted in equation (22) following the vector $\boldsymbol{\tau}$ (i.e., $\mathbf{x} = \boldsymbol{\tau}1 + \boldsymbol{\Lambda}_X \xi_1 + \boldsymbol{\delta}$), conveying that the $\boldsymbol{\tau}$ values can be treated as paths to each measured X from a unit-constant *pseudovariable* (i.e., a variable on which all subjects have a value of 1). This convention is depicted in Figure 17.3 with paths to each of the X variables from a triangle representing this unit constant, thus denoting measured variable intercepts. These structural equations may now be interpreted just as in regression. For example, considering the structural equation for the measured variable X_2, the loading λ_{X21} is a slope relating change in the construct ξ_1 to change in X_2, whereas τ_2 represents the predicted value of X_2 for a subject having a zero value on the construct ξ_1.

The reader will also notice another convention represented in Figure 17.3, that of a path from the unit constant to the construct ξ_1 and labeled as $M(\xi_1)$. Because any score in a set can be written as a deviation from the set's mean, theoretical scores on the construct

ξ_1 can be written as a function of the ξ_1 mean and a residual:

$$\xi_1 = M(\xi_1) + \zeta_1. \quad (23)$$

As described previously, notice that a unit constant may be inserted in equation (23) following the latent mean $M(\xi_1)$ (i.e., $\xi_1 = M(\xi_1)1 + \zeta_1$), conveying that $M(\xi_1)$ can be treated as a path from the unit constant to the latent construct ξ_1. This is conveyed in Figure 17.3 by a path from the unit constant to ξ_1, which makes ξ_1 appear as a dependent construct. However, because the unit constant explains no variance in ξ_1, ξ_1 is technically still exogenous. For this reason, the disturbance ζ_1 shown in equation (23) is also omitted from the diagram.

Based on the structural equations contained in this model, relations are implied regarding population variances, covariances, and also means, as seen in Figure 17.3. Specifically, the three variances and three covariances among the three observed variables are seen to be a function of six parameters requiring estimation: two loadings (λ_{X21}, λ_{X31}), one construct variance ($V(\xi_1)$), and three error variances ($V(\delta_1)$, $V(\delta_2)$, $V(\delta_3)$). These model-implied relations constitute the covariance structure for one of the two populations being compared; the complete covariance structure includes similar equations for the second population as well. As for the mean structure, the last three relations shown in Figure 17.3 express the population means for X_1 through X_3 as a function of four additional parameters to be estimated (i.e., in addition to the two loadings, which would be estimable from the covariance model alone): three intercepts (τ_1, τ_2, τ_3) and the construct mean ($M(\xi_1)$). These model-implied relations constitute the mean structure for one population; the complete mean structure includes similar equations for the second population as well. As the reader may have noticed, though, the mean structure is currently underidentified; that is, there are too few equations for the number of unknowns that must be determined. The solution to this problem lies in the implementation of theoretically meaningful cross-group constraints, as discussed next.

Recall the goal of the methods presented in this chapter: to make inferences about latent means. In SMM, this is done, at least in part, using information provided by the observed variable means. Specifically, any population differences on the observed variables are presumed to be the direct result of population differences on the underlying construct. For this to be a valid presumption, the structural relation between the construct and each observed variable should be alike in both populations. That is, SMM generally requires the equivalence of corresponding loadings across populations (λ_{X21} and λ_{X31} in this example) and of corresponding intercept terms across populations (τ_1 through τ_3 in this example). In practice, this is achieved through the imposition of cross-group constraints on those parameters. These constraints reflect the desired condition that the impact of the construct on the measured variables is *invariant* across both populations and that a zero value (or any specific value) on the construct would yield the same amount of a given variable in either population. If this were not the case, the interpretational equivalence of the constructs across populations could be in question, making a comparison of means on those constructs substantively tenuous. Furthermore, with respect to statistical considerations, constraints involving the intercepts reduce the number of unique parameters to be estimated within the complete means model and thus help to make the model identified (as shown later, however, one additional constraint is still required).

To illustrate this general invariance assumption of intercepts and loadings across populations, we present a diagram for X_2 in Figure 17.4. The ellipses represent hypothetical scatter plots of points for each population on ξ_1 and X_2, through which pass (overlapping) best-fit regression lines. Certainly, if the two populations' scatter plots were identically positioned, they would be invariant in terms of their lines' slope (the loading λ_{X21}) and intercept (τ_2). In the figure, the populations are presented as having different means on the construct ξ_1 and the variable X_2; however, the structural relation between ξ_1 and X_2 (represented by a regression line) is still the same in both groups. Interpretationally, this implies the desired condition that any difference the populations may have in terms of the observed variable is directly attributable to a difference in the underlying construct and not to differences in the nature of the structural relation. In actuality, in some situations, such invariance may not hold, and yet meaningful inferences may be drawn regarding latent means. This issue will be mentioned more fully later; for now, the current discussion will proceed under this commonly assumed invariance condition.

As alluded to earlier, the imposition of one more cross-group constraint is still required so that we can estimate the mean structure's parameters. To understand the necessity of this constraint, consider the model-implied means relations from both populations based on our discussion thus far. These population relations would be as listed in Figure 17.5, where a prescript indicates a specific population and a lack of prescript indicates a parameter that has been constrained to be equal across populations (and therefore

Figure 17.4 Invariance Assumption for Structured Means Modeling (SMM)

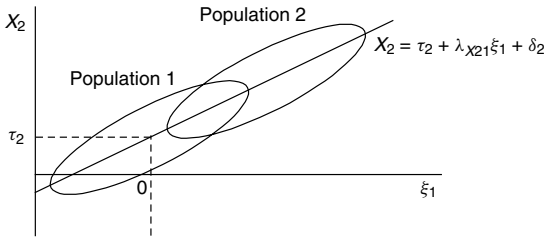


Figure 17.5 Model-Implied Relations for Mean Structure With Loading and Intercept Invariance Constraints

$$
\begin{aligned}
{}_1M(X_1) &= \tau_1 + {}_1M(\xi_1) \\
{}_1M(X_2) &= \tau_2 + \lambda_{X21}[{}_1M(\xi_1)] \\
{}_1M(X_3) &= \tau_3 + \lambda_{X31}[{}_1M(\xi_1)] \\
{}_1M(X_4) &= \tau_4 + \lambda_{X41}[{}_1M(\xi_1)] \\
{}_2M(X_1) &= \tau_1 + {}_2M(\xi_1) \\
{}_2M(X_2) &= \tau_2 + \lambda_{X21}[{}_2M(\xi_1)] \\
{}_2M(X_3) &= \tau_3 + \lambda_{X31}[{}_2M(\xi_1)] \\
{}_2M(X_4) &= \tau_4 + \lambda_{X41}[{}_2M(\xi_1)]
\end{aligned}
$$

requires no such differentiation). Because the loading parameters (λ_{X21}, λ_{X31}) are estimable from the covariance model alone, the additional relations in the mean structure pose no threat to their estimation. Unique to the six equations of the complete mean structure are five unknowns: three intercepts (τ_1 through τ_3, constrained equal across groups) and two construct means (${}_1M(\xi_1)$, ${}_2M(\xi_1)$), with these last two being of key importance to answering the question about group differences. On the surface, with six equations and five parameters to estimate, there would seem to be sufficient information for the mean structure to be identified. Unfortunately, as the reader can illustrate with a little algebra, this is not the case.

Specifically, in attempting to solve the system of relations for either ${}_1M(\xi_1)$ or ${}_2M(\xi_1)$, there is no way to isolate one of these parameters without the other being part of the equation. An expression for the difference $[{}_1M(\xi_1)] - [{}_2M(\xi_1)]$ can be isolated but not for either mean individually. The problem is akin to being told that the difference between two numbers is 10 and then being asked what each number must be—there can be no unique answer without assuming a value for one of those two numbers. The same is true of the construct means. The only way the relations of the means model can be used to estimate the relevant parameters is by fixing one of the construct means to a particular numerical value. It may seem that fixing one of the factor means defeats the purpose of means modeling, but such is not the case. Remember that the goal is actually to evaluate the *difference* between construct means, not the means themselves. Fixing one construct mean to a particular value does not affect the difference between those means, just as fixing one of the two unknown numbers in the above example would not make the difference between the two numbers any more or less than the stated 10 points. Doing so merely allows the determination of a unique value for the other number in question.

In SMM, it is customary to fix one population's construct means to zero. Zero is a most prudent choice because a test of the difference between the free factor mean and the fixed factor mean of zero is precisely accomplished by a one-sample test on the free factor mean, as will be seen later. In the current example, then, Population 2 may be considered the *reference population* by fixing its mean to zero. This simplifies the last four structural equations of the means model shown in Figure 17.5 such that each intercept term is equal to the observed variable mean for that reference population, thus allowing a unique solution to be determined for the mean structure when estimated along with the covariance structure.

Estimating the complete model for both populations, with both covariance and mean structure, involves many model-implied relations and the estimation of many unknown parameters. Before doing so, however, in practice it makes sense to estimate the covariance structure first (with the loading constraints discussed earlier). If there is a poor fit of this model to the data, one or both of the following problems may be present. First, perhaps a single factor model does not adequately describe the relations among the variables in one or both groups; second, it is possible that the loadings are not invariant across populations as constrained. In either case, the meaningfulness of proceeding with the mean structure of the model is in question and is potentially ill-advised; more on this subject follows later.

Assuming the covariance structure fits satisfactorily as described, the covariance and mean structures may then be estimated simultaneously. In the current example (with all loading and intercept constraints imposed as discussed), the full model consists of 18 model-implied variance, covariance, and mean relations over both populations, requiring the estimation of 14 unique parameters: 6 error variances (3 per

population), 2 construct variances (1 per population), 2 loadings ($\lambda_{X21}, \lambda_{X31}$), 3 intercepts ($\tau_1, \tau_2, \tau_3$), and 1 factor mean (${}_1M(\xi_1)$). The model thus has 4 ($18 - 14 = 4$) degrees of freedom across its covariance and mean structures. The 6 model-implied variances and covariances for the $p = 3$ variables appear in matrices $\hat{\boldsymbol{\Sigma}}_1$ and $\hat{\boldsymbol{\Sigma}}_2$ for Population 1 and Population 2, respectively, whereas the observed variances and covariances are contained in matrices $\mathbf{S}_1$ and $\mathbf{S}_2$. The 3 model-implied means appear in vectors $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ for Population 1 and Population 2, respectively, whereas the observed means are contained in vectors $\mathbf{m}_1$ and $\mathbf{m}_2$. In the context of ML estimation specifically, parameter values are selected to imply (reproduce) population variance-covariance matrices and mean vectors from which the observed data's sample variance-covariance matrices and mean vectors have the maximum likelihood of arising by random chance. As presented elsewhere (e.g., Bollen, 1989), this process is operationalized for the two-group case by choosing parameters so as to minimize the multisample fit function

$$G_{\text{ML}} = \sum_{j=1}^{2}(n_j/N)\{[\ln|\hat{\boldsymbol{\Sigma}}_j| + \text{tr}(\mathbf{S}_j\hat{\boldsymbol{\Sigma}}_j^{-1}) - \ln|\mathbf{S}_j| - p] + (\mathbf{m}_j - \hat{\boldsymbol{\mu}}_j)'\hat{\boldsymbol{\Sigma}}_j^{-1}(\mathbf{m}_j - \hat{\boldsymbol{\mu}}_j)\}. \tag{24}$$

After fitting this complete model, one again hopes for an acceptable degree of data-model fit (i.e., that $\hat{\boldsymbol{\Sigma}}_1$ and $\hat{\boldsymbol{\Sigma}}_2$ are sufficiently close to $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively, and that $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ are sufficiently close to $\mathbf{m}_1$ and $\mathbf{m}_2$, respectively). Given that the loading invariance across populations has already been assessed in a preliminary covariance structure analysis, poorness of data-model fit in this model may be the result of stress in the mean structure created by untenable intercept constraints. If so, suspected noninvariant intercepts may have their constraint released before proceeding with the latent means comparison. This precisely parallels the differential test functioning issue discussed previously in the context of MIMIC models.

Once satisfactory data-model fit is achieved when fitting the covariance and mean structures simultaneously, then the primary question of interest may finally be addressed. As mentioned previously, a test of whether the two population means differ is precisely a test of whether the estimated construct mean for Population 1, ${}_1\hat{M}(\xi_1)$, differs statistically significantly from 0 (the value to which the reference population mean was fixed). This is accomplished by a simple z-test using the standard error associated with the Population 1 construct mean estimate, $SE[{}_1\hat{M}(\xi_1)]$, where $z = {}_1\hat{M}(\xi_1)/SE[{}_1\hat{M}(\xi_1)]$. As with MIMIC modeling, statistical significance requires one to interpret which population has more of the construct under investigation. Assuming that the researcher has correctly interpreted the construct by examining signs of loadings on each specific variable, the sign of the value of ${}_1\hat{M}(\xi_1)$ facilitates interpretation—if ${}_1\hat{M}(\xi_1)$ is positive, then Population 1 is inferred to have more of (be higher on) the construct; if ${}_1\hat{M}(\xi_1)$ is negative, then Population 1 has less (is lower).

Finally, as detailed in Hancock (2001), the magnitude of the estimated latent mean difference may be expressed as a standardized effect size estimate $\hat{d}$, where $\hat{d} = |{}_1\hat{M}(\xi_1)|/[\hat{V}(\xi_1)]^{1/2}$, and $\hat{V}(\xi_1)$ is the weighted average of the two samples' factor variances (with respective sample sizes as weights). The value of $\hat{d}$ may again be interpreted as the estimated number of latent standard deviations separating the two populations on the latent continuum of interest and is useful in post hoc power analysis and sample size determination (Hancock, 2001).

17.3.2. Two-Group SMM Example

The contrived reading proficiency data displayed in Table 17.1 (*voc, comp,* and *lang*) are again used, illustrating the SMM approach to latent mean differences. Following the model in Figure 17.3, all free loadings and intercepts were constrained to be equal across populations, the first indicator (*voc*) was selected as the scale indicator in both populations (fixing the loading to 1), and whole language (Population 2) was made the reference population (i.e., by fixing the latent mean to 0). The program EQS 5.7b (Bentler, 1998) was again used to conduct ML estimation. First, as expected, the fit of the model across both groups simultaneously was excellent: $\chi^2(4, N = 1{,}000) = 4.036$, CFI = 1.000, SRMR = .020, and RMSEA = .003, with 90% CI = (.000, .048). Second, key parameter estimates (all with $p < .05$) may be summarized as follows: $\hat{\lambda}_{X21} = 0.843$, $\hat{\lambda}_{X31} = 0.722$, $\hat{\tau}_1 = 69.118$, $\hat{\tau}_2 = 68.036$, $\hat{\tau}_3 = 65.517$, ${}_1\hat{M}(\xi_1) = 1.871$, ${}_1\hat{V}(\xi_1) = 16.454$, and ${}_2\hat{V}(\xi_1) = 16.375$. Because ${}_1\hat{M}(\xi_1)$ is positive and statistically significant, we may again tentatively infer that the phonics population has a higher latent reading proficiency mean than the whole-language population. As for the magnitude of the effect, a pooled factor variance must first be determined: $\hat{V}(\xi_1) = [n_1({}_1\hat{V}(\xi_1)) + n_2({}_2\hat{V}(\xi_1))]/(n_1 + n_2) = 16.415$. The estimated standardized effect size would then be computed as

$\hat{d} = |_1\hat{M}(\xi_1)|/[\hat{V}(\xi_1)]^{1/2} = 1.871/(16.415)^{1/2} = .448$. As with MIMIC modeling, this value implies a medium effect size in which the phonics distribution's latent mean sits almost one half of a standard deviation higher on the latent reading proficiency continuum than that of the whole-language distribution.

17.3.3. Extensions of the Basic SMM Model

Similar to MIMIC modeling, SMM can be extended to accommodate more complex designs. Data from J groups in a one-way design can be modeled separately but simultaneously as was done with two groups in this chapter. Initial cross-group constraints are imposed such that all corresponding loadings are equal, all corresponding intercepts are equal, and one population's construct mean is fixed to zero. All other construct means are solved with respect to this reference population; together with their accompanying standard errors, inferences may be made regarding population differences on the construct of interest. Accommodating designs that are factorial in nature, however, is less straightforward than with a MIMIC model's group code variables; the creative application of a series of cross-group constraints would be required for main effect and interaction inferences to be made within SMM.

As SMM involves the explicit modeling of multiple groups simultaneously, issues about the invariance of that model are relevant to ensure the integrity of the resulting latent mean inference. Many authors have discussed these issues, resulting in a continuum of invariance conditions. At the most extreme end is what Meredith (1993) termed *strict factorial invariance,* which implies identical loadings, error variances and covariances, factor variances, and intercepts across populations. This degree of stringency, however, is not essential for accurate latent mean inference. Meredith defined *strong factorial invariance* as the condition of equal loadings and equal intercepts across populations, a condition whose satisfaction will still preserve the integrity of latent mean inference. The loading and intercept constraints suggested previously within SMM mirror this desired condition. Beneath this exist a variety of weaker conditions, such as *partial measurement invariance* (not all loadings are identical across populations) and *partial intercept invariance* (not all intercepts are identical across populations), as discussed by Byrne, Shavelson, and Muthén (1989). Indeed, if truly noninvariant loadings and/or intercepts exist, and if their associated cross-group constraints are released as a result of theoretical rationale a priori or empirical rationale post hoc (i.e., modification indices), accurate latent mean inference can be preserved. The challenge is in accurately locating such noninvariance, and the failure to do so could induce apparent latent mean differences where none truly exist or attenuate the magnitude (and even sign) of those that truly do (see, e.g., Cole et al., 1993; Hancock, Stapleton, & Berkovits, 1999).

Finally, as with MIMIC modeling, SMM may also be extended to control for covariates. As described by Sörbom (1978), the incorporation of covariates into structured means models can facilitate inference regarding latent mean differences among populations above and beyond those resulting from differences on covariates (measured or latent). In such models, the covariate is an exogenous construct ξ_1, whereas the construct on which latent population mean differences are of interest is now labeled η_1 due to its theoretical dependence on the covariate. Thus, in addition to all loadings and intercepts being assumed invariant across populations, the latent structure of the model for each population may be represented as

$$\eta_1 = \kappa_1 + \gamma_{11}\xi_1 + \zeta_1, \tag{25}$$

where κ_1 is the latent intercept representing the predicted value of η_1 associated with a zero value on the construct ξ_1. Furthermore, given that γ_{11} is assumed (and constrained to be) invariant across populations to mirror ANCOVA's homogeneity of slope or *parallelism* assumption, the expected value of equation (25) may be taken as follows. For Population 1,

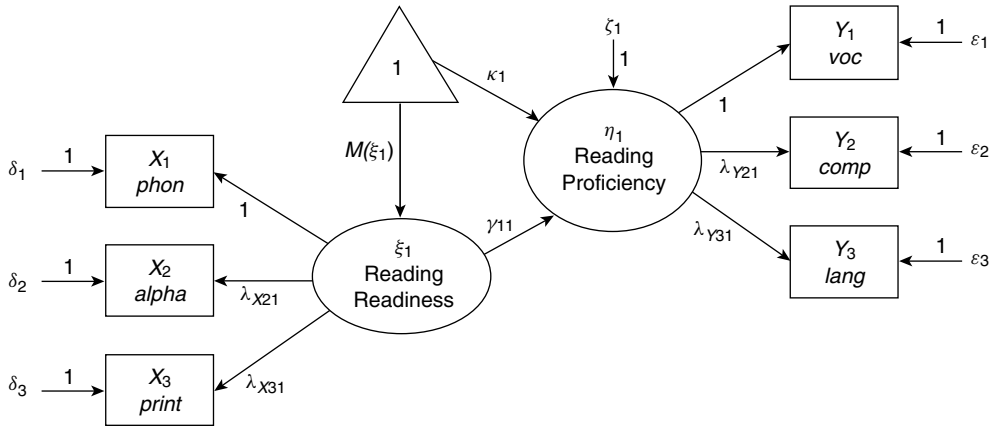
$$_1M(\eta_1) = {}_1\kappa_1 + \gamma_{11}[{}_1M(\xi_1)]. \tag{26}$$

For Population 2,

$$_2M(\eta_1) = {}_2\kappa_1 + \gamma_{11}[{}_2M(\xi_1)]. \tag{27}$$

If, for identification purposes, Population 2 is made the reference population and both of its latent means are constrained to 0 (i.e., $_2M(\xi_1) = 0, {}_2M(\eta_1) = 0$, and thus $_2\kappa_1 = 0$), then only equation (26) becomes useful. Specifically, equation (26) states that there are two reasons why $_1M(\eta_1)$ might differ from $_2M(\eta_1) = 0$: a difference in the mean level of the covariate (i.e., between $_1M(\xi_1)$ and $_2M(\xi_1) = 0$) and a difference beyond the covariate (i.e., between $_1\kappa_1$ and $_2\kappa_1 = 0$). Thus, the test of the parameter estimate $_1\hat{\kappa}_1$ (against 0) is precisely the test of the difference between population means on η_1 above and beyond those explainable by a difference on the covariate ξ_1.

Figure 17.6 Model for the Structured Means Modeling (SMM) Example With Latent Covariate



NOTE: For clarity, variable intercepts are not shown.

Figure 17.6 represents such a model with a three-indicator latent covariate. It is important to note that for clarity, the measured variable intercepts have been omitted from the figure; however, the proper execution of this model requires intercepts for both factors' indicators in both populations. The numerical example paralleling this figure (and paralleling the previous MIMIC strategy with a covariate) appears next.

17.3.4. Two-Group SMM Example, With Latent Covariate

In the previous SMM example, the final tentative inference was that the phonics population was higher in average latent reading proficiency than the whole-language population. Given the intact groups' threat to internal validity, however, we may again wish to employ the statistical control of a latent covariate. Extending the previous SMM example (and paralleling the second MIMIC example), the three indicators of reading readiness from start of kindergarten (*phon, alpha,* and *print*) were incorporated into the model, as shown in Figure 17.6. The model depicted was fit to the two samples' covariance matrices and mean vectors using ML estimation in EQS 5.7b (Bentler, 1998), imposing all cross-group loading and intercept constraints and making whole language the reference population with latent means set to 0. To start, the data-model fit was excellent: $\chi^2(25, N = 1{,}000) = 14.646$, CFI $= 1.000$, SRMR $= .021$, and RMSEA $= .000$, with 90% CI $= (.000, .002)$. Equality-constrained loading parameter estimates for both factors and equality-constrained intercepts for all measured variables were statistically significant ($p < .05$). Key structural parameters were as follows: $\hat{\gamma}_{11} = 2.486$ ($p < .05$), ${}_1\hat{M}(\xi_1) = 0.562$ ($p < .05$), ${}_1\hat{\kappa}_1 = 0.462$ ($p = .088$), ${}_1\hat{V}(\zeta_1) = 12.110$, and ${}_2\hat{V}(\zeta_1) = 13.150$. In these structural parameters, the two effects of using a covariate are again apparent. First, there is a reduction in the endogenous construct's residual variance, from 16.454 and 16.375 for the two-group SMM example without the covariate to 12.110 and 13.150 now. Second, the estimated population mean difference is adjusted from the previous statistically significant 1.871 to a nonsignificant 0.462 with the latent covariate. This result parallels the MIMIC example with covariate, implying that once the differences in the intact samples' reading readiness are taken into account, the null hypothesis that phonics and whole language yield equivalent average reading proficiency appears tenable.

17.4. Summary and Conclusions

As this chapter indicated from the start, the choice of an SEM method for handling multivariate group comparisons rests first and foremost in the nature of the variable system under investigation. Specifically, latent variable systems are more appropriately handled with methods designed to treat the latent variable; applying emergent variable system methods such as MANOVA, although a common approach to dealing with latent variable systems, is theoretically less appropriate and can be statistically problematic (see Cole at al., 1993). Whereas MANOVA methods for assessing group differences involve the creation of

a linear composite of measured variables, thereby incorporating variables' measurement error into the composite, SEM methods use a theoretically error-free construct in tests of group differences. Similarly, SEM methods allow for the inclusion of latent, theoretically error-free covariates derived from measured variables. Finally, SEM methods are more flexible in that they allow for covariance adjustments not just at the construct level but also at the individual variable level. Individual variables' residuals, for example, might have reason to covary above and beyond their variables' common construct; SEM methods have no difficulty accommodating such relations.

In the matter of selecting between MIMIC and SMM approaches, the reader might infer that MIMIC modeling is more desirable based solely on its relative simplicity. It is quite true that SMM is more complex, often requiring good start values to achieve model convergence, involving the estimation of more parameters, and possibly requiring a larger sample size to do so reliably. Within both methods, however, lie assumptions, advantages, and disadvantages that make the choice between SEM methods require considerations beyond apparent simplicity.

The primary assumption implicit in MIMIC modeling is that, because the data from the groups are combined and only one model results, the same measurement model holds in both populations. This includes loadings, construct variance, and error variances. In effect, all sources of covariation among observed variables are assumed to be equal in both populations, making the assumption of identical measurement models tantamount to an assumption of equal variance-covariance matrices (as is actually assumed in MANOVA as well). As discussed, such restrictiveness is not generally required in SMM, in which only the corresponding loadings are commonly constrained across populations in the complete covariance model. Furthermore, additional flexibility may exist to allow for some loading differences across populations under particular configurations of partial measurement invariance (Byrne et al., 1989).

Considering Type I error rate and statistical power for testing latent mean differences, Hancock et al. (2000) used Monte Carlo simulation to show that SMM appears to control Type I error acceptably in many invariant and noninvariant loading scenarios (whether or not loading constraints were in place). The MIMIC approach, on the other hand, controlled the error rate when approximately equal generalized variances (covariance matrix determinants) and/or equal sample sizes were present, but not with both sample size and generalized variance disparities. In short, when the sample with smaller loadings had a larger sample size, the Type I error rate increased. Conversely, if the sample with the smaller loadings had a smaller sample size, this yielded conservatism in Type I error control. Finally, regarding power under loading and sample size scenarios in which both methods controlled Type I error, the authors used population analysis (see also Kaplan & George, 1995) to show overall that the power of both methods to detect true differences in latent means was quite comparable and unilaterally superior to MANOVA methods (see also Kano, 2001).

Finally, in favor of the MIMIC approach is its design flexibility. Specifically, the creative use of group code predictors of the latent construct of interest can fairly easily facilitate inferences that parallel those of more complex ANOVA designs. Although one can imagine adapting SMM to do likewise, as alluded to previously, precise methods for doing so currently remain unarticulated.

In sum, the choice between the MIMIC and SMM approaches may rest with the researcher's specific modeling scenario. Some researchers have already found these methods to fit their research needs: Kinnunen and Leskinen (1989) examining teacher stress; Aiken, Stein, and Bentler (1994) examining treatment for drug addiction; Gallo, Anthony, and Muthén (1994) examining depression; and Dukes, Ullman, and Stein (1995) examining drug abuse resistance education with elementary school children. However, many more opportunities to answer construct-level questions exist throughout the social sciences. Consider data from various social science databases in which, although not truly experimental in design, these SEM methods can assist with latent mean inference. In the *Monitoring the Future: Lifestyles and Values of Youth 1976–1992* database, nationally representative samples of high school seniors were selected each year from 1976 to 1992. A key construct of interest from this database would be risk tendency, the manifest indicators of which could include (but are not limited to) traffic citations and illicit substance use. Population differences in this construct could be investigated across a variety of interesting between-subject dimensions, including sex of student, geographic region of the country, urbanicity of the student's childhood environment, mother's employment status while growing up, and point in time (e.g., 1976–1980, 1981–1985, 1986–1990). In the *National Health Survey: Longitudinal Study of Aging, 70 Years and Over, 1984–1990* database, a sample of 7,527 noninstitutionalized elderly people in the United States were assessed on a variety of dimensions (and at multiple time points). Three critical constructs of interest

from this database are personal care independence (as indicated by ability to perform seven specific personal care skills), home management independence (as indicated by ability to perform six specific home management skills), and medical fragility (with manifest variables such as length of hospital stays, nursing home time, and number of doctor's visits). Latent mean differences could be assessed across such grouping variables as sex, race, or region of the country. Finally, a most interesting database exists in the *National Commission on Children: Parent and Child Study,* in which a national sample of 1,738 parents living with their children was surveyed for data from parent-child pairs. One key construct is family cohesion, as assessed by parents and their children using a variety of measured indicators. Coupling this with between-subject variables such as urbanicity of child's home environment, sex of child, and income classification might lead to some extremely interesting latent mean inferences.

It is hoped that the conceptual introduction offered in this chapter will help to motivate other applied researchers to investigate latent means methods more fully and, ultimately, to find applications of these methods to problems in their own areas of inquiry.

REFERENCES

Aiken, L. S., Stein, J. A., & Bentler, P. M. (1994). Structural equation analyses of clinical subpopulation differences and comparative treatment outcomes: Characterizing the daily lives of drug addicts. *Journal of Consulting and Clinical Psychology, 62*, 488–499.

Bentler, P. M. (1998). *EQS structural equations program.* Encino, CA: Multivariate Software, Inc.

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

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*, 305–314.

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

Cleary, T. A., & Linn, R. L. (1969). Error of measurement and the power of a statistical test. *British Journal of Mathematical and Statistical Psychology, 22*, 49–55.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, P., Cohen, J., Teresi, M., Marchi, M., & Velez, C. N. (1990). Problems in the measurement of latent variables in structural equations causal models. *Applied Psychological Measurement, 14,* 183–196.

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.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.

Dukes, R. L., Ullman, J. B., & Stein, J. A. (1995). An evaluation of D.A.R.E. (Drug Abuse Resistance Education), using a Solomon four-group design with latent variables. *Evaluation Review, 19*, 409–435.

Gallo, J. J., Anthony, J. C., & Muthén, B. O. (1994). Age differences in the symptoms of depression: A latent trait analysis. *Journal of Gerontology: Psychological Sciences, 49*, 251–264.

Hakstian, A. R., & Whalen, T. E. (1976). A *K*-sample significance test for independent alpha coefficients. *Psychometrika, 41*, 219–231.

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

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2000). Type I error and power of latent mean methods and MANOVA in factorially invariant and noninvariant latent variable systems. *Structural Equation Modeling: A Multidisciplinary Journal, 7*, 534–556.

Hancock, G. R., Stapleton, L. M., & Berkovits, I. (1999, April). *Minimum constraints for loading and intercept invariance in covariance and mean structure models.* Paper presented at the 1999 annual meeting of the American Educational Research Association, Montréal, Canada.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances.* Baltimore: Johns Hopkins University Press.

Holland, P., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum.

Hoyle, R. H., & Kenny, D. A. (1999). Sample size, reliability, and tests of statistical mediation. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 195–222). Thousand Oaks, CA: Sage.

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.

Kano, Y. (2001). Structural equation modeling for experimental data. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future—A Festschrift in honor of Karl Jöreskog* (pp. 381–402). Lincolnwood, IL: Scientific Software International.

Kaplan, D. (1999). An extension of the propensity score adjustment method for the analysis of group differences in MIMIC models. *Multivariate Behavioral Research, 34*, 467–492.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions.* Thousand Oaks, CA: Sage.

Kaplan, D., & George, R. (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 2*, 101–118.

Kinnunen, U., & Leskinen, E. (1989). Teacher stress during the school year: Covariance and mean structure analyses. *Journal of Occupational Psychology, 62*, 111–122.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York: Guilford.

Loehlin, J. C. (1998). *Latent variable models* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

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

Mueller, R. O. (1996). *Basic principles of structural equation modeling: An introduction to LISREL and EQS.* New York: Springer-Verlag.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557–585.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction.* Fort Worth, TX: Harcourt Brace.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.

Schumacker, R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling.* Mahwah, NJ: Lawrence Erlbaum.

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

Sörbom, D. (1978). An alternative to the methodology for analysis of covariance. *Psychometrika, 43*, 381–396.

Trochim, W. M. K. (2001). *The research methods knowledge base* (2nd ed.). Cincinnati, OH: Atomic Dog Publishing.

Chapter 18

Applying Dynamic Factor Analysis in Behavioral and Social Science Research

John R. Nesselroade

Peter C. M. Molenaar

18.1. Introduction

Behavioral and social sciences have witnessed a significant increase in the use of multivariate, correlational techniques in the past three decades. Much of this increase has involved applications of the common factor model, either in freestanding factor-analytic applications or as the so-called measurement model in structural equation modeling (SEM) applications. Of particular interest in this chapter is the application of the common factor model to the data obtained when individual participants (one or several) are measured many times on many variables to produce multivariate time series for the purpose of investigating patterns and interrelationships defined in intra-individual variability rather than the more common inter-individual differences framework.

Although it is true that the common factor model has been applied to multivariate time-series data for more than 50 years to represent more effectively process and other kinds of changes (e.g., Cattell, 1963; Cattell, Cattell, & Rhymer, 1947), powerful attempts to remedy the limitations of these earlier representations have only been made available relatively recently. These newer, more powerful conceptualizations appear capable of dealing much more effectively with some of the major challenges presented by multivariate time-series data (e.g., Browne & Nesselroade, in press; Hamaker, Dolan, & Molenaar, 2003; Hershberger, Molenaar, & Corneal, 1996; McArdle, 1982; Molenaar, 1985; Nesselroade, McArdle, Aggen, & Meyers, 2002; Nesselroade & Molenaar, 1999; Wood & Brown, 1994).

We are strongly committed to the idea that the study of changes in behavior with age, for example, is a much more complicated set of activities than the dominant, traditional methods of empirical research allow (e.g., Molenaar & Nesselroade, 2001; Nesselroade, 2002; Nesselroade & Ghisletta, 2000). It seems that mounting challenges to the dominant modes of thought regarding the conduct of developmental research have, at the very least, the potential of strengthening current

AUTHORS' NOTE: This work was supported by the Institute for Developmental and Health Research Methodology at the University of Virginia.

research approaches through self-examination and, at the very most, the possibility of leading to significant steps forward in the articulation and solution of problems.

One of the early, systematic attempts to model the nature of intra-individual change with factor analysis was called *P-technique* factor analysis (Cattell et al., 1947). It involved applying the common factor model to many repeated measurements of one individual with a battery of measures. Despite the fact that particular applications of the model have been controversial (Anderson, 1963; Cattell, 1963; Holtzman, 1963; Molenaar, 1985; Steyer, Ferring, & Schmitt, 1992), the logic underlying its use seems to be sound (e.g., Bereiter, 1963), and the results have been instrumental in the development of important lines of behavioral research such as the trait-state distinction (Cattell, 1957; Cattell & Scheier, 1961; Horn, 1972; Kenny & Zautra, 1995; Nesselroade & Ford, 1985; Steyer et al., 1992). These applications have helped to fuel a long-standing interest in intra-individual variability as a source of measurable individual differences (Baltes, Reese, & Nesselroade, 1977; Cattell, 1957; Eizenman, Nesselroade, Featherman, & Rowe, 1997; Fiske & Maddi, 1961; Fiske & Rice, 1955; Flugel, 1928; Kim, Nesselroade, & Featherman, 1996; Larsen, 1987; Magnusson, 1997; Nesselroade & Boker, 1994; Valsiner, 1984; Wessman & Ricks, 1966; Woodrow, 1932, 1945; Zevon & Tellegen, 1982). In this chapter, we will briefly examine some of the history and key issues of factor analyzing multivariate time series and, to exemplify the methods, will present analyses of some promising recent developments aimed at further improving such applications.

The analytical focus is on building a structural representation of patterns of within-person fluctuation of the variables over time. The intention of Cattell et al. (1947) in introducing this method of analysis was to discover "source traits" at the individual level. Cattell (1966) argued for some congruence between the way people change and the way they differ from each other. He declared that "we should be very surprised if the growth pattern in a trait bore *no* relation to its absolute pattern, as an individual differences structure" (p. 358), thus arguing for a similarity of patterns of intra-individual change and inter-individual differences (Hundleby, Pawlik, & Cattell, 1965). Bereiter (1963) noted that

> correlations between measures over individuals should bear some correspondence to correlations between measures for the same or randomly equivalent individuals over varying occasions, and the study of individual differences may be justifiable as an expedient substitute for the more difficult *P*-technique. (p. 15)

The flip side of this interpretation, which is not so often played, is that some of what are interpreted to be individual differences structures are actually intra-individual variability patterns that are asynchronous across persons and (perhaps erroneously) frozen in time by limiting the observations to a single measurement occasion. A key concern in either case is the degree of convergence between patterns of within-person change and among-person differences. Other authors have discussed this general topic under the label *ergodicity* (e.g., Jones, 1991; Molenaar, Huizenga, & Nesselroade, 2003; Molenaar & Nesselroade, 2001; Nesselroade & Molenaar, 1999). The essential point is that investigation of variation (and covariation) in the individual over time is a meaningful and necessary enterprise, the results of which need to be integrated into the larger framework of behavioral research and theory.

Since 1947, a large number of *P*-technique studies have been conducted (for reviews, see Jones & Nesselroade, 1990; Luborsky & Mintz, 1972). By the early 1960s, neither the proponents of *P*-technique factor analysis, such as Cattell, nor its critics, such as Anderson (1963) and Holtzman (1963), were satisfied with its ability to model the subtleties of intra-individual change. Consider, for example, the matter of influence exerted on observed variables by the unobserved factors. The common factor model, as traditionally applied to individual differences information (e.g., ability test scores), implies that individual differences in the underlying factors are responsible for individual differences in the observed variables. In *P*-technique applications, however, there are no individual differences because only one person is measured. Rather, the differences are in that individual's scores from one occasion to another (i.e., they are changes). Changes in the observed variables are modeled as having been produced by changes in the underlying factors.

The original *P*-technique model implies that the total influence of a factor on an observed variable is exerted instantaneously. Restricting the coupling between factors and variables in this way implies that, on those occasions when the factor score is extreme, the variable score will also tend toward the extreme, and on those occasions when the factor score is moderate, the variable score will also tend to be moderate. The model does not afford explicit representation of more intricate (read: realistic) patterns of influence of factors on variables such as persistence over time (e.g., the gradual dissipation or strengthening of the effects of

extreme factor scores on one occasion on the variables at a later occasion). Moreover, the pattern of effect gradients may differ with different observed variables. Statements of the type "I'm okay now, I just can't seem to stop shaking" illustrate the differences in the rates at which various components of a response pattern (e.g., self-reported internal state and objectively verifiable physical manifestations) return to equilibrium after the organism experiences an extreme in level of anxiety or fear. The basic *P*-technique model simply does not have the ability to represent the rich variety of relationships that we tend to associate with notions of process.

18.2. Dynamic Factor Models

Cattell (1963) himself called for refinements in the *P*-technique model that would allow representation of the effects exerted on the variables by the factors to dissipate or strengthen gradually over time rather than to be merely concurrent. For instance, he wanted it to be possible to represent delayed effects of the factors on the variables. It was not until the 1980s, however, that some key attempts to elaborate the *P*-technique factor model appeared that improved its capacity to represent change processes more veridically (e.g., Engle & Watson, 1981; Geweke & Singleton, 1981; McArdle, 1982; Molenaar, 1985). It is only in the past decade that the implementation of more promising, rigorous approaches to the study of intensively measured intra-individual variability in the single case via multivariate modeling has begun seemingly in earnest.

In the remainder of this chapter, we will briefly identify some alternative models and then focus on one exemplar of these approaches, labeled elsewhere the WNFS (white-noise factor score) model (Nesselroade et al., 2002) and the shock factor analysis model (SFA model) (Browne & Nesselroade, in press). The model was first articulated by Molenaar (1985). We will provide a description of the model and an example of fitting it to empirical data. In so doing, we want to draw further attention to the evolving interest in intra-individual variability phenomena in a wide variety of content domains and identify some research tools that seem particularly promising for rapid advance in these areas. To illustrate the applications concretely, the factor model will be presented, discussed, and compared in the context of fitting it to real data using standard structural equation modeling software (e.g., LISREL 8 by Jöreskog & Sörbom, 1993a).

Nesselroade et al. (2002) and Browne and Nesselroade (in press) distinguished between two promising dynamic factor models for multivariate time-series data. Both were developed explicitly to meet the shortcomings of traditional *P*-technique factor analysis. Browne and Nesselroade, as well as Hamaker et al. (2003), relate the dynamic factor models to the autoregressive and moving average models of time-series analysis.

One of the dynamic factor analysis models, explored by McArdle (1982), featured an auto- and cross-regressive structure at the level of the latent variables or factors. Thus, in this model, "continuity" or process resides at the more abstract level. The factors "drive" the manifest variables concurrently, as in traditional *P*-technique factor analysis, but the values of the factors at a given occasion (t) are influenced by the values of those same factors at earlier occasions ($t - 1, t - 2$, etc.). This allows for the preservation of the "signature" loadings of the factors on the variables in an invariant configuration while allowing for the kind of continuity we tend to associate with the term *process*. As was illustrated by Nesselroade et al. (2002), it is possible to estimate the lagged and cross-regressions of the factors under the constraints of factorial invariance over time in the loading patterns.

The second dynamic factor model we consider is the specification developed and presented by Molenaar (1985). This model, as will be seen in subsequent detail, is also geared toward lagged relationships but identifies them between earlier values of the factors and later values of the manifest variables. Thus, the current values of the variables (t) are "driven" by both the current values of the factors (t) and earlier values of the factors ($t - 1, t - 2$, etc.). The continuity or sense of process is in the patterns of concurrent and lagged loadings. The values of the factors at any given time are system inputs. These ideas will be discussed in detail in the following sections.

18.2.1. Technical Aspects of the Dynamic Factor Model

For this exposition, we will present the dynamic factor model in close parallel to the traditional common factor model with which most investigators in the field of behavioral research have some familiarity. This has the advantage, perhaps, of "demystifying" the procedure to some extent while also letting the reader capitalize on already familiar terms and con-

cepts. One of the easiest ways to grasp the implications of the common factor model is by apprehending the fundamental postulate and its derivative, the fundamental theorem, of common factor analysis. The fundamental postulate can be written as

$$\mathbf{z} = \mathbf{\Lambda} \cdot \boldsymbol{\eta} + \boldsymbol{\varepsilon},$$

where

- $\mathbf{z}$ is a centered (means of 0.0) vector variable of scores on p observed variables,
- $\mathbf{\Lambda}$ is a $p \times k$ matrix of loadings (regression-like weights) of p variables on k common factors,
- $\boldsymbol{\eta}$ is a vector variable of k common factor scores, and
- $\boldsymbol{\varepsilon}$ is a vector variable of p unique factor scores (specific factors + errors). Under the assumptions that

- the common and unique factors do not covary, and
- the covariance matrix of the unique factors ($\mathbf{\Psi}$) is diagonal,

the expected value of $z \cdot z'$ (the covariance matrix of the variables in $\mathbf{z}$) yields a version of the fundamental theorem of factor analysis, namely,

$$\mathbf{\Sigma} = \mathbf{\Lambda} \cdot \mathbf{\Phi} \cdot \mathbf{\Lambda}' + \mathbf{\Psi},$$

where

- $\mathbf{\Sigma}$ is the covariance matrix,
- $\mathbf{\Phi}$ is a factor covariance matrix,
- $\mathbf{\Psi}$ is a diagonal matrix of unique variances, and
- $\mathbf{\Lambda}$ is as defined above.

In operational terms, the fundamental postulate says that the observed scores are linear combinations of the common factors and one unique factor corresponding to each observed variable. The fundamental theorem says that the covariance matrix can be decomposed into the product of the factor loadings times the factor covariance matrix times the transpose of the factor loadings matrix plus the (diagonal) covariance matrix of the unique factors. Older, exploratory factor analysis routines (e.g., the principal axes) operated on the principal of finding sets of matrices that met this description and approximated the observed covariance matrix as closely as possible under the constraints of a particular algorithm. Newer methods (e.g., maximum likelihood techniques) involve estimating the elements in the matrices on the right-hand side of the equation according to the appropriate statistical estimation algorithm.

If one accepts and understands this basic factor analysis model, then the essential dynamic factor (DFA) model follows rather straightforwardly. Indeed, the DFA counterpart of the fundamental postulate of factor analysis is as follows:

$$\mathbf{z}(t) = \mathbf{\Lambda}(0) \cdot \boldsymbol{\eta}(t) + \mathbf{\Lambda}(1) \cdot \boldsymbol{\eta}(t-1) + \cdots + \mathbf{\Lambda}(s) \cdot \boldsymbol{\eta}(t-s) + \boldsymbol{\varepsilon}(t).$$

The big difference is that the terms of the DFA model include time indices.[1] These convey explicitly the time-contingent features of the DFA model. Thus, for example, the observed scores at time t are determined, in part, by the factor scores at earlier times. Clearly, the DFA model has an additional "job to do" over that required of the traditional common factor model. That additional job is to represent and account for lagged relationships in the data as well as the concurrent relationships accounted for by the traditional model.

To make clear how we are using the terms *concurrent* and *lagged relationships* in this context, consider the following empirical example.[2] The data consist of scores for one participant measured on each of six variables for 103 successive days, yielding a 6×103 score matrix. The six variables comprise six adjective rating scales: *active, lively, peppy, sluggish, tired,* and *weary.* These six scales were deliberately selected to mark two factors that might be called *energy* and *fatigue.* When these six scales are inter-correlated over the 103 occasions of measurement, a 6×6 correlation matrix is obtained (see Table 18.1).

One can clearly see in this pattern of intercorrelations a two-factor representation, with the three *energy* markers and the three *fatigue* markers clustering among themselves and the two sets being slightly to moderately negatively correlated. A relatively clean two-factor solution, with the two factors moderately negatively correlated, is a reasonable expectation for a factorial representation of this matrix.

1. Another difference is that Molenaar (1985) recommended truncating the latter terms in the model to some arbitrarily chosen level of precision. The implications of this are explored by Browne and Nesselroade (in press) and Nesselroade, McArdle, Aggen, and Meyers (2002) and will not be detailed here.

2. We wish to thank Dr. Michael A. Lebo for permission to use these data. Some of the same data were used for exemplary purposes by Nesselroade et al. (2002).

Table 18.1 Correlations Among Scales Over Time

	active	*lively*	*peppy*	*sluggish*	*tired*	*weary*
Lag 0						
active	1.00					
lively	.64	1.00				
peppy	.56	.41	1.00			
sluggish	−.48	−.34	−.42	1.00		
tired	−.47	−.42	−.47	.72	1.00	
weary	−.43	−.43	−.44	.64	.83	1.00

Table 18.2 Correlations Among Scales Over Time at a Lag 1 Occasion of Measurement

	active	*lively*	*peppy*	*sluggish*	*tired*	*weary*
active	.06	.03	.18	−.15	−.08	−.17
lively	.10	.08	.16	.03	.01	−.10
peppy	.02	.03	.15	−.07	.05	−.03
sluggish	−.03	.02	−.21	.40	.30	.28
tired	−.15	.02	−.22	.31	.25	.24
weary	−.07	.02	−.08	.27	.17	.21

Table 18.3 Correlations Among Scales Over Time at a Lag 2 Occasion of Measurement

	active	*lively*	*peppy*	*sluggish*	*tired*	*weary*
active	.03	.08	.09	−.18	−.16	−.12
lively	.13	.24	.13	−.15	−.23	−.19
peppy	.01	.07	.08	−.17	−.14	−.15
sluggish	−.10	.03	−.11	.35	.32	.19
tired	−.09	−.01	−.16	.30	.34	.23
weary	−.05	−.01	−.08	.26	.27	.10

The six scales can be lagged by one occasion of measurement on themselves and each other, with the resulting lagged correlations shown Table 18.2.

Note that this matrix is not symmetric because the correlation between *active* and *lively,* for example, is different when *active* lags *lively* by one occasion of measurement versus when *lively* lags *active* by one occasion. The diagonal elements of this matrix are the familiar autocorrelations (lag 1) of the variables. An inspection of the values indicates that the *fatigue* variables do exhibit some systematic predictability from time t to $t + 1$. This is not the case for the *energy* variables.

The six variables can be lagged by two occasions of measurement on themselves and each other, with the lagged correlations shown in Table 18.3.

Similarly to the lag 1 matrix above, this matrix is not symmetric either. Its diagonal elements are the familiar autocorrelations (lag 2) of the variables. There is evidence of "carryover" at two lags also, especially in the *fatigue* variables.

Additional lags are easily computed, but there are two main issues bearing on how many lags to consider in analyzing such a multivariate time series. One issue is determining when additional lags fail to yield useful information. There is no point to computing additional lags once the information has been exhausted.[3] The second is more of a structural consideration. Each lag "costs" an occasion of measurement because it results in an unpaired set of scores. For instance, if the time series is 100 occasions long, for lag 1, the observations at occasions $1, 2, 3, \ldots, 99$ are paired with the observations at occasions $2, 3, 4, \ldots, 100$,

3. Realize, however, that it is possible for relationships to be stronger at lag $t + k$ than they are at lag t if some cyclicity, for example, resides in the time series.

respectively. There are thus only 99 pairings of scores for computing the covariance or correlation. As the functional number of observations is decreasing, the number of variables involved is increasing (e.g., 10 variables at lag 0 becomes 20 variables at lag 1, 30 at lag 2, etc.). One can quickly reach an unfavorable ratio of variables to occasions of measurement, resulting eventually in singular covariance or correlation matrices.

A corresponding representation for the dynamic factor model takes the same form as the traditional common factor model given above but with the additional feature that lagged information is introduced explicitly into the model. For example, consider the covariance matrix, $\mathbf{\Sigma}$. It can be represented as a block-Toeplitz matrix as follows:

$$\begin{bmatrix} \mathbf{\Sigma}_{(0)} & & & & & \\ \mathbf{\Sigma}_{(1)} & \mathbf{\Sigma}_{(0)} & & & & \\ \mathbf{\Sigma}_{(2)} & \mathbf{\Sigma}_{(1)} & \mathbf{\Sigma}_{(0)} & & & \\ \cdots & \mathbf{\Sigma}_{(2)} & \mathbf{\Sigma}_{(1)} & \mathbf{\Sigma}_{(0)} & & \\ \mathbf{\Sigma}_{(t-1)} & \cdots & \mathbf{\Sigma}_{(2)} & \mathbf{\Sigma}_{(1)} & \mathbf{\Sigma}_{(0)} & \\ \mathbf{\Sigma}_{(t)} & \mathbf{\Sigma}_{(t-1)} & \cdots & \mathbf{\Sigma}_{(2)} & \mathbf{\Sigma}_{(1)} & \mathbf{\Sigma}_{(0)} \end{bmatrix},$$

where $\mathbf{\Sigma}_{(0)}$ represents the concurrent covariances (and variances) of the variables included in a time series. $\mathbf{\Sigma}_{(0)}$, of course, is a symmetric matrix. The submatrix $\mathbf{\Sigma}_{(1)}$ represents the asymmetric covariances of the variables lagged by one occasion of measurement on themselves. It is asymmetric because, as was pointed out earlier, the covariance of x lagged one step on y is not likely to be the same value as y lagged one step on x. Correspondingly, submatrix $\mathbf{\Sigma}_{(j)}$ represents the covariance matrix for the variables lagged on themselves on j occasions of measurement. The reason the block-Toeplitz matrix is used, despite the obvious redundancy, is that it provides an overall matrix that both contains all the lagged information and is symmetric, thus making it possible to use conventional software to fit a variety of models to it.

The corresponding DFA factor-loading pattern can be defined to include the lagged information as follows:

$$\begin{bmatrix} \Lambda(0) & \Lambda(1) & \Lambda(2) & \cdots & \Lambda(s-1) & \Lambda(s) & 0 & 0 & 0 & 0 & 0 \\ 0 & \Lambda(0) & \Lambda(1) & \Lambda(2) & \cdots & \Lambda(s-1) & \Lambda(s) & 0 & 0 & 0 & 0 \\ 0 & 0 & \Lambda(0) & \Lambda(1) & \Lambda(2) & \cdots & \Lambda(s-1) & \Lambda(s) & 0 & 0 & 0 \\ 0 & 0 & 0 & \Lambda(0) & \Lambda(1) & \Lambda(2) & \cdots & \Lambda(s-1) & \Lambda(s) & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \Lambda(0) & \Lambda(1) & \Lambda(2) & \cdots & \Lambda(s-1) & \Lambda(s) & 0 \\ 0 & 0 & 0 & 0 & 0 & \Lambda(0) & \Lambda(1) & \Lambda(2) & \cdots & \Lambda(s-1) & \Lambda(s) \end{bmatrix}$$

The corresponding factor covariance matrix can be defined as

$$\begin{bmatrix} \Phi & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \Phi & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \Phi & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \Phi & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & \Phi \end{bmatrix}.$$

And, finally, the corresponding unique covariance matrix, $\mathbf{\Psi}$, can be defined as

$$\begin{bmatrix} diag[C_\varepsilon(0)] & diag[C_\varepsilon(1)] & \cdots & diag[C_\varepsilon(s)] \\ diag[C_\varepsilon(1)] & diag[C_\varepsilon(0)] & \cdots & diag[C_\varepsilon(1)] \\ \cdots & \cdots & \cdots & \cdots \\ diag[C_\varepsilon(s)] & diag[C_\varepsilon(s-1)] & \cdots & diag[C_\varepsilon(0)] \end{bmatrix}.$$

Although the necessary redundancy makes these matrices more awkward looking than their usual counterparts, they can be arrayed in a manner completely analogous to the fundamental theorem of factor analysis presented above to reproduce the block-Toeplitz lagged covariance matrix as follows:

$$\mathbf{\Sigma} = \mathbf{\Lambda} \cdot \mathbf{\Phi} \cdot \mathbf{\Lambda}' + \mathbf{\Psi},$$

where

- $\mathbf{\Sigma}$ is the block-Toeplitz lagged covariance matrix,
- $\mathbf{\Lambda}$ is the super matrix of factor loadings,
- $\mathbf{\Phi}$ is the factor covariance super matrix, and
- $\mathbf{\Psi}$ is the super matrix of diagonal unique covariance submatrices, all as shown above. The point is that, multiplied and added as indicated just above, the matrices and their equation are counterparts to the ordinary common factor model representation.

In summary, then, in conducting a dynamic factor analysis, one achieves a set of factor loadings, a factor covariance or correlation matrix, and a uniqueness covariance matrix that satisfy the fundamental theorem of factor analysis by reproducing the input covariance or correlation matrix. The key difference is, however, that the input covariance or correlation matrix that is being accounted for contains the lagged as well as the concurrent relationships.

18.3. Example Application

In this section, we will demonstrate the fitting of a particular specification of the dynamic factor model. As was pointed out by Molenaar (1985; see also Nesselroade et al., 2002), many alternative specifications are possible and should be explored as suitable from the standpoint of the investigator's hypotheses regarding the nature of change processes.

Table 18.4 Factor Loadings for Lag 0 Correlation Matrix (*P*-Technique Factor Analysis)

	Lag 0 Factor Loading Pattern		
	Energy	*Fatigue*	*Uniqueness*
active	.86	.00	.27
lively	.72	.00	.48
peppy	.65	.00	.58
sluggish	.00	.76	.42
tired	.00	.95	.10
weary	.00	.87	.25

Table 18.5 Factor Intercorrelation for Lag 0 Analysis

	Factor	
	Energy	*Fatigue*
Energy	.00	−.63
Fatigue	−.63	1.00

To do so, we will fit a DFA model to the matrix given above. This will be done in three steps. First, the model will be fitted to the lag 0 correlations, as presented in Table 18.1. Next, the model will be fitted to the lag 0 and lag 1 correlations shown in Tables 18.1 and 18.2. Finally, the model will be fitted to the lag 0, lag 1, and lag 2 correlations given in Tables 18.1, 18.2, and 18.3.

18.3.1. Fitting Lag 0 Correlations

Fitting the lag 0 correlations is analogous to fitting the conventional *P*-technique factor model, depending on whether one explicitly models the autocorrelations of the unique parts. In this case we did not, although that will be done when the modeling includes the lag 1 and lag 2 correlations. The outcome of fitting a common factor model to this matrix is presented in Tables 18.4 and 18.5. Here one sees a clean two-factor solution, with the two factors negatively correlated.

Table 18.6 Factor Loadings for Lag 1 Correlation Matrix

	Factor Loadings				
Variable	*Energy Lag 0*	*Fatigue Lag 0*	*Energy Lag 1*	*Fatigue Lag 1*	*Uniqueness*
active	.80	.00	.24	.00	.28
lively	.71	.00	.08	.00	.45
peppy	.56	.00	.32	.00	.56
sluggish	.00	.72	.00	.37	.40
tired	.00	.92	.00	.31	.10
weary	.00	.82	.00	.30	.25

Table 18.7 Factor Intercorrelation for Lag 1 Analysis

	Factor	
	Energy	*Fatigue*
Energy	1.00	−.65
Fatigue	−.65	1.00

18.3.2. Fitting Lag 0 and Lag 1 Correlations

The outcome of fitting the dynamic factor model to the lag 0 and lag 1 correlations is shown in Tables 18.6 and 18.7. Now there are both concurrent factor loadings and lag 1 factor loadings to consider. The lag 1 loadings for the *fatigue* factor are relatively consistent and in keeping with the pattern of lagged correlations for these variables seen in Table 18.2.

Fitting Lag 0, 1, and 2 Correlations

Fitting the DFA model to the lag 0, 1, and 2 correlations yields the matrices shown in Tables 18.8 and 18.9. There is clearly still interesting information in the lag 2 loadings for the *fatigue* factor. The negative correlation between *energy* and *fatigue* has held steady from one analysis to another.

18.4. Technical Support

Many readers and potential appliers of dynamic factor models are not in a position to write their own model code to fit the models to data. There are two main tasks: (a) developing the lagged covariance or correlation matrix and (b) specifying and fitting a dynamic factor model to this lagged matrix. Since Molenaar (1985) presented the DFA model, Wood and Brown

Table 18.8 Factor Loadings for Lag 2 Correlation Matrix

	Factor Loadings						
Variable	*Energy Lag 0*	*Fatigue Lag 0*	*Energy Lag 1*	*Fatigue Lag 1*	*Energy Lag 2*	*Fatigue Lag 2*	*Uniqueness*
active	**.79**	.00	**.16**	.00	**.08**	.00	.28
lively	**.70**	.00	**.04**	.00	**.10**	.00	.45
peppy	**.57**	.00	**.16**	.00	**.22**	.00	.57
sluggish	.00	**.63**	.00	**.41**	.00	**.34**	.38
tired	.00	**.80**	.00	**.18**	.00	**.48**	.08
weary	.00	**.71**	.00	**.27**	.00	**.31**	.25

Table 18.9 Factor Intercorrelation for Lag 2 Analysis

	Factor	
	Energy	*Fatigue*
Energy	1.00	−.65
Fatigue	−.65	1.00

(1994) made available the SAS code for conducting these analyses.[4]

Nesselroade et al. (2002) provided some examples of LISREL (Jöreskog & Sörbom, 1993b) code for fitting dynamic factor specifications to lagged covariance or correlation matrices. Nesselroade et al. explored two basic specifications: (a) that by Molenaar (1985) and (b) one developed by McArdle (1982). Thus, for those wishing to try fitting DFA models to data, there are already several points of entry available.

18.5. Conclusion

The rigorous modeling of intra-individual variability in different content domains (e.g., cognition, temperament) is becoming more and more prevalent in the behavioral science literature. Compared to earlier approaches to modeling intra-individual variability and change such as *P*-technique factor analysis, dynamic factor analysis models promise much more effective means for extracting information regarding lagged relationships from multivariate time-series data. As the literature is also beginning to reflect, advances in the modeling of intra-individual variability promise to hold the key to developing more powerful nomothetic laws regarding behavior (e.g., Cattell, 1957; Nesselroade & Ford, 1985; Nesselroade & Molenaar, 1999; Shoda, Mischel, & Wright, 1994; Zevon & Tellegen, 1982). With the availability of tools such as the dynamic factor models discussed here, the systematic study of the nature of intra-individual variation becomes ever more productive and feasible. These improvements in modeling capabilities are timely adjuncts to stronger research designs (e.g., measurement "bursts") and more appropriately constructed measurement instruments (e.g., adaptive testing, change-sensitive measures).

The empirical example of short-term, affective variability presented here illustrates in considerable detail how one particular DFA model represents dynamic processes underlying time-series data. First, the asymmetrical cross-correlations at lag 2 (see Table 18.3) are indicative of lead-lag relationships similar to the well-known cross-lagged designs. For instance, the observed cross-correlation is $\mathbf{r}_{[Lively(t),Tired(t-2)]} = -.23$, whereas the counterpart cross-correlation is $\mathbf{r}_{[Tired(t),Lively(t-2)]} = -.01$. In the cross-lagged design framework, this could be interpreted as a (one-way) causal influence of *tired* on *lively.* In factor-analytic terms, such specific lead-lag relationships can only be captured by the dynamic factor models. State-space models cannot portray such specific lead-lag relationships because state-space models restrict the cross-correlations to symmetry.

Second, the dynamic factor analysis makes explicit that the *fatigue* factor series has a longer lasting after effect (dissipates more slowly) than the *energy* factor series in these data. That is, it takes at least 2 consecutive days, $t + 1$ and $t + 2$, before the prediction of the *fatigue* factor series from day t dissipates, whereas this is much less so for the *energy* factor series. In view of the substantial correlation (−.65) between the *fatigue* and *energy* factor series, the aftereffect of the *fatigue* factor series also will explain a substantial part of the sequential correlations of the observed scores loading on the *energy* series. Thus, this modeling approach

4. R. Nabors-Oberg and P. K. Wood have also written an Mx program for dynamic factor modeling. Copies can be obtained from woodpk@missouri.edu.

offers a way to "fill the gap" noted by Cattell (1963) and others regarding lagged relationships between factors and variables.

Third, using these techniques to examine between-individual differences, it becomes feasible to examine how intra-individual change patterns differ not only in terms of "dimensionality" (e.g., number of factors) but also in terms of temporal complexity or organization of behavior in time. Individuals who show very little occasion-to-occasion predictability no doubt differ in important ways from those who manifest a considerable amount of such "continuity" over time. Musher-Eizenman, Nesselroade, and Schmitz (2002), for example, reported such temporal organization differences in intra-individual variability in comparing low- versus high-performing schoolchildren.

Elsewhere, in arguing for informed rather than blind aggregation of multivariate time-series information across multiple individuals, Nesselroade and Molenaar (1999) presented an approach for identifying subsets of individuals whose lagged covariance functions are not different and therefore could be justifiably pooled for dynamic factor analysis. Pooling the lagged covariance information, when so justified, functionally increases the number of observations on which model parameters rest without unduly increasing the measurement burden on individual subjects. This has important implications for optimal design. J. L. Horn (personal communication, December 2000), among others, has raised the question that it might be better to look for similarities and differences at the level of individuals' factor models rather than their lagged covariance matrices because some, but not all, factors might be invariant over individuals. Certainly, there is some merit in this suggestion, and such alternatives can be easily explored within a given set of data provided there are ample occasions for fitting factor models to each individual's data.

In the firm belief that psychology is overdue for concentrating more heavily on concepts and methods reflecting less static and more dynamic properties, we have attempted to provide some insight into a small subset of the modeling possibilities and to point the way to applying such methods for those who are intrigued by the idea but unsure about how to move in that direction. In so doing, we have minimized or ignored a number of technical issues and potential problems (see, e.g., Nesselroade et al., 2002; Nesselroade & Molenaar, 2003). We believe, however, that learning to swim is better done in the water than on the land. With that in mind, we invite the reader to use the traditional factor analysis model as a springboard by which to leap into the deeper end of the pool wherein lie the challenging waters of intra-individual change and variability.

References

Anderson, T. W. (1963). The use of factor analysis in the statistical analysis of multiple time series. *Psychometrika, 28,* 1–24.

Baltes, P. B., Reese, H. W., & Nesselroade, J. R. (1977). *Life-span developmental psychology: Introduction to research methods.* Monterey, CA: Brooks/Cole.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 3–20). Madison: University of Wisconsin Press.

Browne, M. W., & Nesselroade, J. R. (in press). Representing psychological processes with dynamic factor models: Some promising uses and extensions of ARMA time series models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Psychometrics: A festschrift to Roderick P. McDonald.*

Cattell, R. B. (1957). *Personality and motivation structure and measurement.* New York: World Book Co.

Cattell, R. B. (1963). The structuring of change by *P*-technique and incremental *R*-technique. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 167–198). Madison: University of Wisconsin Press.

Cattell, R. B. (1966). Patterns of change: Measurement in relation to state dimension, trait change, lability, and process concepts. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 355–402). Chicago: Rand McNally.

Cattell, R. B., Cattell, A. K. S., & Rhymer, R. M. (1947). *P*-technique demonstrated in determining psychophysical source traits in a normal individual. *Psychometrika, 12*, 267–288.

Cattell, R. B., & Scheier, I. H. (1961). *The meaning and measurement of neuroticism and anxiety.* New York: Ronald Press.

Eizenman, D. R., Nesselroade, J. R., Featherman, D. L., & Rowe, J. W. (1997). Intra-individual variability in perceived control in an elderly sample: The MacArthur Successful Aging Studies. *Psychology and Aging, 12*, 489–502.

Engle, R., & Watson, M. (1981). A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association, 76*, 774–781.

Fiske, D. W., & Maddi, S. R. (Eds.). (1961). *Functions of varied experience.* Homewood, IL: Dorsey.

Fiske, D. W., & Rice, L. (1955). Intra-individual response variability. *Psychological Bulletin, 52*, 217–250.

Flugel, J. C. (1928). Practice, fatigue, and oscillation. *British Journal of Psychology, Monograph Supplement, 4,* 1–92.

Geweke, J. F., & Singleton, K. J. (1981). Maximum likelihood "confirmatory" factor analysis of economic time series. *International Economic Review, 22*, 37–54.

Hamaker, E. L., Dolan, C. V., & Molenaar, P. (2003). *Statistical modeling of the individual: Rational and application of multivariate time series analysis.* Unpublished manuscript, University of Amsterdam, Department of Psychology.

Hershberger, S. L., Molenaar, P. C. M., & Corneal, S. E. (1996). A hierarchy of univariate and multivariate time series models.

In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 159–194). Mahwah, NJ: Lawrence Erlbaum.

Holtzman, W. H. (1963). Statistical models for the study of change in the single case. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 199–211). Madison: University of Wisconsin Press.

Horn, J. L. (1972). State, trait, and change dimensions of intelligence. *British Journal of Educational Psychology, 42(2)*, 159–185.

Hundleby, J. D., Pawlik, K., & Cattell, R. B. (1965). *Personality factors in objective test devices.* San Diego: R. Knapp.

Jones, C. J., & Nesselroade, J. R. (1990). Multivariate, replicated, single-subject designs and *P*-technique factor analysis: A selective review of the literature. *Experimental Aging Research, 16*, 171–183.

Jones, K. (1991). The application of time series methods to moderate span longitudinal data. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 75–87). Washington, DC: American Psychological Association.

Jöreskog, K. G., & Sörbom, D. (1993a). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Hillsdale, NJ: Lawrence Erlbaum.

Jöreskog, K. G., & Sörbom, D. (1993b). *LISREL 8 user's reference guide.* Chicago: Scientific Software International.

Kenny, D. A., & Zautra, A. (1995). The trait-state error model for multiwave data. *Journal of Consulting and Clinical Psychology, 63*, 52–59.

Kim, J. E., Nesselroade, J. R., & Featherman, D. L. (1996). The state component in self-reported world views and religious beliefs in older adults: The MacArthur Successful Aging Studies. *Psychology and Aging, 11*, 396–407.

Larsen, R. J. (1987). The stability of mood variability: A spectral analytic approach to daily mood assessments. *Journal of Personality and Social Psychology, 52*, 1195–1204.

Luborsky, L., & Mintz, J. (1972). The contribution of *P*-technique to personality, psychotherapy, and psychosomatic research. In R. M. Dreger (Ed.), *Multivariate personality research: Contributions to the understanding of personality in honor of Raymond B. Cattell* (p. 387–410). Baton Rouge, LA: Claitor's Publishing Division.

Magnusson, D. (1997). The logic and implications of a person approach. In R. B. Cairns, L. R. Bergman, & J. Kagan (Eds.), *The individual as a focus in developmental research* (pp. 33–63). Thousand Oaks, CA: Sage.

McArdle, J. J. (1982). *Structural equation modeling of an individual system: Preliminary results from "A case study in episodic alcoholism."* Unpublished manuscript, University of Denver, Department of Psychology.

Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika, 50*(2), 181–202.

Molenaar, P. C. M., Huizenga, H. M., & Nesselroade, J. R. (2003). The relationship between the structure of interindividual and intraindividual variability: A theoretical and empirical vindication of developmental systems theory. In U. M. Staudinger & U. Lindenberger (Eds.), *Understanding human development: Dialogues with lifespan psychology* (pp. 339–360). Norwell, MA: Kluwer Academic.

Molenaar, P. C. M., & Nesselroade, J. R. (2001). Rotation in the dynamic factor modeling of multivariate stationary time series. *Psychometrika, 66*, 99–107.

Musher-Eizenman, D. R., Nesselroade, J. R., & Schmitz, B. (2002). Perceived control and academic performance: A comparison of high- and low-performing children on within-person change patterns. *International Journal of Behavioral Development, 26*, 540–547.

Nesselroade, J. R. (2002). Elaborating the different in differential psychology. *Multivariate Behavioral Research, 37*(4), 543–561.

Nesselroade, J. R., & Boker, S. M. (1994). Assessing constancy and change. In T. Heatherton & J. Weinberger (Eds.), *Can personality change?* (pp. 121–147). Washington, DC: American Psychological Association.

Nesselroade, J. R., & Ford, D. H. (1985). *P*-technique comes of age: Multivariate, replicated, single-subject designs for research on older adults. *Research on Aging, 7*, 46–80.

Nesselroade, J. R., & Ghisletta, P. (2000). Beyond static concepts in modeling behavior. In L. R. Bergman & R. B. Cairns (Eds.), *Developmental science and the holistic approach* (pp. 121–135). Mahwah, NJ: Lawrence Erlbaum.

Nesselroade, J. R., McArdle, J. J., Aggen, S. H., & Meyers, J. M. (2002). Alternative dynamic factor models for multivariate time-series analyses. In D. M. Moskowitz & S. L. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data: Advances and techniques* (pp. 235–265). Mahwah, NJ: Lawrence Erlbaum.

Nesselroade, J. R., & Molenaar, P. (2003). Quantitative models for developmental processes. In J. Valsiner & K. Connolly (Eds.), *Handbook of developmental psychology* (pp. 622–639). London: Sage.

Nesselroade, J. R., & Molenaar, P. C. M. (1999). Pooling lagged covariance structures based on short, multivariate time-series for dynamic factor analysis. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 223–250). Thousand Oaks, CA: Sage.

Shoda, Y., Mischel, W., & Wright, J. C. (1994). Intraindividual stability in the organization and patterning of behavior: Incorporating psychological situations into the idiographic analysis of behavior. *Journal of Personality and Social Psychology, 67*, 674–687.

Steyer, R., Ferring, D., & Schmitt, M. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment, 8*, 79–98.

Valsiner, J. (1984). Two alternative epistemological frameworks in psychology: The typological and variational modes of thinking. *Journal of Mind and Behavior, 5*(4), 449–470.

Wessman, A. E., & Ricks, D. F. (1966). *Mood and personality.* New York: Holt, Rinehart, & Winston.

Wood, P., & Brown, D. (1994). The study of intraindividual differences by means of dynamic factor models: Rationale, implementation, and interpretation. *Psychological Bulletin, 116*(1), 166–186.

Woodrow, H. (1932). Quotidian variability. *Psychological Review, 39*, 245–256.

Woodrow, H. (1945). Intelligence and improvement in school subjects. *Journal of Educational Psychology, 36*, 155–166.

Zevon, M., & Tellegen, A. (1982). The structure of mood change: Idiographic/nomothetic analysis. *Journal of Personality and Social Psychology, 43*(1), 111–122.

Chapter 19

LATENT VARIABLE ANALYSIS

Growth Mixture Modeling and Related Techniques for Longitudinal Data

BENGT MUTHÉN

19.1. INTRODUCTION

This chapter gives an overview of recent advances in latent variable analysis. Emphasis is placed on the strength of modeling obtained by using a flexible combination of continuous and categorical latent variables. To focus the discussion and make it manageable in scope, analysis of longitudinal data using growth models will be considered. Continuous latent variables are common in growth modeling in the form of random effects that capture individual variation in development over time. The use of categorical latent variables in growth modeling is, in contrast, perhaps less familiar, and new techniques have recently emerged. The aim of this chapter is to show the usefulness of growth model extensions using categorical latent variables. The discussion also has implications for latent variable analysis of cross-sectional data.

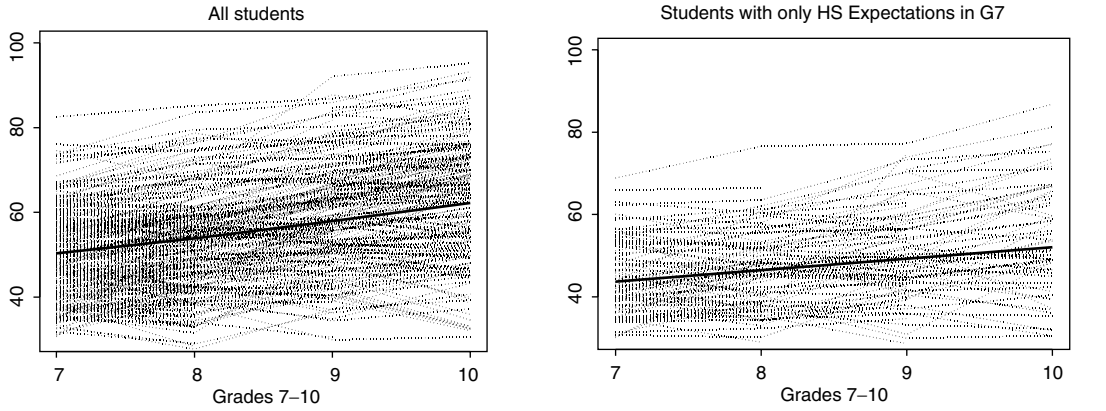
The chapter begins with two major parts corresponding to continuous outcomes versus categorical outcomes. Within each part, conventional modeling using continuous latent variables will be described first, followed by recent extensions that add categorical latent variables. This covers growth mixture modeling, latent class growth analysis, and discrete-time survival analysis. Two additional sections demonstrate further extensions. Analysis of data with strong floor effects gives rise to modeling with an outcome that is part binary and part continuous, and data obtained by cluster sampling give rise to multilevel modeling. All models fit into a general latent variable framework implemented in the Mplus program (Muthén & Muthén, 1998–2003). For overviews of this modeling framework, see Muthén (2002) and Muthén and Asparouhov (2003a, 2003b). Technical aspects are covered in Asparouhov and Muthén (2003a, 2003b).

19.2. CONTINUOUS OUTCOMES: CONVENTIONAL GROWTH MODELING

In this section, conventional growth modeling will be briefly reviewed as a background for the more general growth modeling to follow. To prepare for this

AUTHOR'S NOTE: The research was supported under grant K02 AA 00230 from NIAAA. I thank the Mplus team for software support, Karen Nylund and Frauke Kreuter for research assistance, and Tihomir Asparouhov for helpful comments. Please send correspondence to bmuthen@ucla.edu.

Figure 19.1 LSAY Math Achievement in Grades 7 to 10



transition, the multilevel and mixed linear modeling representation of conventional growth modeling will be related to representations using structural equation modeling and latent variable modeling.

To introduce ideas, consider an example from mathematics achievement research. The Longitudinal Study of Youth (LSAY) is a national sample of mathematics and science achievement of students in U.S. public schools (Miller, Kimmel, Hoffer, & Nelson, 2000). The sample contains 52 schools with an average of about 60 students per school. Achievement scores were obtained by item response theory equating. There were about 60 items per test with partial item overlap across grades. Tailored testing was used so that test results from a previous year influenced the difficulty level of the test of a subsequent year. The LSAY data used here are from Cohort 2, containing a total of 3,102 students followed from Grade 7 to Grade 12 starting in 1987. Individual math trajectories for Grades 7 through 10 are shown in Figure 19.1.

The left-hand side of Figure 19.1 shows typical trajectories from the full sample of students. Approximately linear growth over the grades is seen, with the average linear growth shown as a bold line. Conventional growth modeling is used to estimate the average growth, the amount of variation across individuals in the growth intercepts and slopes, and the influence of covariates on this variation. The right-hand side of Figure 19.1 uses a subset of students defined by one such covariate, considering students who, in seventh grade, expect to get only a high school degree. It is seen that the intercepts and slopes are considerably lower for this group of low-expectation students.

A conventional growth model is formulated as follows for the math achievement development related to educational expectations. For ease of transition between modeling traditions, the multilevel notation of Raudenbush and Bryk (2002) is chosen. For time point t and individual i, consider the variables

y_{ti} = repeated measures on the outcome (e.g., math achievement),
a_{1ti} = time-related variable (time scores) (e.g., Grades 7–10),
a_{2ti} = time-varying covariate (e.g., math course taking),
x_i = time-invariant covariate (e.g., Grade 7 expectations),

and the two-level growth model,

$$\text{Level 1: } y_{ti} = \pi_{0i} + \pi_{1i}\, a_{1ti} + \pi_{2ti}\, a_{2ti} + e_{ti}, \quad (1)$$

$$\text{Level 2: } \begin{cases} \pi_{0i} = \beta_{00} + \beta_{01} x_i + r_{0i} \\ \pi_{1i} = \beta_{10} + \beta_{11} x_i + r_{1i}. \\ \pi_{2i} = \beta_{20} + \beta_{21} x_i + r_{2i} \end{cases} \quad (2)$$

Here, π_{0i}, π_{1i}, and π_{2i} are random intercepts and slopes varying across individuals. The residuals e, r_0, r_1, and r_2 are assumed normally distributed with zero means and uncorrelated with a_1, a_2, and w. The Level 2 residuals r_0, r_1, and r_2 are possibly correlated but uncorrelated with e. The variances of e_t are typically assumed equal across time and uncorrelated across time, but both of these restrictions can be relaxed.[1]

1. The model may alternatively be expressed as a mixed linear model relating y directly to a_1, a_2, and x by inserting (2) into (1). Analogous to a two-level regression, when either a_{ti} or π_{2ti} varies across i, there is variance heteroscedasticity for y given covariates and therefore not a single covariance matrix for model testing.

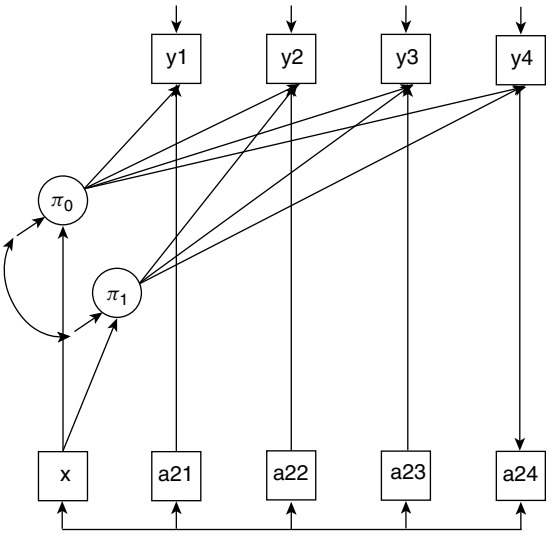
The growth model above is presented as a multilevel, random-effects model. Alternatively, the growth model can be seen as a latent variable model, where the random effects π_0, π_1, and π_2 are latent variables. The latent variables π_0, π_1 will be called growth factors and are of key interest here. As will be shown, the latent variable framework views growth modeling as a single-level analysis. A special case of latent variable modeling is obtained via mean- and covariance-structure structural equation modeling (SEM). Connections between multilevel, latent variable, and SEM growth analysis will now be briefly reviewed.

When there are individually varying times of observation, a_{1ti} in (1) varies across i for given t. In this case, a_{1ti} may be read as data. This means that in conventional multilevel modeling, π_{1i} is a (random) slope for the variable a_{1ti}. When $a_{1ti} = a_{1t}$ for all t values, a reverse view can be taken. In SEM, each a_{1t} is treated as a parameter, where a_{1t} is a slope multiplying the (latent) variable π_{1i}. For example, accelerated or decelerated growth at a third time point may be captured by $a_{1t} = (0, 1, a_3)$, where a_3 is estimated.[2]

Typically in conventional multilevel modeling, the random slope π_{2ti} (1) for the time-varying covariate a_{2t} is taken to be constant across time, $\pi_{2ti} = \pi_{2i}$. It is possible to allow variation across both t and i, although this may be difficult to find evidence for in data. In SEM, however, the slope is not random, $\pi_{2ti} = \pi_{2t}$, because conventional covariance structure modeling cannot handle products of latent and observed continuous variables.

In the latent variable modeling and SEM frameworks, the distinction between Level 1 and Level 2 is not made, but a regular (single-level) analysis is done. This is because the modeling framework considers the T-dimensional vector $\mathbf{y} = (y_1, y_2, \ldots, y_T)'$ as a multivariate outcome, accounting for the correlation across time by the same random effects influencing each of the variables in the outcome vector. In contrast, multilevel modeling typically views the outcome as univariate, accounting for the correlation across time by the two levels of the model. From the latent variable and SEM perspective, (1) may be seen as the measurement part of the model where the growth factors π_0 and π_1 are measured by the multiple indicators y_t. In (2), the structural part of the model relates growth factors and random slopes to other variables. A growth model diagram corresponding to the SEM perspective is shown in Figure 19.2, where circles correspond to latent variables and boxes correspond to observed variables.

Figure 19.2 Growth Model Diagram

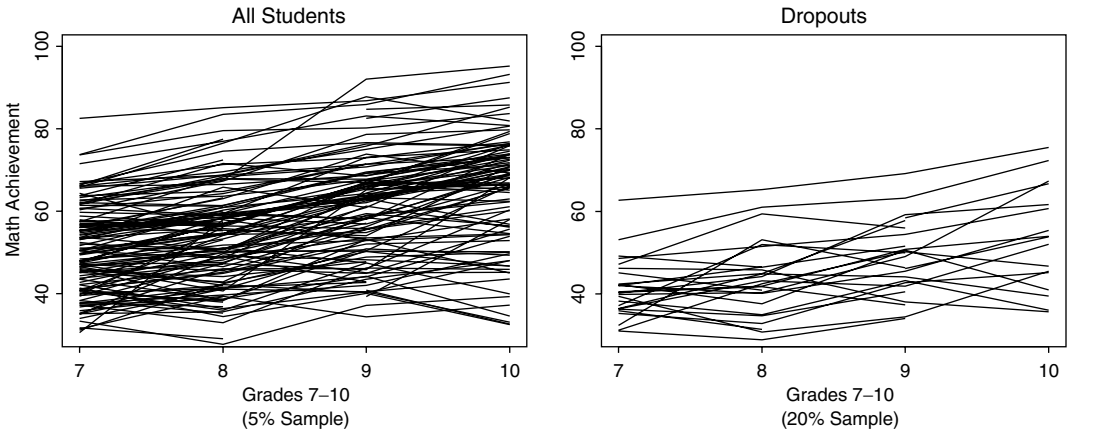


There are several advantages of placing the growth model in an SEM or latent variable context. Growth factors may be regressed on each other—for example, studying growth while controlling for not only observed covariates but also the initial status growth factor. Or, a researcher may want to study growth in a latent variable construct measured with multiple indicators. Other advantages of growth modeling in a latent variable framework include the ease with which to carry out analysis of multiple processes, both parallel in time and sequential, as well as multiple groups with different covariance structures. More generally, the growth model may be only a part of a larger model, including, for instance, a factor analysis measurement part for covariates measured with errors, a mediational path analysis part for variables influencing the growth factors, or a set of variables that are influenced by the growth process (distal outcomes).

The more general latent variable approach to growth goes beyond the SEM approach by handling (1) as stated (i.e., allowing individually varying times of observation and random slopes for time-varying covariates). Here, $a_{1ti} = a_{1t}$ and $\pi_{2ti} = \pi_{2t}$ are allowed as special cases. The latent variable approach thereby combines the strength of conventional multilevel modeling and SEM. An overview showing the advantages of this combined type of modeling is given in Muthén

2. When choosing $a_{11} = 0, \pi_{0i}$ is defined as the initial status of the growth process. In multilevel analysis, a_{1ti} is often centered at the mean (e.g., to avoid collinearity when using quadratic growth), whereas in SEM, parameters may get highly correlated.

Figure 19.3 LSAY Math Achievement in Grades 7 to 10 and High School Dropout



and Asparouhov (2003a), and a technical background is given in Asparouhov and Muthén (2003a). In addition, general latent variable modeling allows modeling with a combination of continuous and categorical latent variables to more realistically represent longitudinal data. This aspect is the focus of the current chapter.

19.3. Continuous Outcomes: Growth Mixture Modeling

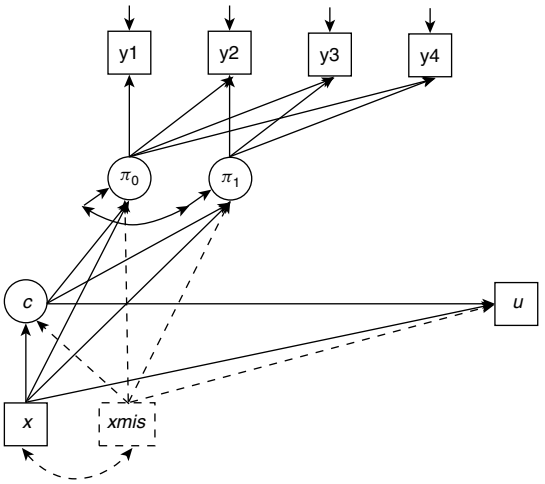
The model in (1) and (2) has two key features. On one hand, it allows individual differences in development over time because the growth intercept π_{0i} and growth slope π_{1i} vary across individuals, resulting in individually varying trajectories for y_{ti} over time. This heterogeneity is captured by random effects (i.e., continuous latent variables). On the other hand, it assumes that all individuals are drawn from a single population with common population parameters. Growth mixture modeling relaxes the single population assumption to allow for parameter differences across unobserved subpopulations. This is accomplished using latent trajectory classes (i.e., categorical latent variables). This implies that instead of considering individual variation around a single mean growth curve, the growth mixture model allows different classes of individuals to vary around different mean growth curves. The combined use of continuous and categorical latent variables provides a very flexible analysis framework. Growth mixture modeling was introduced in Muthén and Shedden (1999) with extensions and overviews in Muthén and Muthén (1998–2003) and Muthén (2001a, 2001b, 2002).

Consider again the math achievement example and the math development shown in the right-hand part of Figure 19.3. This is the development for individuals who are later classified as having dropped out by Grade 12. Note that while Figure 19.1 considers an antecedent of development, Grade 7 expectations, Figure 19.3 considers a consequence of development, high school dropout. It is seen that, with a few exceptions, the high school dropouts typically have a lower starting point in Grade 7 and grow slower than the average students in the left-hand part of the figure. This suggests that there might be an unobserved subpopulation of students who, in Grades 7 through 10, show poor math development and who have a high risk for dropout. In educational dropout research, such a subpopulation is often referred to as "disengaged," where disengagement has many hypothesized predictors. The subpopulation membership is not known during Grades 7 through 10 but is revealed when students drop out of high school. The subpopulation membership can, however, be inferred from the Grade 7 through 10 math achievement development.

19.3.1. Growth Mixture Model Specification

To introduce growth mixture modeling (GMM), consider a latent categorical variable c_i representing the unobserved subpopulation membership for student i, $c_i = 1, 2, \ldots, K$. Here, c will be referred to as a latent class variable or, more specifically, a trajectory class variable. Assume tentatively that in the math achievement example, $K = 2$, representing a disengaged class ($c = 1$) and a normative class ($c = 2$). An example of the different parts of the model is shown

Figure 19.4 GGMM Diagram



in the model diagram in Figure 19.4. The model has covariates x and *xmis*, a latent class variable c, repeated continuous outcomes y, and a distal dichotomous outcome u. For simplicity, time-varying covariates are not included in this example. The covariate x influences c and has direct effects on the growth factors π_0 and π_1, as well as a direct effect on u. In this section, the *xmis* covariate will be assumed to have no role in the model. Its effects will be studied in later sections.

Consider first the prediction of the latent class variable by the covariate x using a multinomial logistic regression model for K classes,

$$P(c_i = k|x_i) = \frac{e^{\gamma_{0k}+\gamma_{1k}x_i}}{\sum_{s=1}^{K} e^{\gamma_{0s}+\gamma_{1s}x_i}}, \qquad (3)$$

with the standardization $\gamma_{0K} = 0$, $\gamma_{1K} = 0$. With a binary $c(c = 1, 2)$, this gives

$$P(c_i = 1|x_i) = \frac{1}{1 + e^{-l_i}}, \qquad (4)$$

where l is the logit (i.e., the log odds),

$$\log[P(c_i = 1|x_i)/P(c_i = 2|x_i)] = \gamma_{01} + \gamma_{11}\, x_i, \qquad (5)$$

so that γ_{11} is the increase in the log odds of being in the disengaged versus the normative class for a unit increase in x. For example, assume that x is dichotomous and scored 0, 1 for females versus males. From (4), it follows that $e^{\gamma_{11}}$ is the odds ratio for being in the disengaged class versus the normative class when comparing males to females. For example, $\gamma_{11} = 1$ implies that the odds of being in the disengaged class versus the normative class is $e^1 = 2.72$ times higher for males than females.

Generalizing (1) and (2), GMM considers a separate growth model for each of the two latent classes. Key differences across classes are typically found in the fixed effects β_{00}, β_{10}, and β_{20} in (2). For example, the disengaged class would have lower β_{00} and β_{10} values (i.e., lower means) than the normative class. Class differences may also be found in the covariate influence, with class-varying β_{01}, β_{11}, and β_{21}. In addition, class-varying variances and covariances for the r residuals may be found. In (1), the type of growth function for Level 1 is perhaps different across class as well. For example, although the disengaged class may be well represented by linear growth, the normative class may show accelerated growth over some of the grades (e.g., calling for a quadratic growth curve). Here, the variance for the e residual may also be class varying.

The basic GMM can be extended in many ways. One important extension is to include an outcome that is predicted from the growth. Such an outcome is often referred to as a *distal outcome*, whereas in this context, the growth outcomes are referred to as *proximal outcomes.* Dropping out of high school is an example of such a distal outcome in the math achievement context. Given that the growth is succinctly summarized by the latent trajectory class variable, it is natural to let the latent trajectory class variable predict the distal outcome. With the example of a dichotomous distal outcome u scored 0, 1, this model part is given as a logistic regression with covariates c and x,

$$P(u_i = 1|c_i = k, x_i) = \frac{1}{1 + e^{\tau_k - \kappa_k x_i}}, \qquad (6)$$

where the main effect of c is captured by the class-varying thresholds τ_k (an intercept with its sign reversed), and κ_k is a class-varying slope for x. For each class, the same odds ratio interpretation given above can be applied also here. Model extensions of this type will be referred to as general growth mixture modeling (GGMM).

19.3.1.1. Latent Class Growth Analysis

A special type of growth mixture model has been studied by Nagin and colleagues (see, e.g., Nagin, 1999; Nagin & Land, 1993; Roeder, Lynch, & Nagin, 1999) using the SAS procedure PROC TRAJ (Jones, Nagin, & Roeder, 2001). See also the 2001 special issue of *Sociological Methods & Research* (Land, 2001). The models studied by Nagin are characterized by having zero variances and covariances for r in (2); that is, individuals within a class are treated as

homogeneous with respect to their development.[3] Analysis with zero growth factor variances and covariances will be referred to as latent class growth analysis (LCGA) in this chapter. As will be discussed in the context of categorical outcomes, the term LCGA is motivated by it being more similar to latent class analysis than growth modeling.

LCGA may be useful in two major ways. First, LCGA may be used to find cut points on the GMM growth factors. A k-class GMM that has within-class variation may have a model fit similar to that of a $k + m$-class LCGA for some $m > 0$. The extra m classes may be a way to objectively find cut points in the within-class variation of a GMM to the extent that such further grouping is substantively useful. This situation is similar to the relationship between factor analysis and latent class analysis, as discussed in Muthén (2001a), where latent classes of individuals were identified along factor dimensions. From a substantive point of view, however, this poses the challenge of how to determine which latent classes represent fundamentally different trajectories and which represent only minor variations. Second, as pointed out in Nagin's work, the latent classes of LCGA may be viewed as producing a nonparametric representation of the distribution of the growth factors, resulting in a semi-parametric model. This view will be further discussed in the next section.

LCGA is straightforward to specify within the general Mplus framework. The zero variance restriction makes LCGA easy to work with, giving relatively fast convergence. If the model fits the data, the simplicity can be a practically useful feature. Also, LCGA can be used in conjunction with GMM as a starting point for analyses. Section 19.3.4.1 discusses the use of LCGA on data that have been generated by a GMM in which covariates have direct influence on the growth factors. This misapplication leads to serious distortions in the formation of the latent classes.

19.3.1.2. Nonparametric Estimation of Latent Variable Distributions

In the GMM described earlier, the normality assumption for the residuals on Level 1 and Level 2 is applied to each class. Within class, the latent variables of π_0, π_1, and π_2 of (2) may have a nonnormal distribution due to the influence of a possibly nonnormal x covariate, and the distribution of y in (1) is further influenced by possibly nonnormal Level 1 covariates. This implies that the distribution of the outcomes y can be nonnormal within class. Strong nonnormality for y is obtained when latent classes with different means and variances are mixed together.

The normality assumption for the residuals is not innocuous in mixture modeling. Alternative distributions would result in somewhat different latent class formations. The literature on nonparametric estimation of random-effect distributions reflects such a concern, especially with categorical and count outcomes already in nonmixture models. Maximum likelihood estimation for logistic models with random effects typically uses Gauss-Hermite quadrature to integrate out the normal random effects. The quadrature uses fixed nodes and weights for a set of quadrature points. As pointed out by Aitkin (1999), a more flexible distributional form is obtained if both the nodes and the weights are estimated, and this approach is an example of mixture modeling. The mixture modeling approximation to a continuous random-effect distribution, such as a random intercept growth factor, is illustrated in Figure 19.5 using an approximately normal distribution as well as a skewed distribution. In both cases, five nodes and weights are used, corresponding to a mixture with five latent classes. Aitkin argues that the mixture approach may be particularly suitable with categorical outcomes in which the usual normality assumption for the random effects has scarce empirical support. For an overview of related work, see also Heinen (1996); for a more recent discussion in the context of logistic growth analysis, see Hedeker (2000).

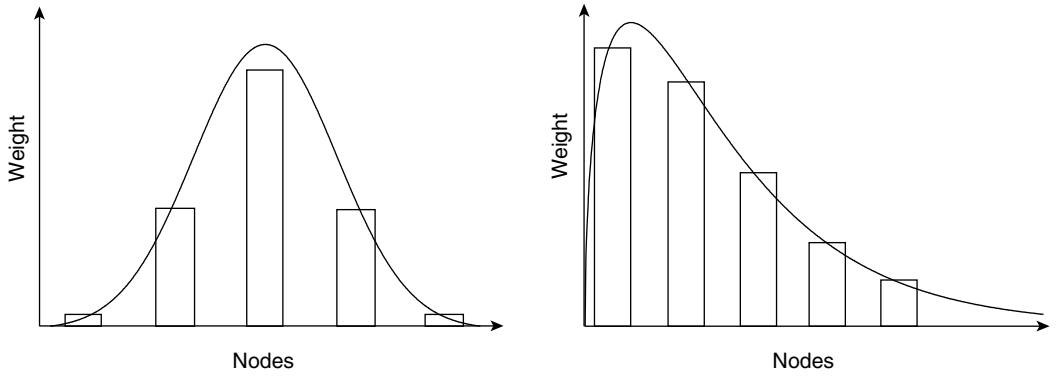
The Mplus latent variable framework can be used for this type of nonparametric approach. Corresponding to Figure 19.5, a random intercept growth factor distribution can be represented by a five-class mixture. Here, the estimation of the nodes is obtained by estimating the growth factor means in the different classes, and the estimation of the weights is obtained by estimating the class probabilities (the growth factor variance parameter is held fixed at zero). If a single-class model is considered, the other parameters of the model are held equal across classes; otherwise, they are not.

19.3.1.3. Growth Mixture Modeling Estimation

The growth mixture model can be estimated by maximum likelihood using an EM algorithm. For a given solution, each individual's probability of

3. Nagin's work focuses on count data using Poisson distributions. As discussed in later sections, modeling with count outcomes and categorical outcomes can also use nonzero variance for r.

Figure 19.5 Random-Effects Distributions Represented by Mixtures



membership in each class can be estimated, as well as the individual's score on the growth factors π_{0i} and π_{1i}. Measures of classification quality can be considered based on the individual class probabilities, such as entropy. This has been implemented in the Mplus program (Muthén & Muthén, 1998–2003). Technical aspects of the modeling, estimation, and testing are given in Technical Appendix 8 of the *Mplus User's Guide* (Muthén & Muthén, 1998–2003), Muthén and Shedden (1999), and Asparouhov and Muthén (2003a, 2003b). Missing data on y are handled using MAR. Muthén, Jo, and Brown (2003) discuss nonignorable missing data modeling using missing data indicators. As with mixture modeling in general, local optima are often encountered in the likelihood. This phenomenon is well known, for example, in latent class analysis, particularly in models with many classes and data that carry limited information about the class membership. Because of this, the use of several different sets of starting values is recommended, and this is automated in Mplus.

19.3.1.4. The LSAY Example

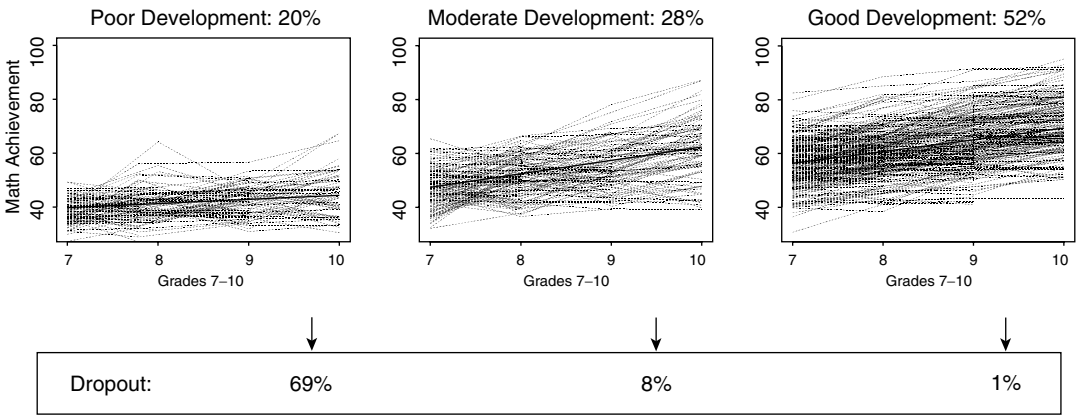
To conclude this section in a concrete way using the LSAY math achievement data, a brief preview of the analyses in Section 3.5 is of interest. Figure 19.6 shows that three latent trajectory classes are found, including their class probabilities, the mean trajectory and individual variation for each class, and the probability of dropping out of high school for each class. Of the students, 20% are found to belong to a disengaged class with poor math development. Membership in the disengaged class dramatically enhances the risk of dropping out of high school, raising the dropout percentage from 1% and 8% to 69%. Section 3.5 presents the covariates predicting latent trajectory class membership, and it is found that having low educational expectations and dropout thoughts already by Grade 7 are key predictors.

Before going through the analysis steps for the LSAY math achievement example, model interpretation, estimation, and model selection procedures will be discussed. Latent variable modeling requires good analysis strategies, and this is even more true in the framework of growth mixture modeling, where both continuous and categorical latent variables are used. Many statistical procedures have been suggested within the related statistical area of finite mixture modeling (see, e.g., McLachlan & Peel, 2000), and some key ideas and new extensions will be briefly reviewed. Both substantive and statistical considerations are critical and will be discussed. Early prediction of class membership is also of interest in growth mixture modeling and will be briefly covered. In the LSAY math achievement example, it is clearly of interest to make such early predictions of risk for high school dropout to make interventions possible.

19.3.2. Substantive Theory and Auxiliary Information for Predicting and Understanding Model Results

GGMM should be investigated using substantively based theory and evidence. Auxiliary information can be used to more fully understand model results even at an exploratory stage, when little theory exists. Once substantive theory has been formulated, it can be used to predict a related set of events that can then be tested.

Figure 19.6 LSAY Math Achievement in Grades 7 to 10 and High School Dropout



Substantive theory building typically does not rely on only a single outcome measured repeatedly, accumulating evidence for a theory only by sorting into classes observed trajectories on a single outcome variable. Instead, many different sources of auxiliary information are used to check the theory's plausibility. Mental health research may find that a pattern of a high level of deviant behavior at ages when this is not typical is often accompanied with a variety of negative social consequences, so that there is a distinct subtype. A good education study of failure in school also considers what else is happening in the student's life, involving predictions of accompanying problems. Gene-environment interaction theories may predict the emergence of problems as a response to adverse life events at certain ages. These are the situations when GGMM is particularly useful. GGMM can include the auxiliary information in the model and test if the classes formed have the characteristics on the auxiliary variables that are predicted by theory. Auxiliary information may take the form of antecedents, concurrent events, or consequences. These are briefly discussed in turn below.

19.3.2.1. Antecedents

Auxiliary information in the form of antecedents (covariates) of class membership and growth factors should be included in the set of covariates to correctly specify the model, find the proper number of classes, and correctly estimate class proportions and class membership. The fact that the "unconditional model" without covariates is not necessarily the most suitable for finding the number of classes has not been fully appreciated and will be discussed below.

An important part of GGMM is the prediction of class membership probabilities from covariates. This gives the profiles of the individuals in the classes. The estimated prediction of class membership is a key feature in examining predictions of theory. If classes are not statistically different with respect to covariates that, according to theory, should distinguish classes, crucial support for the model is absent.

Class variation in the influence of antecedents (covariates) on growth factors or outcomes also provides a better understanding of the data. As a caveat, one should note that if a single-class model has generated the data with significant positive influence of covariates on growth factors, GGMM that incorrectly divides up the trajectories in, say, low, medium, and high classes might find that covariates have lower and insignificant influence in the low class due to selection on the dependent variable. If a GGMM has generated the data, however, the selected subpopulation is the relevant one to which to draw the inference. In either case, GGMM provides considerably more flexibility than what can be achieved with conventional growth modeling. As an example, consider Muthén and Curran's (1997) analysis of a preventive intervention with a strong treatment-baseline interaction. The intervention aimed at changing the trajectory slope of aggressive-disruptive behavior of children in Grades 1 through 7. No main effect was found, but Muthén and Curran used multiple-group latent growth curve modeling to show that the initially more aggressive children benefited from the intervention

in terms of lowering their trajectory slope. The Muthén-Curran technique is not, however, able to capture a nonmonotonic intervention effect that exists for children of medium-range aggression and is absent for the most or least aggressive children. In contrast, such a nonmonotonic intervention effect can be handled using GGMM with the treatment/control dummy variable as a covariate having class-varying slopes (see Muthén et al., 2002). There are probably many cases in which the effect of a covariate is not strong or even present, except in a limited range of the growth factor or outcome.

19.3.2.2. Concurrent Events and Consequences (Distal Outcomes)

Modeling with concurrent events and consequences speaks directly to standard considerations of concurrent and predictive validity. In GGMM, concurrent events can be handled as time-varying covariates that have class-varying effects, as time-varying outcomes predicted by the latent classes, or as parallel growth processes. Consequences can be handled as distal outcomes predicted by the latent classes or as sequential growth processes. Examples of distal outcomes in GGMM include alcohol dependence predicted by heavy drinking trajectory classes (Muthén & Shedden, 1999) and prostate cancer predicted by prostate-specific antigen trajectory classes (Lin, Turnbull, McCulloch, & Slate, 2002).

A very useful feature of GMM, even if a single-class nonnormal growth model cannot be rejected, is that cut points for classification are provided. For instance, individuals in the high class, giving the higher probability for the distal outcome, are identified, whereas this information is not provided by the conventional single-class growth analysis. It is true that this classification is done under a certain set of model assumptions (e.g., within-class conditional normality of outcomes given covariates), but even if the classification is not indisputable, it is nevertheless likely to be useful in practice. In single-class analysis, one may estimate individuals' values on the growth factors and attempt a classification, but it can be very difficult to identify cut points, and the classification is inefficient. The added classification information in GMM versus conventional single-class growth modeling is analogous to the earlier discussion of latent class and latent profile analysis adding complementary information to factor analysis. In addition, GMM classification is an important tool for early detection of likely membership in a problematic class, as will be discussed in the example below.

19.3.3. Statistical Aspects of Growth Mixture Modeling: Studying Model Estimation Quality and Power by Monte Carlo Simulation Studies

Because growth mixture modeling is a relatively new technique, rather little is known about requirements in terms of sample size and the number of time points needed for good estimation and strong power. Monte Carlo studies are useful for gaining understanding about this. Figure 19.4 shows a prototypical growth mixture model with a distal outcome. The following is a brief description of how a Monte Carlo study can be carried out based on this model using Mplus. For background about Monte Carlo studies of latent variable models using Mplus, see Muthén and Muthén (2002). As argued in that article, general rules of thumb are not likely to be dependable, but Monte Carlo studies can be done in settings similar to those of the study at hand.

A total of 100 data sets were generated according to the Figure 19.4 model without the *xmis* covariates, using a sample size of 3,000, similar to that of LSAY. Here, the class percentages are 27% and 73%. Maximum likelihood estimation was carried out and results summarized over the 100 replications. The Mplus output contains average parameter estimates, parameter estimate standard deviations, average standard errors, 95% coverages, and power estimates. Here, *power* refers to the proportion of replications in which the hypothesis that the parameter value is zero is rejected.[4]

The results indicate very good estimation of parameters and standard errors as well as good coverage. The quality is a function of the sample size, the number of time points, the separation between the classes, and the within-class variation. Here, the intercept growth factor means in the two classes are one standard deviation apart. As examples of the power estimates, the regression coefficient for the slope growth factor on the covariate is 0.43 for the smaller class, which has a smaller coefficient, and 1.00 for the larger class, which has a larger coefficient. Changing the sample size to 300, the results are still acceptable, although the power estimates for the slope growth factor regression coefficients are now reduced to 0.11 and 0.83.

The Mplus Monte Carlo facility is quite flexible. For example, to study model misspecification, one could analyze a different model than the one that generated the data. In latent class models, the misspecification may concern the number of classes. For Monte

4. The Mplus input and output for this analysis are given in Example 1 at www.statmodel.com/mplus/examples/penn.html.

Carlo designs that are not offered in Mplus, externally generated data can be analyzed using the RUNALL utility.[5] An extensive Monte Carlo study of growth mixture and related factor mixture models is given in Lubke and Muthén (2003).

19.3.4. Statistical Aspects of Growth Mixture Modeling: Model Selection Procedures

This section gives an overview of strategies and methods for model selection and testing. An emphasis is placed on practical analysis steps and recent testing developments.

19.3.4.1. Analysis Steps

In conventional growth modeling, a common analysis strategy is to first consider an "unconditional model" (i.e., not introducing covariates for the growth factors). This strategy can lead to confusion with growth mixture modeling. Consider the growth mixture model diagram shown earlier in Figure 19.4. Here the model has covariates x and *xmis*, a latent class variable c, repeated continuous outcomes y, and a distal dichotomous outcome u. The covariate x influences c, has direct effects on the growth factors π_0 and π_1, and also has direct effects on u.

Consider first an analysis of this model without u and without the xs. Here, the class formation is based on information from the observed variables y, channeled through the growth factors. A distorted analysis is obtained if the xs are excluded because they have direct effects on the growth factors. This is because the only observed variables, y, are incorrectly related to c if the xs are excluded. The distortion can be understood based on the analogy of a misspecified regression analysis. Leaving out an important predictor, the slope for the other predictor is distorted. In Figure 19.4, the other predictor is the latent class variable c, and the distortion of its effect on the growth factors causes incorrect evaluation of the posterior probabilities in the E step and therefore incorrect class probability estimates and incorrect individual classification. If, on the other hand, the x covariates do not have a direct influence on the growth factors (and no direct influence on y), the "unconditional model" without the xs would be correct, giving correct class probabilities and growth curves for y.

To further explicate the reasoning above, consider a data set generated by the model in Figure 19.4 without the *xmis* covariate, using the Monte Carlo feature of Mplus discussed earlier.[6] Analysis of the generated data by the correct model recovers the population parameters well, as expected. The estimated Class 1 probability of 0.26 is close to the true value of 0.27. The entropy is not large, despite the correctness of the model, 0.57, but this is a function of the degree of separation between the classes and the within-class variation. In line with the discussion above, the influence of the covariate x is of special interest. The model that generated the data has a positive slope for the influence of x on being in the smaller Class 1, positive slopes for the influence on the growth factors, and a positive slope for the influence on u. The estimated class-specific means and variances of the x covariate are 0.63 and 0.79 for Class 1 and −0.20 and 0.82 for Class 2. The higher mean for Class 1 is expected, given the positive slope for the influence on the Class 1 membership. Being in Class 1, in turn, implies higher means for the growth factors. Within class, the growth factor means are higher due to the direct positive influence of x on the growth factors. With x left out of the model, the latent class variable alone needs to account for the differences in growth factor values across individuals. As a result, the class probabilities are misestimated. In the generated data example, the Class 1 probability is now misestimated as 0.35.[7]

Analyzing the Figure 19.4 model excluding u but correctly including x gives the correct answer in terms of class membership probabilities for c and growth curves for y. This is because excluding u does not imply that the observed variables (y or x) are incorrectly related to c. Excluding u simply makes the standard errors larger and worsens the classification precision (entropy). In the generated data example, the Class 1 probability is well estimated as 0.26, whereas the entropy is now lowered to 0.50.[8]

In practice, model estimation with and without a distal outcome u may give different results for the class probabilities and growth curves for two reasons. First, if you include u but misspecify the model by not allowing direct effects from the xs to u, you get distorted parameter estimates (e.g., incorrect class probabilities) by the same regression misspecification analogy given above. In the generated data example,

5. See http://www.statmodel.com/runutil.html.

6. The Mplus input and output for this analysis is given in Example 2 at www.statmodel.com/mplus/examples/penn.html.

7. The Mplus input and output for this analysis is given in Example 3 at www.statmodel.com/mplus/examples/penn.html.

8. The Mplus input and output for this analysis are given in Example 4 at www.statmodel.com/mplus/examples/penn.html.

this misspecification gave the strongly distorted Class 1 probability estimate as 0.40. Second, key covariates may have been left out of the model (i.e., may not have been measured or are missing), causing a model misspecification. The notation *xmis* in Figure 19.4 refers to such a covariate. Consider two cases, both assuming that *xmis* is not available. First, if *xmis* influences only u and not the growth factors, the analysis excluding u gives correct results, but the analysis including u gives incorrect and hence different results. Second, if *xmis* influences both the growth factors and u, the analyses with and without u give incorrect results and are different.

In conclusion, the proper choice of covariates is important in growth mixture modeling. Substantive theory and previous analyses are needed to make a choice that is sufficiently inclusive. The covariates should be allowed to influence not only class membership but also the growth factors directly, unless there are well-motivated reasons not to. An analysis without covariates can be useful to study different growth in different trajectory classes. However, it should not be expected that the class distribution or individual classification remains the same when adding covariates. It is the model with covariates properly included that gives the better answer.

It should also be noted that choosing the correct within-class variance structure is important. The data above were generated from a model with class-varying variances for the residuals of e in (1). Misspecifying the model by holding these variances equal across class leads to an estimated Class 1 probability of 0.23. Larger distortions would be obtained if the growth factor variances differ across classes.

It is instructive to consider model misspecification results if data generated by the growth mixture model are analyzed by a latent class growth analysis. In the generated data example above, LCGA leads to a misspecified model. The misspecification can be studied in two steps, first by restricting the residual (co)variances and second by also not allowing the direct influence from x to the growth factors. In both cases, the distal outcome is u. In the first step, the estimated Class 1 probability is found to be 0.42, a value far off from the true probability of 0.27. In the second step, the estimated Class 1 probability is even more strongly distorted, 0.51. It is noteworthy that the misspecification of not letting x have a direct effect on the growth factors cannot be discovered using LCGA. Note that in the last two analyses, the entropy values are strongly overestimated, 0.80 and 0.85. It is also likely that more than two classes are needed to account for the within-class variation. This implies that some of the classes are merely slight variations on a theme and do not have a substantial meaning.

19.3.4.2. Equivalent Models

With latent variable models in general and mixture models in particular, the phenomenon of equivalent models may be encountered. Here, *equivalent models* means that two or more models fit the same data approximately the same so that there is no statistical basis on which to base a model choice. Consider two psychometric examples. First, in exploratory factor analysis, a rotated solution using uncorrelated factors gives the same estimated correlation matrix as a rotated solution with correlated factors. Second, Bartholomew and Knott (1999, pp. 154–155) point out a well-known psychometric fact that a covariance matrix generated by a latent profile model (a latent class model with continuous outcomes) can be perfectly fitted by a factor analysis model. A covariance matrix from a k-class model can be fitted by a factor analysis model with $k - 1$ factors. Molenaar and von Eye (1994) show that a covariance matrix generated by a factor model can be fitted by a latent class model. This should not be seen as a problem but merely as two ways of looking at the same reality. The factor analysis informs about underlying dimensions and how they are measured by the items, whereas the latent profile analysis sorts individuals into clusters of individuals who are homogeneous with respect to the item responses. The two analyses are not competing but are complementary.

The issue of alternative explanations is classic in finite mixture statistics. Mixtures have two separate uses. One is to simply fit a nonnormal distribution without a particular interest in the mixture components. The other is to capture substantively meaningful subgroups. For a historical overview, see, for instance, McLachlan and Peel (2000, pp. 14–17), who refer to a debate about blood pressure. A classic example concerns data from a univariate (single-class) lognormal distribution that are fitted well by a two-class model that assumes within-class normality and has different means. Bauer and Curran (2003) consider the analogous multivariate case arising with growth mixture modeling.[9] The authors use a Monte Carlo simulation study to show that a multiclass growth mixture model can be arrived at using conventional Bayesian information criterion (BIC) approaches (see below) to determine the number of classes when data, in fact, have been generated by a nonnormal multivariate

9. Multivariate formulas that show equivalence are not given.

distribution that is skewed and kurtotic. Although the authors only consider GMM, the resulting overextraction of classes would be more pronounced for LCGA. Bauer and Curran's study serves as a caution to researchers to not automatically assume that the latent trajectory classes of a growth mixture model have substantive meaning. Their article is followed by three commentaries and a rejoinder that place the discussion in a larger context. Two of the commentaries, including one by Muthén (2003), point out that BIC does not address model fit to data but is a relative fit measure comparing competing models. Muthén discusses new mixture tests that aim to address data fit, which are mentioned below. The use of these alternative models ultimately has to be guided by arguments related to substantive theory, auxiliary information, predictive validity, and practical usefulness.

19.3.4.3. Conventional Mixture Tests

The selection of the number of latent classes has been discussed extensively in the statistical literature on finite mixture modeling (see, e.g., McLachlan & Peel, 2000). The likelihood ratio comparing a $k - 1$ and a k-class model does not have the usual large-sample chi-square distribution due to the class probability parameter being at the border (zero) of its admissible space. A commonly used alternative procedure is the BIC (Schwartz, 1978), defined as

$$\text{BIC} = -2 \log L + p \ln n, \tag{7}$$

where p is the number of parameters and n is the sample size. Here, BIC is scaled so that a small value corresponds to a good model with a large log-likelihood value and not too many parameters.

Consider as an example the generated data example of the previous section. Here, the analysis without the x covariate or the u distal outcome gave the following BIC values for one, two, and three classes: 39,676.166, 39,603.274, and 39,610.785. This points correctly to two classes, despite the fact that the model is misspecified due to not including x and its direct effect on the growth factors. This fortunate outcome cannot be relied on, however.

19.3.4.4. New Mixture Tests

This section briefly describes two new mixture test approaches. A key notion is the need for checking how well the mixture model fits the data, not merely basing a model choice on k classes fitting better than $k - 1$ classes. It should be emphasized that there are many possibilities for checking model fit against data in mixture settings, and the methodology for this is likely to expand considerably in the future. One promising approach is the residual diagnostics based on pseudo-classes, proposed in Wang, Brown, and Bandeen-Roche (2002).

Lo, Mendell, and Rubin (2001) proposed a likelihood ratio–based method for testing $k - 1$ classes against k classes. The Lo-Mendell-Rubin approach has been criticized (Jeffries, 2003), although it is unclear to which extent the critique affects its use in practice. The Lo-Mendell-Rubin likelihood ratio test (LMR LRT) avoids a classic problem of chi-square testing based on likelihood ratios. This concerns models that are nested, but the more restricted model is obtained from the less restricted model by a parameter assuming a value on the border of the admissible parameter space—in the present case, a latent class probability being zero. It is well known that such likelihood ratios do not follow a chi-square distribution. Lo, Mendell, and Rubin consider the same likelihood ratio but derive its correct distribution. A low p-value indicates that the $k - 1$-class model has to be rejected in favor of a model with at least k classes. The Mplus implementation uses the usual Mplus mixture modeling assumption of within-class conditional normality of the outcomes given the covariates. When nonnormal covariates are present, this allows a certain degree of within-class nonnormality of the outcomes. The LMR LRT procedure has been studied for GMMs by Monte Carlo simulations (Masyn, 2002). More investigations of performance in practice are, however, of interest, and readers can easily conduct studies using the Mplus Monte Carlo facility for mixtures.

Muthén and Asparouhov (2002) proposed a new approach for testing the fit of a k-class mixture model for continuous outcomes. As opposed to the LMR LRT, this procedure concerns a test of a specific model's fit against data. The procedure relies on testing if the multivariate skewness and kurtosis (SK) estimated by the model fit the corresponding sample quantities. The sampling distributions of the SK tests are assessed by computing these values over a number of replications in data generated from the estimated mixture model. Obtaining low p-values for skewness and kurtosis indicates that the k-class model does not fit the data. Univariate and bivariate test results are also provided for each variable and pair of variables. These tests may provide a useful complement to the LMR LRT. Currently, the SK tests are not available with missing data. Given the inherent sensitivity to outliers, the SK testing should be preceded by outlier investigations.

The SK procedure needs further investigation but is offered here as an example of the many possibilities of testing a mixture model against data (see also Wang et al., 2002).

19.3.5. The LSAY Math Achievement Example

This section returns to the analysis of the mathematics achievement data from the LSAY data mentioned earlier. Based on the educational literature, the following covariates are included: female; Hispanic; Black; mother's education; home resources; the student's educational expectations, measured in seventh grade (1 = high school only, 2 = vocational training, 3 = some college, 4 = bachelor's degree, 5 = master's degree, 6 = doctorate); the student's thoughts of dropping out, measured in seventh grade; whether the student has ever been arrested, measured by seventh grade; and whether the student has ever been expelled by seventh grade. Corresponding to individuals with complete data on the covariates, the analyses consider a subsample of 2,757 of the total 3,116 individuals. The analyses were carried out by maximum likelihood estimation using Mplus Version 2.13.

19.3.5.1. Statistical Checking

The univariate skewness and kurtosis sample values in the LSAY data are as follows:

$$\text{Skewness} = (0.168 \;\; 0.030 \;\; 0.063 \;\; -0.077), \quad (8)$$

$$\text{Kurtosis} = (-0.551 \;\; -0.338 \;\; -0.602 \;\; -0.559). \quad (9)$$

In line with the earlier discussion of the LMR LRT, due to the low nonnormality in the outcomes, it is plausible that this test is applicable in the LSAY analysis for testing a one-class model versus more than one class. In the LSAY analysis, this test points to at least two classes with a strong rejection ($p = .0000$) of the one-class model. The SK tests carried out on the listwise present subsample of 1,538 reject the one-class model ($p = .0000$ for both multivariate skewness and multivariate kurtosis) but do not reject two classes ($p = .4300$ and .5800). The LMR LRT for two versus three or more classes obtained a high p-value (.6143) in support of two classes. Taken together, the statistical evidence points to at least two classes. Given that the skewness and kurtosis tests found that two- and three-class GMMs fit the data, the LMR LRT is useful for testing the multiclass alternatives against each other.

19.3.5.2. Substantive Checking and Further Statistical Analysis

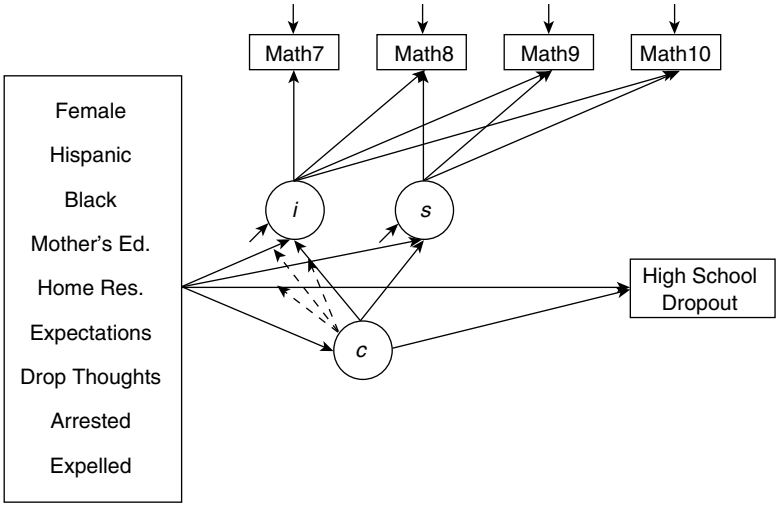
This section compares analysis results using a conventional one-class growth model and different forms of GMMs and discusses substantive meaningfulness based on educational theory, auxiliary information, and practical usefulness. Figure 19.7 shows a diagram of the general model.

19.3.5.2.1. Conventional one-class growth modeling. As a first step, the conventional one-class growth model results are considered. Briefly stated, a linear growth model fits reasonably well and has a positive growth rate mean of about 1 standard deviation across the four grades. The covariates with significant influence (sign in parentheses) on the initial status are as follows: female (+), Hispanic (−), Black (−), mother's education (+), home resources (+), expectations (+), dropout thoughts (−), arrest (−), and expelled (−). The covariates with significant influence (sign in parentheses) on the growth rate are as follows: female (−), Hispanic (−), home resources (+), expectations (+), and expelled (−).

19.3.5.2.2. Two-class GMM. The two-class solution is characterized by a low class of 41%, which, in comparison to the high class, has a lower initial status mean and variance, a lower growth rate mean, and a higher growth rate variance. It is interesting to consider what characterizes these students apart from their poor mathematics achievement development. The multinomial logistic regression for class membership indicates that, relative to the high class, the odds of membership in the low class are significantly increased by being male, being Hispanic, having a mother with a low level of education, having low seventh-grade educational expectations, having had seventh-grade thoughts of dropping out, having been arrested, and having been expelled. The low class appears to be a class of students with problems both in and out of school. The profile of the low class is reminiscent of individuals at risk for dropping out of high school (see, e.g., Rumberger & Larson, 1998, and references therein). Many of these students are "disengaged," to use language from high school dropout theories.

The within-class influence of the covariates on the initial status and growth rate factors varies significantly across class. The low class has no significant predictors of growth rate, whereas the growth rates of the two higher classes are significantly enhanced in well-known ways by being male, having a mother with a high level of education, having high home resources,

Figure 19.7 GGMM Diagram for LSAY Data



and having high expectations. To the extent that the low class has substantive meaning, the findings suggest that different processes are in play for students in the low class.

19.3.5.2.3. Three-class GMM including a distal outcome. To more specifically investigate the data from the high school dropout perspective and further characterize the low class, the distal binary outcome of dropping out of high school, as recorded in Grade 12, was added. The overall dropout rate in the sample is 14.7%, or 458 individuals. Here, class membership in the GMM is, to some extent, also determined by the Grade 12 dropout indicator and not only by the covariates and math achievement development. Adding the distal outcome, the LMR LRT rejected the two-class model in favor of at least three classes ($p = .0060$). The three-class solution produces a more distinct low class of 19%, a middle class of 28%, and a high class of 52%. Here, the low class (estimated as 536 students) has a lower growth rate mean and lower growth rate variance than in the two-class solution without the distal outcome.[10]

The class membership regression part of the model indicates that for the low class relative to the highest class, the same covariates as in the two-class solution are significant, except that Hispanic and mother's education are insignificant, whereas Black and home resources are significant. Interestingly, comparing the middle class to the high class, the disengagement covariates of low educational expectations, seventh-grade dropout thoughts, having been arrested, and having been expelled are no longer significant. This suggests that the low class is now a more distinct class that is more specifically characterized as disengaged and at risk for high school dropout. The two higher classes may or may not make a substantively meaningful distinction among students, but their presence helps to isolate the low class. In a two-class solution including the distal outcome, the low class is not very different from the more unspecific low class of the initial two-class solution without the distal outcome. It is interesting to note that although the LMR LRT does not point to three classes without the distal outcome, the three-class solution without the distal outcome shows a similar low class as in the three-class solution with the distal outcome. As will be shown next, the three-class solution with the distal outcome gets not only statistical support from the LMR LRT but also substantive support from predicting dropout.

Further bolstering the notion that the low class is prone to high school dropout, the probability of dropping out, as estimated from the three-class model, is distinctly different in the low class. The probabilities are .692 for the low class, .076 for the middle class,

10. The Akaike information criterion (AIC) points to at least three classes, whereas the Bayesian information criterion (BIC) points to two classes. The one-class log-likelihood, number of parameters, AIC, and BIC values are as follows: −30,021.955, 27, 60,097.909, and 60,257.791. The two-class log-likelihood, number of parameters, AIC, BIC, and entropy values are as follows: −29,676.457, 63, 59,478.914, 59,851.971, and 0.552. The three-class log-likelihood, number of parameters, AIC, BIC, and entropy values are as follows: −29,566.679, 99, 59,331.359, 59,917.591, and 0.620.

and .006 for the high class. Other concurrent and distal events were also added to the three-class model to further understand the context of the low class, including responses to the following 10th grade question: "How many of your friends will drop out before graduating from high school?" (1 = *none*, 2 = *a few*, 3 = *some*, 4 = *most*). Treating this as an ordered polytomous outcome influenced by class and the covariates resulted in estimated probabilities for response in either of the three highest categories (few, some, most): .259 for the low class, .117 for the middle class, and .030 for the high class. Considerably more students in the low class have friends who are also thinking of dropping out. In contrast, heavy alcohol involvement in Grade 10 was not distinctly different in the low class. The estimated growth curves and individual trajectories can be seen in Figure 19.6.

19.3.5.2.4. Practical usefulness. An educational researcher is likely to find it interesting that the analyses suggest that dropout by Grade 12 can be predicted already by the end of Grade 10 with the help of information on problematic math achievement development. Whether the division into growth mixture classes is meaningful is largely a substantive question. An argument in favor of there being a distinct "failing class" is obtained from the distal outcome of high school dropout. The fact that the dropout percentage is dramatically higher for the low class than for the other two, 69% versus 8% and 1%, suggests that the three classes are not merely gradations on an achievement development scale but that the low class represents a distinct group of students.

From the point of view of intervention, it is valuable to explore whether a dependable classification into the low class can be achieved earlier than Grade 10. GGMM can help answer this question. For example, by Grade 7, the covariates and the first math achievement outcome are available, and given the estimated three-class model, new students can be classified based on the model and their Grade 7 data. GGMM allows the investigation of whether this information is sufficient or if math achievement trend information provided by adding Grade 8 information (or Grades 8 and 9 information) is needed before a useful classification can be made.

19.4. Categorical Outcomes: Conventional Growth Modeling

With categorical outcomes, the Level 1 model part (1) has to be replaced with a model that describes the probability of the outcome at different time points for different individuals. This model has been studied by Hedeker and Gibbons (1994). Here, logistic regression will be used, so that with the example of a binary outcome u scored 0 and 1,

$$P(u_{ti} = 1|a_{1ti}, a_{2ti}, x_i) = \frac{1}{1 + e^{\tau - logit(u_{ti})}}, \qquad (10)$$

$$\text{Level 1 (Within): } \text{logit}(u_{ti}) = \pi_{0i} + \pi_{1i}\, a_{1ti} + \pi_{2ti} a_{2ti} + e_{ti}, \qquad (11)$$

$$\text{Level 2 (Within): } \begin{cases} \pi_{0i} = \beta_{00} + \beta_{01} x_i + r_{0i} \\ \pi_{1i} = \beta_{10} + \beta_{11} x_i + r_{1i}. \\ \pi_{2i} = \beta_{20} + \beta_{21} x_i + r_{2i} \end{cases} \qquad (12)$$

A perhaps more common parameterization is to fix the threshold parameter τ in (10) at zero, which enables the identification of β_{00}.[11] The variance of e is not a free parameter but is fixed in line with logistic regression. With ordered polytomous outcomes, Mplus uses the proportional odds logistic regression model (see, e.g., Agresti, 1990, pp. 322–324). This may be thought of as a threshold model for a latent response variable, so that with C categories, there is a series of $C-1$ ordered thresholds. The thresholds are held equal across time. As a standardization, $\beta_{00} = 0$ may be chosen, or alternatively, the first threshold may be set at zero. Hedeker and Gibbons (1994) describe maximum likelihood estimation and show that this requires heavier computations than with continuous outcomes, calling on numerical integration using quadrature methods. The computational burden is directly related to the number of random effects (i.e., the number of coefficients π for which the variance of r is not fixed at zero).

19.5. Categorical Outcomes: Growth Mixture Modeling

The conventional growth modeling for categorical outcomes given in (11) and (12) can be extended to growth mixture modeling with latent trajectory classes. This is a new technique introduced in Asparouhov and Muthén (2003b), using maximum likelihood estimation based on an EM algorithm with numerical integration. In line with the latent variable approach to

11. The Mplus input and output for these analyses are given in Example 5 at www.statmodal.com/mplus/examples/penn.html.

growth modeling with continuous outcomes discussed in Section 19.2, the Asparouhov-Muthén approach allows a_{1ti} in (11) to be handled as data or as parameters to be estimated. Furthermore, the π_{2ti} slopes can be random for the time-varying covariates a_{2ti}.[12] The Hedeker-Gibbons model is obtained as a special case with a single latent class.

As in (3), the covariate effect on class membership is a multinomial logistic regression,

$$P(c_i = k|x_i) = \frac{e^{\gamma_{0k}+\gamma_{1k}x_i}}{\sum_{s=1}^{K} e^{\gamma_{0s}+\gamma_{1s}x_i}}. \quad (13)$$

The growth mixture extension of (10) is

$$P(u_{ti} = 1|a_{1ti}, a_{2ti}, x_i, c_i = k) = \frac{1}{1 + e^{\tau - logit(u_{tik})}}, \quad (14)$$

where the added conditioning on c and the subscript k emphasize that the growth model for u, as expressed by the logits, varies across classes. In line with the extension for continuous outcomes, the different latent classes have different growth models (11) and (12), with key differences typically found in the β coefficients but also in the (co)variances of the Level 2 residuals r. Typically, the thresholds τ would be time and class invariant to represent measurement invariance, although class invariance is not necessary. Generalizations to including distal outcomes u_d, as in (15), is of interest also here:

$$P(u_{di} = 1|c_i = k, x_i) = \frac{1}{1 + e^{\tau_k - \kappa_k x_i}}, \quad (15)$$

with coefficients varying across classes k.

Model building and testing strategies for categorical outcomes are in line with those discussed earlier for continuous outcomes.

19.5.1. Categorical Outcomes: Latent Class Growth Analysis

Latent class growth analysis (LCGA) for categorical outcomes considers the model in (11) through (13) with the restriction of zero variances and covariances for the residuals r. Background references for LCGA include Nagin (1999), Nagin and Land (1993), and Nagin and Tremblay (2001).

12. Threshold parameters are useful with ordered polytomous outcomes, in which case β_{00} can be fixed at zero, or, alternatively, the first threshold is fixed at zero.

It is instructive to relate LCGA to latent class analysis (LCA). As in LCGA, LCA considers multiple u variables seen as indicators of c and assumed conditionally independent given c. As in LCGA, there are no continuous latent variables to explain further within-class correlation among the u variables. Typically, all outcomes are categorical. Continuous outcomes are, however, possible, giving rise to latent profile analysis. In LCA, the multiple indicators are cross-sectional measures, not longitudinal. When the multiple indicators correspond to repeated measures over time, latent classes may correspond to different trends, and trend structures can be imposed across the indicators' probabilities. To clarify this, consider again (14):

$$P(u_{ti} = 1|a_{1ti}, a_{2ti}, x_i, c_i = k) = \frac{1}{1 + e^{\tau - logit(u_{tik})}}. \quad (16)$$

This means that with, for example, linear growth over T time points, the probabilities of the T u variables are structured according to a logit-linear trend, where the intercept and slope factors have different means across the classes. Note here that τ is held equal across time points. In contrast, LCA considers

$$P(u_{ti} = 1|x_i, c_i = k) = \frac{1}{1 + e^{\tau_{tk}}}, \quad (17)$$

where the τ_{tk} thresholds vary in an unrestricted fashion across the u variables and across the classes. In this way, LCGA gives a more parsimonious description of longitudinal data than LCA.

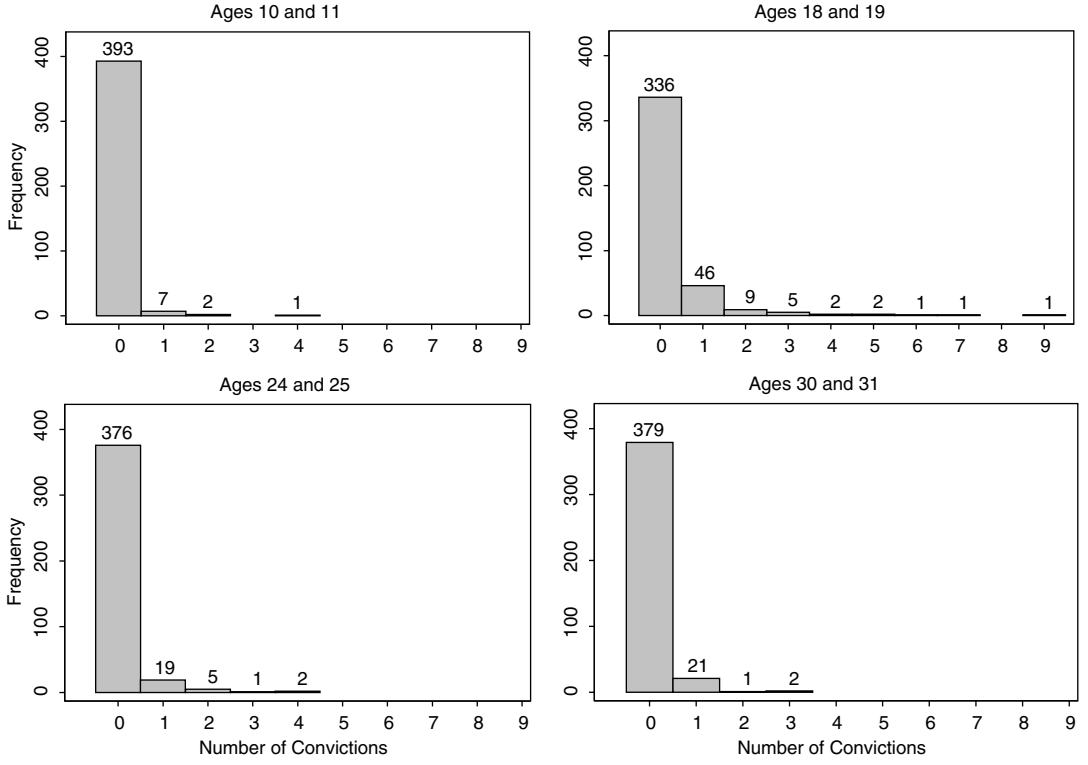
Models with more than one latent class variable are also of interest. Examples of LCGA with multiple-class variables are given in Muthén and Muthén (2000), Muthén (2001a), and Nagin and Tremblay (2001). In this connection, it is useful to consider another important class of growth models, latent transition analysis (LTA). LTA uses time-specific latent class variables measured by multiple indicators at each time point to study class membership change over time.

Both LCA and LTA can be generalized to include random effects as in growth mixture modeling (Asparouhov & Muthén, 2003b). All of these model variations can be captured in a general latent variable modeling framework and are included in Mplus.

19.5.2. Categorical Outcomes: Comparing LCGA and GMM on Delinquency Data

Nagin and Land (1993), Nagin (1999), Roeder et al. (1999), and Jones et al. (2001) used PROC TRAJ

Figure 19.8 Frequency Distributions for Cambridge Data



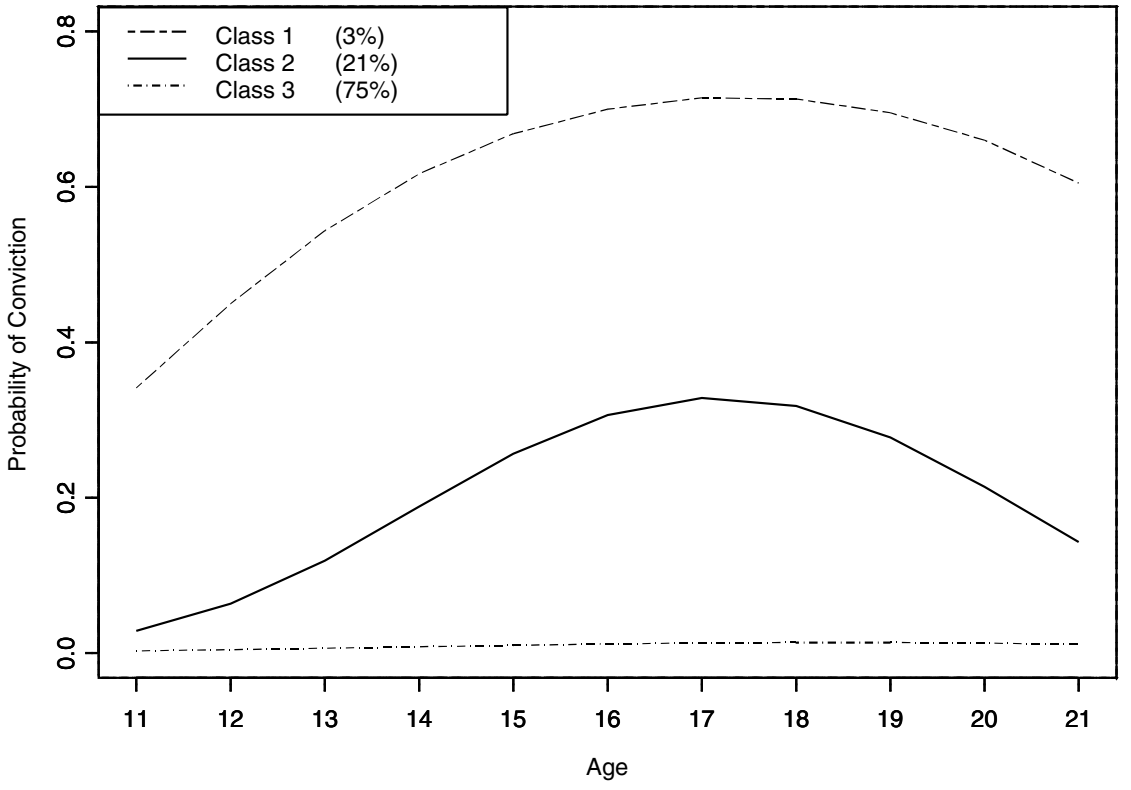
LCGA to study the development of delinquency over ages 10 to 32 in a sample of 411 boys in a working-class section of London (Farrington & West, 1990). These "Cambridge data" were studied from the substantive perspective of the Moffitt (1993) theory of adolescent-limited versus life course–persistent antisocial behavior. This theory suggests two major trajectory classes. Using different ways to aggregate and model the outcomes, Nagin and Land found four classes, Nagin three classes, Roeder et al. four classes, and Jones et al. three classes. Nagin (1999) used 2-year intervals and excluded the 8 boys who died during the study, resulting in 11 time points and $n = 403$. The frequency distributions are shown in Figure 19.8. Only ages 11 to 21 will be used here.

Given that few individuals have more than two convictions in the 2-year interval, data will be coded as 0, 1, and 2 for zero, one, or more convictions; 69% have 0 value at all 11 time points. A logistic ordered polytomous response model will be used, and three types of analyses will be illustrated: latent class growth analysis, conventional growth modeling, and growth mixture modeling. The analyses draw on Muthén, Kreuter, and Asparouhov (2003).

19.5.2.1. Latent Class Growth Analysis of the Cambridge Data

Latent class growth analysis was performed with two, three, and four classes applying a quadratic growth curve for all classes. The corresponding BIC values were 2,230.014, 2,215.251, and 2,227.976. This points to the three-class model as being the best. This model has a log-likelihood value of −1,071.632, 12 parameters, and an entropy of 0.821. The estimated class percentages are 3%, 21%, and 75%, arranging the curves from high to low. The LMR LRT also points to three classes in that the test of the two-class model against the three-class model has a p-value of .0030, suggesting rejection, whereas the three-class model tested against the four-class model has a p-value of .1554. The estimated three-class growth curves for the

Figure 19.9 Three-Class LCGA for Cambridge Data



probability of having at least one conviction are shown in Figure 19.9.[13]

19.5.2.2. Growth and Growth Mixture Analysis of the Cambridge Data

Conventional one-class growth modeling of the ordered polytomous outcome used a centering of the time scale at age 17 and let the intercept and linear slope growth factors be random, and the quadratic slope factor variance was fixed at zero. The intercept and linear slope were allowed to correlate. This one-class growth model resulted in a log-likelihood value of –1,072.396 with seven parameters and a BIC value of 2,186.785.[14] The linear slope variance is not significant and will, for simplicity, be set to zero in subsequent analyses. In the growth mixture analyses to follow, this intercept variance was allowed to vary across the classes.

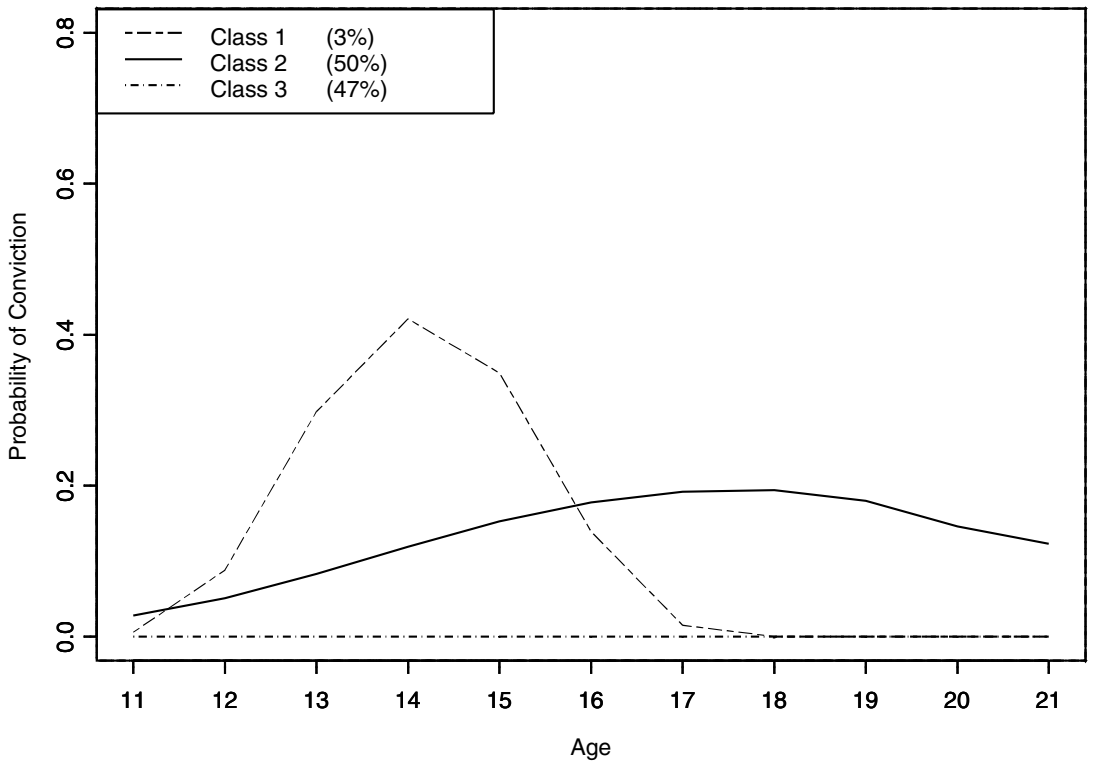
The two-class growth mixture modeling resulted in a log-likelihood value of –1,070.898, a BIC of 2,201.785, 10 parameters, and an entropy of 0.414. The estimated class percentages are 46% and 54%, arranging the classes from high to low. The intercept variance is significant in both classes and lower in the low class. The LMR LRT p-value for one class tested against two classes is .0362, pointing to the need for at least two classes.

A specific three-class growth mixture model was considered next, in which one class was specified to have zero probability of conviction throughout the time period. This zero class corresponds to the notion that some individuals do not get involved in delinquency activities at all. In the other two classes, the intercept variance was allowed to be free to be estimated and different across those classes. This model resulted in a log-likelihood value of –1,066.767, a BIC of 2,199.523, 11 parameters, and an entropy of 0.535. The estimated class percentages are 3%, 50%, and 47%, arranging the classes from high to

13. The Mplus input and output for these analyses are given in Example 6 at www.statmodel.com/mplus/examples/penn.html.

14. The Mplus input and output for these analyses are given in Example 7 at www.statmodel.com/mplus/examples/penn.html. This analysis was carried out by Mplus Version 3.

Figure 19.10 Three-Class LCGA for Cambridge Data



low. The intercept variance is nonsignificant for the highest class but significant for the middle class.[15] An interesting finding is that this three-class GMM, which allows within-class variation, has 1 parameter less than the three-class LCGA but a better fit in terms of log-likelihood and BIC values. The zero class is smaller in the GMM than in the LCGA, 47% versus 75%. The fact that 69% of the individuals have observed values at zero throughout, whereas the GMM zero class has only 47% prevalence, is due to the fact that the individuals who are most likely to be in the low class according to the posterior probabilities have a sizable probability of being in the middle class. It should be noted, however, that the model has several local optima with log likelihood values close to that of the best solution, possibly indicating a weakly defined solution which might not be replicated with new data. The estimated three-class growth curves for the probability of having at least one conviction are shown in Figure 19.10. These curves are clearly different from the LCGA curves in Figure 19.9, with Class 1 and Class 2 peaking at different ages for GMM but not for LCGA. This may lead to different substantive interpretations in the context of Moffitt's (1993) theory.

15. The Mplus input and output for these analyses are given in Example 8 at www.statmodel.com/mplus/examples/penn.html. This analysis was carried out by Mplus Version 3.

19.5.3. Categorical Outcomes: Discrete-Time Survival Analysis

Discrete-time survival analysis (DTSA) uses the categorical variables u to represent events modeled by a logistic hazard function (cf. Muthén & Masyn, in press). For an overview of conventional DTSA, see, for example, Singer and Willett (1993). Consider a set of binary 0/1 variables u_j, $j = 1, 2, \ldots, r$, where $u_{ij} = 1$ if individual i experiences the nonrepeatable event in time period j, and define j_i as the last time period in which data were collected for individual i. The hazard is the probability of experiencing the event in time period j given that it was not experienced prior to j. The hazard is written as

$$h_{ij} = \frac{1}{1 + e^{-(-\tau_j + \kappa_j x_i)}}, \tag{18}$$

where a proportional-odds assumption is obtained by dropping the j subscript for κ_j. Discrete-time survival analysis is fitted into the general mixture model above by noting that the likelihood is the same as for u related to c and x in a single-class model.

The fact that individual i does not have observations on u after time period j_i is handled as missing data. For example, with five time periods ($r = 5$), an individual who experiences the event in Period 4 has the data vector $\mathbf{u}_i'$

$$(0 \quad 0 \quad 0 \quad 1 \quad 999),$$

with 999 representing missing data. An individual who is censored in Period 5 has the data vector $\mathbf{u}_i'$

$$(0 \quad 0 \quad 0 \quad 0 \quad 0),$$

whereas an individual who is censored in Period 4 has the data vector $\mathbf{u}_i'$

$$(0 \quad 0 \quad 0 \quad 999 \quad 999).$$

Muthén and Masyn (in press) also propose general discrete-time survival mixture analysis (DTSMA) models, in which different latent classes have different hazard and survival functions. For example, a growth mixture model for y can be combined with a survival model for u.

19.6. Combination of Categorical and Continuous Outcomes: Modeling With Zeros

In the previous section, it was seen that the u variables need not represent conventional categorical outcomes but can be used as indicators of events. In this section, this idea is taken further by using the u variables as indicators of zero values on a continuous and on a count outcome variable.

Growth mixture modeling is useful for describing growth in outcomes that can be seen as continuous but nonnormally distributed. A type of nonnormality that cannot be well captured by mixtures of normal distributions arises in studies in which a significant number of individuals are at the lowest value of an outcome, for example, representing absence of a behavior. Applications include alcohol, drug, and tobacco use among adolescents. Censored-normal models are often used for outcomes of this kind, including classic Tobit regression analysis (Amemiya, 1985; Tobin, 1958) and LCGA in the PROC TRAJ program (Jones et al., 2001). A recent article by Olsen and Schafer (2001) gives an excellent overview of several related modeling efforts. Censored-normal models have been criticized (see, e.g., Duan, Manning, Morris, & Newhouse, 1983) because of the limitation of assuming that the same set of covariates influences both the decision to engage in the behavior and the amount observed. A two-part modeling approach proposed in Olsen and Schafer avoids this limitation.

To simplify the discussion, the lowest value will be taken to be zero. It is useful to distinguish between two kinds of zero outcomes. First, individuals may have zero values at a given time point because their behavioral activity is low and is zero during certain periods ("random zeros"). Second, individuals may not engage in the activity at all and therefore have zeros throughout all time points of the study ("structural zeros"). Olsen and Schafer (2001) proposed a two-part model for the case of random zeros, whereas Carlin, Wolfe, Brown, and Gelman (2001) considered the case of structural zeros. In both articles, a random-effects logistic regression was used to express the probabilities of nonzeros versus zeros.

Olsen and Schafer (2001) studied alcohol use in Grades 7 through 11. To capture the changing zero status across time, they expressed the logistic regressions for each time point as a random-effects growth model. The term *two-part model* refers to having both a logistic model part to model the probability of nonzero versus zero outcomes (Part 1) and a continuous-normal or lognormal model part for the values of the nonzero outcomes (Part 2). In Olsen and Schafer, the two parts have correlated random effects. The two parts are also allowed to have different covariates, avoiding the limitation of censored-normal modeling.

Carlin et al. (2001) studied cigarette smoking among adolescents. A two-class model was used with a "zero class" (structural zeros) representing individuals not susceptible to regular smoking (also referred to as "immunes"). As pointed out in Carlin et al., an individual with zeros throughout the study does not necessarily belong to the zero class but may show zeros by chance. In their analysis, the estimated proportion of immunes was 69%, whereas the empirical proportion with all zeros was 77%. Because of this, an ad hoc analysis based on deleting individuals with all zeros may lead to distorted results.

Inspired by Olsen and Schafer (2001) and Carlin et al. (2001), Muthén (2001b) proposed a generalization of growth mixture modeling to handle both random and structural zeros in a two-part model. Multiple latent classes are used to represent the growth in the probability of nonzero values in Part 1 as well as

the growth in the nonzero outcomes in Part 2. For the Part 1 modeling of the probability of nonzero values, Muthén considered a latent class growth alternative to the random-effects modeling of Olsen and Schafer (2001) and Carlin et al. (2001)—that is, a model in line with Nagin (1999). The use of latent classes for the Part 1 modeling of the probability of nonzero values may be seen as a semi-parametric alternative to a random-effects model in line with Aitkin (1999). In addition to accounting for random zeros as in Olsen and Schafer, Muthén's Part 1 approach incorporates Carlin et al.'s concept of a zero class that has zero probability of nonzero values throughout the study. A further advantage of the proposed approach is that covariates are allowed to have a different influence in different classes. For the Part 2 modeling of the nonzero outcomes, the proposed modeling extends the Olsen-Schafer growth model to a growth mixture model. The Olsen-Schafer model, the mixture version of Olsen-Schafer, the Carlin et al. model, and the Muthén two-part growth mixture model can all be fitted into the general latent variable modeling framework of Mplus.

The question of the proper treatment of zeros also arises with count variables. Roeder et al. (1999) considered zero-inflated Poisson modeling (ZIP) (Lambert, 1992) in the context of LCGA. When a count outcome is modeled by ZIP, it is assumed that a zero value can be observed for two reasons. The ZIP model is a two-class mixture model, similar in spirit to that of Carlin et al. (2001). First, if an individual is in the zero class, a zero count has probability 1. Second, if an individual is in the nonzero class, the probability of a zero count is expressed by the Poisson distribution. The probability of being in the zero class can be modeled by covariates that are different from those that predict the counts for the nonzero class. In longitudinal data, this probability can be modeled to vary across time. The model by Roeder et al. considered an LCGA for the nonzero part.

19.7. Multilevel Growth Mixture Modeling

This final section returns to the analysis of the LSAY math achievement example. Longitudinal data are often collected through cluster sampling. This was the case in the LSAY study, in which students were observed within randomly sampled schools. This gives rise to three-level data with variation across time on Level 1, variation across individuals on Level 2, and variation across clusters on Level 3. This section discusses three-level growth modeling and its new extension to three-level growth mixture modeling. Due to lack of space, details of the modeling will not be discussed here, but an analysis of the LSAY example will instead be discussed in general terms. The reader is referred to Asparouhov and Muthén (2003b) for technical details.

The model diagram of Figure 19.11 is useful for understanding the general ideas of the multilevel growth mixture modeling. This is the LSAY math achievement example discussed in Section 19.3.5. In Figure 19.11, the observed math variable rectangles at the top of the figure represent the Level 1 variation across time. The latent variable circles, labeled *i* and *s*, represent the Level 2 variation in the intercept and slope growth factors across students. The *ib*, *cb*, *sb*, and *hb* latent variable circles represent the Level 3 variation across schools. Here, *b* refers to between-school variation. One aim of three-level growth modeling is the decomposition of the intercept variance into *i* and *ib* variation and the decomposition of the slope variance into *s* and *sb* variation. Furthermore, it is of interest to describe part of this variation by school-level covariates, as shown at the bottom of the diagram.

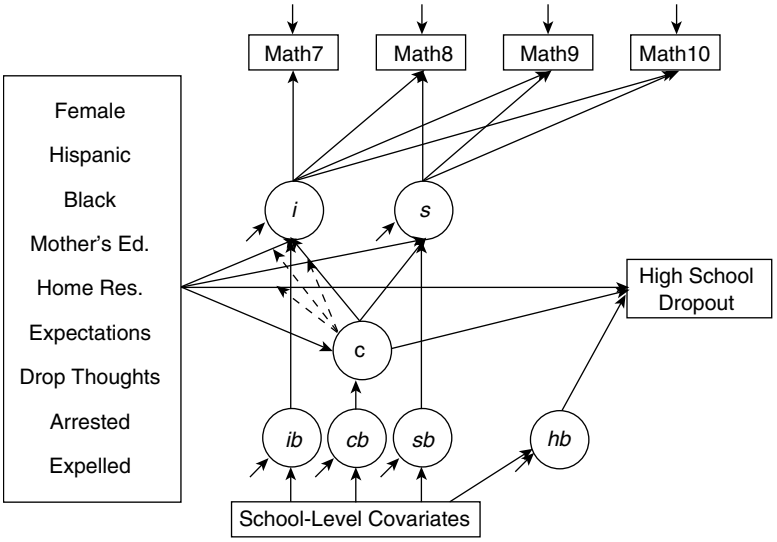
Figure 19.11 also includes a distal outcome of high school dropout and considers across-school variation in its intercept *hb* (there may also be across-school variation in some of the slopes). The intercept variation is again described by school-level covariates. This model part is analogous to two-level logistic regression (see, e.g., Hedeker & Gibbons, 1994). In Figure 19.11, a new feature is that the two-level logistic regression has as one of its predictors a latent categorical variable *c*, the latent trajectory class variable.

A key new feature in Figure 19.11 is the across-school variation *cb* in the individual-level latent class variable *c*. This part of the model makes it possible to study the influence of school-level variables on the class member probability for the students. This corresponds to multinomial logistic regression with random effects, except that the dependent variable is latent.

The model in Figure 19.11 was analyzed using maximum likelihood estimation in Mplus.[16] A key school-level variable used in the modeling was a school poverty index, measured as the percentage of the student body receiving full school lunch support. It was found that this school poverty index did not have a significant effect on the probability of dropping out

16. The Mplus input and output for these analyses are given in Example 9 at www.statmodel.com/mplus/examples/penn.html. This analysis was carried out by Mplus Version 3.

Figure 19.11 Multilevel GGMM for LSAY Data



of high school. It did, however, have a significant influence on *c* in the sense that a high index value resulted in a higher probability of being a member of the class with a poor math achievement trajectory in Grades 7 through 10. The growth mixture analyses reported on earlier showed that membership in the failing class gave a very high risk of dropping out of high school. In this way, the multilevel growth mixture modeling implies that school poverty does not influence dropout directly but indirectly, in that it influences achievement trajectory class, which in turn influences dropout. This is an interesting new type of mediational process, whereby the mediator is not only categorical but also latent.

The general latent variable modeling framework considered here allows multilevel modeling, such as three-level growth modeling, not only for continuous outcomes but also for categorical outcomes. In this way, multilevel modeling is available in Mplus for GGMM, LCGA, LCA, LTA, and DTSMA.

19.8. Conclusions

This chapter has shown how modeling, using a combination of continuous and categorical latent variables, provides an extremely flexible analysis framework. Different traditions such as growth modeling, latent class analysis, and survival analysis are brought together using the unifying theme of latent variable modeling. New developments in these areas have been presented. Not only does this create more interesting analysis options in each area, but the combination of model parts that is possible leads to even further opportunities for investigating data. Several such combinations were not discussed but include the following (see also Muthén, 2001a, 2002; Muthén & Asparouhov, 2003a, 2003b):

- Multiple-process growth mixture modeling
  Parallel (dual) processes: studying relations between concurrent outcomes
  Sequential processes: predicting later growth from earlier growth
- Multiple-group growth mixture modeling: studying similarities and differences across known groups
- Multiple indicator growth mixture modeling: studying growth in a latent variable construct
- Embedded growth mixture modeling: combining the growth model with LCA, factor analysis, path analysis, and SEM components
- Combined growth mixture and discrete-time survival modeling: predicting survival from trajectory classes and vice versa

Mplus covers these models for outcomes that are continuous, binary, ordered polytomous, two-part, zero-inflated Poisson, or combinations thereof, allowing both missing data and cluster data.

REFERENCES

Agresti, A. (1990). *Categorical data analysis.* New York: John Wiley.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics, 55,* 117–128.

Amemiya, T. (1985). *Advanced econometrics.* Cambridge, MA: Harvard University Press.

Asparouhov, T., & Muthén, B. (2003a). *Full-information maximum-likelihood estimation of general two-level latent variable models.* Manuscript in preparation.

Asparouhov, T., & Muthén, B. (2003b). *Maximum-likelihood estimation in general latent variable modeling.* Manuscript in preparation.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London: Arnold.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for over-extraction of latent trajectory classes. *Psychological Methods, 8,* 338–363.

Carlin, J. B., Wolfe R., Brown, C. H., & Gelman, A. (2001). A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. *Biostatistics, 2,* 397–416.

Duan, N., Manning, W. G., Morris, C. N., & Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business and Economic Statistics, 1,* 115–126.

Farrington, D. P., & West, D. J. (1990). The Cambridge study in delinquent development: A prospective longitudinal study of 411 males. In H.-J. Kernere & G. Kaiser (Eds.), *Criminality: Personality, behavior, and life history.* New York: Springer-Verlag.

Hedeker, D. (2000). *A fully semi-parametric mixed-effects regression model for categorical outcomes.* Paper presented at the Joint Statistical Meetings, Indianapolis, IN.

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics, 50,* 933–944.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences.* Thousand Oaks, CA: Sage.

Jeffries, N. O. (2003). A note on "Testing the number of components in a normal mixture." *Biometrika, 90,* 991–994.

Jones, B. L., Nagin, D. S., & Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods & Research, 29,* 374–393.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34,* 1–13.

Land, K. C. (2001). Introduction to the special issue on finite mixture models. *Sociological Methods & Research, 29,* 275–281.

Lin, H., Turnbull, B. W., McCulloch, C. E., & Slate, E. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association, 97,* 53–65.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88,* 767–778.

Lubke, G., & Muthén, B. (2003). *Performance of factor mixture models.* Manuscript submitted for publication.

Masyn, K. (2002, June). *Latent class enumeration revisited: Application of Lo, Mendell, and Rubin to growth mixture models.* Paper presented at the meeting of the Society for Prevention Research, Seattle, WA.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models.* New York: John Wiley.

Miller, J. D., Kimmel, L., Hoffer, T. B., & Nelson, C. (2000). *Longitudinal study of American youth: User's manual.* Evanston, IL: Northwestern University, International Center of the Advancement of Scientific Literacy.

Moffitt, T. E. (1993). Adolescence-limited and life-course persistent antisocial behavior. *Psychological Review, 100,* 674–701.

Molenaar, P. C., & von Eye, A. (1994). On the arbitrary nature of latent variables. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis* (pp. 226–242). Thousand Oaks, CA: Sage.

Muthén, B. (2001a). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1–33). Mahwah, NJ: Lawrence Erlbaum.

Muthén, B. (2001b). *Two-part growth mixture modeling.* Draft.

Muthén, B. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29,* 81–117.

Muthén, B. (2003). Statistical and substantive checking in growth mixture modeling. *Psychological Methods, 8,* 369–377.

Muthén, B., & Asparouhov, T. (2002). *Mixture testing using multivariate skewness and kurtosis.* Manuscript in preparation.

Muthén, B., & Asparouhov, T. (2003a). *Advances in latent variable modeling, part I: Integrating multilevel and structural equation modeling using Mplus.* Manuscript in preparation.

Muthén, B., & Asparouhov, T. (2003b). *Advances in latent variable modeling, part II: Integrating continuous and categorical latent variable modeling using Mplus.* Manuscript in preparation.

Muthén, B., Brown, C. H., Masyn, K., Jo, B., Khoo, S. T., Yang, C. C., et al. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics, 3,* 459–475.

Muthén, B., & Curran, P. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2,* 371–402.

Muthén, B., Jo, B., & Brown, H. (2003). Comment on the Barnard, Frangakis, Hill & Rubin article, Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *Journal of the American Statistical Association, 98,* 311–314.

Muthén, B., Kreuter, F., & Asparouhov, T. (2003). *Applications of growth mixture modeling to non-normal outcomes.* Manuscript in preparation.

Muthén, B., & Masyn, K. (in press). Mixture discrete-time survival analysis. *Journal of Educational and Behavioral Statistics.*

Muthén, B., & Muthén, L. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling

with latent trajectory classes. *Alcoholism: Clinical and Experimental Research, 24,* 882–891.

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55,* 463–469.

Muthén, L., & Muthén, B. (1998–2003). *Mplus user's guide.* Los Angeles: Author.

Muthén, L. K., & Muthén, B. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4,* 599–620.

Nagin, D. S. (1999). Analyzing developmental trajectories: A semi-parametric, group-based approach. *Psychological Methods, 4,* 139–157.

Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model. *Criminology, 31,* 327–362.

Nagin, D. S., & Tremblay, R. E. (2001). Analyzing developmental trajectories of distinct but related behaviors: A group-based method. *Psychological Methods, 6,* 18–34.

Olsen, M. K., & Schafer, J. L. (2001). A two-part random effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association, 96,* 730–745.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Roeder, K., Lynch, K. G., & Nagin, D. S. (1999). Modeling uncertainty in latent class membership: A case study in criminology. *Journal of the American Statistical Association, 94,* 766–776.

Rumberger, R. W., & Larson, K. A. (1998). Student mobility and the increased risk of high school dropout. *American Journal of Education, 107,* 1–35.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464.

Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics, 18,* 155–195.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica, 26,* 24–36.

Wang, C. P., Brown, C. H., & Bandeen-Roche, K. (2002). *Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior.* Manuscript submitted for publication.

Section VI

Foundational Issues

Chapter 20

Probabilistic Modeling With Bayesian Networks

Richard E. Neapolitan

Scott Morris

20.1. Introduction

Given a set of random variables, *probabilistic modeling* consists of acquiring properties of a joint probability distribution of the variables and thereby representing that distribution. These properties can be very important because they often enable us to succinctly represent a distribution and to do inference with the variables. For example, we may be able to concisely represent a joint probability distribution of diseases and manifestations in a medical application and, using this representation, compute the probability that a patient has certain diseases given the patient has some manifestations. First, Section 20.2 gives a brief philosophical overview of the notion of a probability as a relative frequency, which probabilistic modeling using data presupposes. Then, Section 20.3 introduces Bayesian networks and Bayesian network models (also called directed acyclic graph [DAG] models). Next, Section 20.4 discusses learning DAG models. Finally, Section 20.5 shows applications of learning DAG models.

20.2. Philosophical Background

The focus of this chapter is on learning DAG models from data. The enterprise of learning something about a probability distribution from data relies on the notion of a probability as a relative frequency. So we first review the relative frequency approach to probability, and then we discuss its relationship to another approach to probability, called *subjective* or *Bayesian.*

20.2.1. The Relative Frequency Approach to Probability

In 1919, Richard von Mises developed the relative frequency approach to probability, which concerns repeatable identical experiments. First we describe relative frequencies, and then we discuss how we can learn something about them from data.

20.2.1.1. Relative Frequencies

von Mises (1928/1957) formalized the notion of repeatable identical experiments as follows:

> The term is "the *collective,*" and it denotes a sequence of uniform events or processes which differ by certain observable attributes, say colours, numbers, or anything else. (p. 12, emphasis added)

The classical example of a collective is an infinite sequence of tosses of the same coin. Each time we toss

the coin, our knowledge about the conditions of the toss is the same (assuming we do not sometimes "cheat" by, for example, holding it close to the ground and trying to flip it just once). Of course, something is different in the tosses (e.g., the distance from the ground, the torque we put on the coin, etc.) because otherwise, the coin would always land heads or always land tails. But we are not aware of these differences. Our knowledge concerning the conditions of the experiment is always the same. von Mises (1928/1957) argued that, in such repeated experiments, the fraction of occurrence of each outcome approaches a limit, and he called this limit the probability of the outcome. It has become standard to call this limit a *relative frequency* and to use the term *probability* in a more general sense.

Note that the collective (infinite sequence) only exists in theory. We never will toss the coin indefinitely. Rather, the theory assumes that there is a *propensity* for the coin to land heads, and as the number of tosses approaches infinity, the fraction of heads approaches this propensity. For example, if m is the number of times we toss the coin, S_m is the number of heads, and p is the true value of the propensity for the coin to land heads, then

$$p = \lim_{m \to \infty} \frac{S_m}{m}. \tag{1}$$

Because the propensity is a physical property of the coin, it is also called a *physical probability.* In 1946, J. E. Kerrich conducted many experiments using games of chance (e.g., coin tosses) indicating that the fraction does appear to approach a limit.

Note further that a collective is only defined relative to a *random process,* which, in the von Mises theory, is defined to be a repeatable experiment for which the infinite sequence of outcomes is assumed to be a random sequence. Intuitively, a *random sequence* is one that shows no regularity or pattern. For example, the finite binary sequence "1011101100" appears random, whereas the sequence "1010101010" does not because it has the pattern "10" repeated five times. There is evidence that experiments such as coin tossing and dice throwing are indeed random processes. Namely, Iversen, Longcor, Mosteller, Gilbert, and Youtz (1971) ran many experiments with dice indicating that the sequence of outcomes is random. It is believed that unbiased sampling also yields a random sequence and is therefore a random process. See van Lambalgen (1987) for a thorough discussion of this matter, including a formal definition of *random sequence.* Neapolitan (1990) provides a more intuitive, less mathematical treatment. We close here with an example of a nonrandom process. One of the authors prefers to exercise at his health club on Tuesday, Thursday, and Saturday. However, if he misses a day, he usually makes up for it the following day. If we track the days he exercises, we will find a pattern because the process is not random.

Under the assumption that the fraction approaches a limit and that a random sequence is generated, in 1928, von Mises was able to derive the rules of probability theory and the result that the trials are probabilistically independent. In terms of relative frequencies, what does it mean for the trials to be independent? The following example illustrates what it means. Suppose we develop many sequences of length 20 (or any other number), where each sequence represents the result of 20 coin tosses. Then we separate the set of all these sequences into disjoint subsets such that the sequences in each subset all have the same outcome on the first 19 tosses. Independence means that the fraction of heads on the 20th toss is the same in all the subsets (in the limit).

A common way to define probability in applications such as games of chance is to assign the same probability to all possible elemental outcomes. For example, in the draw of the top card from an ordinary deck of cards, each elemental outcome is assigned a probability of 1/52 because there are 52 different cards. Such probabilities are called *ratios.* We say we are using the *principle of indifference* (a term popularized by J. M. Keynes in 1921/1948) when we assign probabilities this way. The probability of a set of elemental outcomes is the sum of the probabilities of the outcomes in the set. For example, the probability of a king is 4/52 because there are four kings. How are relative frequencies related to ratios? Intuitively, we would expect that if, for example, we repeatedly shuffled a deck of cards and drew the top card, the ace of spades would come up about 1 out of every 52 times. In the experiment performed by J. E. Kerrich in 1946 (discussed above), the principle of indifference seemed to apply, and the limit was indeed the value obtained via the principle of indifference.

20.2.1.2. Sampling

Sampling techniques estimate a relative frequency for a given collective from a finite set of observations. In accordance with standard statistical practice, we use the term *random sample* (or simply *sample*) to denote the set of observations, and we call a collective a *population.* Note the difference between a *collective* and a *finite population.* There are currently a finite number of smokers in the world. The fraction of them with lung cancer is the probability (in the sense of a ratio) of a current smoker having lung cancer. The propensity

(relative frequency) of a smoker having lung cancer may not be exactly equal to this ratio. Rather, the ratio is just an estimate of that propensity. When doing statistical inference, we sometimes want to estimate the ratio in a finite population from a sample of the population, and other times we want to estimate a propensity from a finite sequence of observations. For example, TV raters ordinarily want to estimate the actual fraction of people in a nation watching a show from a sample of those people. On the other hand, medical scientists want to estimate the propensity with which smokers have lung cancer from a finite sequence of smokers. One can create a collective from a finite population by returning a sampled item back to the population before sampling the next item. This is called *sampling with replacement.* In practice, it is rarely done, but ordinarily, the finite population is so large that statisticians make the simplifying assumption that sampling is done with replacement. That is, they do not replace the item, but they still assume the finite population is unchanged for the next item sampled. In this chapter, we are always concerned with propensities rather than current ratios, so this simplifying assumption does not concern us.

Estimating a relative frequency from a sample seems straightforward. That is, we simply use S_m/m as our estimate, where m is the number of trials and S_m is the number of successes. However, there is a problem in determining our confidence in the estimate. That is, the von Mises theory only says the limit in Equality 1 physically exists and is p. It is not a mathematical limit in that, given an $\varepsilon > 0$, it offers no means for finding an $M(\varepsilon)$ such that

$$\left| p - \frac{S_m}{m} \right| < \varepsilon \quad \text{for } m > M(\varepsilon).$$

Mathematical probability theory enables us to determine confidence in our estimate of p. First, if we assume the trials are probabilistically independent and the probability for each trial is p, we can prove that S_m/m is the *maximum likelihood* (ML) value of p. That is, if $\mathbf{d}$ is a set of results of m trials, and $P(\mathbf{d} : \hat{p})$ denotes the probability of $\mathbf{d}$ if the probability of success were $\hat{p}$, then S_m/m is the value of $\hat{p}$ that maximizes $P(\mathbf{d} : \hat{p})$. Furthermore, we can prove the weak and strong laws of large numbers. The weak law says the following. Given $\varepsilon, \delta > 0$,

$$P\left(\left| p - \frac{S_m}{m} \right| < \varepsilon\right) > 1 - \delta \quad \text{for} \quad m > \frac{1}{4\delta\varepsilon^2}.$$

So mathematically, we have a means of finding an $M(\varepsilon, \delta)$.

The weak law is not applied directly to obtain confidence in our estimate. Rather, we obtain a confidence interval using the following result, which is obtained in a standard statistics text such as Brownlee (1965). Suppose we have m independent trials, and the probability of success on each trial is p, and we have k successes. Let

$$0 < \beta < 1,$$

$$\alpha = (1 - \beta)/2,$$

$$\theta_1 = \frac{kF_\alpha(2k, 2[m - k + 1])}{m - k + 1 + kF_\alpha(2x, 2[m - k + 1])},$$

$$\theta_2 = \frac{k}{(m - k + 1)F_{1-\alpha}(2[m - k + 1], 2k) + (k)},$$

where F is the F distribution. Then,

$$(\theta_1, \theta_2) \text{ is a } \beta\% \text{ confidence interval for } p.$$

This means that $\beta\%$ of the time, the interval generated will contain p.

Example 1. Suppose we toss a thumbtack 30 times and it lands heads (i.e., on its head) 8 times. Then the following is a 95% confidence interval for p, the probability of heads:

$$(.123,\ .459).$$

Because 95% of the time we will obtain an interval that contains p, we are pretty confident p is in this interval.

One should not conclude that mathematical probability theory somehow proves S_m/m will be close to p and that therefore we have no need for the von Mises theory. Without some assumption about S_m/m approaching p, the mathematical result would say nothing about what is happening in the world. For example, without some such assumption, our explanation of confidence intervals would become the following: Suppose we have a sample space determined by m identically distributed independent discrete random variables, where p is the probability that each of them assumes its first value. Consider the random variable whose possible values are the probability intervals obtained using the method for calculating a $\beta\%$ confidence interval. Then, β is the probability that the value of this random variable is an interval containing p. This result says nothing about what will happen when, for example, we toss a thumbtack m times. However, if we assume that the probability (relative frequency) of an event is the limit of the ratio of occurrences of the event in the world, this means that if we repeatedly did

the experiment of tossing the thumbtack m times, in the limit, 95% of the time we will generate an interval containing p, which is how we described confidence intervals above.

Some probabilists find fault with the von Mises theory because it assumes that the observed relative frequency definitely approaches p. For example, Ash (1970) says,

> An attempt at a frequency definition of probability will cause trouble. If S_n is the number of occurrences of an event in n independent performances of an experiment, we expect physically that the relative frequency S_n/n should converge to a limit; however, we cannot assert that the limit exists in a mathematical sense. In the case of the tossing of an unbiased coin, we expect $S_n/n \to 1/2$, but a conceivable outcome of the process is that the coin will keep coming up heads forever. In other words, it is possible that $S_n/n \to 1$, or that $S_n/n \to$ any number between 0 and 1, or that S_n/n has no limit at all. (p. 2)

As mentioned previously, in 1946, J. E. Kerrich conducted many experiments using games of chance indicating that the relative frequency does appear to approach a limit. However, even if it is only most likely that a limit is approached, Kerrich's experiments may indicate that this happens. So to resolve the objection posed by Ash, in 1992, R. E. Neapolitan obtained von Mises's results concerning the rules of probability by assuming $S_m/m \to p$ only in the sense of the weak law of large numbers.

20.2.2. The Subjective/Bayesian Approach to Probability

Next we discuss another approach to probability called the *subjective* or *Bayesian approach.* First we describe the approach and then show how its proponents use Bayes's theorem; finally, we discuss its relevance to relative frequencies.

20.2.2.1. Subjective Probabilities

We start with an example.

Example 2. If you were going to bet on an upcoming basketball game between the Chicago Bulls and the Detroit Pistons, you would want to ascertain how probable it was that the Bulls would win. This probability is certainly not a ratio, and it is not a relative frequency because the game cannot be repeated many times under the exact same conditions (actually, with your knowledge about the conditions the same). Rather the probability only represents your belief concerning the Bulls' chances of winning.

A probability such as the one illustrated in the previous example is called a *degree of belief* or *subjective probability.* There are a number of ways for ascertaining such probabilities. One of the most popular methods is the following, which was suggested by D. V. Lindley in 1985. This method says an individual should liken the uncertain outcome to a game of chance by considering an urn containing white and black balls. The individual should determine for what fraction of white balls the individual would be indifferent between receiving a small prize if the uncertain outcome happened (or turned out to be true) and receiving the same small prize if a white ball was drawn from the urn. That fraction is the individual's probability of the outcome. Such a probability can be constructed using binary cuts. If, for example, you were indifferent when the fraction was .75, for you, $P(\{bullswin\}) = .75$. If someone else were indifferent when the fraction was .6, for that individual, $P(\{bullswin\}) = .6$. Neither individual is right or wrong. Subjective probabilities are unlike ratios and relative frequencies in that they do not have objective values on which we all must agree. Indeed, that is why they are called *subjective*. Neapolitan (1996) discusses the construction of subjective probabilities further.

When we are able to compute ratios or estimate relative frequencies, the probabilities obtained agree with most individuals' beliefs. For example, most individuals would assign a subjective probability of 1/13 to the top card being an ace because they would be indifferent between receiving a small prize if it were the ace and receiving that same small prize if a white ball were drawn from an urn containing 1 white ball out of 13 total balls.

20.2.2.2. Using Bayes's Theorem

The subjective probability approach is called *Bayesian* because its proponents use Bayes's theorem to infer unknown probabilities from known ones. The following example illustrates this.

Example 3. Suppose Joe has a routine diagnostic chest X ray required of all new employees at Colonial Bank, and the X ray comes back positive for lung cancer. Joe then becomes certain he has lung cancer and panics. But should he? Without knowing the accuracy of the test, Joe really has no way of knowing how probable it is that he has lung cancer. When he discovers the test is not absolutely conclusive, he decides

to investigate its accuracy and he learns that it has a false-negative rate of .4 and a false-positive rate of .02. We represent this accuracy as follows. First we define these random variables:

Variable	*Value*	*When the Variable Takes This Value*
Test	Positive	X-ray is positive
	Negative	X-ray is negative
LungCancer	Present	Lung cancer is present
	Absent	Lung cancer is absent

We then have these conditional probabilities:

$$P(Test = positive|Lung\ cancer = present) = .6.$$

$$P(Test = positive|Lung\ cancer = absent) = .02.$$

Given these probabilities, Joe feels a little better. However, he then realizes he still does not know how probable it is that he has lung cancer. That is, the probability of Joe having lung cancer is $P(Lung\ cancer = present|Test = positive)$, and this is not one of the probabilities listed above. Joe finally recalls Bayes's theorem and realizes he needs yet another probability to determine the probability of his having lung cancer. That probability is $P(Lung\ cancer = present)$, which is the probability of his having lung cancer before any information on the test results was obtained. Even though this probability is not based on any information concerning the test results, it is based on some information. Specifically, it is based on all information (relevant to lung cancer) known about Joe before he took the test. The only information about Joe, before he took the test, was that he was one of a class of employees who took the test routinely required of new employees. So, when he learns only 1 out of every 1,000 new employees has lung cancer, he assigns .001 to $P(Lung\ cancer = present)$. He then employs Bayes's theorem as follows:

$$\begin{aligned} &P(present|positive) \\ &\quad = \frac{P(positive|present)\,P(present)}{P(positive|present)\,P(present) \\ &\qquad + P(positive|absent)\,P(absent)} \\ &\quad = \frac{(.6)(.001)}{(.6)(.001) + (.02)(.999)} = .029. \end{aligned}$$

So Joe now feels that the probability of his having lung cancer is only about .03, and he relaxes a bit while waiting for the results of further testing.

A probability such as $P(Lung\ cancer = present)$ is called a *prior probability* because, in a particular model, it is the probability of some event prior to updating the probability of that event, within the framework of that model, using new information. Do not mistakenly think it means a probability prior to any information because $P(Lung\ cancer = present)$ is based on some information obtained from past experience. A probability such as $P(Lung\ cancer = present|Test = positive)$ is called a *posterior probability* because it is the probability of an event after its prior probability has been updated, within the framework of some model, based on new information.

A strict frequentist (e.g., von Mises) could not infer the probability of Joe having lung cancer using Bayes's theorem. That is, from the data used to obtain the false-negative rate, false-positive rate, and prior probability, a strict frequentist could obtain confidence intervals for the actual relative frequencies and maximum likelihood values. However, because strict frequentists do not have subjective probabilities, they cannot obtain subjective probabilities of the test results and of the presence of lung cancer from the data and then use these subjective probabilities to compute the subjective probability of Joe having lung cancer. On the other hand, using a subjective approach, Joe can obtain beliefs from the data, regardless of the size of the samples, and proceed using Bayes's theorem.

A statistician who uses Bayes's theorem is sometimes called a *Bayesian.* I. J. Good (1983) shows that there are 46,656 different Bayesian interpretations (he notes that von Mises's view is not one of these). He bases this on 11 different facets of the approach on which Bayesians can differ. Briefly, there is a descriptive Bayesian interpretation that maintains that humans reason using subjective probabilities and Bayes's theorem, there is a normative Bayesian interpretation that says humans should reason that way, and there is an empirical Bayesian interpretation that says, based on data, we can update our beliefs concerning a relative frequency using Bayes's theorem. Mulaik, Raju, and Harshman (1997) discuss and criticize the first two views. The methods presented in this chapter concern only the third view, which we briefly illustrated in Example 3. In the next subsection, we discuss that view further.

Before that, we note that one of the facets distinguishing types of Bayesians is the concept of *physical probability,* as developed in Section 20.2.1. The three categories for this facet are (a) assumed to exist, (b) denied, and (c) used as if they exist but without philosophical commitment. Both Good (1983) and ourselves are in Category 3. Briefly, in applications such as sampling individuals and determining whether they have lung cancer, a limit seems to be approached. However, an infinite sequence of sampled items only

exists as an idealization. That is, the circumstances of the experiment (e.g., pollution, changes in health care, etc.) change as time goes. Even a coin's composition changes as we toss it. In these situations, it seems philosophically difficult to maintain that a physical probability, accurate to an arbitrary number of digits, exists at any given point in time. It appears that precise relative frequencies exist in very few applications. These include applications such as the repeated drawing of a top card from a deck of cards, the estimation of the ratio in a finite population from a sample of the population (see Section 20.2.1.2), and perhaps some applications in physics such as statistical mechanics. In most real-world applications, the notion of a relative frequency is an idealization, which can be used as a model for a subjective approach to probability.

20.2.2.3. Bayesian Learning of Relative Frequencies

At the end of Section 20.2.1.2, we discussed how a frequentist learns something about a relative frequency from data by obtaining a confidence interval for the relative frequency. We illustrate the issue with the following example:

Example 4. Suppose we sampled 100 American males and the average height turned out to be 4 feet. Using a confidence interval, we would become highly confident that the average height of American males is close to 4 feet.

Good (1983) would say that we are "sweeping our prior experience under the carpet" to reach the absurd conclusion in the previous example. He maintains that we should instead assign a prior probability distribution to the average height based on our prior knowledge, and then we should update this distribution based on the data. On the other hand, Mulaik et al. (1997) criticize the use of prior probabilities when they state, "at the outset, subjective/personal Bayesian inference based on these subjective/personal probabilities does not give what is just the evidentiary reasons to believe in something and is unable to separate in its inference what is subjective from what is objective and evidentiary." Both authors have their points. However, it seems their points concern different circumstances. Example 4 showed one situation in which we would not want to sweep our prior knowledge under the carpet. As another example, suppose we see a given coin land heads five times in a row. We would not want to sweep our prior knowledge about coins under the carpet and bet according to the belief that it will most likely land heads on the next toss. Rather, we would want to update our prior belief with the data consisting of the five heads. Of more practical interest, consider the development of a medical expert system, which is a system used to diagnose illnesses based on symptoms. Suppose we had access to the knowledge of a renowned medical authority. We would not want to sweep the authority's knowledge about the probabilistic relationships among the domain variables under the carpet and only use information from a database in our system. Rather, we would want to develop our system based both on the authority's knowledge and on what could be learned from the database. On the other hand, suppose a pharmaceutical company is testing the effectiveness of a new drug by performing a treatment study, and it wants to communicate its result to the scientific community. The scientific community would not be interested in the company's prior belief concerning the drug's effectiveness but only in what the data had to say. So in this case, even if its prior belief was that the drug was effective, it would not be acceptable to base the stated result partly on that belief.

When we obtain an updated probability distribution for a relative frequency, we can obtain, for example, a 95% probability interval for the relative frequency. The probability interval is the Bayesian's counterpart to the confidence interval. Neapolitan (2003) shows that, in many cases, they are mathematically identical.

20.3. BAYESIAN NETWORK (DAG) MODELS

First we introduce Bayesian networks; then we discuss Bayesian network (DAG) models.

20.3.1. Bayesian Networks

Suppose we have a joint probability distribution P of the random variables in some set $\mathbf{V}$, and a DAG $G = (\mathbf{V}, \mathbf{E})$. We say that (G, P) satisfies the *Markov condition* if, for each variable, $X \in \mathbf{V}$, X is conditionally independent of the set of all its nondescendents given the set of all its parents. We call (G, P) a *Bayesian network* if (G, P) satisfies the Markov condition. If (G, P) satisfies the Markov condition, it is possible to show that P is the product of its conditional distributions in G, and this is the way P is always represented in a Bayesian network. Furthermore, if we specify a DAG G and any discrete conditional distribution (and many continuous ones), the probability distribution that is the product of

the conditional distributions satisfies the Markov condition with the DAG, and so we obtain a Bayesian network. This is the way Bayesian networks are constructed in practice. See Neapolitan (2003) for proofs and further discussion of these facts.

If, for example, $\{X\}$ and $\{Y, W\}$ are conditionally independent given $\{Z\}$ in probability distribution P, this means for all values of x, y, w, and z, we have

$$P(x|y, w, z) = P(x|z).$$

We represent this conditional independency succinctly by $I_P(\{X\}, \{Y, W\}|\{Z\})$. If there is no conditioning bar, it means they are conditionally independent given the empty set of variables, which means they are simply independent. If a set only contains one element, we sometimes do not show curly braces.

Example 5. Suppose we have the following random variables:

Variable	*Value*	*When the Variable Takes This Value*
S	$s1$	There is a history of smoking
	$s2$	There is no history of smoking
B	$b1$	Bronchitis is present
	$b2$	Bronchitis is absent
L	$l1$	Lung cancer is present
	$l2$	Lung cancer is absent
F	$f1$	Fatigue is present
	$f2$	Fatigue is absent
X	$x1$	Chest X-ray is positive
	$x2$	Chest X-ray is negative

Figure 20.1 shows a Bayesian network containing those variables in which the conditional distributions were estimated from actual data. The probability distribution P in the Bayesian network is the product of the conditional distributions. For example,

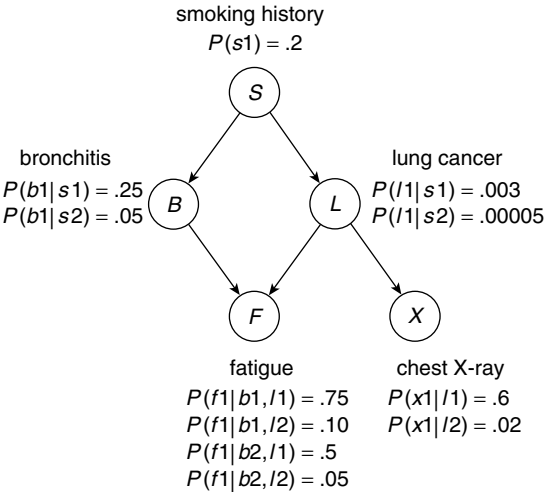
$$\begin{aligned} &P(f1, c1, b1, l1, s1) \\ &\quad = P(f1|b1, l1)P(c1|l1)P(b1|s1)P(l1|s1)P(s1) \\ &\quad = (.75)(.6)(.25)(.003)(.2) = .0000675. \end{aligned}$$

Note that there are only 11 parameter values in the Bayesian network, but there are 32 values in the joint probability distribution. When the DAG in a Bayesian network is sparse, a Bayesian network is a very succinct way to represent a probability distribution.

Each variable in the network is conditionally independent of its nondescendents given its parents. For example, we have

$$I_P(B, \{L, X\}|\{S\}).$$

Figure 20.1 A Bayesian Network



Approximately assuming that relative frequencies exist (see the end of Section 20.2.2.2), there is some actual relative frequency distribution F (for frequency) of the five variables in Example 5. It is argued that if we draw a causal DAG, F satisfies the Markov condition with that DAG (see Spirtes, Glymour, & Scheines, 1993, 2000). By a *causal DAG,* we mean a DAG in which each edge represents a direct causal influence. The DAG in Figure 20.1 is a causal DAG. The following example briefly illustrates why a causal DAG should satisfy the Markov condition with F. Smoking causes both lung cancer and bronchitis. So the presence of lung cancer makes it more probable the person is a smoker. Because smoking also causes bronchitis, this increased likelihood of smoking makes it more probable that the person has bronchitis. So lung cancer and smoking are not independent. However, if we know that the person is a smoker, that person has a certain probability of having bronchitis based on this information. Because lung cancer can now no longer increase the likelihood of smoking (we know the person smokes), it cannot increase the likelihood of bronchitis through this chain. So bronchitis and lung cancer are conditionally independent given smoking, as entailed by the Markov condition.

The conditional distributions in Figure 20.1 were obtained from data and are only estimates of the actual conditional relative frequency distributions. So their product P is only an estimate of the actual joint relative frequency distribution F. Nevertheless, their product still satisfies the Markov condition with the DAG because, as mentioned above, if

we specify any discrete conditional probability distributions, their product satisfies the Markov condition with the DAG. So we have a Bayesian network that contains an estimate of the relative frequency distribution.

There are two main applications of Bayesian networks. The first is in expert systems (see Neapolitan, 1990). An *expert system* is a system that is capable of making the inferences and possibly the decisions of an expert. For example, we might include the Bayesian network in Figure 20.1 in an expert system whose purpose is to diagnose and treat respiratory problems. The system would need to perform probabilistic inference such as the computation of $P(l1|s1, c1)$. Inference algorithms for Bayesian networks have been developed, which are efficient for a large class of networks (see Castillo, Gutiérrez, & Hadi, 1997; Neapolitan, 1990; Pearl, 1988). However, Cooper (1990) has shown that the problem of inference in a Bayesian network is NP-hard. Bayesian networks that are augmented with decision nodes and a value node are called *influence diagrams* (see Clemen, 2000). An influence diagram can recommend decisions such as treatment options in the medical domain.

Initially, Bayesian networks for expert systems were constructed using the knowledge of domain experts. However, in the 1990s, a great deal of research was done on learning both the structure (the DAG) and the parameter values (the conditional probabilities) from data. For example, if we had data on the five variables in Example 5, we might learn the Bayesian network in Figure 20.1.

The second application of Bayesian networks only concerns learning. In these applications, we try to learn something about the causal relationships among the variables from data (see Spirtes et al., 1993, 2000).

20.3.2. Modeling With Bayesian Networks

A *probabilistic model* M for a set $\mathbf{V}$ of random variables is a set of joint probability distributions of the variables. Ordinarily, a model is specified using a parameter set $\mathbf{F}$ and combinatoric rules for determining the joint probability distribution from the parameter set. Each member of the model is then obtained by assigning values to the members of $\mathbf{F}$ and applying the rules. If probability distribution P is a member of model M, we say that P is *included* in M. If the probability distributions in a model are obtained by assignments of values to the members of a parameter set $\mathbf{F}$, this means there is some assignment of values to the parameters that yield the probability distribution. A conditional independency common to all probability distributions included in model M is said to be *in* M. An example of a probabilistic model follows.

Example 6. Suppose we are going to toss a die and a coin, neither of which are known to be fair. Let X be a random variable whose value is the outcome of the die toss, and let Y be a random variable whose value is the outcome of the coin toss. Then the space of X is $\{1, 2, 3, 4, 5, 6\}$, and the space of Y is $\{heads, tails\}$. The following is a probabilistic model M for the joint probability distribution of X and Y:

1. $\mathbf{F} = \{f_{11}, f_{12}, f_{13}, f_{14}, f_{15}, f_{16}, f_{21}, f_{22}\}$, $0 \leq f_{ij} \leq 1$, $\sum_{j=1}^{6} f_{1_j} = 1$, $\sum_{j=1}^{2} f_{2j} = 1$.
2. For each permissible combination of the parameters in $\mathbf{F}$, obtain a member of M as follows:

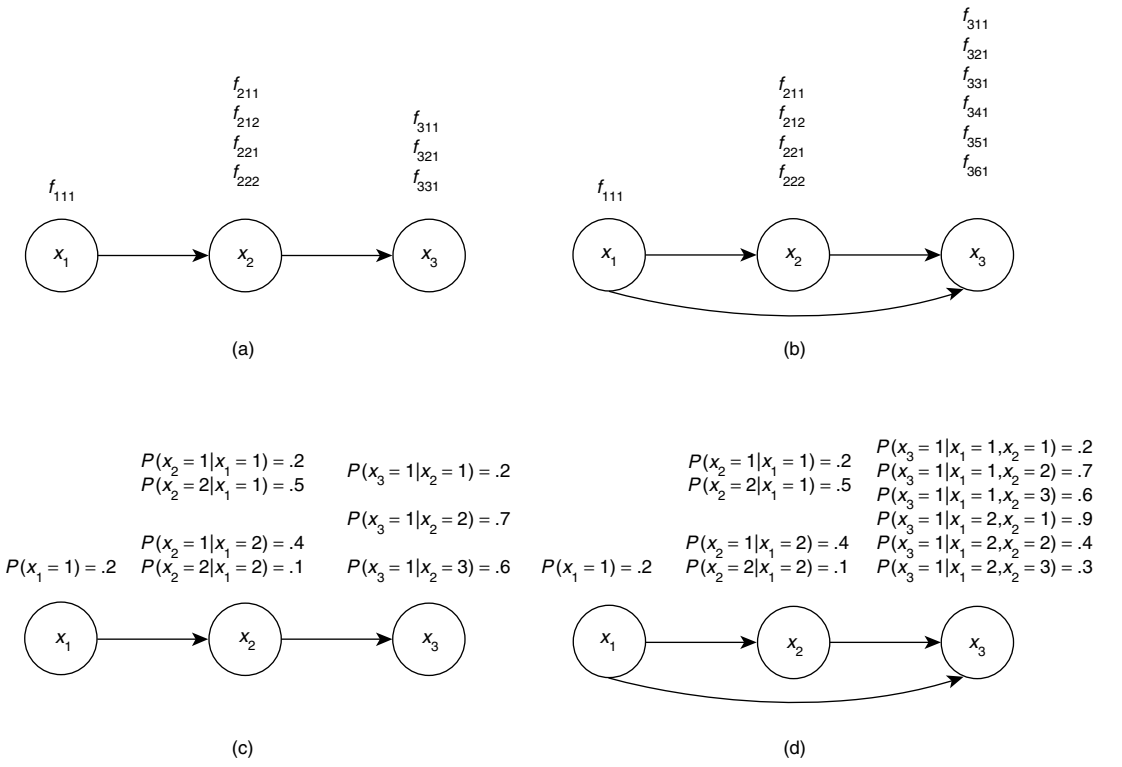
$$P(X = i, Y = heads) = f_{1i} f_{21},$$
$$P(X = i, Y = tails) = f_{1i} f_{22}.$$

The conditional independency $I_P(X, Y)$ is in M. Any probability distribution of X and Y for which X and Y are independent is included in M; any probability distribution of X and Y for which X and Y are not independent is not included in M.

A *Bayesian network model* (or *DAG model*) consists of a DAG $G = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is a set of random variables, and a parameter set $\mathbf{F}$ whose members determine conditional probability distributions for the DAGs, such that for every permissible assignment of values to the members of $\mathbf{F}$, the joint probability distribution of $\mathbf{V}$ is given by the product of these conditional distributions, and this joint probability distribution satisfies the Markov condition with the DAG. Owing to the discussion at the beginning of Section 20.3.1, if $\mathbf{F}$ determines discrete probability distributions (and many continuous ones), the product of the conditional distributions will satisfy the Markov condition. For simplicity, we ordinarily denote a Bayesian network model using only G (i.e., we do not show $\mathbf{F}$).

Example 7. Bayesian network models appear in Figure 20.2a, b. The conditional independency $I_P(X_3, X_1|\{X_2\})$ is in the model in Figure 20.2a; no conditional independencies are in the model in Figure 20.2b. The probability distribution contained in the Bayesian network in Figure 20.2c is included in both models, whereas the one in the Bayesian network in Figure 20.2d is included only in the model in Figure 20.2b. That is, even though there are no conditional independencies in the model in Figure 20.2b, the distribution in Figure 20.2c is in that model.

Figure 20.2 Bayesian Network Models



NOTE: Bayesian network models appear in (a) and (b). The probability distribution in the Bayesian network in (c) is included in both models, whereas the one in (d) is included only in the model in (b).

A set of models, each of which is for the same set of random variables, is called a *class* of models.

Example 8. The set of Bayesian network models containing the same discrete random variables is a class of models. We call this class a *multinomial Bayesian network model class*. Figure 20.2 shows two models from the class when $\mathbf{V} = \{X_1, X_2, X_3\}$, X_1 and X_3 are binary, and X_2 has space size 3.

Given some class of models, if M_2 includes probability distribution P and there exists no M_1 in the class, such that M_1 includes P, and M_1 has smaller dimension than M_2, then M_2 is called a *parameter optimal map.* In the case of DAG models, the *dimension* of the model is the number of parameters in the models. However, as discussed in Neapolitan (2003), this is not always the case. The dimension of the model in Figure 20.2a is 8, whereas the dimension of the one in Figure 20.2b is 11. The model in Figure 20.2a is a parameter optimal map of the probability distribution in the Bayesian networks in Figure 20.2c.

20.4. Learning DAG Models

In general, the problem of *model selection* is to find a concise model that, based on a random sample of observations from the population that determines a probability (relative frequency) distribution P, includes P. So given a class of models, we would want to find a parameter optimal map of P. We use $\mathbf{d}$ to represent the set of values (data) in the sample. After discussing the Bayesian method for learning DAG models, we illustrate the constraint-based method.

20.4.1. Bayesian Method

Although we develop the theory using binary variables, the theory extends to multinomial and multivariate normally distributed variables (see Neapolitan, 2003).

One way to perform model selection is to develop a scoring function *score* (called a *scoring criterion*) that assigns a value $score(\mathbf{d}, M)$ to each model under

consideration based on the data. We have the following definition concerning scoring criteria:

Definition 1. Let $\mathbf{d}_M$ be a set of values (data) of a set of M mutually independent random vectors, each with probability distribution P, and let P_M be the probability function determined by the joint distribution of the M random vectors. Furthermore, let *score* be a scoring criterion over some class of models for the random variables that constitute each vector. We say *score* is *consistent* for the class of models if the following two properties hold:

1. If M_1 includes P and M_2 does not, then

$$\lim_{M\to\infty} P_M(score(\mathbf{d}_M, M_1) > score(\mathbf{d}_M, M_2)) = 1.$$

2. If M_1 and M_2 both include P, and M_1 has a smaller dimension than M_2, then

$$\lim_{M\to\infty} P_M(score(\mathbf{d}_M, M_1) > score(\mathbf{d}_M, M_2)) = 1.$$

We call P the *generative distribution.* The limit, as the size of the data set approaches infinity, of the probability of a consistent scoring criterion choosing a parameter optimal map of P is 1.

The Bayesian scoring criterion $score_B$, which is the probability of the data given the DAG, is a consistent scoring criterion. Before showing that criterion, we need to discuss quantifying our belief concerning a relative frequency.

20.4.1.1. Quantifying Our Prior Belief

First we present the beta density function.

Definition 2. The *beta density function* with parameters $a, b, N = a + b$, where a and b are real numbers > 0, is

$$\rho(f) = \frac{\Gamma(N)}{\Gamma(a)\Gamma(b)} f^{a-1}(1-f)^{b-1} \qquad 0 \le f \le 1.$$

A random variable F that has this density function is said to have a *beta distribution.*

We refer to the beta density function as $beta(f; a, b)$.

Γ is the gamma function. If x is an integer ≥ 1, it is possible to show $\Gamma(x) = (x-1)!$. The uniform density function is $beta(f; 1, 1)$. Figures 20.3 and 20.4 show the $beta(f; 3, 3)$ and $beta(f; 18, 2)$ density functions.

Figure 20.3 The $beta(f; 3, 3)$ Density Function

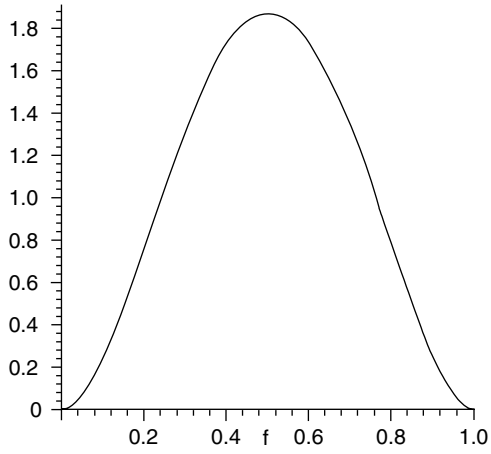
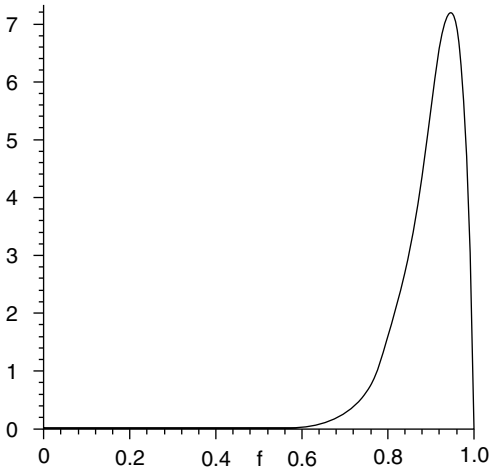


Figure 20.4 The $beta(f; 18, 2)$ Density Function



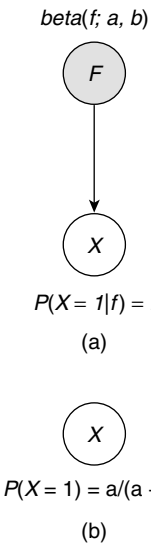
We discuss these figures further after we introduce the method for quantifying our prior belief concerning a relative frequency.

In general, the Bayesian approach is to assume we can represent our belief concerning the relative frequency with which X equals 1 using a random variable F whose space is the interval $[0, 1]$. We further assume our beliefs are such that

$$P(X = 1|f) = f.$$

That is, if we knew for a fact that the relative frequency with which X equals 1 was f, our belief concerning the occurrence of 1 in the first execution of the experiment would be f. This situation is

Figure 20.5 An Augmented Bayesian Network (a) and the Network That It Embeds (b)



represented by the Bayesian network in Figure 20.5a, in which we have assumed that F has the $beta(f; a, b)$ density function. We stress that the theory does not require that we use the beta density function for F. Rather, it is just that this function is commonly used and is convenient for this overview. We call such a Bayesian network an *augmented Bayesian network* because it augments another Bayesian network (in this case, one containing the single node X) with node(s) representing our beliefs about relative frequencies. The Bayesian network containing the single node X and its marginal distribution is said to be *embedded* in the augmented Bayesian network. This embedded Bayesian network appears in Figure 20.5b. Note in that network that $P(X = 1) = a/(a + b)$. The following theorem obtains this result.

Theorem 1. Suppose we have an augmented Bayesian network containing nodes X and F, and F has the $beta(f; a, b)$ density function. Then the marginal distribution of X is given by

$$P(X = 1) = E(F) = \frac{a}{a + b},$$

where E is the expected value.

Proof. The proof appears in Neapolitan (2003).

The beta density function is often used to quantify a belief concerning a relative frequency. Briefly, the reason is as follows: Notice in Figures 20.3 and 20.4 that the larger the values of a and b are, the more the mass is concentrated around $a/(a + b)$. For this and other reasons, when a and b are integers, we often say the values of a and b are such that the probability assessor's experience is equivalent to having seen the first outcome occur a times in $a + b$ trials. Zabell (1982) discusses this matter more and proves that, if we make certain assumptions about an individual's beliefs, then that individual must use the beta density function to quantify any prior beliefs about a relative frequency. Zabell's theorem actually concerns the Dirichlet distribution, which is a generalization of the beta distribution to more than two outcomes. The Dirichlet distribution is used to quantify prior beliefs concerning relative frequencies when we have multinomial variables.

20.4.1.2. The Bayesian Scoring Criterion

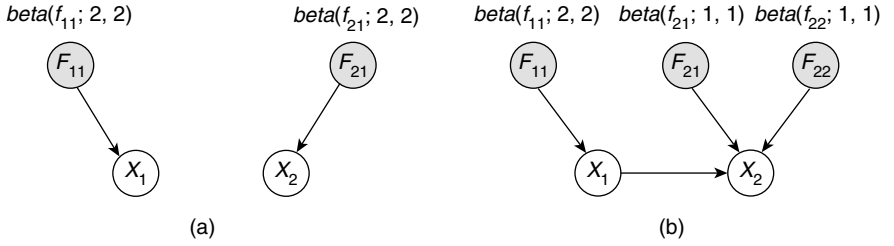
Now we can present the Bayesian scoring criterion, which is given by

$$\begin{aligned} score_B(\mathbf{d}, G) &= P(\mathbf{d}|G) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{q_i^{(G)}} \frac{\Gamma(N_{ij}^{(G)})}{\Gamma(N_{ij}^{(G)} + M_{ij}^{(G)})} \\ &\quad \cdot \frac{\Gamma(a_{ij}^{(G)} + s_{ij}^{(G)})\Gamma(b_{ij}^{(G)} + t_{ij}^{(G)})}{\Gamma(a_{ij}^{(G)})\Gamma(b_{ij}^{(G)})} \end{aligned} \quad (2)$$

where n is the number of variables in the DAG; q_i is the number of different instantiations of the parents of X_i; F_{ij} is a random variable representing our belief concerning the relative frequency with which X_i equals 1, given that the parents of X_i are in the jth instantiation; F_{ij} has the $beta(f_{ij}; a_{ij}, b_{ij})$ density function, $N_{ij} = a_{ij} + b_{ij}$; M_{ij} is the number of cases in which X_i's parents are in their jth instantiation; and of these M_{ij} cases, s_{ij} is the number in which X_i is equal to 1, and t_{ij} is the number in which it is equal to 2. This scoring criterion is the binary-variable special case of the multinomial-variable scoring criterion first developed in Cooper and Herskovits (1992). Geiger and Heckerman (1994) have developed a Bayesian scoring criterion in the case of Bayesian networks containing multivariate normally distributed variables.

Neapolitan (2003) shows that $score_B$ is consistent for Bayesian networks containing multinomial distributed variables. We say P admits a *faithful DAG representation* if there exists a DAG G for which the Markov condition entails all and only the conditional independencies in P. In this case, we say P and G are faithful to each other. Neapolitan (2003) shows further

Figure 20.6 Prior Augmented Bayesian Networks



that if the generative distribution P admits a faithful DAG representation (not all P do), the limit, as the size of the data set approaches infinity, of the probability of a consistent scoring criterion choosing a DAG faithful to P is 1.

Next we present two examples of scoring using the Bayesian scoring criterion.

Example 9. Suppose we have the data **d** in the following table:

Case	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
X_1	1	1	1	1	2	2	2	2
X_2	1	1	1	1	2	2	2	2

Let G_I be the DAG with no edges and G_D be $X_1 \rightarrow X_2$. Then, using the prior augmented Bayesian networks in Figure 20.6a, b to quantify our prior beliefs, we have

$$\begin{aligned} &score_B(\mathbf{d}, G_I) \\ &\quad = \left(\frac{\Gamma(4)}{\Gamma(4+8)}\frac{\Gamma(2+4)\Gamma(2+4)}{\Gamma(2)\Gamma(2)}\right) \\ &\qquad \cdot \left(\frac{\Gamma(4)}{\Gamma(4+8)}\frac{\Gamma(2+4)\Gamma(2+4)}{\Gamma(2)\Gamma(2)}\right), \\ &\quad = 4.6851 \times 10^{-6} \end{aligned}$$

$$\begin{aligned} &score_B(\mathbf{d}, G_D) \\ &\quad = \left(\frac{\Gamma(4)}{\Gamma(4+8)}\frac{\Gamma(2+4)\Gamma(2+4)}{\Gamma(2)\Gamma(2)}\right) \\ &\qquad \cdot \left(\frac{\Gamma(2)}{\Gamma(2+4)}\frac{\Gamma(1+4)\Gamma(1+0)}{\Gamma(1)\Gamma(1)}\right) \\ &\qquad \cdot \left(\frac{\Gamma(2)}{\Gamma(2+4)}\frac{\Gamma(1+0)\Gamma(1+4)}{\Gamma(1)\Gamma(1)}\right) \\ &\quad = 8.658 \times 10^{-5}. \end{aligned}$$

An attractive feature of the Bayesian scoring criterion is that it enables us to incorporate prior probabilities (beliefs) into our posterior beliefs concerning the models. For example, if we assign

$$P(G_I) = P(G_D) = .5,$$

then owing to Bayes's theorem,

$$\begin{aligned} P(G_I|\mathbf{d}) &= \alpha P(\mathbf{d}|G_D)P(G_D) \\ &= \alpha(4.6851 \times 10^{-6})(.5) \end{aligned}$$

and

$$\begin{aligned} P(G_D|\mathbf{d}) &= \alpha P(\mathbf{d}|G_I)P(G_I) \\ &= \alpha(8.658 \times 10^{-5})(.5), \end{aligned}$$

where α is a normalizing constant equal to $1/P(\mathbf{d})$. Eliminating α, we have

$$P(G_I|\mathbf{d}) = .05134$$

and

$$P(G_D|\mathbf{d}) = .94866.$$

Notice that we become highly confident that the DAG is the one with the dependency because in the data, the variables are deterministically related.

Example 10. Suppose we have the data **d** in the following table:

Case	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
X_1	1	1	1	1	2	2	2	2
X_2	1	1	2	2	1	1	2	2

Then using the prior augmented Bayesian networks in Figure 20.6a, b, we have

$$\begin{aligned} &score_B(\mathbf{d}, G_I) \\ &\quad = \left(\frac{\Gamma(4)}{\Gamma(4+8)}\frac{\Gamma(2+4)\Gamma(2+4)}{\Gamma(2)\Gamma(2)}\right) \\ &\qquad \cdot \left(\frac{\Gamma(4)}{\Gamma(4+8)}\frac{\Gamma(2+4)\Gamma(2+4)}{\Gamma(2)\Gamma(2)}\right) \\ &\quad = 4.6851 \times 10^{-6}, \end{aligned}$$

$$\begin{aligned} &score_B(\mathbf{d}, G_D) \\ &\quad = \left(\frac{\Gamma(4)}{\Gamma(4+8)}\frac{\Gamma(2+4)\Gamma(2+4)}{\Gamma(2)\Gamma(2)}\right) \\ &\qquad \cdot \left(\frac{\Gamma(2)}{\Gamma(2+4)}\frac{\Gamma(1+2)\Gamma(1+2)}{\Gamma(1)\Gamma(1)}\right) \\ &\qquad \cdot \left(\frac{\Gamma(2)}{\Gamma(2+4)}\frac{\Gamma(1+2)\Gamma(1+2)}{\Gamma(1)\Gamma(1)}\right) \\ &\quad = 2.405 \times 10^{-6}. \end{aligned}$$

If we assign

$$P(G_I) = P(G_D) = .5,$$

then proceeding as in the previous example, we obtain

$$P(G_I|\mathbf{d}) = .66079$$

and

$$P(G_D|\mathbf{d}) = .33921.$$

Notice that we become fairly confident that the DAG is the one with the independency because in the data, the variables are independent.

Although we illustrated the method using just two variables, Equality 2 scores DAGs containing an arbitrary number of variables, and so clearly the method applies to the general case of n variables.

20.4.1.3. Data Compression Scoring Criteria

As an alternative to the Bayesian scoring criterion, Rissanen (1987), Lam and Bacchus (1994), and Friedman and Goldszmidt (1996) developed and discussed a scoring criterion called the minimum description length (MDL). The MDL principle frames model learning in terms of data compression. The MDL objective is to determine the model that provides the shortest description of the data set. You should consult the references above for the derivation of the MDL scoring criterion. This scoring criterion is also consistent for Bayesian networks containing multinomial and multivariate normally distributed variables.

Wallace and Korb (1999) developed a data compression scoring criterion called the minimum message length (MML), which more carefully determines the message length for encoding the parameters in the case of Bayesian networks containing multivariate normally distributed variables.

20.4.1.4. DAG Learning Is NP-Hard

In general, to find a DAG that maximizes a scoring criterion by the brute-force method of considering all DAGs is computationally unfeasible when the number of variables is not small. For example, using a recurrence established by Robinson (1977), it is possible to show there are 4.2×10^{18} DAGs containing just 10 nodes. Furthermore, Chickering (1996) proved that for certain classes of prior distributions, the problem of finding a DAG that maximizes the Bayesian score is NP-hard. One way to handle a problem such as this is to develop a heuristic search algorithm. Heuristic search algorithms are algorithms that search for a solution that is not guaranteed to be optimal; rather, they often find solutions that are reasonably close to optimal. A number of heuristic greedy search algorithms have been developed for approximately finding the DAG that maximizes the Bayesian score. Perhaps most notable of these is the greedy equivalent search (GES) algorithm, developed by Meek in 1997. In 2002, Chickering proved that if the generative distribution P admits a faithful DAG representation, then the limit, as the size of the data set approaches infinity, of the probability of GES yielding a DAG faithful to P is 1.

20.4.2. Constraint-Based Method

In the previous subsection, we assumed we had a set of variables with an unknown relative frequency distribution, and we developed a method for learning the DAG structure from data by computing the probability of the data given different DAGs. Here we take a different approach. Given the set $\mathbf{IND}_P$ of conditional independencies in a probability distribution P, we try to find a DAG for which the Markov condition entails all and only those conditional independencies. That is, we try to find a DAG faithful to P. This is called *constraint-based learning.* We illustrate the technique with two simple examples. The technique was first developed in Spirtes et al. (1993, 2000) and is discussed in detail in Neapolitan (2003).

Example 11. Suppose P is a joint probability distribution of three variables—X, Y, and Z—and

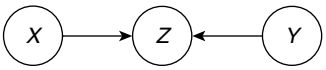
the set $\mathbf{IND}_P$ of conditional independencies in P is given by

$$\mathbf{IND}_P = \{I_P(X, Y)\}.$$

Then the DAG faithful to P is the one in Figure 20.7. There must be an edge between X and Z because otherwise, the Markov condition would entail that they are independent given some set (possibly empty) of variables. Similarly, there must be an edge between Z and Y. There can be no edge between X and Y because otherwise, the Markov condition would not entail that they are independent. The edges connecting X and Y to Z must both be directed toward Z for the following reason. If the edges had any other direction, the Markov condition would entail $I_P(X, Y|Z)$, and this conditional independency is not present.

If we are learning a causal DAG and we assume that there are no hidden common causes and that selection bias is not present, we could conclude that X causes Z and Y causes Z. Causal learning is discussed in detail in Neapolitan (2003).

Figure 20.7 This DAG Is Faithful to P When $\mathbf{IND}_P = \{I_P(X, Y)\}$



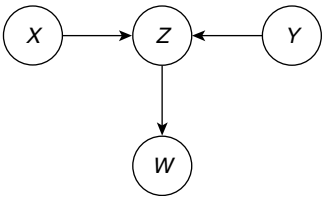
Example 12. Suppose P is a joint probability distribution of four variables—X, Y, Z, and W—and

$$\mathbf{IND}_P = \{I_P(X, Y), I_P(X, W|\{Z\}),$$
$$I_P(Y, W|\{Z\}), I_P(\{X, Y\}, W|\{Z\})\}.$$

Then the DAG faithful to P is the one in Figure 20.8. The links (edges without regard to direction) must be the ones shown for the reason discussed in the previous example. The edges connecting X and Y to Z must both be directed toward Z also for the reasons discussed in the previous example. The edge between Z and W cannot be $W \rightarrow Z$ for the following reason: If it were, the Markov condition would entail $I_P(X, W)$, and we do not have that independency.

If we are learning a causal DAG and we assume that there are no hidden common causes and that selection bias is not present, we could conclude that X causes Z and Y causes Z. Neapolitan (2003) shows that even if we do not make these assumptions, we can conclude that Z causes W. Briefly, if we replace the edge $Z \rightarrow W$ by $Z \leftarrow H \rightarrow W$, where H is a hidden common cause, the Markov condition would entail $I_P(X, W)$, and we do not have that independency.

Figure 20.8 This DAG Is Faithful to P When P Has the Conditional Independencies in Example 12



On the basis of considerations such as those illustrated in the previous examples, Spirtes et al. (1993, 2000) developed an algorithm that finds the DAG pattern faithful to P, when P admits a faithful DAG representation, from the conditional independencies in P. In 1995, Meek proved the correctness of the algorithm. Meek also developed an algorithm that determines whether P admits a faithful DAG representation.

The constraint-based method requires knowledge of the conditional independencies in a probability distribution. Given data, statistical tests can be used to estimate which conditional independencies are present. Spirtes et al. (1993, 2000) and Neapolitan (2003) each describe the statistical tests used in Tetrad II (Scheines, Spirtes, Glymour, & Meek, 1994), a system that contains implementations of the algorithms developed in Spirtes et al. (1993, 2000).

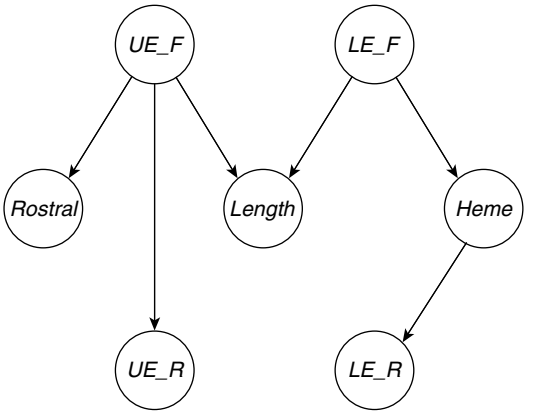
20.5. Applications

First we show an application that learns an expert system. Then we show an example of causal learning. Finally, we offer a cautionary note concerning applying the theory.

20.5.1. Learning an Expert System: Cervical Spinal Cord Trauma

Physicians face the problem of assessing cervical spinal cord trauma. To learn a Bayesian network that could assist physicians in this task, Herskovits and Dagher (1997) obtained a database from the Regional Spinal Cord Injury Center of the Delaware Valley. The database consisted of 104 cases of patients with

Figure 20.9 The Structure Learned by Cogito for Assessing Cervical Spinal Cord Trauma



spinal cord injuries who were evaluated acutely and at a 1-year follow-up. Each case consisted of the following seven variables:

Variable	*What the Variable Represents*
UE_F	Upper extremity functional score
LE_F	Lower extremity functional score
Rostral	Most superior point of cord edema as demonstrated by (MRI)
Length	Length of cord edema as demonstrated by MRI
Heme	Cord hemorrhage as demonstrated by MRI
UE_R	Upper extremity recovery at one year
LE_R	Lower extremity recovery at one year

They discretized the data and used the Bayesian network learning program Cogito™ to learn a Bayesian network containing these variables. Cogito, which was developed by Herskovits and Dagher, does model selection using the Bayesian method discussed in Section 20.4.1. The structure learned is shown in Figure 20.9.

Herskovits and Dagher (1997) compared the performance of their learned Bayesian network to that of a regression model that had independently been developed by other researchers from the same database (Flanders, Spettell, Tartaglino, Friedman, & Herbison, 1996). The other researchers did not discretize the data but rather assumed that they followed a normal distribution. The comparison consisted of evaluating 40 new cases not present in the original database. They entered the values of all variables except the outcomes variables, which are *UE_R* (upper extremity recovery at 1 year) and *LE_R* (lower extremity recovery at 1 year), and used the Bayesian network inference program Ergo™ (Beinlich & Herskovits, 1990) to predict the values of the outcome variables. They also used the regression model to predict these values. Finally, they compared the predictions of both models to the actual values for each case. They found that the Bayesian network correctly predicted the degree of upper extremity recovery three times as often as the regression model. They attributed part of this result to the fact that the original data did not follow a normal distribution, which the regression model assumed. An advantage of Bayesian networks is that they need not assume any particular distribution and therefore can accommodate unusual distributions.

20.5.2. Causal Learning: University Student Retention

Using the data collected by the *U.S. News and World Report* magazine for the purpose of college ranking, Druzdzel and Glymour (1999) analyzed the influences that affect university student retention rate. By *student retention rate,* we mean the percentage of entering freshmen who end up graduating from the university at which they initially matriculate. Low student retention rate is a major concern at many American universities as the mean retention rate over all American universities is only 55%.

The database provided by the *U.S. News and World Report* magazine contains records for 204 U.S. universities and colleges identified as major research institutions. Each record consists of more than 100 variables. The data were collected separately for the years 1992 and 1993. Druzdzel and Glymour (1999) selected the following eight variables as being most relevant to their study:

Variable	*What the Variable Represents*
grad	Fraction of entering students who graduate from the institution
rejr	Fraction of applicants who are not offered admission
tstsc	Average standardized score of incoming students
tp10	Fraction of incoming students in the top 10% of high school class
acpt	Fraction of students who accept the institution's admission offer
spnd	Average educational and general expenses per student
sfrat	Student/faculty ratio
salar	Average faculty salary

From the 204 universities, they removed any universities that had missing data for any of these variables.

Table 20.1 Records for Six Universities

Univ.	*grad*	*rejr*	*tstsc*	*tp10*	*acpt*	*spnd*	*sfrat*	*salar*
1	52.5	29.47	65.06	15	36.89	9855	12.0	60800
2	64.25	22.31	71.06	36	30.97	10527	12.8	63900
3	57.00	11.30	67.19	23	40.29	6601	17.0	51200
4	65.25	26.91	70.75	42	28.28	15287	14.4	71738
5	77.75	26.69	75.94	48	27.19	16848	9.2	63000
6	91.00	76.68	80.63	87	51.16	18211	12.8	74400

This resulted in 178 universities in the 1992 study and 173 universities in the 1993 study. Table 20.1 shows exemplary records for 6 of the universities.

Druzdzel and Glymour (1999) used Tetrad II (Scheines et al., 1994) to learn causal influences from the data. Tetrad II uses the constraint-based method to learn DAG models and allows the user to specify a "temporal" ordering of the variables. If variable Y precedes X in this order, the algorithm assumes that there can be no path from X to Y. It is called a temporal ordering because in applications to causality, if Y precedes X in time, we would assume that X could not cause Y. Druzdzel and Glymour (1999) specified the following temporal ordering for the variables in this study:

1st: *spnd*, *sfrat*, *salar*
2nd: *rejr*, *acpt*
3rd: *tstsc*, *tp10*
4th: *grad*

Their reasons for this ordering are as follows: They believed that the average spending per student (*spnd*), the student/teacher ratio (*sfrat*), and faculty salary (*salar*) are determined based on budget considerations and are not influenced by any of the other five variables. They noted that rejection rate (*rejr*) and the fraction of students who accept the institution's admission offer (*acpt*) precede the average test scores (*tstsc*) and class standing (*tp10*) in time because the values of these latter two variables are only obtained from matriculating students. Finally, they assumed that graduate rate (*grad*) does not cause any of the other variables.

Tetrad II allows the user to enter a significance level. A significance level of α means that the probability of rejecting a conditional independency hypothesis, when it is true, is α. Therefore, the smaller the value α, the less likely we are to reject a conditional independency, and therefore the sparser our resultant graph. Figure 20.10 shows the graphs that Druzdzel and Glymour (1999) learned from *U.S. News and World Report*'s 1992 database using significance levels of .2, .1, .05, and .01. In those graphs, an edge $X \rightarrow Y$ indicates that either X has a causal influence on Y or X and Y have a hidden common cause, an edge $X \leftrightarrow Y$ indicates that X and Y have a hidden common cause, and an edge $X \rightarrowtail Y$ indicates that X has a causal influence on Y.

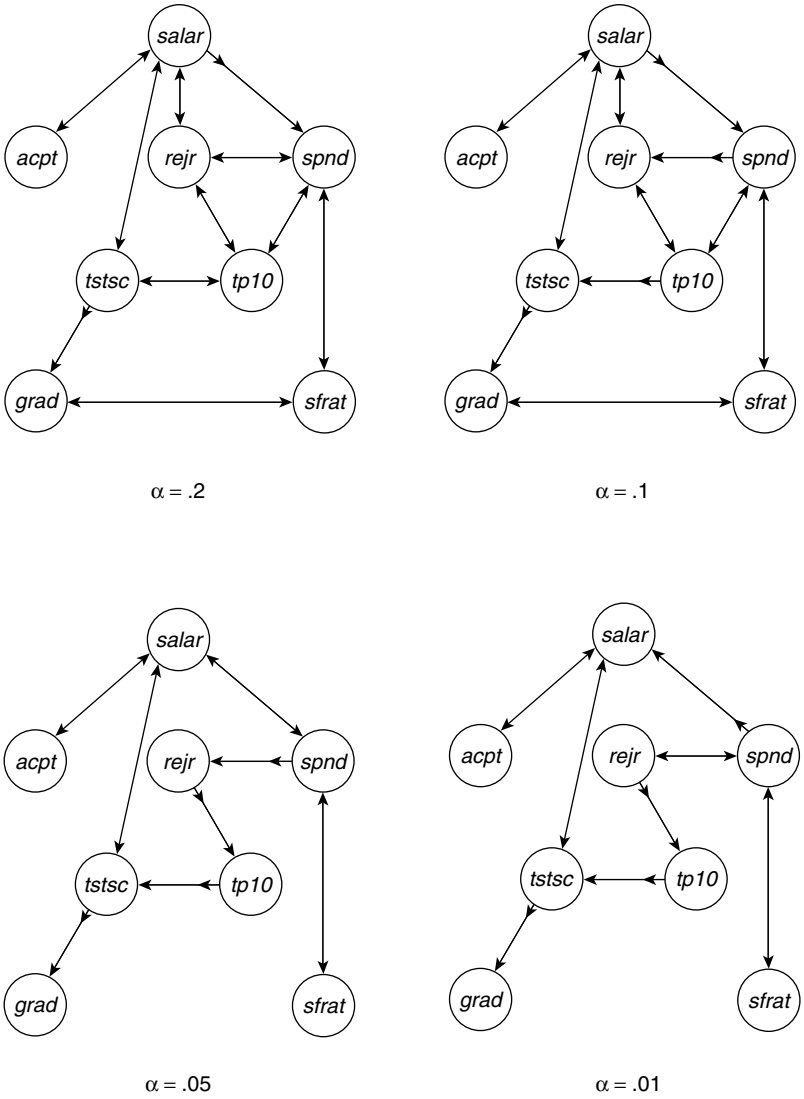
Although different graphs were obtained at different levels of significance, all the graphs in Figure 20.10 show that the average standardized test score (*tstsc*) has a direct causal influence on graduation rate (*grad*), and no other variable has a direct causal influence on *grad*. The results for the 1993 database were not as overwhelming, but they too indicated *tstsc* to be the only direct causal influence on *grad*.

To test whether the causal structure may be different for top research universities, Druzdzel and Glymour (1999) repeated the study using only the top 50 universities according to the ranking of *U.S. News and World Report*. The results were similar to those for the complete databases.

These results indicate that, although factors such as spending per student and faculty salary may have an influence on graduation rates, they do this only indirectly by affecting the standardized test scores of matriculating students. If the results correctly model reality, retention rates can be improved by bringing in students with higher test scores in any way whatsoever. Indeed, in 1994, Carnegie Mellon changed its financial aid policies to assign a portion of its scholarship fund on the basis of academic merit. Druzdzel and Glymour (1999) note that this resulted in an increase in the average test scores of matriculating freshman classes and an increase in freshman retention.

Before closing, we note that the notion that the average test score has a causal influence on graduation rate does not fit into common notions of causation such as the one concerning manipulation (see Neapolitan, 2003). For example, if we manipulated a university's average test score by accessing the testing agency's database and changing the scores of the university's

Figure 20.10 The Graphs Tetrad II Learned From *U.S. News and World Report*'s 1992 Database



students to much higher values, we would not expect the university's graduation rate to increase. Rather, this study indicates that test score is a near-perfect indicator of some other variable, which we can call *graduation potential.*

20.5.3. A Cautionary Note

Next we present another example concerning learning causes from data obtained from a survey, which illustrates problems one can encounter when using such data to infer causation.

Scarville, Button, Edwards, Lancaster, and Elig (1999) provide a database obtained from a survey in 1996 of experiences of racial harassment and discrimination of military personnel in the U.S. Armed Forces. Surveys were distributed to 73,496 members of the U.S. Army, Navy, Marine Corps, Air Force, and Coast Guard. The survey sample was selected using a nonproportional stratified random sample to ensure adequate representation of all subgroups. Usable surveys were received from 39,855 service members (54%). The survey consisted of 81 questions related to experiences of racial harassment and discrimination and job attitudes. Respondents were asked

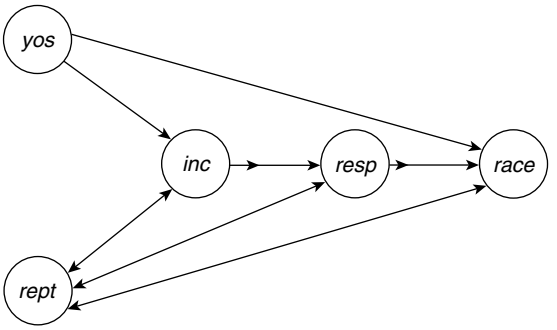
to report incidents that had occurred during the previous 12 months. The questionnaire asked participants to indicate the occurrence of 57 different types of racial/ethnic harassment or discrimination. Incidents ranged from telling offensive jokes to physical violence and included harassment by military personnel as well as the surrounding community. Harassment experienced by family members was also included.

We used Tetrad III to attempt learning causal influences from the database. For our analysis, 9,640 records (13%) were selected that had no missing data on the variables of interest. The analysis was initially based on eight variables. Similar to the situation discussed in the last subsection concerning university retention rates, we found one causal relationship to be present regardless of the significance level. That is, we found that whether the individual held the military responsible for the racial incident had a direct causal influence on the race of the individual. Because this result made no sense, we investigated which variables were involved in Tetrad III learning this causal influence. The five variables involved are the following:

Variable	*What the Variable Represents*
race	Respondent's race/ethnicity
yos	Respondent's years of military service
inc	Whether the respondent experienced a racial incident
rept	Whether the incident was reported to military personnel
resp	Whether the respondent held the military responsible for the incident

The variable *race* consisted of five categories: White, Black, Hispanic, Asian or Pacific Islander, and Native American or Alaskan Native. Respondents who reported Hispanic ethnicity were classified as Hispanic, regardless of race. Respondents were classified based on self-identification at the time of the survey. Missing data were replaced with data from administrative records. The variable *yos* was classified into four categories: 6 years or less, 7 to 11 years, 12 to 19 years, and 20 years or more. The variable *inc* was coded dichotomously to indicate whether any type of harassment was reported on the survey. The variable *rept* indicates responses to a single question concerning whether the incident was reported to military and/or civilian authorities. This variable was coded 1 if an incident had been reported to military officials. Individuals who experienced no incident, did not report the incident, or only reported the incident to civilian officials were coded 0. The variable *resp* indicates responses to a single question concerning whether the respondent believed the military to be responsible for an incident of harassment. This variable was coded 1 if the respondent indicated that the military was responsible for some or all of a reported incident. If the respondent indicated no incident, unknown responsibility, or that the military was not responsible, the variable was coded 0.

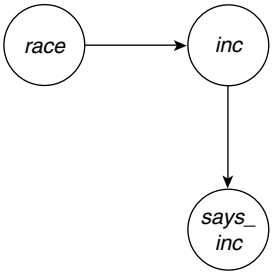
Figure 20.11 The Graph Tetrad III Learned From the Racial Harassment Survey at the .01 Significance Level



We reran the experiment using only these five variables, and again at all levels of significance, we found that *resp* had a direct causal influence on *race*. In all cases, this causal influence was learned because *rept* and *yos* were found to be probabilistically independent, and there was no edge between *race* and *inc*. That is, the causal connection between *race* and *inc* is mediated by other variables. Figure 20.11 shows the graph obtained at the .01 significance level. The edges *yos* → *inc* and *rept* → *inc* are directed toward *inc* because *yos* and *rept* were found to be independent. The edge *yos* → *inc* resulted in the edge *inc* → *resp* being directed the way it was, which in turn resulted in *resp* → *race* being directed the way it was. If there had been an edge between *inc* and *race*, the edge between *resp* and *race* would not have been directed.

It seems suspicious that no direct causal connection between *race* and *inc* was found. Recall, however, that these are the probabilistic relationships among the responses; they are not necessarily the probabilistic relationships among the actual events. There is a problem with using responses on surveys to represent occurrences in nature because subjects may not respond accurately. This is called *response bias*. Let's assume that race is recorded accurately. The actual

Figure 20.12 Possible Causal Relationships Among Race, Incidence of Harassment, and Saying That There Is an Incident of Harassment



causal relationship between *race*, *inc*, and *says_inc* may be as shown in Figure 20.12. By *inc*, we now mean whether there really was an incident, and by *says_inc*, we mean the survey response. It could be that races that experienced higher rates of harassment were less likely to report the incident, and the causal influence of *race* on *says_inc* through *inc* was negated by the direct influence of *race* on *inc*. The previous conjecture is substantiated by another study. Stangor, Swim, Van Allen, and Sechrist (2002) examined the willingness of people to attribute a negative outcome to discrimination when there was evidence that the outcome might be influenced by bias. They found that minority members were more likely to attribute the outcome to discrimination when responses were recorded privately but less likely to report discrimination when they had to express their opinion publicly and there was a member of the nonminority group present. This suggests that although minorities are more likely to perceive the situation as due to discrimination, they are less likely to report it publicly. Although the survey of military personnel was intended to be confidential, minority members in the military may have felt uncomfortable reporting incidents of discrimination.

As noted in the previous subsection, Tetrad II (and III) allows the user to enter a temporal ordering. So we could have put *race* first in such an ordering to avoid it being an effect of another variable. However, one should do this with caution. The fact that the data strongly support that race is an effect indicates there is something wrong with the data, which means we should be dubious of drawing any conclusions from the data. In the present example, Tetrad III actually informed us that we could not draw causal conclusions from the data when we make *race* a root. That is, when we made *race* a root, Tetrad III concluded that there is no consistent orientation of the edge between *race* and *resp*.

REFERENCES

Ash, R. B. (1970). *Basic probability theory.* New York: John Wiley.

Beinlich, I. A., & Herskovits, E. H. (1990). A graphical environment for constructing Bayesian belief networks. In M. Henrion, R. D. Shachter, L. N. Kanal, & J. F. Lemmer (Eds.), *Uncertainty in artificial intelligence* (Vol. 5). Amsterdam: North Holland.

Brownlee, K. A. (1965). *Statistical theory and methodology.* New York: John Wiley.

Castillo, E., Gutiérrez, J. M., & Hadi, A. S. (1997). *Expert systems and probabilistic network models.* New York: Springer-Verlag.

Chickering, D. (1996). Learning Bayesian networks is NP-complete. In D. Fisher & H. Lenz (Eds.), *Learning from data* (pp. 121–130). New York: Springer-Verlag.

Chickering, D. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research, 3*, 507–554.

Clemen, R. T. (2000). *Making hard decisions.* Boston: PWS-KENT.

Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence, 33*, 393–405.

Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*, 309–347.

Druzdzel, M. J., & Glymour, C. (1999). Causal inferences from databases: Why universities lose students. In C. Glymour & G. F. Cooper (Eds.), *Computation, causation, and discovery* (pp. 521–539). Menlo Park, CA: AAAI Press.

Flanders, A. E., Spettell, C. M., Tartaglino, L. M., Friedman, D. P., & Herbison, G. J. (1996). Forecasting motor recovery after cervical spinal cord injury: Value of MRI. *Radiology, 201*, 649–665.

Friedman, N., & Goldszmidt, M. (1996). Building classifiers and Bayesian networks. In *Proceedings of the National Conference on Artificial Intelligence* (Vol. 2, pp. 1227–1284). Menlo Park, CA: AAAI Press.

Geiger, D., & Heckerman, D. (1994). Learning Gaussian networks. In R. L. de Mantras & D. Poole (Eds.), *Uncertainty in artificial intelligence: Proceedings of the Tenth Conference* (pp. 235–243). San Mateo, CA: Morgan Kaufmann.

Good, I. J. (1983). *Good thinking.* Minneapolis: University of Minnesota Press.

Herskovits, E. H., & Dagher, A. P. (1997). Applications of Bayesian networks to health care (Tech. Rep. No. NSI-TR-1997–02). Baltimore: Noetic Systems Incorporated.

Iversen, G. R., Longcor, W. H., Mosteller, F., Gilbert, J. P., & Youtz, C. (1971). Bias and runs in dice throwing and recording: A few million throws. *Psychometrika, 36*, 1–19.

Kerrich, J. E. (1946). *An experimental introduction to the theory of probability.* Copenhagen: Einer Munksgaard.

Keynes, J. M. (1948). *A treatise on probability.* London: Macmillan. (Original work published 1921)

Lam, W., & Bacchus, F. (1994). Learning Bayesian belief networks: An approach based on the MDL principle. *Computational Intelligence, 10*, 269–294.

Lindley, D. V. (1985). *Introduction to probability and statistics from a Bayesian viewpoint.* Cambridge, UK: Cambridge University Press.

Meek, C. (1995). Causal influence and causal explanation with background knowledge. In P. Besnard & S. Hanks (Eds.), *Uncertainty in artificial intelligence: Proceedings of the Eleventh Conference* (pp. 403–410). San Mateo, CA: Morgan Kaufmann.

Meek, C. (1997). *Graphical models: Selecting causal and statistical models.* Unpublished doctoral dissertation, Carnegie Mellon University.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65–115). Mahwah, NJ: Lawrence Erlbaum.

Neapolitan, R. E. (1990). *Probabilistic reasoning in expert systems.* New York: John Wiley.

Neapolitan, R. E. (1992). A limiting frequency approach to probability based on the weak law of large numbers. *Philosophy of Science, 59*(3).

Neapolitan, R. E. (1996, May). Is higher-order uncertainty needed? *IEEE Transactions on Systems, Man, and Cybernetics: Special Issue on Higher Order Uncertainty,* pp. 294–302.

Neapolitan, R. E. (2003). *Learning Bayesian networks.* Upper Saddle River, NJ: Prentice Hall.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* San Mateo, CA: Morgan Kaufmann.

Rissanen, J. (1987). Stochastic complexity (with discussion). *Journal of the Royal Statistical Society, Series B, 49*, 223–239.

Robinson, R. W. (1977). Counting unlabeled acyclic digraphs. In C. H. C. Little (Ed.), *Lecture notes in mathematics, 622: Combinatorial mathematics V* (pp. 28–43). New York: Springer-Verlag.

Scarville, J., Button, S. B., Edwards, J. E., Lancaster, A. R., & Elig, T. W. (1999). *Armed Forces 1996 Equal Opportunity Survey* (DMDC Rep. No. 97–027). Arlington, VA: Defense Manpower Data Center.

Scheines, R., Spirtes, P., Glymour, C., & Meek, C. (1994). *Tetrad II: User manual.* Hillsdale, NJ: Lawrence Erlbaum.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search.* New York: Springer-Verlag.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). Cambridge: MIT Press.

Stangor, C., Swim, J. K., Van Allen, K. L., & Sechrist, G. B. (2002). Reporting discrimination in public and private contexts. *Journal of Personality and Social Psychology, 82*, 69–74.

van Lambalgen, M. (1987). *Random sequences.* Unpublished doctoral dissertation, University of Amsterdam.

von Mises, R. (1919). Grundlagen der Wahrscheinlichkeitsrechnung. (The bases of probability calculations). *Mathematische Zeitschrift, 5*, 52–99.

von Mises, R. (1957). *Probability, statistics, and truth.* London: Allen & Unwin. (Original work published 1928)

Wallace, C. S., & Korb, K. (1999). Learning linear causal models by MML sampling. In A. Gammerman (Ed.), *Causal models and intelligent data management* (pp. 89–111). New York: Springer-Verlag.

Zabell, S. L. (1982). W. E. Johnson's "Sufficientness" postulate. *The Annals of Statistics, 10*(4), 1091–1099.

Chapter 21

THE NULL RITUAL

What You Always Wanted to Know About Significance Testing but Were Afraid to Ask

GERD GIGERENZER

STEFAN KRAUSS

OLIVER VITOUCH

> No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas.
>
> —Ronald A. Fisher (1956, p. 42)

> It is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail.
>
> —A. H. Maslow (1966, pp. 15-16)

One of us once had a student who ran an experiment for his thesis. Let us call him Pogo. Pogo had an experimental group and a control group and found that the means of both groups were exactly the same. He believed it would be unscientific to simply state this result; he was anxious to do a significance test. The result of the test was that the two means did not differ significantly, which Pogo reported in his thesis.

In 1962, Jacob Cohen reported that the experiments published in a major psychology journal had, on average, only a 50-50 chance of detecting a medium-sized effect if there was one. That is, the statistical power was as low as 50%. This result was widely cited, but did it change researchers' practice? Sedlmeier and Gigerenzer (1989) checked the studies in the same journal, 24 years later, a time period that should allow for change. Yet only 2 out of 64 researchers mentioned power, and it was never estimated. Unnoticed, the average power had decreased (researchers now used alpha adjustment, which shrinks power). Thus, if there had been an effect of a medium size, the researchers would have had a better chance of finding it by throwing a coin rather than conducting their experiments. When we checked the years 2000 to 2002, with some 220 empirical articles, we finally found 9 researchers who computed the power of their tests. Forty years after Cohen, there is a first sign of change.

Editors of major journals such as A. W. Melton (1962) made null hypothesis testing a necessary

AUTHORS' NOTE: We are grateful to David Kaplan and Stanley Mulaik for helpful comments and to Katharina Petrasch for her support with journal analyses.

condition for the acceptance of papers and made small p-values the hallmark of excellent experimentation. The Skinnerians found themselves forced to start a new journal, the *Journal of the Experimental Analysis of Behavior,* to publish their kind of experiments (Skinner, 1984, p. 138). Similarly, one reason for launching the *Journal of Mathematical Psychology* was to escape the editors' pressure to routinely perform null hypothesis testing. One of its founders, R. D. Luce (1988), called this practice a "wrongheaded view about what constituted scientific progress" and "mindless hypothesis testing in lieu of doing good research: measuring effects, constructing substantive theories of some depth, and developing probability models and statistical procedures suited to these theories" (p. 582).

The student, the researchers, and the editors had engaged in a statistical ritual rather than statistical thinking. Pogo believed that one always ought to perform a null hypothesis test, without exception. The researchers did not notice how small their statistical power was, nor did they seem to care: Power is not part of the null ritual that dominates experimental psychology. The essence of the ritual is the following:

1. Set up a statistical null hypothesis of "no mean difference" or "zero correlation." Don't specify the predictions of your research hypothesis or of any alternative substantive hypotheses.
2. Use 5% as a convention for rejecting the null. If significant, accept your research hypothesis.
3. Always perform this procedure.

The null ritual has sophisticated aspects we will not cover here, such as alpha adjustment and ANOVA procedures, but these do not change its essence. Typically, it is presented without naming its originators, as statistics per se. Some suggest that it was authorized by the eminent statistician Sir Ronald A. Fisher, owing to the emphasis on null hypothesis testing (not to be confused with the null ritual) in his 1935 book. However, Fisher would have rejected all three ingredients of this procedure. First, *null* does not refer to a zero mean difference or correlation but to the hypothesis to be "nullified," which could postulate a correlation of .3, for instance. Second, as the epigram illustrates, by 1956, Fisher thought that using a routine 5% level of significance indicated lack of statistical thinking. Third, for Fisher, null hypothesis testing was the most primitive type in a hierarchy of statistical analyses and should be used only for problems about which we have very little knowledge or none at all (Gigerenzer et al., 1989, chap. 3). Statistics offers a toolbox of methods, not just a single hammer. In many (if not most) cases, descriptive statistics and exploratory data analysis are all one needs. As we will see soon, the null ritual originated neither from Fisher nor from any other renowned statistician and does not exist in statistics proper. It was instead fabricated in the minds of statistical textbook writers in psychology and education.

Rituals seem to be indispensable for the self-definition of social groups and for transitions in life, and there is nothing wrong about them. However, they should be the subject rather than the procedure of social sciences. Elements of social rituals include (a) the repetition of the same action, (b) a focus on special numbers or colors, (c) fears about serious sanctions for rule violations, and (d) wishful thinking and delusions that virtually eliminate critical thinking (Dulaney & Fiske, 1994). The null ritual has each of these four characteristics: a repetitive sequence, a fixation on the 5% level, fear of sanctions by editors or advisers, and wishful thinking about the outcome (the p-value) combined with a lack of courage to ask questions.

Pogo's counterpart in this chapter is a curious student who wants to understand the ritual rather than mindlessly perform it. She has the courage to raise questions that seem naive at first glance and that others do not care or dare to ask.

21.1. Question 1: What Does a Significant Result Mean?

What a simple question! Who would not know the answer? After all, psychology students spend months sitting through statistics courses, learning about null hypothesis tests (significance tests) and their featured product, the p-value. Just to be sure, consider the following problem (Haller & Krauss, 2002; Oakes, 1986):

> Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say, 20 subjects in each sample). Furthermore, suppose you use a simple independent means t-test and your result is significant ($t = 2.7$, $df = 18$, $p = .01$). Please mark each of the statements below as "true" or "false." *False* means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.

1. You have absolutely disproved the null hypothesis (i.e., there is no difference between the population means). ☐ True ☐ False

2. You have found the probability of the null hypothesis being true. ☐ True False ☐
3. You have absolutely proved your experimental hypothesis (that there is a difference between the population means). ☐ True False ☐
4. You can deduce the probability of the experimental hypothesis being true. ☐ True False ☐
5. You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision. ☐ True False ☐
6. You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions. ☐ True False ☐

Which statements are true? If you want to avoid the I-knew-it-all-along feeling, please answer the six questions yourself before continuing to read. When you are done, consider what a p-value actually is: A p-value is the probability of the observed data (or of more extreme data points), given that the null hypothesis H_0 is true, defined in symbols as $p(D|H_0)$.This definition can be rephrased in a more technical form by introducing the statistical model underlying the analysis (Gigerenzer et al., 1989, chap. 3). Let us now see which of the six answers are correct:

Statements 1 and 3: Statement 1 is easily detected as being false. A significance test can never disprove the null hypothesis. Significance tests provide probabilities, not definite proofs. For the same reason, Statement 3, which implies that a significant result could prove the experimental hypothesis, is false. Statements 1 and 3 are instances of the illusion of certainty (Gigerenzer, 2002).

Statements 2 and 4: Recall that a p-value is a probability of data, not of a hypothesis. Despite wishful thinking, $p(D|H_0)$ is not the same as $p(H_0|D)$, and a significance test does not and cannot provide a probability for a hypothesis. One cannot conclude from a p-value that a hypothesis has a probability of 1 (Statements 1 and 3) or that it has any other probability (Statements 2 and 4). Therefore, Statements 2 and 4 are false. The statistical toolbox, of course, contains tools that allow estimating probabilities of hypotheses, such as Bayesian statistics (see below). However, null hypothesis testing does not.

Statement 5: The "probability that you are making the wrong decision" is again a probability of a hypothesis. This is because if one rejects the null hypothesis, the only possibility of making a wrong decision is if the null hypothesis is true. In other words, a closer look at Statement 5 reveals that it is about the probability that you will make the wrong decision, that is, that H_0 is true. Thus, it makes essentially the same claim as Statement 2 does, and both are incorrect.

Statement 6: Statement 6 amounts to the replication fallacy. Recall that a p-value is the probability of the observed data (or of more extreme data points), given that the null hypothesis is true. Statement 6, however, is about the probability of "significant" data per se, not about the probability of data if the null hypothesis were true. The error in Statement 6 is that $p = 1\%$ is taken to imply that such significant data would reappear in 99% of the repetitions. Statement 6 could be made only if one knew that the null hypothesis was true. In formal terms, $p(D|H_0)$ is confused with $1 - p(D)$. The replication fallacy is shared by many, including the editors of top journals. For instance, the former editor of the *Journal of Experimental Psychology,* A. W. Melton (1962), wrote in his editorial, "The level of significance measures the confidence that the results of the experiment would be repeatable under the conditions described" (p. 553). A nice fantasy, but false.

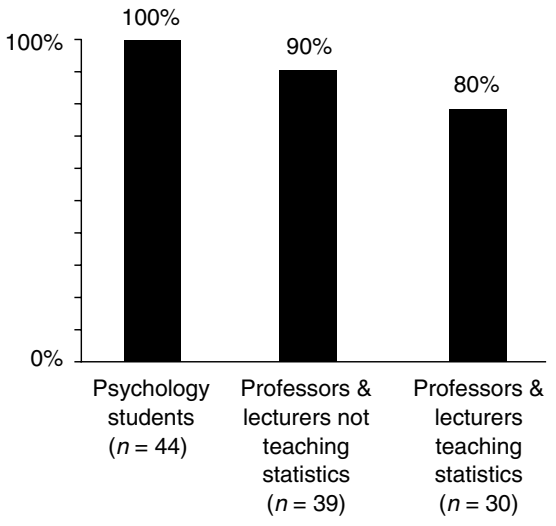
To sum up, all six statements are incorrect. Note that all six err in the same direction of wishful thinking: They overestimate what one can conclude from a p-value.

21.1.1. Students' and Teachers' Delusions

We posed the question with the six multiple-choice answers to 44 students of psychology, 39 lecturers and professors of psychology, and 30 statistics teachers, who included professors of psychology, lecturers, and teaching assistants. All students had successfully passed one or more statistics courses in which significance testing was taught. Furthermore, each of the teachers confirmed that he or she taught null hypothesis testing. To get a quasi-representative sample, we drew the participants from six German universities (Haller & Krauss, 2002).

How many students and teachers noticed that all of the statements were wrong? As Figure 21.1 shows, none of the students did. Every student endorsed one or more of the illusions about the meaning of a p-value. One might think that these students lack the right genes for statistical thinking and are stubbornly resistant to education. A glance at the performance of their teachers, however, indicates that wishful thinking might not be entirely their fault. Ninety percent of the

Figure 21.1 The Amount of Delusions About the Meaning of "$p = .01$"



NOTE: The percentages refer to the participants in each group who endorsed one or more of the six false statements (based on Haller & Krauss, 2002).

professors and lecturers also had illusions, a proportion almost as high as among their students. Most surprisingly, 80% of the statistics teachers shared illusions with their students. Thus, the students' errors might be a direct consequence of their teachers' wishful thinking. Note that one does not need to be a brilliant mathematician to answer the question, "What does a significant result mean?" One only needs to understand that a p-value is the probability of the data (or more extreme data), given that the H_0 is true.

If students "inherited" the illusions from their teachers, where did the teachers acquire them? The illusions were right there in the first textbooks introducing psychologists to null hypothesis testing more than 60 years ago. Guilford's *Fundamental Statistics in Psychology and Education,* first published in 1942, was probably the most widely read textbook in the 1940s and 1950s. Guilford suggested that hypothesis testing would reveal the probability that the null hypothesis is true. "If the result comes out one way, the hypothesis is probably correct, if it comes out another way, the hypothesis is probably wrong" (p. 156). Guilford's logic was not consistently misleading but wavered back and forth between correct and incorrect statements, as well as ambiguous ones that can be read like Rorschach inkblots. He used phrases such as "we obtained directly the probabilities that the null hypothesis was plausible" and "the probability of extreme deviations from chance" interchangeably for referring to the same thing: the level of significance. Guilford is no exception. He marked the beginning of a genre of statistical texts that vacillate between the researchers' desire for probabilities of hypotheses and what significance testing can actually provide. Early authors promoting the illusion that the level of significance would specify the probability of hypothesis include Anastasi (1958, p. 11), Ferguson (1959, p. 133), and Lindquist (1940, p. 14). But the belief has persisted over decades: for instance, in Miller and Buckhout (1973; statistical appendix by Brown, p. 523), Nunally (1975, pp. 194–196), and in the examples collected by Bakan (1966), Pollard and Richardson (1987), Gigerenzer (1993), Nickerson (2000), and Mulaik, Raju, and Harshman (1997).

Which of the illusions were most often endorsed, and which relatively seldom? Table 21.1 shows that Statements 1 and 3 were most frequently detected as being false. These claim certainty rather than probability. Still, up to a third of the students and an embarrassing 10% to 15% of the group of teachers held this illusion of certainty. Statements 4, 5, and 6 lead the hit list of the most widespread illusions. These errors are about equally prominent in all groups, a collective fantasy that seems to travel by cultural transmission from teacher to student. The last column shows that these three illusions were also prevalent among British academic psychologists who answered the same question (Oakes, 1986). Just as in the case of statistical power cited in the introduction, in which little learning was observed after 24 years, knowledge about what a significant result means does not seem to have improved since Oakes. Yet a persistent blind spot for power and a lack of comprehension of significance are consistent with the null ritual.

Statements 2 and 4, which put forward the same type of error, were given different endorsements. When a statement concerns the probability of the experimental hypothesis, it is much more accepted by students and teachers as a valid conclusion than one that concerns the probability of the null hypothesis. The same pattern can be seen for British psychologists (see Table 21.1). Why are researchers and students more likely to believe that the level of significance determines the probability of H_1 rather than that of H_0? A possible reason is that the researchers' focus is on the experimental hypothesis H_1 and that the desire to find the probability of H_1 drives the phenomenon.

Did the students produce more illusions than their teachers? Surprisingly, the difference was only slight. On average, students endorsed 2.5 illusions, their professors and lecturers who did not teach statistics

Table 21.1 Percentages of False Answers (i.e., Statements Marked as True) in the Three Groups of Figure 21.1

	Germany 2000			*United Kingdom 1986*
Statement (Abbreviated)	*Psychology Students*	*Professors and Lecturers: Not Teaching Statistics*	*Professors and Lecturers: Teaching Statistics*	*Professors and Lecturers*
1. H_0 is absolutely disproved	34	15	10	1
2. Probability of H_0 is found	32	26	17	36
3. H_1 is absolutely proved	20	13	10	6
4. Probability of H_1 is found	59	33	33	66
5. Probability of wrong decision	68	67	73	86
6. Probability of replication	41	49	37	60

NOTE: For comparison, the results of Oakes's (1986) study with academic psychologists in the United Kingdom are shown in the right column.

approved of 2.0 illusions, and those who taught significance testing endorsed 1.9 illusions.

Could it be that these collective illusions are specific to German psychologists and students? No, the evidence points to a global phenomenon. As mentioned above, Oakes (1986) reported that 97% of British academic psychologists produced at least one illusion. Using a similar test question, Falk and Greenbaum (1995) found comparable results for Israeli students, despite having taken measures for debiasing students. Falk and Greenbaum had explicitly added the right alternative ("None of the statements is correct"), whereas we had merely pointed out that more than one or none of the statements might be correct. As a further measure, they had made their students read Bakan's (1966) classic article, which explicitly warns against wrong conclusions. Nevertheless, only 13% of their participants opted for the right alternative. Falk and Greenbaum concluded that "unless strong measures in teaching statistics are taken, the chances of overcoming this misconception appear low at present" (p. 93). Warning and reading by itself does not seem to foster much insight. So what to do?

21.2. Question 2: How Can Students Get Rid of Illusions?

The collective illusions about the meaning of a significant result are embarrassing to our profession. This state of affairs is particularly painful because psychologists—unlike natural scientists—heavily use significance testing yet do not understand what its product, the p-value, means. Is there a cure?

Yes. The cure is to open the statistical toolbox. In statistical textbooks written by psychologists and educational researchers, significance testing is typically presented as if it were an all-purpose tool. In statistics proper, however, an entire toolbox exists, of which null hypothesis testing is only one tool among many. As a therapy, even a small glance into the contents of the toolbox can be sufficient. One quick way to overcome some of the illusions is to introduce students to Bayes's rule.

Bayes's rule deals with the probability of hypotheses, and by introducing it alongside null hypothesis testing, one can easily see what the strengths and limits of each tool are. Unfortunately, Bayes's rule is rarely mentioned in statistical textbooks for psychologists. Hays (1963) had a chapter on Bayesian statistics in the second edition of his widely read textbook but dropped it in the subsequent editions. As he explained to one of us (GG), he dropped the chapter upon pressure from his publisher to produce a statistical cookbook that did not hint at the existence of alternative tools for statistical inference. Furthermore, he believed that many researchers are not interested in statistical thinking in the first place but solely in getting their papers published (Gigerenzer, 2000).

Here is a short comparative look at two tools:

1. Null hypothesis testing computes the probability $p(D|H_0)$. The form of conditional probabilities makes it clear that with null hypothesis testing, (a) only statements concerning the probability of data D can be obtained, and (b) the null hypothesis H_0 functions as the reference point for the conditional statement. In other words, any correct answer to the question of what a significant result means must include the conditional

phrase "... *given H_0 is true*" or an equivalent expression.

2. Bayes's rule computes the probability $p(H_1|D)$. In the simple case of two hypotheses, H_1 and H_2, which are mutually exclusive and exhaustive, Bayes's rule is the following:

$$p(H_1|D) = \frac{p(H_1)p(D|H_1)}{p(H_1)p(D|H_1) + p(H_2)p(D|H_2)}.$$

For instance, consider HIV screening for people who are in no known risk group (Gigerenzer, 2002). In this population, the a priori probability $p(H_1)$ of being infected by HIV is about 1 in 10,000, or .0001. The probability $p(D|H_1)$ that the test is positive (D) if the person is infected is .999, and the probability $p(D|H_2)$ that the test is positive if the person is not infected is .0001. What is the probability $p(H_1|D)$ that a person with a positive HIV test actually has the virus? Inserting these values into Bayes's rule results in $p(H_1|D) = .5$. Unlike null hypothesis testing, Bayes's rule can actually provide a probability of a hypothesis.

Now let us approach the same problem with null hypothesis testing. The null is that the person is not infected. The observation is a positive test, and the probability of a positive test given that the null is true is $p = .0001$, which is the exact level of significance. Therefore, the null hypothesis of no infection is rejected with high confidence, and the alternative hypothesis that the person is infected is accepted. However, as the Bayesian calculation showed, given a positive test, the probability of an HIV infection is only .5. HIV screening illustrates how one can reach quite different conclusions with null hypothesis testing or Bayes's rule. It also clarifies some of the possibilities and limits of both tools. The single most important limit of null hypothesis testing is that there is only one statistical hypothesis—the null, which does not allow for comparative hypotheses testing. Bayes's rule, in contrast, compares the probabilities of the data under two (or more) hypotheses and also uses prior probability information. Only when one knows extremely little about a topic (so that one cannot even specify the predictions of competing hypotheses) might a null hypothesis test be appropriate.

A student who has understood the fact that the products of null hypothesis testing and Bayes's rule are $p(D|H_0)$ and $p(H_1|D)$, respectively, will note that the Statements 1 through 5 are all about probabilities of hypotheses and therefore cannot be answered with significance testing. Statement 6, in contrast, is about the probability of further significant results, that is, about probabilities of data rather than hypotheses. That this statement is wrong can be seen from the fact that it does not include the conditional phrase " . . . if H_0 is true."

Note that the above two-step course does not require in-depth instruction in Bayesian statistics (see Edwards, Lindman, & Savage, 1963; Howson & Urbach, 1989). This minimal course can be readily extended to a few more tools, for instance, by adding Neyman-Pearson testing, which computes the likelihood ratio $p(D|H_1)/p(D|H_2)$. Psychologists know Neyman-Person testing in the form of signal detection theory, a cognitive theory that has been inspired by the statistical tool (Gigerenzer & Murray, 1987). The products of the three tools can be easily compared:

(a) $p(D|H_0)$ is obtained from null hypothesis testing.
(b) $p(D|H_1)/p(D|H_2)$ is obtained from Neyman-Pearson hypotheses testing.
(c) $p(H_1|D)$ is obtained by Bayes's rule.

For null hypothesis testing, only the likelihood $p(D|H_0)$ matters; for Neyman-Pearson, the likelihood ratio matters; and for Bayes, the posterior probability matters. By opening the statistical toolbox and comparing tools, one can easily understand what each tool delivers and what it does not. For the next question, the fundamental difference between null hypothesis testing and other statistical tools such as Bayes's rule and Neyman-Pearson testing is that in null hypothesis testing, only one hypothesis—-the null—-is precisely stated. With this technique, one is not able to compare two or more hypotheses in a symmetric or "fair" way and might draw wrong conclusions from the data.

21.3. Question 3: Can the Null Ritual Hurt?

But it's just a little ritual. It may be a bit silly, but it can't hurt, can it? Yes, it can. Consider a study in which the authors had two precisely formulated hypotheses, but instead of specifying the predictions of both hypotheses for their experimental design, they performed the null ritual. The question was how young children judge the area of rectangles, and the two hypotheses were the following: Children add height plus width, or children multiply height times width (Anderson & Cuneo, 1978). In one experiment, 5- to 6-year-old children rated the joint area of two rectangles (not an easy task). The reason for having them rate the area of two rectangles rather than one was to disentangle the integration rule (adding vs. multiplying)

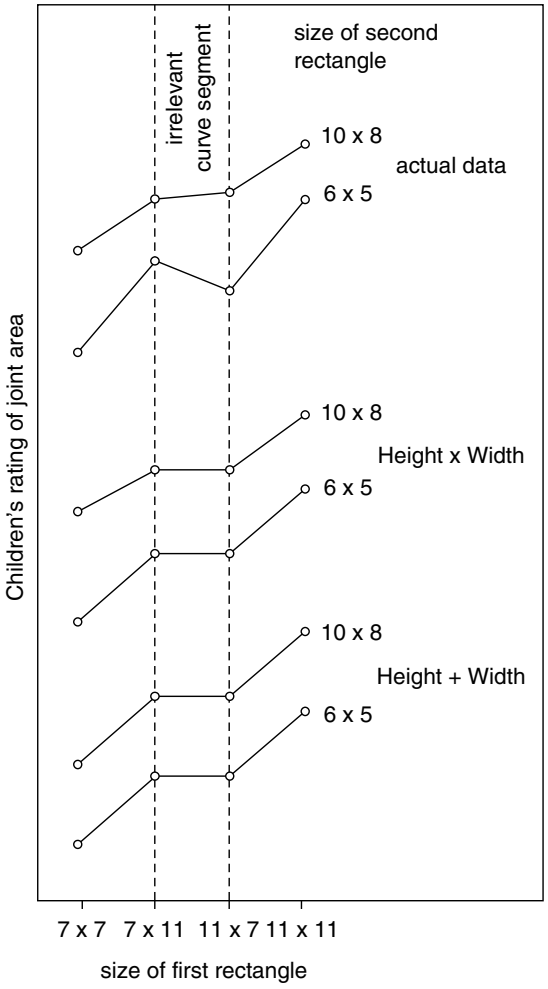
from the response function (linear vs. logarithmic). Suffice to say that the idea for the experiment was ingenious. The Height + Width rule was identified with the null hypothesis of no linear interaction in a two-factorial analysis of variance. The prediction of the second hypothesis, the Height × Width rule, was never specified, as it never is with null hypothesis testing. The authors found that the "curves are nearly parallel and the interaction did not approach significance, $F(4, 56) = 1.20$" (p. 352). They concluded that this and similar results would support the Height + Width rule and disconfirm the multiplying rule. In Anderson's (1981) words, "Five-year-olds judge area of rectangles by an adding, Height + Width rule" (p. 33).

Testing a null, however, is a weak argument if one has some ideas about the subject matter, as Anderson and Cuneo (1978) did. So let us derive the actual predictions of both of their hypotheses for their experimental design (for details, see Gigerenzer & Murray, 1987). Figure 21.2 shows the prediction of the Height + Width rule and that of the Height × Width rule. There were eight pairs of rectangles, shown by the two curves. Note that the middle segment (the parallel lines) does not differentiate between the two hypotheses, as the left and the right segments do. Thus, only these two segments are relevant. Here, the Height + Width rule predicts parallel curves, whereas the Height × Width rule predicts converging curves (from left to right). One can see that the data (top panel) actually show the pattern predicted by the multiplying rule and that the curves converge even more than predicted. If either of the two hypotheses is supported by the data, then it is the multiplying rule (this was supported by subsequent experimental research in which the predictions of half a dozen hypotheses were tested; see Gigerenzer & Richter, 1990). Nevertheless, the null ritual misled the researchers into concluding that the data would support the Height + Width rule.

Why was the considerable deviation from the prediction of the Height + Width rule not statistically significant? One reason was the large amount of error in the data: Asking young children to rate the joint area of two rectangles produced highly unreliable responses. This contributed to the low power of the statistical tests, which was consistently below 10% (Gigerenzer & Richter, 1990)! That is, the experiments were set up so that the chance of accepting the Height × Width rule if it is true was less than 1 in 10.

But doesn't the alternative hypothesis always predict a significant result? As Figure 21.2 illustrates, this is not the case. Even if the data had coincided exactly with the prediction of the multiplying rule, the result would not have been significant (because the even larger deviation of the actual data was not significant either). In general, a hypothesis predicts a value or a curve but not significance or nonsignificance. The latter is the joint product of several factors that have little to do with the

Figure 21.2 How to Draw the Wrong Conclusions by Using Null Hypothesis Testing



NOTE: Anderson and Cuneo (1978) asked which of two hypotheses, Height + Width or Height × Width, describes young children's judgments of the joint area of rectangle pairs. Following null hypothesis testing, they identified the Height + Width rule with nonsignificance of the linear interaction in an analysis of variance and the Height × Width rule with a significant interaction. The result was not significant; the Height × Width rule was rejected and the Height + Width rule accepted. When one instead specifies the predictions of both hypotheses (Gigerenzer & Murray, 1987), the Height + Width rule predicts the parallel curves, and the Height × Width rule predicts the converging curves. One can see that the data are actually closer to the pattern predicted by the Height × Width rule (see text).

hypothesis, including the number of participants, the error in the data, and the statistical power.

This example is not meant as a critique of specific authors but as an illustration of how routine null hypothesis testing can hurt. It teaches two aspects of statistical thinking that are alien to the null ritual. First, it is important to specify the predictions of more than one hypothesis. In the present case, descriptive statistics and mere eyeballing would have been better than the null ritual and analysis of variance. Second, good statistical thinking is concerned with minimizing the real error in the data, and this is more important than a small p-value. In the present case, a small error can be achieved by asking children for paired comparisons—which of two rectangles (chocolate bars) is larger? Unlike ratings, comparative judgments generate highly reliable responses, clear individual differences, and allow researchers to test hypotheses that cannot be easily expressed in the "main-effect plus interaction" language of analysis of variance (Gigerenzer & Richter, 1990).

21.4. Question 4: Is the Level of Significance the Same Thing as Alpha?

Let us introduce Dr. Publish-Perish. He is the average researcher, a devoted consumer of statistical methods. His superego tells him that he ought to set the level of significance before an experiment is performed. A level of 1% would be impressive, wouldn't it? Yes, but . . . there is a dilemma. He fears that the p-value calculated from the data could turn out slightly higher, such as 1.1%, and he would then have to report a nonsignificant result. He does not want to take that risk. Then there is the option of setting the level at a less impressive 5%. But what if the p-value turned out to be smaller than 1% or even .1%? Then he would regret his decision deeply because he would have to report this result as $p < .05$. He does not like that either. So he thinks the only choice left is to cheat a little and disobey his superego. He waits until he has seen the data, rounds the p-value up to the next conventional level, and reports that the result is significant at $p < .001, .01$, or .05, whatever is next. That smells of deception, and his superego leaves him with feelings of guilt. But what should he do when everyone else seems to play this little cheating game?

Dr. Publish-Perish does not know that his moral dilemma is caused by a mere confusion, a product of textbook writers who failed to distinguish the three main interpretations of the level of significance and mixed them all up.

21.4.1. Interpretation 1: Mere Convention

So far, we have mentioned only in passing the statisticians who have created and shaped the ideas we are talking about. Similarly, most statistical textbooks for psychology and education are generally mute about these eminent people and their ideas, which is remarkable for a field where authors are cited compulsively, and no shortage of competing theories exists.

The first person to introduce is Sir Ronald A. Fisher (1890–1962), one of the most influential statisticians ever, who also made first-rate contributions to genetics and was knighted for his achievements. Fisher spent most of his career at University College, London, where he held the chair of eugenics. His publications include three books on statistics. For psychology, the most influential of these was the second one, *The Design of Experiments,* first published in 1935. In the *Design,* Fisher suggested that we think of the level of significance as a *convention:* "It is usual and convenient for experimenters to take 5 per cent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard" (p. 13). Fisher's assertion that 5% (in some cases, 1%) is a convention to be adopted by all experimenters and in all experiments, whereas nonsignificant results are to be ignored, became part of the null ritual. For instance, the 1974 *Publication Manual* of the American Psychological Association instructed experimenters to make mechanical decisions using a conventional level of significance:

> Caution: Do not infer trends from data that fail by a small margin to meet the usual levels of significance. Such results are best interpreted as caused by chance and are best reported as such. Treat the result section like an income tax return. Take what's coming to you, but no more. (p. 19; this passage was deleted in the 3rd edition [American Psychological Association, 1983])

In a recent defense of what he calls NHSTP (null hypothesis significance testing procedure), Chow (1998) still proclaims that null hypothesis tests should be interpreted mechanically, using the conventional 5% level of significance. This view reminds us of a maxim regarding the critical ratio, the predecessor of the significance level: "A critical ratio of three, or no Ph.D."

21.4.2. Interpretation 2: Alpha

The second eminent person we would like to introduce is the Polish mathematician Jerzy Neyman, who worked with Egon S. Pearson (the son of Karl Pearson) at University College in London and later, when the tensions between Fisher and himself grew too heated, moved to Berkeley, California. Neyman and Pearson criticized Fisher's null hypothesis testing for several reasons, including that no alternative hypothesis is specified, which in turn does not allow computation of the probability β of wrongly rejecting the alternative hypothesis (Type II error) or of the power of the test $(1 - \beta)$ (Gigerenzer et al., 1989, chap. 3). In Neyman-Pearson theory, the meaning of a level of significance such as 3% is the following: If the hypothesis H_1 is correct, and the experiment is repeated many times, the experimenter will wrongly reject H_1 in 3% of the cases. Rejecting the hypothesis H_1 if it is correct is called a Type I error, and the probability of rejecting H_1 if it is correct is called alpha (α). Neyman and Pearson insisted that one must specify the level of significance *before* the experiment to be able to interpret it as α. The same holds for β, which is the rate of rejecting the alternative hypothesis H_2 if it is correct (Type II error). Here we get the second classical interpretation of the level of significance: the error rate α, which is determined before the experiment, albeit not by mere convention but by cost-benefit calculations that strike a balance between α, β, and sample size n (Cohen, 1994).

21.4.3. Interpretation 3: The Exact Level of Significance

Fisher had second thoughts about his proposal of a conventional level and stated these most clearly in the mid-1950s. In his last book, *Statistical Methods and Scientific Inference* (1956, p. 42), Fisher rejected the use of a conventional level of significance and ridiculed this practice as "absurdly academic" (see epigram). Fisher's primary target, however, was the interpretation of the level of significance as α, which he rejected as unscientific. In science, Fisher argued, unlike in industrial quality control, one does not repeat the same experiment again and again, as is assumed in Neyman and Pearson's interpretation of the level of significance as an error rate in the long run. What researchers should do instead, according to Fisher's second thoughts, is publish the *exact level of significance,* say, $p = .02$ (not $p < .05$), and communicate this result to their fellow researchers.

Thus, the phrase *level of significance* has three meanings:

1. the conventional level of significance, a common standard for all researchers (early Fisher);
2. the α level, that is, the relative frequency of wrongly rejecting a hypothesis in the long run if it is true, to be decided jointly with β and the sample size before the experiment and independently of the data (Neyman & Pearson);
3. the exact level of significance, calculated from the data after the experiment (late Fisher).

The basic difference is this: For Fisher, the exact level of significance is a property of the data, that is, a relation between a body of data and a theory; for Neyman and Pearson, α is a property of the test, not of the data. Level of significance and α are not the same thing. The practical consequences are straightforward:

1. *Conventional level:* You specify only one statistical hypothesis, the null. You always use the 5% level and report whether the result is significant or not; that is, you report $p < .05$ or $p > .05$, just like in the null ritual. If the result is significant, you reject the null; otherwise, you do not draw any conclusion. There is no way to confirm the null hypothesis. The decision is asymmetric.

2. *Alpha level:* You specify two statistical hypotheses, H_1 and H_2, to be able to calculate the desired balance between α, β, and the sample size n. If the result is significant (i.e., if it falls within the alpha region), the decision is to reject H_1 and to act as if H_2 were true; otherwise, the decision is to reject H_2 and to act as if H_1 were true. (We ignore here, for simplicity, the option of a region of indecision.) For instance, if $\alpha = \beta = .10$, then it does not matter whether the exact level of significance is .06 or .001. The level of significance has no influence on α. Unlike in null hypothesis testing with a conventional level, the decision is symmetric.

3. *Exact level of significance:* You calculate the exact level of significance from the data. You report, say, $p = .051$ or $p = .048$. You do not use statements of the type "$p < .05$" but report the exact (or rounded) value. There is no decision involved. You communicate information; you do not make yes-no decisions.

These three interpretations of the level of significance are conflated in most textbooks used in

psychology and education. This confusion is a direct consequence of the sour fact that these textbooks do not teach the toolbox and competing statistical theories but instead only one apparently monolithic form of "statistics"—a mishmash that does not exist in statistics proper (Gigerenzer, 1993, 2000).

Now let us go back to Dr. Publish-Perish and his moral conflict. His superego demands that he specify the level of significance before the experiment. We now understand that this doctrine is part of the Neyman-Pearson theory. His ego personifies Fisher's theory of calculating the exact level of significance from the data but is conflated with Fisher's earlier idea of making a yes-no decision based on a conventional level of significance. The conflict between his superego and his ego is the source of his guilt feelings, but he does not know that. Never having heard that there are different theories, he has a vague feeling of shame for doing something wrong. Dr. Publish-Perish does not follow any of the three different conceptions. Unknowingly, he tries to satisfy all of them and ends up presenting an exact level of significance as if it were an alpha level, yet first rounding it up to one of the conventional levels of significance, $p < .05$, $p < .01$, or $p < .001$. The result is not α, nor an exact level of significance, nor a conventional level. It is an emotional and intellectual confusion.

Figure 21.3 A Freudian Analogy for the Unconscious Conflicts in the Minds of Researchers

The Unconscious Conflict

Superego

(Neyman-Pearson)

Two or more hypotheses; alpha and beta determined before the experiment; compute sample size; no statements about the truth of hypotheses . . .

Ego

(Fisher)

Null hypothesis only; significance level computed after the experiment; beta ignored; sample size by rule of thumb; gets papers published but left with feelings of guilt

Id

(Bayes)

Desire for probabilities of hypotheses

21.5. Question 5: What Emotional Structure Sustains the Null Ritual?

Dr. Publish-Perish is likely to share some of the illusions demonstrated in the first section. Recall that most of these illusions involve the confusion of the level of significance with the probability of a hypothesis. Yet every person of average intelligence can understand the difference between $p(D|H)$ and $p(H|D)$, suggesting that the issue is not an intellectual but a social and emotional one. Following Gigerenzer (1993; see also Acree, 1978), we will continue to use the Freudian language of unconscious conflicts as an analogy to analyze why intelligent people surrender to statistical rituals rather than engage in statistical thinking.

The Neyman-Pearson theory serves as the superego of Dr. Publish-Perish's statistical thinking, demanding in advance the specification of precise alternative hypotheses, significance levels, and power to calculate the sample size necessary, as well as teaching the doctrine of repeated random sampling (Neyman, 1950, 1957). Moreover, the frequentist superego forbids the interpretation of levels of significance as the degree of confidence that a particular hypothesis is true or false. Hypothesis testing, in its view, is about decision making (i.e., acting as if a hypothesis were true or false) but not about epistemic statements (i.e., believing in a hypothesis).

The Fisherian theory of significance testing functions as the ego. The ego gets things done in the laboratory and papers published. The ego determines the level of significance after the experiment, and it does not specify power or calculate the sample size necessary. The ego avoids precise predictions from its research hypothesis and instead claims support for it by rejecting a null hypothesis. The ego makes abundant epistemic statements about particular results and hypotheses. But it is left with feelings of guilt and shame for having violated the rules.

The Bayesian posterior probabilities form the id of this hybrid logic. These probabilities of hypotheses are censored by both the frequentist superego and the pragmatic ego. However, they are exactly what the Bayesian id wants, and it gets its way by wishful thinking and blocking the intellect from understanding what a level of significance really is.

The Freudian analogy (see Figure 21.3) illustrates the unconscious conflicts in the minds of the average student, researcher, and editor and provides a way to understanding why many psychologists cling to null hypothesis testing like a ritual and why they

do not seem to want to understand what they easily could. The analogy brings the anxiety and guilt, the compulsive behavior, and the intellectual blindness associated with the hybrid logic into the foreground. It is as if the raging personal and intellectual conflicts between Fisher and Neyman and Pearson, as well as between these frequentists and the Bayesians, were projected into an "intra-psychic" conflict in the minds of researchers. In Freudian theory, ritual is a way of resolving unconscious conflict.

Textbook writers, in turn, have tried to resolve the conscious conflict between statisticians by collective silence. You will rarely find a textbook for psychologists that points out even a few issues in the heated debate about what is good hypotheses testing, which is covered in detail in Gigerenzer et al. (1989, chaps. 3, 6). The textbook method of denial includes omitting the names of the parents of the various ideas—that is, Fisher, Neyman, and Pearson—except in connection with trivialities such as an acknowledgment for permission to reproduce tables. One of the few exceptions is Hays (1963), who mentioned in one sentence in the second edition that statistical theory made cumulative progress from Fisher to Neyman and Pearson, although he did not hint at their differing ideas or conflicts. In the third edition, however, this sentence was deleted, and Hays fell back to common standards. When one of us (GG) asked him why he deleted this sentence, he gave the same reason as for having removed the chapter on Bayesian statistics: The publisher wanted a single-recipe cookbook, not names of statisticians whose theories might conflict. The fear seems to be that a statistical toolbox would not sell as well as one truth or one hammer.

Many textbook writers in psychology continue to spread confusion about statistical theories, even after they have learned otherwise. For instance, in response to Gigerenzer (1993), Chow (1998) acknowledges that different logics of statistical inference exist. But a few lines later, he falls back into the "it's-all-the-same" fable when he asserts, "To K. Pearson, R. Fisher, J. Neyman, and E. S. Pearson, NHSTP was what the empirical research was all about" (p. xi). Calling the heroes of the past to justify the null ritual (to which NHSTP seems to amount) is bewildering. Each of these statisticians would have rejected NHSTP. Neyman and Pearson spent their careers arguing against null hypothesis testing, against a magical 5% level, and for the concept of Type II error (which Chow declares not germane to NHSTP). Chow's confusion is not an exception. NHSTP is the symptom of the unconscious conflict illustrated in Figure 21.3. Laying open the conflicts between major approaches rather than denying them would be a first step to understanding the underlying issues, a prerequisite for statistical thinking.

21.6. Question 6: Who Keeps Psychologists Performing the Null Ritual?

Ask graduate students, and they likely point to their advisers. The students do not want problems with their thesis. When we meet them again as post-docs, the answer is that they need a job. After getting their first job, they still feel restricted because there is a tenure decision in a couple of years. When they are safe as associate or full professors, it is still not their fault because they believe the editors of the major journals will not publish their papers without the null ritual. There is always someone else to blame, rather than one's own lack of having the courage to know. But fears about punishment for rule violations are not entirely unfounded. For instance, Melton (1962) insisted on the null ritual and also made it clear in his editorial that he wants to see $p < .01$, not just $p < .05$. The reasons he gave were two of the illusions listed in Section 21.1. He misleadingly asserted that the lower the p-value, the higher the confidence that the alternative hypothesis is true and the higher the probability that a replication will find a significant result. Nothing beyond p-values is mentioned in the editorial: Precise hypotheses, good descriptive statistics, confidence intervals, effect sizes, and power do not appear in his statement about good research. Thus, the null ritual seems to be enforced by editors.

The story of a recent editor, however, reveals that the truth is not as simple as that. In his "On the Tyranny of Hypothesis Testing in the Social Sciences," Geoffrey Loftus (1991) reviewed *The Empire of Chance* (Gigerenzer et al., 1989), which presented one of the first analyses of how psychologists mishmashed ideas of Fisher and also Neyman and Pearson into one hybrid logic. When Loftus (1993) became the editor of *Memory & Cognition,* he made it clear in his editorial that he did not want authors to submit papers in which p-, t-, or F-values are mindlessly being calculated and reported. Rather, he asked researchers to keep it simple and report figures with error bars, following the proverb that "a picture is worth more than a thousand p-values." We admire Loftus for having had the courage to take this step. Years after, one of us (GG) asked Loftus about the success of his crusade against thoughtless significance testing. Loftus

bitterly complained that most researchers actually refused the opportunity to escape the ritual. Even when he asked in his editorial letter to get rid of dozens of p-values, the authors insisted on keeping them in. There is something deeply engrained in the minds of many researchers that makes them repeat the same action over and over again.

21.7. Question 7: How Can We Advance Statistical Thinking?

There is no single recipe for promoting statistical thinking, but there are several good heuristics. We sketch a few of these, which the readers can use to construct their own program or curriculum.

21.7.1. *Hypotheses* Is in the Plural

If there is one single severe problem with the null ritual, then it is the fact that *hypothesis* is in the singular. Hypotheses testing should always be competitive; that is, the predictions of several hypotheses should be specified. Figure 21.2 gives an example of how the predictions of two hypotheses can be specified graphically. Rieskamp and Hoffrage (1999), for instance, test eight competing hypotheses about how people predict the profit of companies, and Gigerenzer and Hoffrage (1995) test the predictions of six cognitive strategies in problem solving. One advantage of multiple hypotheses is the analysis of individual differences: For instance, one can show that people systematically follow different problem-solving strategies.

21.7.2. Minimize the True Error

Statistical thinking does not simply involve measuring the error and inserting the value into the denominator of the t-ratio. Good statistical thinking is about how to minimize the real error. By *real error,* we refer to the true variability of measurements or observations, not the variance divided by the square root of the number of observations. W. S. Gosset, who published the t-test in 1908 under the pseudonym "Student," wrote, "Obviously the important thing . . . is to have a low real error, not to have a 'significant' result at a particular station. The latter seems to me to be nearly valueless in itself" (quoted in Pearson, 1939, p. 247). Methods of minimizing the real error include proper choice of task (e.g., paired comparison instead of rating) (see Gigerenzer & Richter, 1990), proper choice of experimental environment (e.g., testing participants individually rather than in large classrooms), proper motivation (e.g., by performance-contingent payment rather than flat sums), instructions that are unambiguous rather than vague, and the avoidance of unnecessary deception of participants about the purpose of the experiment, which can lead to second-guessing and increased variability of responses (Hertwig & Ortmann, 2001).

21.7.3. Think of a Toolbox, Not of a Hammer

Recall that the problem of inductive inference has no single best solution—-it has many good solutions. Statistical thinking involves analyzing the problem at hand and then selecting the best tool in the statistical toolbox or even constructing such a tool. No tool is best for all problems. For instance, there is no single best method of representing a central tendency: Whether to report the mean, the median, the mode, or all three of these needs to be decided by the problem at hand. The toolbox includes, among others, descriptive statistics, methods of exploratory data analysis, confidence intervals, Fisher's null hypothesis testing, Neyman-Pearson hypotheses testing, Wald's sequential analysis, and Bayesian statistics.

The concept of a toolbox has an important consequence for teaching statistics. *Stop teaching the null ritual or what is called NHSTP* (see, e.g., Chow, 1998; Harlow, 1997). Teach statistics in the plural: the major statistical tools together with good examples of problems they can solve. For instance, the logic of Fisher's (1956) null hypothesis testing can easily be made clear in three steps:

1. Set up a statistical null hypothesis. The null need not be a nil hypothesis (zero difference).
2. Report the exact level of significance (e.g., $p = .011$ or .051). Do not use a conventional 5% level (e.g., $p < .05$), and do not talk about accepting or rejecting hypotheses.
3. Use this procedure only if you know very little about the problem at hand.

Note that Fisher's null hypothesis testing is, at each step, unlike the null ritual (see introduction). One can see that statistical power has no place in Fisher's framework—one needs a specified alternative hypothesis to compute power. In the same way, one can explain the logic of Neyman-Pearson hypotheses testing, which we illustrate for the case of two hypotheses and a binary decision criterion as follows:

1. Set up two statistical hypotheses, H_1 and H_2, and decide about α, β, and sample size before the experiment, based on subjective cost-benefit considerations. These define a rejection region for each hypothesis.
2. If the data falls into the rejection region of H_1, accept H_2; otherwise, accept H_1. Note that accepting a hypothesis does not imply that you believe in it; it only means that you act as if it were true.
3. The usefulness of the procedure is limited to situations in which you have a disjunction of hypotheses (e.g., either $\mu = 8$ or $\mu = 10$ is true) and in which the scientific context can provide the utilities that enter the choice of alpha and beta.

A typical application of Neyman-Pearson testing is in quality control. Imagine a manufacturer of metal plates that are used in medical instruments. She considers a mean diameter of 8 mm (H_1) as optimal and 10 mm (H_2) as dangerous to the patients and hence unacceptable. From past experience, she knows that the random fluctuations of diameters are approximately normally distributed and that the standard deviations do not depend on the mean. This allows her to determine the sampling distributions of the mean for both hypotheses. She considers accepting H_1 while H_2 is true (Type II error) to be the most serious error because it may cause harm to patients and to the firm's reputation. She sets its probability as $\beta = 0.1\%$ and $\alpha = 10\%$. Now she calculates the required sample size n of plates that must be sampled every day to test the quality of the production. When she accepts H_2, she acts as if there were a malfunction and stops production, but this does not mean that she believes that H_2 is true. She knows that she must expect a false alarm in 1 out of 10 days in which there is no malfunction (Gigerenzer et al., 1989, chap. 3).

The basic logic of other statistical tools can be taught in the same way, and examples for their usefulness and limits can be provided.

21.7.4. Know and Show Your Data

Descriptive statistics and exploratory data analysis are typically more informative than the null ritual, specifically in the presence of multiple hypotheses. For instance, the plot of the three curves shown in Figure 21.2 is more informative than the result of the analysis of variance that the data do not deviate significantly from the predictions of the null. Showing in addition the individual data points around the means of the data curve, or at least the error bars, would be even more informative. Similarly, a scatter plot showing the data points is more informative than a correlation coefficient, for each scatter plot corresponds to one correlation, whereas a correlation of .5, for example, corresponds to many and strikingly different scatter plots. Wilkinson and the Task Force on Statistical Inference (1999) give examples for informative graphs.

21.7.5. Keep It Simple

A statistical analysis should be transparent to its author and the readership. Each statistical method consists of a sequence of mathematical operations, and to understand what the end product (factor scores, regression weights, nonsignificant interactions) means, one needs to check the meaning of each operation at each step. Transparency allows the reader to follow each step and to understand or criticize the analysis. The best vehicle for transparency is simplicity. If a point can be made by a simple analysis, such as plotting the means and standard deviations, one should stick with it rather than using a less transparent method, such as factor analysis or path analysis. The purpose of a statistical analysis is not to impress others with a complex method they do not fully understand. We have witnessed painful talks whereby the audience actually insisted on clarification, only to learn that the author did not understand his fancy method either. Never use a statistical method that is not entirely transparent to you.

21.7.6. *p*-Values Want Company

If you wish to report a *p*-value, remember that it conveys very limited information. Thus, report *p*-values together with information about effect sizes, or power, or confidence intervals. Recall that the null hypothesis that defines the *p*-value need not be a nil hypothesis (e.g., zero difference); any hypothesis can be a null, and many different nulls can be tested simultaneously (e.g., Gigerenzer & Richter, 1990).

21.8. Question 8: How Can We Have More Fun With Statistics?

Many students experience statistics as dry, dull, and dreary. It certainly need not be; real-world

examples (as in Gigerenzer, 2002) can make statistical thinking exciting. Here are several other ways of turning students into statistics addicts, or at least of making them think. The first heuristic is to draw a red thread from the past to the present. We understand the aspirations and fears of a person better if we know his or her history. Knowing the history of a statistical concept can create a similar feeling of intimacy.

21.8.1. Connecting to the Past

The first test of a null hypothesis was by John Arbuthnot in 1710. His aim was to give an empirical proof of divine providence, that is, of an active God. Arbuthnot observed that "the external accidents to which males are subject (who must seek their food with danger) do make a great havock of them, and that this loss exceeds far that of the other sex" (p. 188). To repair this loss, he argued, God brings forth more males than females, year after year. He tested this hypothesis of divine purpose against the null hypothesis of mere chance, using 82 years of birth records in London. In every year, the number of male births was larger than that of female births. Arbuthnot calculated the "expectation" of these data if the hypothesis of blind chance were true. In modern terms, the probability of these data if the null hypothesis were true was

$$p(D|\mathrm{H}_0) = (1/2)^{82}.$$

Because this probability was so small, he concluded that it is divine providence, not chance, that rules:

> *Scholium.* From hence it follows, that Polygamy is contrary to the Law of Nature and Justice, and to the Propagation of the human Race; for where Males and Females are in equal number, if one Man takes Twenty Wifes, Nineteen Men must live in Celibacy, which is repugnant to the Design of Nature; nor is it probable that Twenty Women will be so well impregnated by one Man as by Twenty. (qtd. in Gigerenzer & Murray, 1987, pp. 4–5)

Arbuthnot's proof of God highlights the limitations of null hypothesis testing. The research hypothesis (God's divine intervention) is not stated in statistical terms. Nor is a substantial alternative hypothesis stated in statistical terms (e.g., 3% of female newborns are abandoned immediately after birth). Only the null hypothesis ("chance") is stated in statistical terms—a nil hypothesis. A result that is unlikely if the null were true (a low p-value) is taken as "proof" of the unspecified research hypothesis.

Arbuthnot's test was soon forgotten. The specific techniques of null hypothesis testing, such as the t-test (devised by Gosset in 1908) or the F-test (F for Fisher, e.g., in analysis of variance), were first applied in the context of agriculture. The examples in Fisher's first book on statistics (1925) smelled of manure, potatoes, and pigs. In his second book (1935), Fisher had cleaned out this odor, as well as much of the mathematics, so that social scientists could bond with the new statistics. The first applications of these tests in psychology were mostly in parapsychology and education.

A striking change in research practice, which was named the *inference revolution* in psychology (Gigerenzer & Murray, 1987), happened from approximately 1940 to 1955 in the United States. It led to the institutionalization of the null ritual as *the* method of scientific inference in university curricula, textbooks, and the editorials of major journals. Before 1940, null hypothesis testing using analysis of variance or the t-test was practically nonexistent: Rucci and Tweney (1980) found a total of only 17 articles published from 1934 to 1940 that used it. By the early 1950s, half of the psychology departments in leading U.S. universities had made inferential statistics a graduate program requirement (Rucci & Tweney, 1980). By 1955, more than 80% of the empirical articles in four leading journals used null hypothesis testing (Sterling, 1959). Today, the figure is close to 100%. Despite decades of critique of the null ritual, it is still practiced and defended by the majority of psychologists. For instance, it is often argued that if we can strip routine null hypothesis testing of the mental confusion associated with it, something of limited but important use is left: "deciding whether or not research data can be explained in terms of chance influences" (Chow, 1998, p. 188). We are back to Arbuthnot: The focus is on chance; to test substantive alternative hypotheses is not an issue. Arbuthnot, it should be said to his defense, was a step ahead—he did not recommend his procedure as a routine.

Materials to connect with the past can be drawn from two seminal books by Stephen Stigler (1986, 1999). His writing is so clear and entertaining that it feels as though one had grown up with statistical thinking. Danziger (1987), Gigerenzer (1987, 2000), and Gigerenzer et al. (1989) tell the story of the institutionalization of the null ritual in psychology.

21.8.2. Controversies and Polemics

Statistics has plenty of controversies. These stories of conflict can provide highly motivating material

for students, who learn that—unlike in their textbooks—statistics is about real people and their struggles with ideas and with one another. Because of Fisher's remarkable talent for polemics, his writings can serve as a starting point. Here are a few highlights.

Fisher once congratulated the Reverend Thomas Bayes for his insight to withhold his treatise from publication (it was published posthumously in 1763/1963). Why did Fisher say that? Bayes's rule presupposes the availability of a prior probability distribution over the possible hypotheses, and Fisher insisted that such a distribution is only meaningful when it can be verified by sampling from a population. Such distributional data are available in the case of HIV testing (see Question 2) but obviously uncommon for scientific hypotheses. Fisher believed that the Bayesians are wrong in assuming that all uncertainties can be expressed in terms of probabilities (see Gigerenzer et al., 1989, pp. 92–93).

Bayes's rule and subjective probabilities were not the only target for Fisher. He branded Neyman's position as "childish" and "horrifying [for] the intellectual freedom of the west." Indeed, he likened Neyman to

> Russians [who] are made familiar with the ideal that research in pure science can and should be geared to technological performance, in the comprehensive organized effort of a five-year plan for the nation ... [whereas] in the U.S. also the great importance of organized technology has I think made it easy to confuse the process appropriate for drawing correct conclusions, with those aimed rather at, let us say, speeding production, or saving money. (Fisher, 1955, p. 70)

Why did Fisher link the Neyman-Pearson theory to Stalin's 5-year plans? Why did Fisher also compare them to the Americans, who confuse the process of gaining knowledge with speeding up production and saving money? It is probably not an accident that Neyman was born in Russia and, at the time of Fisher's comment, had moved to the United States. What Fisher believed was that cost-benefit calculations, Type I error rates, Type II error rates, and accept-reject decisions had nothing to do with gaining knowledge but instead with technology and making money, as in quality control in industry. Researchers do not accept or reject hypotheses; rather, they communicate the exact level of significance to fellow researchers, so that others can freely make up their minds. In Fisher's eyes, free communication was a sign of the freedom of the West, whereas being told a decision was a sign of communism. For him, the concepts of α, β, and power $(1 - \beta)$ have nothing to do with testing scientific hypotheses. They are defined as long-run frequencies of errors in repeated experiments, whereas in science, there are no experiments repeated again and again.

Fisher (1956) drew a bold line between his null hypothesis tests and Neyman-Pearson's tests, which he ridiculed as originating from "the phantasy of circles [i.e., mathematicians] rather remote from scientific research" (p. 100). Neyman, for his part, responded that some of Fisher's tests "are in a mathematically specifiable sense 'worse than useless'" (Hacking, 1965, p. 99). What did Neyman have in mind with this verdict? Neyman had estimated the power of some of Fisher's tests, including the famous Lady-tea-tasting experiment in Fisher (1935), and found that the power was sometimes smaller than α.

Polemics can motivate students to ask questions and to understand the competing ideas underlying the tools in the toolbox. For useful material, see Fisher (1955, 1956), Gigerenzer (1993), Gigerenzer et al. (1989, chap. 3), Hacking (1965), and Neyman (1950).

21.8.3. Playing Detective

Aside from motivating examples, history, and polemics, a further way to engage students is to challenge them to find the errors of others. For instance, assign your students the task of looking up the section on the logic of hypothesis testing in textbooks for statistics in psychology and checking for wishful thinking, as in Table 21.1. Table 21.2 shows the result for a widely read textbook whose author, as usual, did not spell out the differences between Fisher, Neyman and Pearson, and the Bayesians but mixed them all up. The price for this was confusion and wishful thinking about the omnipotence of the level of significance. Table 21.2 shows quotes from three pages of the textbook, in which the author tries to explain to the reader what a level of significance means. For instance, the first three assertions are unintelligible or plainly wrong and suggest that a level of significance would provide information about the probability of hypotheses, and the fourth amounts to the replication fallacy.

Over the years, textbooks writers in psychology have learned to avoid obvious errors but still continue to teach the null ritual. For instance, the 16th edition of a very influential textbook, Gerrig and Zimbardo's (2002) *Psychology and Life,* contains sections on "inferential statistics" and "becoming a wise consumer of statistics" (pp. 37–46), which are pure guidelines for the null ritual. The ritual is portrayed as statistics per se and named the "backbone of psychological research" (p. 46). Our detective student will find that the names

Table 21.2 What Does "Significant at the 5% Level" Mean?

- "If the probability is low, the null hypothesis is improbable"
- "The *improbability* of observed results being due to error"
- "The probability that an observed difference is real"
- "The *statistical confidence* . . . with odds of 95 out of 100 that the observed difference will hold up in investigations"
- Degree to which experimental results are taken "seriously"
- "The danger of accepting a statistical result as real when it is actually due only to error"
- Degree of "faith [that] can be placed in the reality of the finding"
- "The investigator can have 95 percent confidence that the sample mean actually differs from the population mean"
- "All of these are different ways to say the same thing"

SOURCE: Nunally (1975).
NOTE: Within three pages of text, the author of a widely read textbook explained to the reader that "level of significance" means all of the above (Nunally, 1975, pp. 194–196). Smart students will be confused, but they may misattribute their confusion to their own lack of understanding.

of Fisher, Bayes, Neyman, and Pearson are not mentioned, nor are concepts such as power, effect size, or confidence intervals. She may also stumble upon the prevailing oracular language: "Inferential statistics indicate the probability that the particular sample of scores obtained are actually related to whatever you are attempting to measure or whether they could have occurred by chance" (p. 44). Yet in the midst of unintelligible and nonsensical explanations such as these appear moments of deep insight: "Statistics can also be used poorly or deceptively, misleading those who do not understand them" (p. 46).

21.9. Question 9: What If There Were No Significance Tests?

This question has been asked in a series of articles in Harlow, Mulaik, and Steiger (1997) and in similar debates, which are summarized in the superb review by Nickerson (2000). However, there are actually two different questions: What if there were no null hypothesis testing (significance testing), as advocated by Fisher? What if there were no null ritual (or NHSTP)?

If eminent psychologists have anything in common, it is their distaste for mindless null hypothesis testing—which contrasts with the taste of the masses. You will not catch Jean Piaget testing a null hypothesis. Piaget worked out his logical theory of cognitive development, Wolfgang Köhler the Gestalt laws of perception, I. P. Pavlov the principles of classical conditioning, B. F. Skinner those of operant conditioning, and Sir Frederick Bartlett his theory of remembering and schemata—all without rejecting a null hypothesis. Moreover, F. Bartlett, R. Duncan Luce, Herbert A. Simon, B. F. Skinner, and S. S. Stevens explicitly protested in their writings against the null ritual (Gigerenzer, 1987, 1993; Gigerenzer & Murray, 1987).

So what if there were no null ritual or NHST? Nothing would be lost, except confusion, anxiety, and a platform for lazy theoretical thinking. Much could be gained, such as knowledge about different statistical tools, training in statistical thinking, and a motivation to deduce precise predictions from one's hypotheses. Should we ban the null ritual? Certainly—it is a matter of intellectual integrity. Every researcher should have the courage not to surrender to the ritual, and every editor, textbook writer, and adviser should feel obliged to promote statistical thinking and reject mindless rituals.

What if there were no null hypothesis testing, as advocated by Fisher? Not much would be lost, except in situations in which we know very little, where a p-value by itself can contribute something. Note that this question is a different one: Fisher's null hypothesis testing is one tool in the statistical toolbox, not a ritual. Should we ban null hypothesis testing? No, there is no reason to do so; it is just one small tool among many. What we need is to educate the next generation to dare to think and free themselves from compulsive hand-washing, anxiety, and feelings of guilt.

References

Acree, M. C. (1978). *Theories of statistical inference in psychological research: A historicocritical study.* Ann Arbor, MI: University Microfilms International. (University Microfilms No. H790 H7000)

American Psychological Association. (1974). *Publication manual.* Baltimore, MD: Garamond/Pridemark.

American Psychological Association. (1983). *Publication manual* (3rd ed.). Baltimore, MD: Garamond/Pridemark.

Anastasi, A. (1958). *Differential psychology* (3rd ed.). New York: Macmillan.

Anderson, N. H. (1981). *Foundations of information integration theory.* New York: Academic Press.

Anderson, N. H., & Cuneo, D. (1978). The height + width rule in children's judgments of quantity. *Journal of Experimental Psychology: General, 107,* 335–378.

Arbuthnot, J. (1710). An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society, 27,* 186–190.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423–437.

Bayes, T. (1963). An essay towards solving a problem in the doctrine of chances. In W. E. Deming (Ed.), *Two papers by Bayes.* New York: Hafner. (Original work published 1763)

Chow, S. L. (1998). Précis of "Statistical significance: Rationale, validity, and utility." *Behavioral and Brain Sciences, 21,* 169–239.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145–153.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Danziger, K. (1987). Statistical methods and the historical development of research practice in American psychology. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences* (pp. 35–47). Cambridge, MA: MIT Press.

Dulaney, S., & Fiske, A. P. (1994). Cultural rituals and obsessive-compulsive disorder: Is there a common psychological mechanism? *Ethos, 22,* 243–283.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard. *Theory & Psychology, 5,* 75–98.

Ferguson, L. (1959). *Statistical analysis in psychology and education.* New York: McGraw-Hill.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh, UK: Oliver & Boyd.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh, UK: Oliver & Boyd.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society, Series B, 17,* 69–77.

Fisher, R. A. (1956). *Statistical methods and scientific inference.* Edinburgh, UK: Oliver & Boyd.

Gerrig, R. J., & Zimbardo, P. G. (2002). *Psychology and life* (16th ed.). Boston: Allyn & Bacon.

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Krüger, G. Gigerenzer & M. Morgan (Eds.), *The probabilistic revolution: Vol. II. Ideas in the sciences* (pp. 11–33). Cambridge, MA: MIT Press.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, NJ: Lawrence Erlbaum.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world.* New York: Oxford University Press.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you.* New York: Simon & Schuster.

Gigerenzer, G. (2003). Reckoning with risk: Learning to live with uncertainty. London: Penguin.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684–704.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Lawrence Erlbaum.

Gigerenzer, G., & Richter, H. R. (1990). Context effects and their interaction with development: Area judgments. *Cognitive Development, 5,* 235–264.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and every day life.* Cambridge, UK: Cambridge University Press.

"Student" [W. S. Gosset] (1908). The probable error of a mean. *Biometrica, 6,* 1–25.

Guilford, J. P. (1942). *Fundamental statistics in psychology and education.* New York: McGraw-Hill.

Hacking, I. (1965). *Logic of statistical inference.* Cambridge, UK: Cambridge University Press.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research–Online [Online serial], 7*(1), 1–20. Retrieved June 10, 2003, from www.mpr-online.de

Harlow, L. L. (1997). Significance testing: Introduction and overview. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 1–17). Mahwah, NJ: Lawrence Erlbaum.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum.

Hays, W. L. (1963). *Statistics for psychologists* (2nd ed.). New York: Holt, Rinehart & Winston.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences, 24,* 383–403.

Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach.* La Salle, IL: Open Court.

Lindquist, E. F. (1940). *Statistical analysis in educational research.* Boston: Houghton Mifflin.

Loftus, G. R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology, 36,* 102–105.

Loftus, G. R. (1993). Editorial comment. *Memory & Cognition, 21,* 1–3.

Luce, R. D. (1988). The tools-to-theory hypothesis: Review of G. Gigerenzer and D. J. Murray, "Cognition as intuitive statistics." *Contemporary Psychology, 33,* 582–583.

Maslow, A. H. (1966). *The psychology of science.* New York: Harper & Row.

Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology, 64,* 553–557.

Miller, G. A., & Buckhout, R. (1973). *Psychology: The science of mental life.* New York: Harper & Row.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and a place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65–115). Mahwah, NJ: Lawrence Erlbaum.

Neyman, J. (1950). *First course in probability and statistics.* New York: Holt.

Neyman, J. (1957). Inductive behavior as a basic concept of philosophy of science. *International Statistical Review, 25,* 7–22.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.

Nunally, J. C. (1975). *Introduction to statistics for psychology and education.* New York: McGraw-Hill.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* Chichester, UK: Wiley.

Pearson, E. S. (1939). "Student" as statistician. *Biometrika, 30,* 210–250.

Pollard, P., & Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 102,* 159–163.

Rieskamp, J., & Hoffrage, U. (1999). When do people use simple heuristics and how can we tell? In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Simple*

heuristics that make us smart (pp. 141–167). New York: Oxford University Press.

Rucci, A. J., & Tweney, R. D. (1980). Analysis of variance and the "second discipline" of scientific psychology: A historical account. *Psychological Bulletin, 87,* 166–184.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316.

Skinner, B. F. (1984). *A matter of consequences.* New York: New York University Press.

Sterling, R. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54,* 30–34.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: Belknap Press of Harvard University Press.

Stigler, S. M. (1999). *Statistics on the table: The history of statistical concepts and methods*. Cambridge, MA: Harvard University Press.

Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Chapter 22

On Exogeneity

David Kaplan

22.1. Introduction

When using linear statistical models to estimate substantive relationships, a distinction is made between endogenous variables and exogenous variables. Alternative names for these variables are *dependent* and *independent,* or *criterion* and *predictor.* In the case of multiple linear regression, one variable is designated as the endogenous variable, and the remaining variables are designated as exogenous. In multivariate regression, a set of endogenous variables is chosen and related to one or more exogenous variables. In the case of structural equation modeling, there is typically a set of endogenous variables that are related to each other and also related to a set of exogenous variables.

More often than not, the choice of endogenous and exogenous variables is guided by the research question of interest, with little consideration given to statistical consequences of that choice. Moreover, an inspection of standard textbooks in the social and behavioral sciences reveals confusing definitions of endogenous and exogenous variables. For example, Cohen and Cohen (1983) write,

> Exogenous variables are measured variables that are not caused by any other variable in the model except (possibly) other exogenous variables. They have essentially the same meaning as independent variables in ordinary regression analysis except that they explicitly include the assumption that they are not causally dependent on the endogenous variables in the model. Endogenous variables are, in part, effects of exogenous variables and do not have a causal effect on them. (p. 375)

In another example, Bollen (1989) writes,

> The terms exogenous and endogenous are model specific. It may be that an exogenous variable in one model is endogenous in another. Or, a variable shown as exogenous, in reality, may be influenced by a variable in the model. Regardless of these possibilities, the convention is to refer to variables as exogenous or endogenous based on their representation in a particular model. (p. 12)

And finally, from an econometric perspective, Wonnacott and Wonnacott (1979) write with regard to treating income (denoted as I in their definition) as an exogenous variable,

> An important distinction must be made between two kinds of variables in our system. By assumption, I is an exogenous variable. Since its value is determined from *outside* the system, it will often be referred to as *predetermined;* however it should be recognized that a predetermined variable may be either fixed *or* random. The essential point is that its values are determined elsewhere. (pp. 257–258)

AUTHOR'S NOTE: This research was supported by a fellowship from the American Educational Research Association, which receives funds for its "AERA Grants Program" from the National Science Foundation, the National Center for Education Statistics, and the Office of Educational Research and Improvement (U.S. Department of Education) under NSF Grant # REC-9980573. Opinions reflect those of the author and do not necessarily reflect those of the granting agencies. The author is grateful to Professor Aris Spanos for valuable comments on an earlier draft of this chapter.

The above definitions of exogenous variables are typical of those found in most social science statistics textbooks.[1] Nevertheless, these and other similar definitions are problematic for a number of reasons. First, these definitions do not articulate precisely what it means to say that exogenous variables are not dependent on the endogenous variables in the model. For example, with the Cohen and Cohen (1983) definition, if an exogenous variable is possibly dependent on other exogenous variables, then these "dependent" exogenous variables are actually endogenous, and what is being described by this definition is a system of structural equations. Second, Bollen's (1989) definition, although accurately describing the standard convention, seems to confuse a variable's representation in a model with exogeneity. However, the location of a variable in a model does not necessarily render the variable exogenous. In other words, Bollen's definition implies that simply stating that a variable is exogenous makes it so. Moreover, Bollen defines *exogeneity* with respect to a model and not with respect to the statistical structure of the data used to test the model. Third, in the Wonnacott and Wonnacott (1979) definition, the notion of "outside the system" is never really developed. Implied by these different definitions is a confusion between theoretical exogeneity versus statistical exogeneity and the consequences for the former when the latter does not hold.

From our discussion so far, it is clear that these common definitions of exogeneity do not provide a complete picture of the subtleties or seriousness of the problem. A more complete study of the problem of exogeneity comes from the work of Richard (1982) and his colleagues within the domain of econometrics. This chapter, therefore, provides a didactic introduction to the econometric notion of exogeneity as it pertains to linear regression with a brief discussion of the problem with respect to structural equation modeling, multilevel modeling, and growth curve modeling. It is the goal of this chapter to highlight the seriousness of examining exogeneity assumptions carefully when specifying statistical models—particularly if models are to be used for prediction or the evaluation of policies or interventions. Attention will focus primarily on the concept of weak exogeneity and informal methods for testing whether weak exogeneity holds. The more restrictive concept of strong exogeneity will be similarly introduced along with the notion of Granger noncausality, which will require incorporating a dynamic component into the simple linear regression model. Super exogeneity will be introduced along with related concepts of parameter constancy and invariance. Methods for testing strong and super exogeneity will be outlined. Weak, strong, and super exogeneity will be linked to the uses of a statistical model for inference, forecasting, and policy analysis, respectively.

The organization of this chapter is as follows. In Section 22.2, the general problem of exogeneity is introduced. In Section 22.3, the concept of weak exogeneity will be defined in the case of simple linear regression. Here, the auxiliary concepts of *parameters of interest* and *variation freeness* will be introduced. This section will also discuss exogeneity in the context of structural equation modeling. In Section 22.4, we will consider the conditions under which weak exogeneity can be assumed to hold, as well as conditions where it is likely to be violated. We will also consider three indirect but related tests of weak exogeneity. In Section 22.5, we will introduce a temporal component to the model that will lead to the concept of Granger noncausality and, in turn, to strong exogeneity. We will discuss these concepts as they pertain to the use of statistical models for prediction. In Section 22.6, we will consider the problem of super exogeneity and the concepts of parameter constancy and invariance. We will consider these concepts in light of their implications for evaluating interventions or policies. Finally, Section 22.7 will conclude with a discussion of the implications of the exogeneity assumption for the standard practice of statistical modeling, briefly touching on the implications of the exogeneity assumption for two other popular statistical methodologies in the social and behavioral sciences. Throughout this chapter, concepts will be grounded in substantive problems within the field of education and education policy.

22.2. The Problem of Exogeneity

It was noted in Section 22.1 that definitions of exogenous and endogenous variables encountered in standard social science statistics textbooks are often confusing. In this section, we consider the problem of defining exogeneity more carefully, relying on work in econometric theory. A collection of seminal papers on the problem of exogeneity can be found in Ericsson and Irons (1994), and a brief discussion of the problem was introduced to the structural equation modeling literature by Kaplan (2000).

To begin, it is typical to invoke the heuristic that an exogenous variable is one whose cause is determined from "outside the system under investigation." This heuristic is implied in the Wonnacott and

1. It is also quite common to find that textbooks avoid a definition of exogenous variables altogether.

Wonnacott (1979) definition of an exogenous variable given above. Usually, the notion of a variable being generated from "outside the system" is another way of stating that there is zero covariance between the regressor and the disturbance term. However, such a heuristic is problematic upon close inspection because it does not explicitly define what "outside the system" actually means.

As a way of demonstrating the problem with this heuristic, consider the counterexample given by Hendry (1995) of a fixed-regressor model. To provide a substantive motivation for these ideas, consider the problem of estimating the relationship between reading proficiency in young children as a function of parental reading activities (e.g., how often each week parents read to their children). We may represent this relationship by the simple model

$$y_t = \beta x_t + u_t, \tag{1}$$

where y represents reading proficiency, x represents the parental reading activities, β is the regression coefficient, and u is the disturbance term, which is assumed to be $NID(0, \sigma_u^2)$. The subscript t denotes the particular time point of measurement—a distinction that might be needed with the analysis of panel data.

Typically, parental reading activities are treated as fixed. That is, at time t, levels of parental involvement in reading are assumed to be set and remain the same from that point on. If this assumption were true, then conditional estimation of reading proficiency given parental involvement in reading activities would be valid. However, it is probably not the case in practice that parental reading activities are fixed but rather are likely to be a function of past parental reading activities. That is, perhaps the mechanism that generates parental reading activities at time t is better represented by a first-order autoregressive model,

$$x_t = \gamma x_{t-1} + v_t, \tag{2}$$

where we will assume that $|\gamma| < 1$, ensuring a stable autoregressive process. Even if it were the case that the model in equation (2) generated parental reading activities *prior* to generating reading proficiency, that is still not a sufficient condition to render parental reading activities exogenous in this example. The reason is that such a condition does not preclude current disturbances in equation (1) to be related to past disturbances in equation (2)—namely,

$$u_t = \varphi v_{t-1} + \varepsilon_t. \tag{3}$$

If equation (3) holds for $\varphi \neq 0$, then

$$\begin{aligned} E(x_t, u_t) &= E[(\gamma x_{t-1} + v_t)(\varphi v_{t-1} + \varepsilon_t)] \\ &= \gamma\varphi\sigma_v^2, \end{aligned} \tag{4}$$

and, therefore, x_t is correlated with u_t and hence is not exogenous.

This simple counterexample serves to illustrate the subtleties of the problem of exogeneity. Despite treating parental reading activities as a fixed regressor and assuming that it is generated "from outside the system," the fact is that the true mechanism that generates current values of the regressor yields a model in which the regressor is correlated with the disturbance term, suggesting that it is generated from inside the system as far as the model is concerned. Therefore a rigorous definition of exogeneity is required that does not depend on the particular model under study but rather is based on the true structure of the system under investigation (Hendry, 1995).

22.3. Weak Exogeneity

Having shown that the concept of exogeneity is more subtle than standard definitions imply, we can begin our formal discussion of the problem by introducing the concept of weak exogeneity, which will serve to set the groundwork for subsequent discussions of other forms of exogeneity. To fix ideas, consider a matrix of variables denoted as $\mathbf{z}$ of order $N \times r$, where N is the sample size and r is the number of variables. Under the assumption of independent observations, the joint distribution of $\mathbf{z}$ is given as

$$f(\mathbf{z}|\boldsymbol{\theta}) = f(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N|\boldsymbol{\theta}) = \prod_{i=1}^{N} f(\mathbf{z}_i|\boldsymbol{\theta}), \tag{5}$$

where $\boldsymbol{\theta}$ is a vector of parameters of the joint distribution of $\mathbf{z}$. Most statistical modeling requires a partitioning of $\mathbf{z}$ into endogenous variables to be modeled and exogenous variables that are assumed to account for the variation and covariation in the endogenous variables. Denote by $\mathbf{y}$ the $N \times p$ matrix of endogenous variables and denote by $\mathbf{x}$ an $N \times q$ matrix of exogenous variables where $r = p + q$. We can rewrite equation (1) in terms of the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ and the marginal distribution of $\mathbf{x}$. That is, equation (1) can be related to the conditional distribution in the following decomposition:

$$f(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) = f(\mathbf{y}|\mathbf{x}, \boldsymbol{\omega}_1) f(\mathbf{x}, \boldsymbol{\omega}_2), \tag{6}$$

where $\boldsymbol{\omega}_1$ are the parameters associated with the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$, and $\boldsymbol{\omega}_2$ are the parameters associated with the marginal distribution of $\mathbf{x}$. The parameter spaces of $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ are denoted as $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$, respectively.

It is clear that factoring the joint distribution in equation (5) into the product of the conditional distribution and marginal distribution in equation (6) presents no loss of information. However, standard statistical modeling almost always focuses on the conditional distribution in equation (6). Indeed, the conditional distribution is often referred to as the *regression function.* That being the case, then focusing on the conditional distribution assumes that the marginal distribution can be taken as given (Ericsson, 1994). The issue of exogeneity concerns the implications of this assumption for the parameters of interest.

22.3.1. Variation Freeness

Another important concept as it relates to the problem of exogeneity is that of *variation freeness.* Specifically, variation freeness means that for any value of $\boldsymbol{\omega}_2$ in $\boldsymbol{\Omega}_2$, $\boldsymbol{\omega}_1$ can take on any value in $\boldsymbol{\Omega}_1$ and vice versa (Spanos, 1986). In other words, it is assumed that the pair $(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ belong to the product of their respective parameter spaces—namely, $(\boldsymbol{\Omega}_1 \times \boldsymbol{\Omega}_2)$—and that the parameter space $\boldsymbol{\Omega}_1$ is not restricted by $\boldsymbol{\omega}_2$ and vice versa. Thus, knowing the value of a parameter in the marginal model provides no information regarding the range of values that a parameter in the conditional model can take. Alternatively, restricting $\boldsymbol{\omega}_2$ in any way that ensures that $\boldsymbol{\omega}_2$ is in $\boldsymbol{\Omega}_2$ does not restrict $\boldsymbol{\omega}_1$ in any way that does not allow it to take all possible values in $\boldsymbol{\Omega}_1$.

As an example of variation freeness, consider a simple regression model with one endogenous variable y and one exogenous variable x. The parameters of interest of the conditional distribution are $\boldsymbol{\omega}_1 \equiv (\beta_0, \beta_1, \sigma_u^2)$, and the parameters of the marginal distribution are $\boldsymbol{\omega}_2 \equiv (\mu_x, \sigma_x^2)$. Furthermore, note that $\beta_1 = \sigma_{xy}/\sigma_x^2$, where σ_{xy} denotes the covariance of x and y. Following Ericsson (1994), if σ_{xy} varies proportionally with σ_x^2, then σ_x^2, which is in $\boldsymbol{\omega}_2$, carries no information relevant for the estimation of $\beta_1 = \sigma_{xy}/\sigma_x^2$, which is in $\boldsymbol{\omega}_1$. Therefore, $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ are variation free. An example in which variation freeness could be violated is in cases where a parameter in the conditional model is constrained to be equal to a parameter in the marginal model—however, such cases are rare in the social and behavioral sciences. Below we will show an example in which the condition of variation freeness does not hold.

22.3.2. Parameters of Interest

Variation freeness does not guarantee that one can ignore the marginal model when interest centers on the parameters of the conditional model. As in Ericsson (1994), if interest centers on estimating the conditional and marginal means, then both the conditional and marginal models are needed.[2] This requires us to focus the issue of variation freeness on the *parameters of interest*—namely, those parameters that are a function of the parameters of the conditional model only. More formally, the parameters of interest $\boldsymbol{\Psi}$ are a function of $\boldsymbol{\omega}_1$; that is, $\boldsymbol{\Psi} = g(\boldsymbol{\omega}_1)$.

22.3.3. A Definition of Weak Exogeneity

The above concepts of factorization, parameters of interest, and variation freeness lead to a definition of weak exogeneity. Specifically, following Richard (1982; see also Ericsson, 1994; Spanos, 1986), a variable x is weakly exogenous for the parameters of interest (say, $\boldsymbol{\Psi}$) if and only if there exists a reparameterization of $\boldsymbol{\theta}$ as $\boldsymbol{\omega}$ with $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$, such that

(i) $\boldsymbol{\Psi} = g(\boldsymbol{\omega}_1)$—that is, $\boldsymbol{\Psi}$ is a function of $\boldsymbol{\omega}_1$ only—and
(ii) $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ are variation free—that is, $(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2) \in \boldsymbol{\Omega}_1 \times \boldsymbol{\Omega}_2$.

22.3.4. Weak Exogeneity and the Problem of Nominal Regressors[3]

It is quite common in the social and behavioral sciences for models to contain regressor variables whose scales are nominal. Examples of such variables include demographic features of individuals such as gender or race. In other cases, nominal variables may represent orthogonal components of an experimental design, such as assignment to a treatment or control group. In both cases, the regressors are fixed, nonstochastic constants to be contrasted with stochastic random variables such as socioeconomic status or the amount of parental reading activities. In both cases, data are often submitted to some regression analysis software package for estimation. In the case of experimental design variables, data are

2. In point of fact, however, one can "recover" the marginal mean of $\mathbf{x}$ from the constant in a regression.

3. The author is grateful to Professor Aris Spanos for clarifying this issue.

often submitted to an analysis-of-variance (ANOVA) package. Experimental design textbooks often include a discussion of how ANOVA can be viewed as a special case of the "linear regression" model (see, e.g., Kirk, 1995). The similarity of ANOVA and the linear regression model generates a problem with respect to our discussion of exogeneity. Specifically, given our discussion of exogeneity to this point, a fair question may be to what extent nominal variables, such as gender, race, or experimental design arrangements, are "exogenous" for statistical estimation. In what sense are these variables generated from "outside the system"?

That the question of the "exogeneity" of nominal regressors is raised at all is suggestive of a conflation of ideas typically represented in statistical textbooks in the social sciences—specifically, the merging of the so-called *Gauss linear model* and the *linear regression model* (Spanos, 1999). Indeed, the similarity of the notation of both models contributes to the confusion.

Briefly, the origins of the Gauss linear model came about as an attempt to explain lawful relationships in planetary orbits using less than perfectly accurate measuring instruments. In that context, the Gauss linear model represented an "experimental design" situation in which the xs were fixed, nonstochastic constants albeit subject to observational error. Only the outcome variable y was considered to be a random variable. Indeed, according to Spanos (1999), the original linear model, as proposed by Legendre (1805), did not rest on any formal probabilistic arguments whatsoever. Rather, probabilistic arguments regarding the structure of the errors were added by Gauss and Laplace to justify the statistical optimality of the least squares approach to parameter estimation. Specifically, if it could be assumed that the errors were normal, independent, and identically distributed, then the least squares approach attained certain optimal properties. Later, Fisher applied the Gauss linear model to experimental designs and added the idea of randomization.

What is important for our discussion is that the Gauss linear model was not explicitly rooted in probabilistic notions of random variables, leading, in turn, to notions of conditional versus marginal distributions. It was Galton, with assistance from Karl Pearson, who later proposed the linear regression model, unaware that it was in any way related to the Gauss linear model. The hope was to use the rigorous "lawlike" modeling ideas of Gauss to support Galton's emerging theories of heredity and eugenics (Spanos, 1999). However, it was G. U. Yule (1897) who demonstrated that the same method of least squares used to estimate the Gauss linear model could also be used to estimate Galton's linear regression model (Mulaik, 1985). In this case, y and x were assumed to be jointly normal random variables, and βx was defined as the conditional expectation of y given x, where x is the realization of a stochastic random variable X.

Defining the conditional expectation requires being able to factor the joint distribution into the conditional and marginal distributions, and this requires stochastic random regressors (Spanos, 1999). Therefore, from the standpoint of our discussion of exogeneity, nominal regressors such as race, gender, or experimental design variables do not lead to any conceptual difficulty. When such variables are of interest, one has specified a Gauss linear model. The notion of the conditional distribution does not enter into the discussion because factoring the joint distribution into the conditional and marginal distributions is only possible in the case of stochastic random regressors. In the context of the linear regression model, however, nonstochastic variables enter the conditional mean via the marginal means of the stochastic variables; that is, the constant term is a function of the nonstochastic variables and is therefore not constant.[4]

22.3.5. An Extension to Structural Equation Modeling

It may be of interest to examine how the problem of weak exogeneity extends to structural equation modeling. We focus on structural equation modeling because it had its origins primarily in econometrics (see Kaplan, 2000, for a brief history), and certain aspects of its development are relevant to our discussion of exogeneity. We consider the problem of exogeneity with reference to other methodologies in Section 22.7.

To examine the relevance of weak exogeneity for structural equation models, we should revisit the distinction between the *structural form* and the *reduced-form* specifications of a structural equation model. The structural form of the general structural equation model is denoted as (e.g., Jöreskog, 1973)

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{B}\mathbf{y} + \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta}, \tag{7}$$

4. To see this, consider the addition of a nonstochastic variable (say, gender) to a regression model with other stochastic regressors. Heterogeneity in the mean of y and the mean of x induced by gender can be modeled as $\mu_y = a(gender)$, and $\mu_x = d(gender)$, where a and d are parameters. Expressed in terms of the regression function, $\mu_y = \beta_0 + \beta_1\mu_x$. After substitution, $a(gender) = \beta_0 + \beta_1 d(gender)$, from which we obtain $\beta_0 = (a - \beta_1 d)gender$. Thus, the constant term is a function of a nonstochastic variable.

where $\mathbf{y}$ is a vector of endogenous variables, $\boldsymbol{\alpha}$ is a vector of structural intercepts, $\mathbf{B}$ is a matrix of coefficients relating endogenous variables to each other, $\boldsymbol{\Gamma}$ is a matrix relating endogenous variables to exogenous variables, $\mathbf{x}$ is a vector of exogenous variables, and $\boldsymbol{\zeta}$ is a vector of disturbance terms. In structural equation modeling, the structural parameters of interest are $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is the covariance matrix of the disturbance terms.

As noted above, equation (7) represents the structural form of the model. The specification of fixed or freed elements in $\mathbf{B}$ and/or $\boldsymbol{\Gamma}$ denotes a priori restrictions, presumably reflecting an underlying hypothesis regarding the mechanism that yields values of $\mathbf{y}$. The standard approach to structural equation modeling requires that certain assumptions be met for application of standard estimation procedures such as maximum likelihood. Specifically, it is generally assumed that the conditional distribution of the endogenous variables, given the exogenous variables, is multivariate normally distributed. Violations of this assumption can, in principle, be addressed via alternative estimation methods that explicitly capture the nonnormality of the data, such as Browne's asymptotic distribution-free estimator (Browne, 1984) or Muthén's weighted least squares estimator for categorical data (Muthén, 1984). If this or other assumptions are violated, then the standard likelihood ratio chi-square test, estimates, and standard errors will be incorrect. A fuller discussion of the assumptions of structural equation modeling can be found in Kaplan (2000).

With regard to the assumption of exogeneity, a perusal of extant textbooks and substantive literature on structural equation modeling suggests that the exogeneity of the predictor variables, as defined above, is not formally addressed—an exception being Kaplan (2000). Indeed, the extant literature reveals that only theoretical considerations are given when delimiting a variable as "exogenous."[5] Assessing exogeneity in terms of the statistical structure of the data requires that we revisit the reduced-form specification of a structural equation model.

22.3.6. The Reduced-Form Specification Revisited

In classic econometric treatments of structural equation modeling, the reduced form plays a central role in establishing the identification of structural parameters. The reduced-form specification of a structural model is derived from rewriting the structural form so that the endogenous variables are on one side of the equation, and the exogenous variables are on the other side. Specifically, considering equation (7), we have

$$\begin{aligned}\mathbf{y} &= \boldsymbol{\alpha} + \mathbf{B}\mathbf{y} + \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta},\\ &= (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\alpha} + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\mathbf{x} + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\zeta},\\ &= \boldsymbol{\Pi}_0 + \boldsymbol{\Pi}_1\mathbf{x} + \boldsymbol{\zeta}^*, \qquad (8)\end{aligned}$$

where it is assumed that $(\mathbf{I} - \mathbf{B})$ is non-singular. In equation (8), $\boldsymbol{\Pi}_0$ is the vector of reduced-form intercepts, $\boldsymbol{\Pi}_1$ is the matrix of reduced-form slope coefficients, and $\boldsymbol{\zeta}^*$ is the vector of reduced-form disturbances, where $\mathrm{Var}(\boldsymbol{\zeta}^*) = \boldsymbol{\Psi}^*$. Establishing the identification of the structural parameters requires determining if they can be solved uniquely from the reduced-form parameters (Fisher, 1966). An inspection of equation (8) reveals that the reduced form is nothing more than the multivariate general linear model. From here, equation (8) can be used to assess weak exogeneity. Specifically, from the context of the reduced form of the model, the parameters of the conditional model are $\boldsymbol{\omega}_1 \equiv (\boldsymbol{\Pi}_0, \boldsymbol{\Pi}_1, \boldsymbol{\Psi}^*)$, and the parameters of the marginal model are $\boldsymbol{\omega}_2 \equiv (\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_x)$, where $\boldsymbol{\mu}_x$ is the mean vector of $\mathbf{x}$, and $\boldsymbol{\Sigma}_{\mathbf{x}}$ is the covariance matrix of $\mathbf{x}$.

22.4. Assessing Weak Exogeneity

Recall that weak exogeneity concerns the extent to which the parameters of the marginal distribution of the exogenous variables are related to the parameters of the conditional distribution. In this section we consider three inextricably related ways in which the assumption of weak exogeneity can be violated: (a) violation of the joint normality of variables; (b) violation of the linearity assumption; and (c) violation of the assumption of homoskedastic errors.

22.4.1. Assessing Joint Normality

For simplicity, consider once again the simple linear regression model discussed in Section 22.2. It is known that within the class of elliptically symmetric multivariate distributions, the bivariate normal distribution possesses a conditional variance (*skedasticity*) that can be shown not to depend on the exogenous variables (Spanos, 1999). To see this, consider the bivariate normal distribution for two random variables y and x.

5. See, for example, Bollen's (1989) definition discussed earlier.

The conditional and marginal densities of the bivariate normal distribution can be written respectively as

$$\begin{aligned}(y|x) &\cong N((\beta_0 + \beta_1 x), \sigma_u^2),\\ x &\cong N[\mu_x, \sigma_x^2],\\ \beta_0 &= \mu_y - \beta_1 \mu_x, \quad \beta_1 = \frac{\sigma_{xy}}{\sigma_x^2},\\ \sigma_u^2 &= \sigma_y^2 - \left(\frac{\sigma_{xy}}{\sigma_x^2}\right)^2, \end{aligned} \tag{9}$$

where $\beta_0 + \beta_1 \mu_x$ is the conditional mean of y given x, σ_u^2 is the conditional variance of y given x, μ_x is the marginal mean of x, and σ_x^2 is the marginal variance of x. Let

$$\begin{aligned}\boldsymbol{\theta} &= (\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_{xy}),\\ \boldsymbol{\omega}_1 &= (\beta_0, \beta_1, \sigma_u^2),\\ \boldsymbol{\omega}_2 &= (\mu_x, \sigma_x^2). \end{aligned} \tag{10}$$

Note that for the bivariate normal distribution (and, by extension, the multivariate normal distribution), x is weakly exogenous for the estimation of the parameters in $\boldsymbol{\omega}_1$ because the parameters of the marginal distribution contained in the set $\boldsymbol{\omega}_2$ do not appear in the set of the parameters for the conditional distribution $\boldsymbol{\omega}_1$. In other words, the choice of values of the parameters in $\boldsymbol{\omega}_2$ does not restrict in any way the range of values that the parameters in $\boldsymbol{\omega}_1$ can take.

The bivariate normal distribution, as noted above, belongs to the class of elliptically symmetric distributions. Other distributions in this family include the Student's t, the logistic, and the Pearson Type III distributions. To demonstrate the problem with violating the assumption of bivariate normality, we can consider the case in which the joint distribution can be characterized by a bivariate Student's t-distribution (i.e., symmetric but leptokurtic). The conditional and marginal densities under the bivariate Student's t can be written as (see Spanos, 1999)

$$\begin{aligned}(y|x) \cong St\Bigg(&(\beta_0 + \beta_1 x),\\ &\frac{\nu \sigma_u^2}{\nu - 1}\left\{1 + \frac{1}{\nu \sigma_x^2}[x - \mu_x]^2\right\}\nu + 1\Bigg),\\ x \cong St[&\mu_x, \sigma_x^2; \nu], \end{aligned} \tag{11}$$

where ν are the degrees of freedom. Let

$$\begin{aligned}\boldsymbol{\theta} &= (\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_{xy}),\\ \boldsymbol{\omega}_1 &= (\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_u^2),\\ \boldsymbol{\omega}_2 &= (\mu_x, \sigma_x^2). \end{aligned} \tag{12}$$

Notice that the parameters of the marginal distribution $\boldsymbol{\omega}_2$ appear with the parameters of conditional distributions $\boldsymbol{\omega}_1$. Thus, by definition, x is not weakly exogenous for the estimation of the parameters in $\boldsymbol{\omega}_1$.

From this discussion, it is clear that one simple test of exogeneity is to assess the assumption of joint normality of y and x by using, say, Mardia's coefficient of multivariate skewness and kurtosis (Mardia, 1970). If the joint distribution is something other than normal, then parameter estimation must occur under the correct distributional form, and hence proper inferences may require estimation of the parameters of the marginal distribution as well as the conditional distribution. Because it is probably the case that joint normality does not hold in practice, this last point is extremely critical for the standard approach to statistical modeling in the behavioral sciences and will be taken up in more detail in Section 22.7.

22.4.2. Assessing the Assumption of Linearity

Joint normality of y and x is clearly central to establishing weak exogeneity. A consequence of the joint normality assumption is that the regression function $E(y|x, \theta) = \beta_0 + \beta_1' x$ is linear in x (Spanos, 1986). This follows from two properties of the normal distribution: (a) that a linear transformation of a normally distributed random variable is normal and (b) that a subset of normally distributed random variables is normal (Spanos, 1986). Therefore, deviations from linearity indirectly point to violations of normality and hence to violations of the weak exogeneity of x. Nonlinear relationships that cannot be transformed into linear relationships through well-behaved transformations will result in biased and inconsistent estimates of the parameters of the regression model. Assessing linearity can be accomplished through informal inspection of plots or more formally by using Kolmogorov-Gabor polynomials or the RESET method, both described in Spanos (1986). Should linearity be rejected, it may be possible to address the problem through normalizing transformations on y and/or x.

22.4.3. Assessing the Assumption of Homoskedastic Errors

The assumption of the joint normality of y and x also implies the assumption of homoskedastic errors. This is because, from the properties of the normal distribution, the conditional variance (skedasticity) function $\text{Var}(y|x) = \sigma_y^2 - \sigma_{xy}^2/\sigma_x^2$ is free of x, where σ_{xy}^2 is the squared covariance of y and x. Thus,

heteroskedasticity calls into question the assumption of weak exogeneity of x because it implies a relationship between the parameters of the marginal distribution and the conditional distribution. In addition, ordinary least squares estimation that ignores heteroskedasticity will result in unbiased but inefficient estimates of the regression coefficients. Most software packages contain easy-to-use options for obtaining residual scatter plots to assess the assumption of homoskedasticity. A direct test of the hypothesis of homoskedasticity was proposed by White (1980) and is available in many statistical software packages. Assessing the assumption of homoskedasticity in the context of structural equation modeling and multilevel modeling introduces additional complexities that will be addressed in Section 22.7.

22.5. Granger Noncausality and Strong Exogeneity

Our discussion of weak exogeneity in Section 22.3 did not specify a temporal structure for the data. Although the concept of weak exogeneity can be motivated by using models with lagged variables (Ericsson, 1994), it is not necessary to do so. The concept of weak exogeneity is applicable to cross-sectional data as well as to temporal data. However, to introduce the concepts of Granger noncausality and strong exogeneity, we must expand our models to account for the dynamic structure of the phenomenon under study. These extensions have important consequences for the statistical analysis of panel data when one wishes to properly model dynamic relationships and to use these models for forecasting or prediction.

To begin, consider an extension of our substantive problem of estimating the relationship between reading proficiency and parental involvement in reading activities. Let $\mathbf{z}_t$ be the vector of variables y_t and x_t. The basic problem now is that there is a dependence of current values of $\mathbf{z}$ on past values of $\mathbf{z}$, denoted as $\mathbf{z}_{t-1}$ with elements y_{t-1} and x_{t-1}. Therefore, the decomposition in equation (5) is no longer valid given the true dynamic structure of the process. Instead, we now need to condition on the past history of the process—namely,

$$f(\mathbf{z}_t|\mathbf{z}_{t-1}; \boldsymbol{\Theta}). \tag{13}$$

The conditioning in equation (13) leads to a decomposition represented as a first-order vector autoregressive model of the form

$$\mathbf{z}_t = \boldsymbol{\pi}\mathbf{z}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{14}$$

from which it follows that

$$y_t = \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 y_{t-1} + u_t, \tag{15}$$

$$x_t = \pi_1 x_{t-1} + \pi_2 y_{t-1} + v_t. \tag{16}$$

From our substantive perspective, equation (15) models current reading scores as a function of current and past parental involvement as well as past reading scores. Equation (16) models current parental involvement as a function of past parental involvement and past reading scores.

The above specification in equations (15) and (16) makes sense substantively insofar as feedback from previous reading scores might influence the amount of current parental involvement in reading activities. In other words, parents may notice improvement in their child's reading proficiency and feel reinforced for their reading activities. The question here, however, concerns whether parental involvement can be considered exogenous to reading proficiency and be used to predict future reading proficiency. In this case, we observe that weak exogeneity is not sufficient for the conditional model to be used to develop predictions of y because, as in our counterexample in Section 22.2, past values of y predict current values of x unless $\pi_2 = 0$. The condition that $\pi_2 = 0$ yields the condition of Granger noncausality (Granger, 1969). Granger noncausality essentially means that only lagged values of x enter into equation (15).

Weak exogeneity along with Granger noncausality yields the condition of *strong exogeneity*. The condition of strong exogeneity allows x_t (parental reading activities) to be treated as fixed at time t for the prediction of future values of y (reading proficiency) using the model in equation (15). Should Granger noncausality not hold (i.e., $\pi_2 \neq 0$), then valid prediction of future values of y would require the joint analysis of the conditional model in equation (15) and the marginal model in equation (16). In other words, the feedback inherent in the model when $\pi_2 \neq 0$ would have to be taken into account when interest centers on prediction.

22.5.1. Testing Strong Exogeneity and Granger Noncausality

Testing for strong exogeneity is relatively straightforward. First, it should be noted again that strong exogeneity requires weak exogeneity. Thus, if weak exogeneity does not hold, then neither does strong exogeneity. However, strong exogeneity also requires Granger noncausality. Thus, should y Granger cause x, then strong exogeneity does not hold. The simple

test for Granger noncausality is given in equation (16), where the null hypothesis of Granger noncausality is given by $\pi_2 = 0$.[6]

22.6. Super Exogeneity

An important application of statistical models in the social and behavioral sciences is in the evaluation of interventions or policies related to the exogenous variables. For example, consider the question of the relationship between per pupil class time spent using Internet technology and classroom-level academic achievement. If interest centers on achievement as a function of time spent using Internet technology, then it is assumed that the parameters of the achievement equation (the conditional model) are invariant to changes in the parameters of the marginal distribution of classroom Internet access time.

One set of policies related to classroom Internet connections and access time may have to do with the so called *e-rate*. The e-rate initiative was put forth during the Clinton administration as a means of providing discounted telecommunication services to schools and libraries. A specific goal of the program was to ameliorate the so-called "digital divide" that separates suburban middle- to upper-middle-class schools from lower-middle-class and inner-city poor schools with respect to access to technology in the classroom. Changes in e-rate policy should, if successful, induce shifts in the distribution of classroom Internet connections. The question is whether a shift in the parameters of the marginal distribution of classroom Internet connections changes the fundamental relationship between the number of classroom Internet connections and classroom achievement.

Formally, invariance concerns the extent to which the parameters of the conditional distribution do not change when there are changes in the parameters of the marginal distribution. As pointed out by Ericsson (1994), *invariance* is not to be confused with *variation freeness*, as discussed under the topic of weak exogeneity. Using the e-rate example, let $\boldsymbol{\omega}_1$ be the parameters of the conditional model describing the relationship between classroom achievement and time spent on classroom Internet activities, and let $\boldsymbol{\omega}_2$ be the parameters of the marginal distribution of time spent on Internet activities. Following Engle and Hendry (1993), assume for simplicity that two scalar parameters are related via the function

$$\omega_{1t} = \varphi\omega_{2t}, \tag{17}$$

where φ is an unknown scalar. Variation freeness suggests that over the period where ω_2 is constant, there is no information in ω_2 that is helpful in the estimation of ω_1. However, it can be seen that ω_1 is not invariant to changes in ω_2—that is, shifts in the parameters of the marginal distribution over some period of time lead to shifts in the parameters of the conditional distribution. By contrast, invariance implies that

$$\omega_1 = \varphi_t\omega_{2t}, \quad \forall t. \tag{18}$$

In terms of our substantive example, equation (18) implies that changes in the parameters of the marginal distribution of classroom time spent on Internet activities due to, say, e-rate policy changes do not change its relationship to academic achievement. Invariance of these parameters, combined with the assumption of weak exogeneity, yields the condition of *super exogeneity*.[7]

22.6.1. Testing Super Exogeneity

There are two common tests for super exogeneity (Ericsson, 1994), but note that super exogeneity also requires that the assumption of weak exogeneity holds. Thus, if weak exogeneity is shown not to hold, then super exogeneity is refuted. The first of the two common tests for super exogeneity is to establish the constancy of $\boldsymbol{\omega}_1$ (the parameters of the conditional model) and the nonconstancy of $\boldsymbol{\omega}_2$ (the parameters of the marginal model). Parameter constancy simply means that the parameters of interest take on the same value over time. *Parameter constancy* is to be contrasted with *invariance* as discussed above, which refers to parameters that do not change as a function of changes in a policy or changes due to interventions.

Continuing, if the parameters of the conditional model remain constant regardless of the nonconstancy of the parameters of the marginal model, then super exogeneity holds. Methods for establishing constancy have been given by Chow (1960). Briefly, the Chow test requires deciding on a possible breakpoint of interest over the period of the analysis based on substantive considerations. Once that breakpoint is decided, then

6. Clearly, this hypothesis will not hold exactly. Issues of power and the size of the alternative hypothesis $\pi_2 \neq 0$ become relevant as they pertain to the accuracy of forecasts when Granger noncausality does not hold.

7. Strong exogeneity is not a precondition for super exogeneity (see Hendry, 1995).

a regression model for the series prior to and after the breakpoint is specified. Let β_1 and β_1 and $\sigma^2_{u_1}$ and $\sigma^2_{u_2}$ be the regression coefficients and disturbance variances for the models before and after the breakpoint, respectively. The Chow test is essentially an F-type test of the form

$$CH = \left(\frac{RSS_T - RSS_1 - RSS_2}{RSS_1 + RSS_2}\right)\left(\frac{T - 2k}{k}\right), \quad (19)$$

where T is the number of time periods, k is the number of regressors, and RSS_T, RSS_1, and RSS_2 are the residual sum of squares for the total sample period, subperiod 1, and subperiod 2, respectively. The test in equation (19) can be used to test $H_0 : \beta_1 = \beta_2$ and $\sigma^2_{u_1} = \sigma^2_{u_2}$ and is distributed under H_0 as $CH \approx F(k, T - 2k)$. Limitations with the Chow test have been discussed in Spanos (1986).

The second test extends beyond the first in the following way. Here, the goal is to model the marginal process in such a way as to render it empirically constant over time (Ericsson, 1994). This can be accomplished by adding dummy variables that account for "seasonal" changes or interventions occurring over time in the marginal process. This exercise amounts to changing or intervening with the marginal process. Once these additional variables are shown to render the marginal model constant, they are then added to the conditional model. If the variables that rendered the marginal model constant are found to be nonsignificant in the conditional model, then this demonstrates the invariance of the conditional model to changes in the process of the marginal model (Engle & Hendry, 1993; Ericsson, 1994).

Returning to the e-rate example, consider the simple model that relates the number of Internet connections to academic achievement. Here we wish to test super exogeneity because we would like to use the measure of Internet connections as a policy variable for forecasting changes in academic achievement as a function of changes in the number of Internet connections over time. To begin, we must test for the weak exogeneity of the number of Internet connections because weak exogeneity is necessary for super exogeneity to hold. Next, we would use, for example, a Chow test to establish the constancy of the conditional model parameters of interest to the nonconstancy of the marginal parameters. This is then followed by developing a model for the change in the number of Internet connections over time, by adding variables that describe this change. These could be dummy variables that measure points in time in which the e-rate policy was enacted or other variables that would describe how the average number of Internet connections in the classroom would have changed over time. These variables are then added to the model relating achievement to the number of Internet connections. Should these new variables be nonsignificant in the conditional model, then this demonstrates how the parameters relating achievement to the number of Internet connections are invariant to changes in the parameters of the marginal model.

22.6.2. An Aside: Inverted Regression and Super Exogeneity

Consider the hypothetical situation in which an investigator wishes to regress science achievement scores on attitudes toward science, both measured on a sample of eighth-grade students using the model in equation (1). Suppose further that both sets of scores are reliable and valid and that, for the sake of this example, the attitude measure is super exogenous for the achievement equation. This implies that the measure of attitudes toward science satisfies the assumption of weak exogeneity and that the parameters of interest are constant and invariant to changes in the marginal distribution of attitudes toward science. Now, suppose the investigator wishes to change the question and estimate the regression of attitudes toward science on science achievement scores. In this case, it would be a simple matter of inverting the regression coefficient, obtaining $1/\beta$ as the inverted regression coefficient. The question is whether the inverted model still retains the property of super exogeneity.

To answer this question, we need to consider the density function for the inverted model. Following Ericsson (1994), let the bivariate density for the inverted regression model of two random variables x and y be defined as

$$\begin{aligned}(x_t|y_t) &\approx N[(c + \delta y_t, \tau^2)],\\ y_t &\approx N(\mu_y, \sigma^2_y),\end{aligned} \quad (20)$$

where $\delta = \sigma_{xy}/\sigma^2_y$, $c = \mu_x - \pi\mu_y$, and $\tau^2 = \sigma^2_x - \sigma^2_{xy}/\sigma_y$. The model in equation (20) can be expressed in model form as

$$\begin{aligned}x_t &= c + \delta y_t + \nu_{2t} \qquad \nu_{2t} \approx N(0, \tau^2),\\ y_t &= \mu_y + \varepsilon_{yt} \qquad \varepsilon_{yt} \approx N(0, \sigma^2_y),\end{aligned} \quad (21)$$

where the usual regression assumptions hold for this model. When equation (20) is written in line with the factorization of density functions, the result is the form

$$F(\mathbf{z}_t|\boldsymbol{\theta}) = F_{x|y}(x_t|y_t, \boldsymbol{\varphi}_1)F_y(y_t|\boldsymbol{\varphi}_2), \quad (22)$$

where $\boldsymbol{\varphi} \equiv (\varphi_1', \varphi_2') = h(\boldsymbol{\theta})$, a one-to-one function. To see the problem with inverted regression, we need to recognize that there is a one-to-one mapping between the parameters of the un-inverted model $\boldsymbol{\omega}$ from Section 22.3 and the inverted model. Specifically, we note that because $\beta = \sigma_{xy}/\sigma_x^2$ and $\sigma_u^2 = \sigma_y^2 - \sigma_{xy}^2/\sigma_x^2$, then after some algebra, it can be shown that

$$\delta = \frac{\beta\sigma_x^2}{\tau^2 + \beta^2\sigma_x^2}. \tag{23}$$

It can be seen from equation (23) that $\delta \neq 1/\beta$ unless $\sigma_u^2 = 0$. Moreover, from Ericsson (1994), we note that if x_t is super exogenous for β and σ_u^2, then even if β is constant, δ will vary due to variation in the marginal process of x_t via the parameter σ_x^2. In other words, super exogeneity is violated because the parameters of the inverted model are nonconstant even when the parameters of the uninverted model are constant (Ericsson, 1994, p. 18).

22.6.3. Super Exogeneity, the Lucas Critique, and Their Relevance for the Social and Behavioral Sciences

Super exogeneity plays an important philosophical role in economics and economic policy analysis. Specifically, super exogeneity protects economic policy analysis from the so-called "Lucas critique." It is beyond the scope of this chapter to delve into the history and details of the Lucas critique. Suffice to say that the Lucas critique concerns the use of econometric models for policy analysis because econometric models contain information that changes as a function of changes in the very phenomenon under study. The following quote of Lucas (1976) illustrates the problem:

> Given that the structure of an econometric model consists of optimal decision rules for economic agents, and that optimal decision rules vary systematically with changes in the structure of the series relevant to the decision maker, it follows that any change in policy will systematically alter the structure of econometric models. (quoted in Hendry, 1995, p. 529)

In other words, "a model cannot be used for policy if implementing the policy would change the model on which that policy was based, since then the outcome of the policy would not be what the model had predicted" (Hendry, 1995, p. 172).

The types of models considered in econometric policy analysis differ in important ways from those considered in the other social and behavioral sciences. For example, typical models used in, say, sociology or education do not consist of specific representations of the optimal decision behavior of "agents" and so do not lend themselves to the exact problem described by the Lucas critique. Also, models used in the social and behavioral sciences do not specify "technical" equations of the output of the system under investigation. Nevertheless, because the Lucas critique fundamentally suggests a denial of the property of invariance (Hendry, 1995), it may still be relevant to models used for policy analysis in domains other than economics. For instance, returning to the example of the e-rate and its role in educational achievement, the Lucas critique would claim that the parameters representing the relationship between Internet connections and educational achievement are not invariant to changes in the marginal process induced by the e-rate policy. However, tests of super exogeneity outlined above are tests of the Lucas critique, and so it is possible to empirically evaluate the seriousness of this problem for policy analysis.

22.7. Summary and Implications

Close examination of typical social and behavioral science definitions of exogenous variables shows that they are fraught with ambiguities. Yet, exogeneity is clearly of such vital importance to applied statistical modeling that a much more rigorous conceptualization of the problem is required, including guidance as to methods of testing exogeneity. The purpose of this chapter was to provide a didactic introduction to econometric notions of exogeneity, motivating these concepts from the standpoint of simple linear regression and its extension to structural equation modeling. The problem of exogeneity, as developed in the econometrics literature, provides a depth of conceptualization and rigor that is argued in this chapter to be of value to the other social and behavioral sciences.

To summarize, each form of exogeneity relates to a particular use of a statistical model. Table 22.1 reviews the different forms of exogeneity, their specific requirements, and informal tests. To review, weak exogeneity relates to the use of a model for purposes of inference. It concerns the extent to which the parameters of the marginal distribution of the exogenous variable can be ignored when focusing on the conditional distribution of the endogenous variable given the exogenous variable. Should weak exogeneity not hold, then estimation must account for both the marginal and conditional distributions. Strong exogeneity

Table 22.1 Summary of Different Forms of Exogeneity

Form of Exogeneity	*Implications for*	*Assumptions*	*Informal/Formal Tests*
Weak exogeneity	Inference	Multivariate normality of the joint distribution; homoskedasticity; linearity	Mardia's measures; homoskedasticity and linearity tests
Strong exogeneity	Forecasting and prediction	Weak exogeneity and Granger noncausality	Weak exogeneity tests; test of coefficient on lagged endogenous variable (see equation (16))
Super exogeneity	Policy analysis	Weak exogeneity, parameter constancy, and parameter invariance	Chow test; nonsignificance in conditional model of variables that describe policy changes in the marginal model

supplements the requirement of weak exogeneity with the notion of Granger noncausality so that exogenous variables can be treated as fixed for purposes of forecasting and prediction. Should Granger noncausality not hold, then prediction and forecasting must account for the dynamic structure underlying the exogenous variables. Super exogeneity requires weak exogeneity to hold and concerns the invariance of the parameters of the conditional distribution given real-world changes in the parameters of the marginal distribution. If an intervention or policy leads to changes in the distribution of the marginal process but does not change the relationship described by the conditional model, then the exogenous variable is super exogenous for policy or intervention analysis.

22.7.1. Implications for Standard Statistical Practice

The impact of the exogeneity assumption on standard statistical practice in the social and behavioral sciences is profound. To begin, it is clear that the exogeneity problem is not unique to linear regression and structural equation models. Indeed, the problem is present in all statistical models in which a distinction is made between exogenous and endogenous variables, resulting in a partitioning of the joint distribution into the conditional and marginal distributions.

It is worth considering briefly how the problem of exogeneity might arise in other statistical models. Here we consider *multilevel modeling* (including growth curve modeling), a methodology that is enjoying widespread popularity in the social and behavioral sciences (see, e.g., Raudenbush & Bryk, 2002). Multilevel modeling is a powerful analytic methodology for the study of hierarchically organized social systems such as schools or businesses. In education, for example, multilevel modeling has yielded a much greater understanding of the organizational structure of schools as they support student learning. In this methodology, the so-called "Level 1" variables constitute endogenous outcomes such as student achievement that can be modeled as a function of student-level exogenous variables. Parameters of the Level 1 model include the intercept and the slope(s) that are allowed to vary over so-called "Level 2" units such as classrooms. Classroom level variation in the Level 1 coefficients can be modeled as a function of classroom exogenous variables such as measures of the amount of teacher training in specific subject matter skills. Variation over Level 3 units such as schools is also possible, and school-level variables can be included to explain this component of variation.

Future research should examine the problem of exogeneity in multilevel models. Suffice to say here that exogeneity enters into multilevel models at each level of the system. Statistical theory underlying multilevel modeling shows that these models have built-in heteroskedasticity problems that are resolved by specialized estimation methods. Yet, what remains to be determined is if the parameters of interest in multilevel models can be shown to be variation free with respect to the parameters of the student-level and school-level exogenous variables. Because multilevel models are used to supplement important discussions

of education policy, assessing the weak exogeneity of policy-relevant variables is crucial.

A special case of multilevel modeling is *growth curve modeling,* a methodology that is also enjoying tremendous popularity in the social sciences and directly accounts for the dynamic features of panel data. In such models, the Level 1 endogenous variable is an outcome such as a reading proficiency score for a particular student measured over multiple occasions. This score is modeled as a function of a time dimension such as grade level, as well as possibly time-varying covariates such as parent involvement in reading activities. The parameters of the Level 1 model constitute the initial level and rate of change, and these are allowed to vary randomly over individuals, who are in turn modeled as a function of time-invariant exogenous variables such as race/ethnicity, gender, or perhaps experience in an early childhood intervention program. Variation in average initial level and rate of change can also be modeled as a function of Level 3 units such as classrooms or schools. The power of this methodology is that it allows one to study individual and group contributions to individual growth over time.

The problem of exogeneity enters growth curve models in a variety of ways. First, repeated measures on individuals can be a function of time-invariant variables. For example, in estimating growth in reading proficiency in the younger grades, time-invariant variables might include the IQ of the children (assumed to be stable over time), the income of the parents, and so on. Again, these variables are assumed to be exogenous.

Second, the repeated outcomes can be modeled as a function of time-varying covariates. Each time-varying covariate is presumed to be exogenous to its respective outcomes and is used to help explain, for example, seasonal trends in the data. However, time-varying variables can also be allowed to have a lagged effect on later outcomes. For example, a time-varying covariate such as parental reading activities at time t can be specified to influence reading achievement at time t as well as reading achievement at time $t + 1$. This represents the introduction of a lagged exogenous variable into the full-growth curve model, and so issues of strong exogeneity and Granger noncausality may be of relevance. In other words, the Level 1 model that characterizes achievement at time t as a function of time-varying covariates assumes that the time-varying covariate at time t is not a function of achievement at time $t - 1$. If this assumption does not hold, then the time-varying covariate is not strongly exogenous.

In addition to the fact that exogeneity represents an issue in a wide range of statistical models, it must also be recognized that most statistical software packages estimate the parameters of statistical models under the untested assumption that weak exogeneity holds. In other words, software packages that engage in *conditional estimation* (e.g., conditional maximum likelihood), conditional on the set of exogenous variables, do so assuming that there is no information in the marginal process that is relevant for the estimation of the conditional parameters. However, as noted above, weak exogeneity is only valid if the joint distribution of the variables is multivariate normal—a heroic assumption at best. Therefore, it is likely in practice that estimates derived under conditional estimation are incorrect. The only situation in which this is not a problem is in estimation of the Gauss linear model with nonstochastic regressors. Future research and software development should explore methods of estimation that account for the parameters of the marginal distribution along with the conditional distribution for a given specification of the form of the joint distribution of the data.

In the context of simple linear regression, informal testing of weak exogeneity via assessing joint normality and homoskedasticity is relatively straightforward. Indeed, most standard statistical software packages provide various direct and indirect tests of these assumptions. In the context of structural equation modeling, however, although considerable attention has been paid to the normality assumption (see, e.g., Kaplan, 2000, for a review), scant attention has been paid to assessing assumptions of linearity and homoskedasticity. This may be due to the fact that textbook treatments of structural equation modeling motivate the methodology from the viewpoint of the structural form of the model, and therefore it is not directly obvious how homoskedasticity could be assessed. However, if attention turns to the reduced form of the model as described in equation (8), then standard methods for assessing the normality assumption—including homoskedasticity and linearity—would be relatively easy to implement. Therefore, users of structural equation modeling should be encouraged to study plots and other diagnostics associated with the multivariate linear model to assess weak exogeneity.

The issue raised here is not so much how to assess weak exogeneity but rather how to proceed if the assumption of weak exogeneity does not hold. Recognition of the seriousness of the exogeneity assumption should lead to fruitful research that focuses on estimation methods under alternative specifications

of the joint distribution of the data. In attempting to characterize the joint distribution of the data, all means of data exploration should be encouraged. There should be no concern about "finding a model in the data" because the joint distribution of the data is theory free[8] (Spanos, 1986). Theory information only becomes a problem when there is a factoring of the joint distribution into the conditional and marginal distributions insofar as that is the point in the modeling process, in which a substantive distinction is made between endogenous and exogenous variables and where parameters of interest are defined (see Spanos, 1999).

The implications of the strong exogeneity assumption for statistical practice are relevant if models are used for prediction and forecasting. In this case, weak exogeneity is still a necessary requirement, but in addition, it is imperative that Granger noncausality be established. Similarly, implications of the super-exogeneity assumption are relevant when models are used for policy or intervention evaluations. Super exogeneity also forces us to consider the requirement of parameter constancy and invariance—issues that have not received as much attention in the social and behavioral sciences as they should. Focusing on parameter constancy and invariance also forces us to consider whether there exist invariants in social and behavioral processes. Moreover, as pointed out by Ericsson (1994), parameter constancy is a central assumption of most estimation methods and hence is of vital importance to statistics generally.

22.7.2. Concluding Remarks

Our discussion throughout this chapter leads to the recognition that *exogeneity* is an adjective describing an assumed characteristic of a variable that is being chosen for theoretical reasons to be an exogenous variable. Weak exogeneity is the necessary condition underlying all forms of exogeneity, and hence this assumption is fundamental and requires empirical confirmation to ensure valid inferences. Additional assumptions are required to yield valid predictions or evaluations of policies or interventions.

Exogeneity resides at the nexus of the actual data-generating process (DGP) and the statistical model used to understand that process. In the simplest terms, the actual DGP is the real-life mechanism that generated the observed data. It is the reference point for both the theory and the statistical model. In the former case, the theory is put forth to explain the reality under investigation—for example, the organizational structure of schooling that generates student achievement. In the latter case, the statistical model is designed to capture the statistical features of that aspect of the actual DGP that we choose to study and measure (Spanos, 1986; see also Kaplan, 2000).

In addition to the role that exogeneity plays with regard to fundamental distinctions between theory, the DGP, and statistical models, exogeneity raises a number of other important philosophical questions that are central to the practice of statistical modeling in the social and behavioral sciences. One issue, for example, concerns the proper place of data mining as a premodeling strategy. We find that when attention focuses on characterizing the joint distribution of the data, then data mining has a central role to play. Another issue arising from our study of exogeneity concerns the dynamic reality of the phenomenon under investigation. Granger noncausality and strong exogeneity force us to consider exogenous variables as possibly being responsive to their own dynamic structure and that this must be correctly modeled to obtain accurate estimates for prediction and forecasting. Super exogeneity reminds us that our models are sensitive to real-life changes in the process under investigation. Finally, serious consideration of the problem of exogeneity forces us to reexamine statistical textbooks in the social and behavioral sciences to clarify ambiguous concepts and historical developments. It is hoped that reflecting on the importance of the exogeneity assumption will lead to a critical assessment of the methods of statistical modeling in the social and behavioral sciences.

8. The exception being that theory enters into the choice of the variable set as well as methods of measurement. These issues are not trivial but are not central to our discussion of the role of theory as it pertains to the separation of variables into endogenous and exogenous variables.

REFERENCES

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

Browne, M. W. (1984). Asymptotic distribution free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62–83.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica, 28,* 591–605.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum.

Engle, R. F., & Hendry, D. F. (1993). Testing super exogeneity and invariance in regression models. *Journal of Econometrics, 56,* 119–139.

Ericsson, N. R. (1994). Testing exogeneity: An introduction. In N. R. Ericsson & J. S. Irons (Eds.), *Testing exogeneity* (pp. 3–38). Oxford, UK: Oxford University Press.

Ericsson, N. R., & Irons, J. S. (Eds.). (1994). *Testing exogeneity.* Oxford, UK: Oxford University Press.

Fisher, F. (1966). *The identification problem in econometrics.* New York: McGraw-Hill.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica, 37,* 424–438.

Hendry, D. F. (1995). *Dynamic econometrics.* Oxford, UK: Oxford University Press.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85–112). New York: Academic Press.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions.* Thousand Oaks, CA: Sage.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences.* Pacific Grove, CA: Brooks/Cole.

Legendre, A. M. (1805). *Nouvelles méthods pour la détermination des orbites des comètes* (New methods for determining the orbits of comets). Paris: Firmin Didot.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Journal of Monetary Economics, 1*(Suppl.), 19–46.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57,* 519–530.

Mulaik, S. A. (1985). Exploratory statistics and empiricism. *Philosophy of Science, 52,* 410–430.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115–132.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousands Oaks, CA: Sage.

Richard, J.-F. (1982). Exogeneity, causality, and structural invariance in econometric modeling. In G. C. Chow & P. Corsi (Eds.), *Evaluating the reliability of macro-economic models* (pp. 105–118). New York: John Wiley.

Spanos, A. (1986). *Statistical foundations of econometric modeling.* Cambridge, UK: Cambridge University Press.

Spanos, A. (1999). *Probability theory and statistical inference.* Cambridge, UK: Cambridge University Press.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48,* 817–838.

Wonnacott, R. J., & Wonnacott, T. H. (1979). *Econometrics* (2nd ed.). New York: John Wiley.

Yule, G. U. (1897). On the theory of correlation. *Journal of the Royal Statistical Society, 60,* 812–854.

Chapter 23

Objectivity in Science and Structural Equation Modeling

Stanley A. Mulaik

23.1. Introduction

Objectivity is a core concept of science. To show what it means, how that comes to be, and how it plays out in science in general and in methodological practices such as structural equation modeling, in particular, is the aim of this chapter. *Objectivity* is the noun that refers to the state of being objective. *Objective,* in turn, is simply an adjective formed from the noun *object* with the suffix *-ive,* meaning "of or pertaining to." So *objectivity* has something to do with objects. More specifically, *Merriam-Webster's Collegiate Dictionary* (1993) defines *objective* as "of, relating to, or being an object, phenomenon, or condition in the realm of sensible experience independent of individual thought and perceptible by all observers : having reality independent of the mind." It further lists an additional related meaning, which has methodological implications: "expressing or dealing with facts or conditions as perceived without distortion by personal feelings, prejudices, or interpretations." In this respect, *objective* is often contrasted with *subjective,* which *Webster's* cites as "relating to or being experience or knowledge as conditioned by personal mental characteristics or states . . . , peculiar to a particular individual . . . , modified or affected by personal views, experience, or background." So, *subject* and *object* are often viewed as inextricably linked in a relation of dialectical opposition, each to the other. *Objective* is frequently identified with knowledge of what is real and "external," independent of the mind or the observer. *Subjective* is identified with distortions in knowledge that are produced by and perhaps unique to the knower, the knower's perspective, thought processes, methods of observation, or motives. Illusions are subjective interpretations of what is presented in external reality. Another aspect of objectivity is that it has a social component, *inter-subjectivity,* that is, agreement between observers as to what is perceived. In some accounts of objectivity, agreement is the only basis for objectivity, and so objectivity is but a social concept and has no psychological basis. Other accounts stress certain perceptual features of objectivity, the perception of invariants across different points of view.

23.2. Early Developments in the Concept of Objectivity

Now these concepts of objectivity did not arise all at once. And many were not directly connected

to a concept of an object but stood on their own. Inter-subjectivity, for example, is a principle that is seen in the requirement of the Paris Académie des Sciences of 1699 that experiments be performed before an assembled company or, at the minimum, several academicians (Daston, 1994).

Inter-subjectivity may already have been a principle recognized in English law in the 17th century, when ordinary citizens were given the right to trial by a jury of peers. Replicability of experiments was also automatically required by these scientific academies and societies but often was difficult to achieve (Daston, 1994). Because bitter personal rivalries and bickering became rampant among 17th-century English and French scientists, threatening the ability of these societies to function, rules of decorum and impersonality and impartiality were imposed on their members (Daston, 1994). But again, no conscious linkage of this requirement to a concept of an object is made.

In fact, knowledge of objects given by the senses was suspect and subject to illusions in the view of the 17th-century French scientist, mathematician, and philosopher René Decartes (1596–1650). It was always logically possible, Descartes (1637/1901, 1641/1901) held, that what appears to him (and others) is an illusory reality constructed by some evil demon. So he sought certain knowledge in what he could clearly and distinctly apprehend in immediate intuition without doubt. This led him to propose a method for finding certain and incorrigible knowledge in philosophy and reliable knowledge in matters of experience. His study of geometry had led him to the method that the ancient Greek geometer Pappas recommended for the solution of problems in geometry: the method of analysis and synthesis. *Analysis* meant separating or breaking up a whole into its parts, whereas *synthesis* meant combining parts or elements into a whole. Descartes saw the method as basic for solving any problem: First break the problem down (analysis) into its component truths and ideas, breaking these down further, if need be, into more and more elementary truths until one arrives at fundamental, elemental truths. Then reverse the process by bringing together (synthesis) the various component truths until one completely reconstructs the thing to be understood or proven while apprehending at each step how the components are combined to achieve larger and larger wholes. He believed that the mind functions in terms of these two principles. He called the mental activity of analysis *intuition,* for it sought to visualize the components of a thing or problem in terms of clear and distinct ideas that were self-evidently true. The activity of synthesis was *deduction*—not strictly or necessarily deduction, as in syllogistic logic, but a leading of the mind from elementary truths to compositions of them in greater wholes. The body of an animal may be understood by dissecting it first into its component organs, which one would see clearly and distinctly, and then seeing how they are combined together into the body as a whole and how they might function with respect to one another. The other aspect of his method was procedural doubt: Doubt everything and anything until you know intuitively that something is self-evidently true. By his method in philosophy, he claimed to have identified certain innate ideas that were indubitable and not derived from experience, such as causality, the infinite, negation, and number. Decartes is known as one of the first "foundationalists" in modern philosophy. He sought incorrigible truths and knowledge by seeking to base knowledge on an undoubtable self-evident foundation, such as what he could directly and immediately discern both clearly and distinctly as true and indubitable. Reason could then proceed from indubitable first principles. Rationalism was born. However, Descartes's method of solving problems by analysis and synthesis became, in some form, a fundamental methodological principle in all philosophy, science, and intellectual discourse that was to follow (Mulaik, 1987; Schouls, 1980). But having introduced the idea that what is given to the knower by the senses may be illusory and not backed up by any reality, he made it possible for British philosophers to conceive of a mind that knows only its own thoughts or the sense impressions of which it is directly aware.

John Locke (1632–1704) was the first of these British philosophers, who were known as the British empiricists. Influenced by Descartes's method of analysis and synthesis, Locke (1694/1962) rejected Descartes's idea of innate ideas and sought to ground certain knowledge on that which is given immediately and directly in experience or in reflection. The mind is like a blank slate (a *tabula rasa*) on which experience is written. All ideas arise in experience. This led to an analysis of experience into fundamental, simple ideas, such as cold and warmth, hardness and softness, solidity, space, figure, and motion. These are known as being clear and distinct from one another. He believed that an external reality caused these to appear to the mind. The order, frequency, and manner in which simple ideas were given to the mind either via the senses or reflection determined how they were combined (synthesized) into complex ideas. External reality drove the formation of complex ideas from simple ideas. He critically examined the concept of substance, traditionally regarded as that to which properties such as color and weight adhered, declaring that

a substance is but a certain complex of simple ideas coexisting together, but nothing stands under them as substance per se.

Already Locke laid the groundwork for the skeptical empiricists who were to follow, who were to be forever skeptical of things such as objects, substances, necessary relations of cause and effect, external reality, and even the self.

George Berkeley (1685–1753) rejected (Berkeley, 1710/1901, 1713/1901) the need to postulate an external reality behind the sense impressions as superfluous. The only reality was the mind and its contents. But he continued the empiricist program begun by Locke. The problem for empiricism was to account for how the mind constructed complex ideas from the simple ideas of experience. This was accomplished by postulating the existence of the associative processes of the mind.

At the hands of David Hume (1711–1776), empiricism was pushed to its ultimate logical limits. Hume argued (1739/1968, 1748/1977) that the mind experiences lively and vivid simple perceptual impressions such as colors and sounds arriving in certain spatial and temporal configurations or order. These, in turn, are registered as simple ideas, which are fainter and less vivid. The mind was driven, he said, by the impressions given to it to synthesize complex ideas from simple ideas by means of the associative processes of (a) resemblance, (b) contiguity, and (c) cause and effect. Similar sets of impressions tended to be joined into kinds. Similar impressions that co-occurred in the same spatial configurations contiguous to one another became our ideas of certain kinds of things in space. The regular succession of certain kinds of impressions gave rise to the ideas of cause and effect. But echoing Locke's skeptical analysis of substance, Hume said that however much he sought to know by direct experience what connects these impressions into kinds, objects, substances, and causes and effects, he could not detect anything. There was nothing in experience (the only reality) behind an object or a causal connection other than a habitual expectation of a contiguous collage of impressions or a regular succession of them. There was no logical necessity for the regular succession or the contiguous collage of impressions to occur in the future either, for he could conceive logically of their not occurring. So, empiricism that had developed the idea of reasoning by induction—that is, generalizing from particulars of experience—could reach no necessary and incorrigible conclusions from experience. There were no necessary connections. In fact, when introspecting his own mind, Hume said that all he ever encountered were the impressions of his senses and ideas, but no self, no knower that possesses them. With that, Hume dispensed with a mind that thinks or contains the impressions and ideas, as well as the idea of an external reality with necessary causal connections. British empiricism had collapsed into an absurd skepticism.

23.3. Kant Formulates the Modern Conception of Objectivity

It is against the backdrop of Descartes's rationalism and British empiricism's rejection, among other things, of necessary ideas, external objects, substance, causal connections, and the self that Immanuel Kant (1724–1804) developed his critical philosophy and a new conception of objectivity grounded in the judgments of objects (Kant, 1787/1996). He accepted the fact that rationalism's attempt to understand the world deductively from self-evident, innate ideas had failed. On the other hand, although empiricism was able to generate new knowledge via experience, the justification of that knowledge was scandalized by the skeptical conclusions that seemed to be an inevitable consequence of its assumptions. Kant accepted the legitimacy of concepts such as substance, identity, cause and effect, and number as not derived from experience. These were a priori categories, that is, not derived a posteriori from experience. They were the forms by which experience was synthesized in the mind. But unlike the associative processes of British empiricism, the mind's ordering and organizing of material from the senses was spontaneous and not determined by the senses. Echoing Aristotle, Kant argued that the mind provided the forms a priori and the senses the matter or substance a posteriori of experience. Without the senses, no object would be given to the mind, and without the a priori categories of the mind, no object could be thought. "Thoughts without content are empty; intuitions without concepts are blind" (Kant, 1787/1996, A52, B76). The problem, however, was to justify the use of the categories in the face of Humean skepticism. Kant rejected, however, Locke's attempt to give them legitimacy by tracing them "physiologically" in experience to external things so that they were properties of things as the things are in themselves.

23.4. Deductions of Legitimacy

The question of legitimacy is never something that is resolved simply on the basis of experience. This is especially so in this case because the argument is

Table 23.1 Kant's Table of Categories

1. Quantity	*2. Quality*	*3. Relation*	*4. Modality*
Unity	Reality	Inherence and subsistence (*substantia et accidens*)	Possibility-Impossibility
Plurality	Negation	Causality and dependence (cause and effect)	Existence-Nonexistence
Allness (totality)	Limitation	Community (interaction between agent and patient)	Necessity-Contingency

that these are concepts that do not originate in what is given to the senses. Furthermore, legitimacy concerns a norm and a community or sovereign that grants the norm. Their legitimacy, therefore, demanded a different "deduction," but Kant did not mean a syllogistic argument. He referred to deductions as forms of legal argument to establish rights. A deduction in the law courts of his time was not a syllogistic argument but a laying out of the legal basis by which a right can be acquired via various intermediary transfers of that right, originating with the sovereign power to grant that right (Henrich, 1989). The facts of the case, which depended on experience, only concerned the who, what, where, and when of how the right was claimed to have been acquired. These were not presented in or as a part of the deduction but separately. And who was the sovereign to grant the rights of legitimacy to these concepts that do not originate in experience—what Kant called a priori or *transcendental* concepts? It would be the intellectual community—although Kant does not explicitly say this, which is why his deductions seem obscure. Kant's transcendental deductions of the a priori categories follow the strategy of clearly describing how they are used throughout the thought of objects but are limited just to the function of providing forms to experience in thought and are unable to provide knowledge by logical deductions independent of experience, as the rationalists believed. Hence the title of his great book, the *Critique of Pure Reason.* Kant shows the a priori categories to be an indispensable part of reasoning about objects of experience but, at the same time, limited to this function and incapable of giving incorrigible and certain knowledge independent of experience and only corrigible and provisional knowledge from experience. As a result, the controversial role played by the a priori ideas or categories in deductive metaphysical speculation is denied and the skeptic's hostility to them defused. Their legitimacy is thus self-evident because they are part of everyone's thought, including the skeptic's, and the community has a right to sanction the forms of thought acceptable to it because these are the forms of thought of everyone (see Mulaik, 1994a, 1994b).

23.4.1. Kant's A Priori Categories

Turning now to Kant's conception of object, we will focus briefly on those categories that he declared were intrinsically involved in the thinking of objects and that are currently useful in elucidating a contemporary conception of objectivity. I will use them to illustrate the ideas of synthesis. Kant provided a table of categories that he said were root categories of that aspect of the mind involved in discursive thought, the understanding. These categories were a priori categories of pure synthesis for putting presentations to the mind given in sensible intuition into combinations. I reproduce a variant of this table in Table 23.1.

Each of the entries in Table 23.1 represents a form of synthesis. Unfortunately, Kant sidestepped offering detailed definitions of each, saying that to do so would divert from his purpose. However, Kant notes that they were not gathered together haphazardly but developed systematically.

One is to note, he said, that the four classes could be joined into two divisions:

> The concepts in the first two classes [quantity and quality] are directed to objects of intuition both pure [mental] and empirical; those in the second division [relation and modality] are directed to the existence of these objects (these objects being referred either to each other or to the understanding). (Kant, 1787/1996, B110)

Those concepts in the first division, furthermore, are not listed with "correlates" (i.e., associated concepts), whereas those in the second are. But a most intriguing observation is why the categories under each class come in threes. The answer is that, given the first category in each class, it must be accompanied by a category that represents its contradiction or something that contrasts with it, which is the second category in

the class. The third category, then, is a concept that is a synthesis of the first two categories.

> This fact, however, must by no means lead us to think that the third category is a mere derivative concept, rather than a root concept, of pure understanding. For combining the first and second, categories, in order to produce the third concept, requires that the understanding perform a special act that is not the same as the act it performs in the case of the first and second concepts. (Kant, 1787/1996, CPR B111)

This is the schema made famous by Hegel and later Marx of first a thesis, then an antithesis, followed by a synthesis of the first two.

23.4.1.1. Class of Quantity

To elucidate, consider that under *quantity,* we first find *unity,* which represents the compression or synthesis of numerous things in perception into a unity, to be treated by the mind as a unitary concept and counted as one. For example, numerous observations from different points of view may be synthesized conceptually into observations of a single thing. In contrast with the operation of unity, one may only be able to synthesize observations into a number of distinct unities, and hence one has not one unity but a plurality. This corresponds to seeing things clearly and *distinctly* as different from one another but principally in terms of their quantity. But then, by a third movement of the mind, one may take a plurality of synthesized unities and think of them as a totality of unities, as an exhaustive set, a new unity.

23.4.1.2. Class of Quality

Again, under *quality,* we may make a judgment that what we intuit is real. Its contradiction or negation is that something is not real. Now a synthesis of these two concepts, which contains them both, would be the concept of limitation. Something is limited in its reality as sensed, in space, in time, or both.

23.4.1.3. Class of Modality

The next two classes of categories concern the existence of objects, either as they are with respect to one another or with respect to their consideration in the understanding. I will skip over the categories of relation for the moment to deal more thoroughly with them after we consider modality. Modality concerns judgments of existence as entertained in thought. We may consider that the existence of a thing is possible or impossible. That may contrast with a definite judgment that the thing actually exists or does not exist, which rules out the tentativeness of considering the existence as possible or impossible. The concept of something's necessary existence is that its existence is guaranteed by its possibility.

23.4.2. Class of Relation

23.4.2.1. Inherence

Let us turn now to the third class, *relation.* This is of considerable interest to us in scientific work and statistics. The first category is inherence. The meaning of the term may not be immediately recognized, but the concept is something we deal with every day, and the operation is enshrined in syntax of language. *Webster's* defines *inherence* as the relationship of a quality to an object or substance (*Merriam-Webster's Collegiate Dictionary,* 1993). That is likely still a bit abstract. But it is a relationship that you use all the time. If you say, "Jane Doe is 62 inches tall," you are using the relation of inherence: Jane Doe is the object, and she *has* the quality of being 62 inches tall. By saying that she *has* this quality, we join the quality to Jane Doe. Inherence is the relationship that connects qualities or attributes to things. Our language uses this relationship all the time in forming noun phrases and verbal phrases (see Figure 23.1).

In "The green giant lifts the little boy," *green* is an attribute of *giant, little* is an attribute of *boy,* and *lifts* is a verb that modifies the noun phrase *green giant* in a sentence, whereas *little boy* is a noun phrase object of the verb *lifts* that qualifies *lifts.* A sentence is just a synthesis, a putting together of noun phrases and verbal phrases. Each of the inverted Vs represents a putting together, a synthesis, some of which bring about a relation of inherence of an object (subject) with a quality. The synthesis of *green/giant, little/boy,* and *giant/lifts* shows cases of inherence. Of course, adjectives represent currently static qualities of the object, whereas verbs represent currently dynamic qualities of the object.

If we describe someone as having the attribute of "blonde," we do not refer to something that is constantly changing before our eyes. To be "blonde" cannot be something varying over the time of observation; otherwise, what will we attribute to the object for that time? Someone can be blonde today and brunette tomorrow if that person dyes his or her hair. So an attribute is assigned on a given occasion or instant.

Figure 23.1 A Sentence Diagram Illustrating the Joining of Predicates to Things

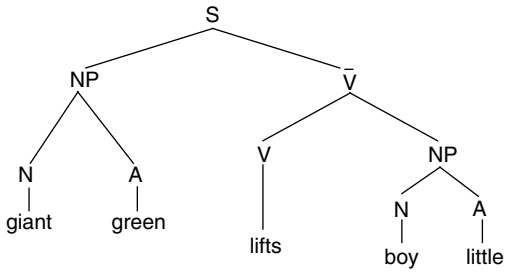
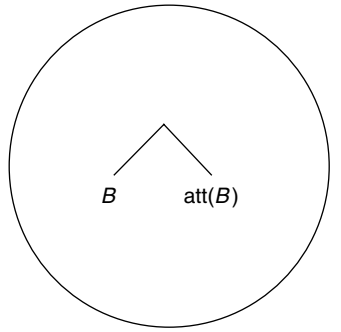


Figure 23.2 Schematic Diagram of the Inherence of Attribute att(B) in Object B



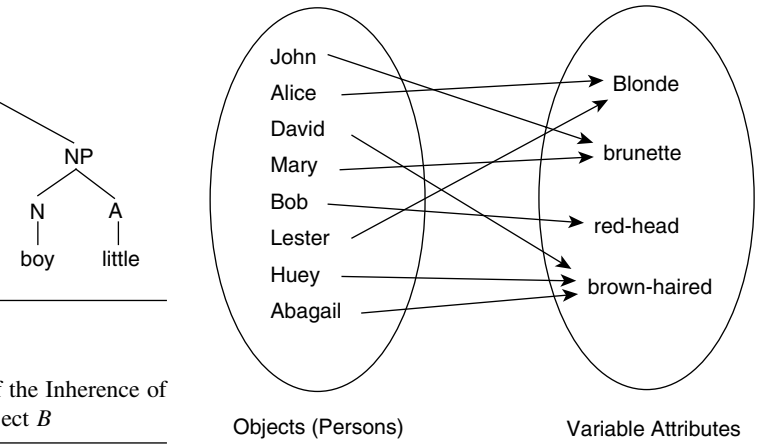
Some of these attributes endure over long durations and become almost "essential" to the person.

We illustrate schematically the concept of inherence in Figure 23.2. Here we have an object B and an attribute att(B). They are joined (synthesized) so that B has the attribute att(B), and the synthesis is represented by the inverted V in the diagram.

23.4.2.2. Inherence and Variables

Let us digress from Kant for a moment to consider the implications of inherence for the formation of variables. The idea of a variable is built from the idea of inherence. For example, at any given time, the attribute of "hair color" will vary *across individuals,* with "blonde" being just one of the *values* of this variable, along with "brunette," "redhead," "brown haired," and so on. *A variable is a set of values, only to one of which can an object ever be assigned at any one time.* ("Blonde," "brunette," "redhead," and "brown haired" are all values of hair color, and we only assign

Figure 23.3 Mapping of a Set of Persons to a Set Containing the Values of a Variable



a person to one of these at any one time in saying what color hair the person has.) "Value" designates an attribute in a set of attributes that define a variable. So, what we are describing is a mapping from a set of objects (in this case, persons) to a set of attribute-values (hair color).

In Figure 23.3, each arrow represents an inherence relation. What makes the set of attributes a variable is the constraint that each object can be assigned to only one value of the set of attribute-values. This constraint does not prevent an object from being assigned to different attribute-values on more than one variable. But the constraint allows us to sort out what attributes go together in a set as values of a variable: They are attributes only to one of which an object may be assigned at any one time. Now, some variables have just a finite number of possible values, such as "sex," with "male," "female," "undecided" as its values. Other variables may have a countably infinite number of values (representing a discrete variable), whereas others may have an uncountably infinite number of possible values (like continuous quantities). But two variables can have the same set of possible values and still be distinct because of the manner in which their values are assigned across a collection of individual objects (persons). This leads to what some philosophers call an *extensional definition of a variable* as the set of assignments of objects to its values. A variable is defined by how its values are extended across a set of objects. Two variables with the same set of possible values differ if, for the same set of objects, the objects are assigned to the values of the variables in different ways. So, two variables, A and B,

Figure 23.4 Spreadsheet Showing Assignment of Values of Variables to Persons

	name	weight	height	iq	gpa	sex
1	Jane Doe	110.00	62.00	142.00	3.80	2
2	Bill Smi	145.00	72.00	121.00	2.87	1
3	Bob Jone	185.00	73.00	110.00	1.79	1
4	Mary Pew	115.00	64.00	105.00	1.90	2
5	Jack Cle	195.00	69.00	121.00	2.55	1
6	David In	155.00	69.00	131.00	3.70	1
7						

representing responses to different questions, may take on the possible values of {1, 2, 3, 4, 5}, representing responses on a 5-point rating scale, but they differ if variable *A* (response to Question A) is assigned different values than variable *B* (response to Question B) across a set of rating subjects:

	A	*B*
John	1	3
Alice	2	5
Mary	3	1
Bob	4	2
Lester	5	4

In this case, the act of checking off a response to a rating scale is the act of assigning a value of the variable to the subject. John checks 1 on variable *A*, but he checks 3 on variable *B*. Alice checks 2 on variable *A*, but she checks 5 on variable *B*, and so on.

In multivariate statistics, we assume we have observations of the values of several variables for each subject. A convenient way to represent observations and the values of variables is in a spreadsheet format (see Figure 23.4).

The rows of the table in Figure 23.4 correspond to subjects (objects) and the columns the variables. Along any given row, we may read off the values that the subject has on each respective variable.

Now, multivariate statistics studies relationships between many variables. For example, we may compute either the covariance or correlation between a pair of variables. We may seek the mean of each variable and place them in a row at the bottom of the table. We can also compute the variance of scores within each column and put each of the variances in a row at the bottom of the table. We can test whether different groups of subjects have the same means on several variables. We can compute a regression equation for predicting one variable from a number of other variables. We can also compute linear combinations of variables and get correlations between these, as in canonical correlation.

In statistics, the variables are presumed to be at least numeric, and often when they are not, they can be coded numerically and preferably should represent quantities. Quantities are things such as the number of answers correct in a test of many questions, temperature, weight, height, and length. Sometimes, we like to think that we can measure psychological attributes quantitatively, as with an IQ score, a score to measure how strong a person's preference is toward something, or a score to represent a measure of how much a subject knows about something. But establishing that one truly has a quantity requires satisfying certain axioms (Michell, 1990), which is not easy to do with psychological variables. But we will not go into that here.

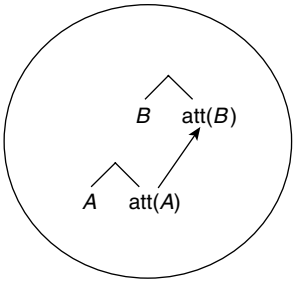
A big problem for graduate students is to learn to think of what quantity a variable measures. It should be thought of as unidimensional. Very often, the literature of social psychology and industrial psychology names variables in ways that make it very difficult to understand what precisely is the quantity measured. Often, what is put forth as a variable is not a variable but a process, a relationship, a substance, or something that would require several variables to describe. The theorist needs to be more specific in identifying the quantity that varies. Recently, I came across a study in which the researcher wanted to measure "leader-follower exchange." What is the quantity implied by that name? I asked. There are numerous variables one could focus on in exchanges between leaders and followers: *How many* orders does the leader give to

the follower? To *what degree* does the subordinate feel he or she has freedom to decide what to do? To *what extent* does the subordinate say negative things about his or her supervisor? To *what extent* does the subordinate say negative things about the boss's orders directly to the boss? To what extent does the follower like the leader? (Notice the use of "to what degree," "to what extent," and "how many" in these sentences. They force you to think in terms of unidimensional quantities when thinking of the variable measured.) When you learn to cut through the obscurity of social psychological and industrial psychological jargon to focus on specific quantities, you will be in a better position to think of what may be causes of these quantities, for the causes will also have to be represented by variables. This will help you better to select possible predictors of quantitative dependent variables. If I had my way, I would never let a student use a word or phrase to name a theoretical variable that does not include, at the beginning of the phrase, one of the following phrases: "the extent to which . . ," "the number of . . .," "the quantity of . . .," "the degree to which . . .," "the amount of . . .," "the score on . . .," or a similar phrase.

23.4.2.3. Cause and Effect

Returning now to Kant's class of categories of relation, we see that his second category under this class is causality and dependence. It is important not to lose sight of its connection with the first category of inherence. The second category must contradict or introduce something that contrasts with the first category. The way we will do it to achieve the concept of causation is to consider a first object with its inherent attribute and introduce another object with its inherent attribute and say that the first object's attribute is conditioned on, dependent on, or determined by the second object's attribute. In Figure 23.5, we schematically represent a cause-effect relationship as Kant envisaged it between the attributes of two objects, A and B. The attribute att(A) of object A is shown to be a cause of the attribute att(B) of object B by use of an arrow to show causal connection and direction. (Inverted Vs still represent relations of inherence.) Now it is quite possible for objects A and B to be the same object so that one of the attributes of the object determines another attribute of the object. Kant's conception of causality explicitly requires that causes always be thought of in connection with objects, for it is the objects that bear the attributes that are causally connected. An implication of this in scientific studies is that if one has a conception of a causal connection between variables, one must always apply it to objects that conform to that conception. We study the causal connection across a collection of objects (research subjects) and, in that way, see how different values of the causal variable tied to different objects determine different values of the effect variable. But we must have reason to believe that all of the objects (subjects) studied are homogeneous in some way in their attributes, conforming to the same functional relation between the causal and the effect variables. Otherwise, if attributes are connected differently across different pairs of objects—that is, by different functional relations, say—then one does not have the same causal connection between each pair of objects, and the functional relation that one finds may not really exist among the objects. This is the problem of selecting objects to study that are causally homogeneous.

Figure 23.5 Schematic Diagram Illustrating That an Attribute att(A) of Object A Is a Cause of the Attribute att(B) of Object B



Kant's category of *causation* lays down a common feature found in a variety of conceptions of causation. Causal relations concern how the attribute of an object is conditioned on, determined by, or dependent on some other attribute, often in another object. Lakoff and Johnson (1999) argue that causation is a radial concept, a concept with a central core but numerous divergent features that serve to make a variety of causal concepts. Kant's category can be seen as a bare-bones schema that would fit well as the common core of most conceptions of causation. But causation has been thought of in numerous ways, according to Lakoff and Johnson. It has been thought of as a force that changes the attributes of an object, as a making of things, as a progeneration whereby what one is depends on one's ancestors, as a necessary temporal precedence, as functional relationship, as a determination of the probability with which values of a variable occur (Mulaik, 1995), as reasons and explanations

for things, or sometimes as just a correlation between things (Pearson, 1911). Many of these different ways of thinking of causation involve metaphors, all of which have the core structure seen in Kant's category of causation.

23.4.2.4. Community

Kant's category of *community* concerns visualizing a community or collection of objects whose attributes are reciprocally determined. One moves up from the level of just considering a causal connection that goes in just one direction from one object to another to consider a whole system of objects whose attributes are mutually and reciprocally determined by all the objects in the system.

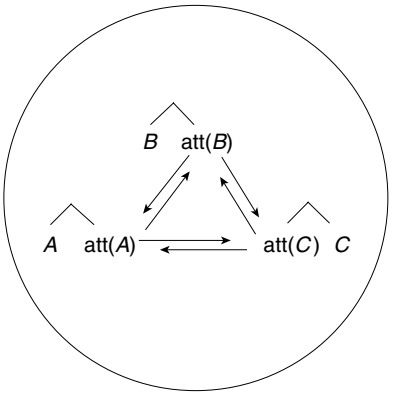
We illustrate schematically the idea of community in Figure 23.6.

The concept of community was a conceptual advance because it introduced the notion of a system as a whole. To see the implications of this concept, consider that behaviorism posed the problem of understanding behavior as that of establishing a causal relation from the environment to the organism. Such a focus may be too narrow if the organism brings distinct properties that moderate its behavior uniquely, which then changes the environment, which in turn modifies subsequent behavior as well, which in turn changes the environment again, ad infinitum. Organism and environment need to be studied as a reciprocally interacting system.

The categories of relation are perhaps the most comprehensible and still useful contribution of Kant's account of the categories. They provide a perspicuous framework for understanding the conceptual interconnections among the relations involving objects and their attributes as used in science. From these concepts, one can construct the abstract idea of a variable and a conceptual blend involving objects and variable attributes by a mapping of objects to values of variables, and from there, one can consider functional relations between the properties of objects as causal relations. For a version of causality as a functional relation applied to probabilistic causality, see Mulaik (1986).

Kant also claimed that one could obtain derivative synthetic a priori concepts by combining the various categories with themselves and/or with a priori modes of the sensibility, but he did not illustrate this claim in the *Critique of Pure Reason.* But he envisioned that complex a priori concepts could be generated in this way. Thus, the material of sensible intuition could be organized and synthesized in complex ways in thought.

Figure 23.6 Schematic Representation of Kant's Conception of Community, Which Concerns How the Attributes of a Community of Objects, A, B, and C, Are Reciprocally Determined



23.4.3. Concept of Object Unifies According to a Rule

So much now for Kant's categories. We need to consider how additional ideas of Kant further shaped the concept of objectivity, for they greatly influenced the elaboration of that concept among German philosophers and scientists in the 19th century (Megill, 1994). In this same period, 19th-century British empiricists generally had great difficulty in making sense of Kant and his notions of objectivity and subjectivity, for they were inclined to believe that all knowledge is basically subjective to begin with and that objects were only habitually anticipated configurations of sense impressions. This tended to be true as British empiricism evolved into logical positivism and logical empiricism in the first 60 years of the 20th century. In the meantime, workbench scientists were developing their own methods and ideas of objectivity.

Kant, as has been said, believed that the categories were combined further into complex concepts for synthesizing the information acquired from the senses. He argued that indeed what is given to us by the senses in "appearances" is subjective. Although we may want to think of an external object "x" behind the appearances, it is nothing to us but just our idea of an object stripped of all its attributes (i.e., an a priori idea). What is something to us are the appearances,

and it is by means of the a priori syntheses of the appearances that we must fabricate objects for us. And the first rule of a concept of an object is that it is a concept that unites synthetically (through the knower's active thought) many appearances according to some particular prior rule of synthesis. The rule must be presumed to be universally valid and necessary. (Not that empirical concepts will turn out to be universally valid and necessary but that objective concepts be asserted as such.) Kant, however, is exceedingly frustrating because just as one thinks that he is about to state the idea of asserting and testing a hypothesis to establish the objectivity of the hypothesis, against future data, he says nothing of this. For Kant, it is sufficient that the manifold appearances are united according to a prior rule. The idea of testing hypotheses probably occurred to various scientists prior to Kant, such as Spallanzani, who conducted controlled experiments to test and refute the hypothesis of spontaneous generation in the latter third of the 1700s, but the idea did not enter into a theory of ideal scientific method until a 19th-century British Kantian and historian of science, William Whewell (1966), introduced the idea of *consilience*—that one could evaluate an induced theory by showing that it is supported by other data not used in its formation.

23.4.4. Inter-subjectivity

However, Kant introduces one last element of the idea of objectivity toward the end of the *Critique of Pure Reason.* If a judgment of an object is found only in a single subject, then that is only "persuasion" and is an illusion, for it has only private validity. "Truth, however," Kant said,

> rests on agreement with the object; consequently, in regard to the object the judgments of every understanding [across a number of individuals] must be in agreement. . . . Thus, whether assent is conviction [of the truth] or mere persuasion, its touchstone externally is the possibility of communicating the assent and of finding it to be valid for every human being's reason. For then there is at least a presumption that the agreement of all the judgments, despite the difference among the subjects, will rest on the common basis, viz., the object, and that hence the judgments will all agree with the object and will thereby prove the truth of the [joint] judgment. (Kant, 1787/1996, A820 821, B848–849)

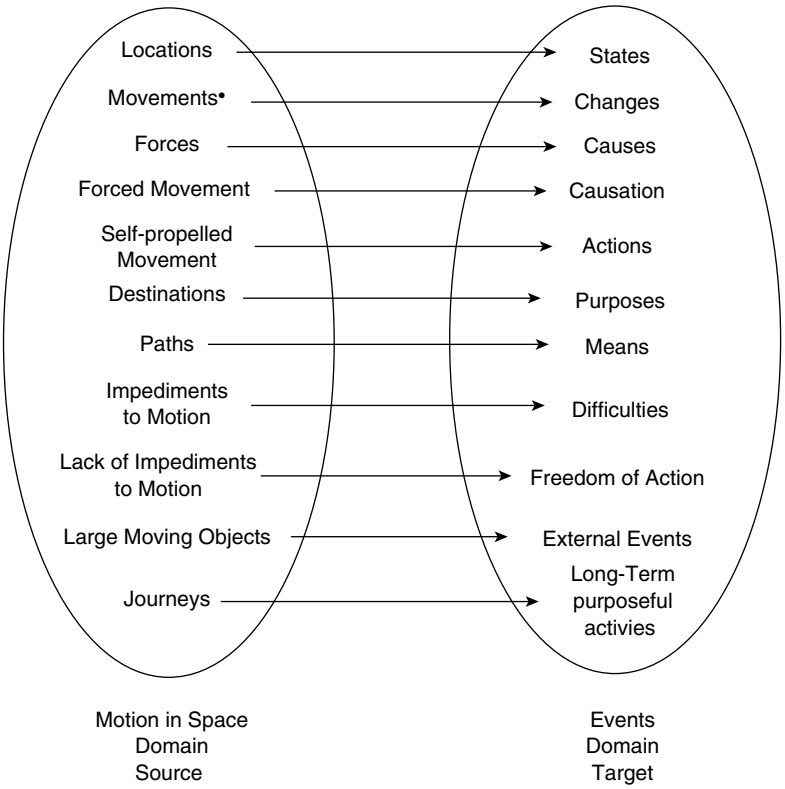
However, he claims that inter-subjective agreement is not an absolute determination of truth but simply a way of detecting possible private validity. He presumes that all humans have the same cognitive apparatus that functions with the same a priori categories and that this makes agreement possible.

23.5. The Cognitive Science of Objectivity

Philosophers are interested in rationally and critically understanding the nature of the world, how we humans know what we know, and how we are to live the good life. Historically, they have tended to ground their reasonings on what to them appeared to be self-evident truths and to avoid empirical studies or borrowing too heavily from the sciences. To do so, they believed, would in many cases rest their deliberations on the presumptions of scientists, which the philosophers felt should be subject to critical philosophical analysis and not the foundation of such an analysis. But in the latter half of the 20th century, philosophers tended increasingly to question the idea of incorrigible foundations for thought and knowledge and to rest content to work with sound but fallible knowledge given by the sciences. Because thinking is itself a central object of philosophical thought, it is understandable, then, that philosophers turned to psychologists, neuroscientists, linguists, and cognitive scientists to help them better understand the nature of thinking. So, we saw philosophers of science, such as Ronald Giere (1988), advocating a cognitive approach to understanding the way scientists think and do science. But in the succeeding decade, there suddenly burst upon the scene a group of linguists, literary theorists, and philosophers with training in mathematics and cognitive science who claimed to also be cognitive scientists, saying that they are now doing not only the cognitive science of ethics (Johnson, 1993), the cognitive science of mathematics (Lakoff & Nuñez, 2000), the cognitive science of language (Fauconnier, 1994, 1997; Fauconnier & Turner, 2002; Lakoff, 1987), the cognitive science of literature and poetry (Lakoff & Turner, 1989; Turner, 1996), the cognitive science of politics (Lakoff, 1996), and the cognitive science of social science (Turner, 2001) but also the cognitive science of philosophy itself (Lakoff & Johnson, 1999). And they challenged fundamental assumptions and beliefs in each of these fields. They have not yet written the book on the cognitive science of science, but it is something we can expect, and ultimately it seems that it is inevitable that there will be a cognitive science of cognitive science.

The basic message of this new school is that cognitive science has achieved a level of sophistication and

Figure 23.7 The Location-Event Structure Metaphor as a Mapping



a set of analytic tools that allows us now to use the methods and findings from that science to illuminate the thinking processes of human beings in each of these basic fields of human activity. And what they find is that many of the theories and beliefs about the way people think that are held by those working in those fields are wrong. For example, many scientists believe that most scientific and mathematical thought is literal and that metaphor is simply a linguistic thing involving flowery language and only for poets and writers. Wrong. Most thinking, especially abstract thinking, even in the sciences, is metaphoric, whereas literal thinking is confined to the concrete, immediate, here and now (Lakoff & Johnson, 1999). Furthermore, metaphor is not about language—although it is shown in language—but about thinking. Another belief is that thought is disembodied, computational, symbolic, and formal. Wrong. Thought is embodied.

> Conceptual structure arises from our sensorimotor experience and the neural structures that give rise to it. The very notion of "structure" in our conceptual system is characterized by such things as image schemas and motor schemas. . . . Mental structures are intrinsically meaningful by virtue of their connections to our bodies and our embodied experience. They cannot be characterized adequately by meaningless symbols. . . . Our brains are structured so as to project activation patterns from sensorimotor areas to higher cortical areas. These constitute what we have called *primary metaphors.* Projections of this kind allow us to conceptualize abstract concepts on the basis of inferential patterns used in sensorimotor processes that are directly tied to the body. (Lakoff & Johnson, 1999, p. 77)

23.5.1. Metaphor

Metaphor is a basic analytic concept used by Lakoff to understand people's reasoning. He and his students and colleagues have cataloged hundreds of metaphors, many of which repeatedly turn up in different situations. Lakoff and Johnson (1999) define *metaphor* as a (partial) mapping from a source domain to a target domain so that inference patterns of the source domain may be applied to the target domain. For example, a commonly used metaphoric structure is the location-event structure metaphor used to think about events and causes. The source domain for the

metaphor is the domain of motion of objects in space. The target domain is the domain of events. A number of metaphoric concepts follow from this metaphor.

For example, states are treated as bounded locations in space: "I'm *in* a funk." "He's *on the edge* of disaster." "We are not *out of* danger." The prepositions are spatial indications of location with respect to the location in question, which represents a state.

Lakoff and Johnson (1999) argue that most metaphors are taken from the image and motor schemas of embodied perception and action. Being in and out of a region is a primary body experience that we easily visualize. So, abstract states such as a "funk," "disaster," and "danger" can be represented by the spatial metaphor of a location, and one's being in such a state can be thought of as being in the location.

Changes are movements: "I'm *going into* a depression." "It *went* from hot to cold in an hour." "I think we are *moving closer* to success."

Means are paths: "If you follow these rules, you will be on the road to success."

Purposive action is movement along a path toward a goal: "Let's get this show on the road!" "We've been working very hard *toward* bringing you the products you need." "We are *moving ahead* with our program in this *way* and expect *to reach* our goals within a week."

Difficulties are impediments to movement: "We've gotten *mired down* in details in working toward our goal."

Stopping purposive action is blocking movement along the path to the goal: "If it were not for the accountants who blocked access to the funds, we would have been able to reach our quota." "We need to close off all avenues to their getting their funds and weapons."

Long-term activities are journeys: "A love relationship is a journey through life." "We've reached a crossroad in our marriage." "It's been a long, bumpy road, but our marriage has survived."

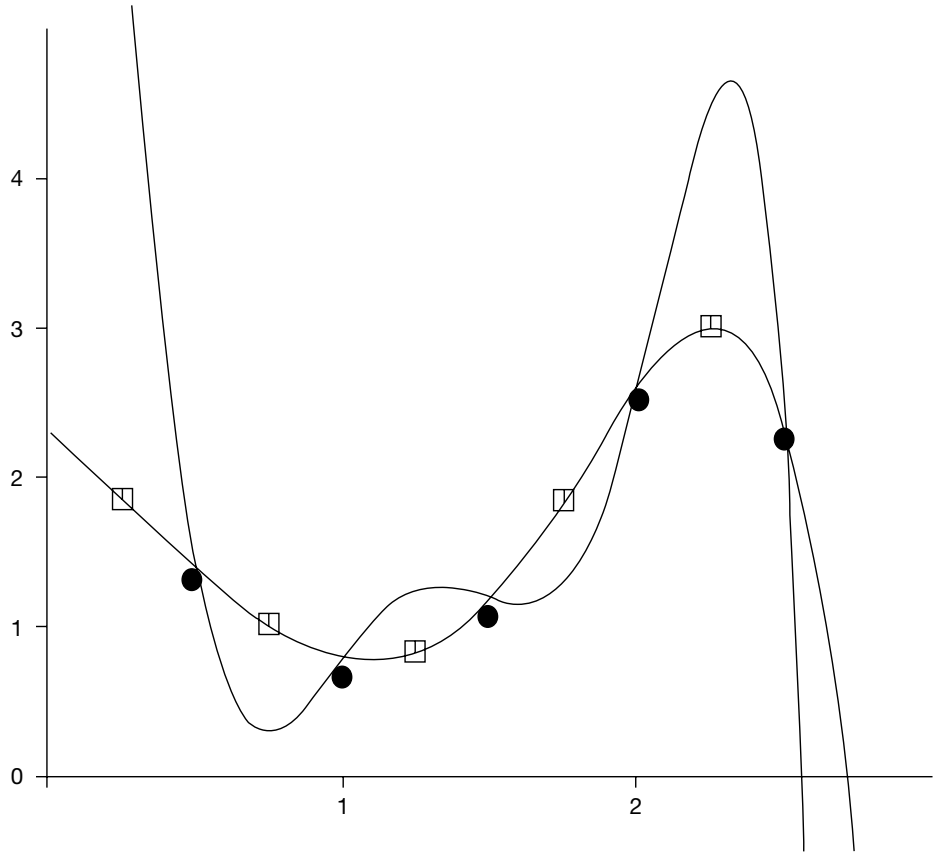
Much of Lakoff and Johnson's (1999) analyses are deconstructive but not destructive. By exposing the metaphoric structures in metaphysical, mathematical, and political thought, they expose the conceptualization to critique. The critique is not that the concepts are merely metaphor. Metaphor is unavoidable in abstract thought. Rather, metaphors for understanding something are not unique—often more than one metaphor might be applicable and necessary to achieve full understanding—and furthermore may not represent important aspects of the target domain so that erroneous inferences may be made.

23.5.2. Conceptual Blending

Lakoff and Johnson (1999) work with exposing the metaphors in everyday and abstract thought, whereas Fauconnier and Turner (2002) study more complex forms of what they call "conceptual blending." Metaphor is an intermediate form of conceptual blending. By mapping elements of a source domain to a target domain, the metaphor allows one to transfer inference patterns from the source domain to the corresponding elements of the target domain. The result is that the target domain becomes a new blended domain operating with the selected inferences from the source domain. But Fauconnier and Turner consider that one may have two input domains—which they call mental spaces—and select certain elements from each of the two domains to include in a new domain. Inference patterns from the original input spaces may also be transferred to the new domain to operate on the corresponding elements. Metaphoric mappings from select elements in each domain to some in the other may also bring some of the elements from one domain under the inference patterns of the other domain. Emergent structure will arise with new rules to allow the elements from the two inputs to function together. For example, blends occur throughout mathematics. A vector space is a blend of elements and axioms of an abelian group with elements and axioms of a field, together with emergent structure in new axioms to govern how the elements of the group function with elements of the field. Additional elaborations of the vector-space blend can involve additional axioms for some of the elements, such as axioms for a scalar product to be applied to the elements of the group, which results in a unitary vector space.

Fauconnier and Turner's (2002) method is to analyze a conceptual blend into its input spaces and then reverse the process to show how the blend was achieved. (This is analysis and synthesis again.) They have done this with numerous cases and have developed a standard way of diagramming the mental spaces, mapping aspects of each to the other spaces, and creating a formation of the blends to make the process perspicuous. However, it would divert us from our purpose to develop all of that here. It will be important, in any case, to recognize that metaphor and conceptual blending are all forms of cognitive synthesis. And what these cognitive scientists are doing is reminiscent of Kant's program to reveal the synthetic operations by which the mind thinks. In fact, some of their blending operations have counterparts in Kant's categories. But whereas Kant located his a priori operations of synthesis in the understanding, the faculty of discursive

Figure 23.8 Graph Showing Two Distinct Curves That Pass Through the Same Data Points



NOTE: Two or more curves can be constructed to fit a given set of data points (black dots). The curves unite the points according to rules but are not unique. The squares represent additional data not used in fitting the curves and may be used to test whether any of the curves that fit to the original points also fit the new points.

thought, Lakoff and Johnson (1999) and Fauconnier and Turner (2002) locate (relatively) a priori structures in embodied perception and motor activity.

23.5.3. An Object as a Concept That Unites Observations According to a Rule

We now need to return to the development of the concept of objectivity where Kant left it. But we will seek to use some of the methods developed by Lakoff and Johnson (1999) and Fauconnier and Turner (2002) to show how objectivity arises out of schemas of object perception. Recall how, for Kant, an objective concept was a concept that united numerous intuitions or percepts into a unity according to a rule. But that rule, as we shall demonstrate, may not be the only one that will do this. Consider the graph in Figure 23.8, in which the plotted round dots correspond to a set of data.

Now each of the curves in the diagram was constructed using a sixth-degree polynomial equation, with different constraints on the coefficients of the equation introduced to identify a solution. As many parameters as points to fit were freed and estimated in such a way as to make the curves each fit the data points. In other words, each curve represented a different saturated model. In fact, an endless number of sixth-degree polynomial curves can be found to fit these five points. If we regard each curve as a way of uniting the points according to a rule, we see that there is no unique rule by which this could be done. If each of several persons has a different rule with different but saturated equations and different sets of just-identifying constraints, the rules are no longer

objective but subjective by being linked to particular persons with corresponding constraints.

But suppose we obtain additional data points that are not used in formulating the curves but that we believe represent data generated by the same process. If we plot these points as squares in the diagram, we can see if either of the curves we originally formulated fits these additional points as well. We began with 5 points and estimated five parameters, ending up with both curves fitting the points perfectly, of necessity, because the points were used in determining the curves. With 5 additional points, however, we now have 10 points in all and use only 5 of them to formulate a curve. But we have 5 additional points against which to test our curves. We see, in this case, that one of the curves in Figure 23.8 also fits the additional points, whereas the other curve does not. It would seem that with more and more additional points fitting one of the curves, the support for one of the curves against that of the other becomes overwhelming. However, it is quite possible that neither of the curves will fit the additional points, but we now have a paradigm case for establishing objective concepts.

23.5.4. Objectivity Implies Hypothesis Testing

Kant's idea that we regard a rule (e.g., a curve) that unites diverse intuitions (think "data points") as universally valid implies that additional intuitions (percepts) assumed to be produced by the same object should conform to the same rule also. It is not that they necessarily will. No inductive generalizations based on experience necessarily hold against additional experience. But thinking by concepts is thinking by means of necessary relations. Seeking to reconcile with experience the necessity with which we think of things occurring according to rules is what establishing objective knowledge is about. This is the basis for hypothesis testing. A hypothesis asserts a universal, invariant rule, and the rule is tested by seeing if it unites or conforms to data not used in the formulation of the rule. If the rule is upheld by the test, that gives a provisional objectivity to the rule. The rule is provisional because additional data presumed under the same rule may not conform to the rule. Concepts of experience thus are "objectively valid." They do not possess "absolute truth." Objectivity is not about absolute truth. It is about a way in which we validate concepts against experience, and the validity is only provisional. Much confusion and metaphysical debate arises out of thinking that theories and hypotheses are "true" in an absolute sense. One is projecting "truth" from formal logic and mathematics onto experience because we use logic and mathematics to reason about experience. But mathematics and logic are but metaphors for dealing with apparent regularities in experience. Treating experience with formal logic and mathematics is a conceptual blend, with the inference patterns of logic and mathematics projected onto our experiences. Or one may imagine as did the philosopher-mathematician Peirce that science at any point may have only provisional knowledge but is still converging to a final truth in the distant future as an absolute limit by means of self-corrections (Peirce, 1931–1958). But again, this projects a concept of a final limit from mathematics onto experience, and yet we have no necessary reason to believe that science will not go on forever revising its results with newer and more encompassing concepts as humans encounter ever newer experiences.

23.5.5. The Metaphor of Science as Knowledge of Objects

But we still have not established why Kant's conception of objective knowledge as knowledge that integrates experience according to rules is so intuitively compelling. To provide a reason, I will now draw upon the cognitive science of objectivity. I have already mentioned that Lakoff, Johnson, Fauconnier, and Turner argue that most metaphors are taken from the primary experiences of embodied perception and action. So, what we need is to provide objectivity with a metaphor for objective knowledge in science that is intuitively compelling because it arises out of embodied object perception. "Science is the knowledge of objects" is the metaphor.

Now, some might object to saying that this is a metaphor because historically science has, at least in the physical sciences, always been about objects. But if we speak here only about objects that we directly perceive, that is false. Electrons, photons, and quarks of physical science are physical objects, but they are not directly perceived but are conceived by integrating a great many perceptual experiences. And turning to the social and behavioral sciences, we have numerous concepts that we treat as objective that do not refer to things directly perceived but are conceptual integrations of many observations (e.g., "inner locus of control," "intelligence," and "extroversion"). So, to ascribe objective status to these concepts, we do not rely just on directly perceiving these things but on using these concepts to unite in thought numerous percepts, often reconstructed from memory. But we

achieve objective status for them, in part, by hypothesis testing.

The account here of objectivity as a metaphor taken from object perception was originally developed in Mulaik (1995) and draws heavily on it: Most of the members of the conceptual integration school posit a philosophy of "embodied realism," which is that humans experience things as embodied beings. As embodied beings, they move around in a world of other bodies and perceive the world directly as consisting of extended, textured, surfaced objects and substances. They react to the world in ways that the human species has evolved to effectively, if not optimally, survive. They are able to directly perceive things like animals, from the size of mites to the size of mammoths, but they cannot directly perceive microbes or viruses, nor can they directly perceive a mountain as a whole, except at a distance, and cannot perceive planets around distant stars, which are only points of light in the night sky, with the naked eye. They do not perceive the earth as round. Nor do they perceive their own perceptual and cognitive processes directly. Much of what they perceive and think involves neural syntheses and integrations that occur quite outside of conscious awareness. So, human cognition is limited and at a scale somewhat comparable to human size. Observational aides, such as microscopes and telescopes, only serve to bring things up to human scale. And what are seen through such devices are often described metaphorically in terms of things familiar to humans at the human scale, such as Leeuwenhoek's characterization of protozoa and microbes, seen through his water-drop microscope, as "animalcules" (little animals) or Schleiden and Schwann's "cells" (small rooms or enclosures), seen through the microscope in the tissues of plants and animals.

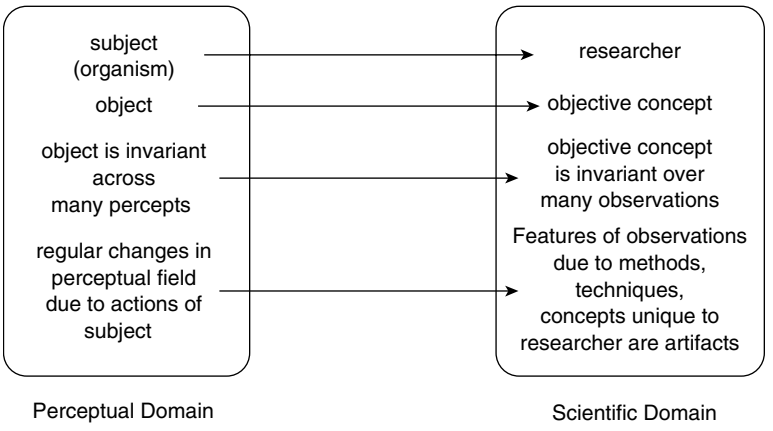
23.5.5.1. Gibson's Theory of Perception

It is natural, then, that Lakoff (1987) recognizes that many of his views about embodied realism and cognition go together well with the views of J. J. Gibson (1966, 1979, 1982), the psychologist of perception who argued that humans directly perceive the world as extended, textured, surfaced objects as opposed to constructing them in conscious awareness from more fundamental categories in awareness. Now, a fundamental concept of Gibson is that we only perceive when the stimulus information presented varies. Our perceptual apparatus is tuned to resonate to variation in the stimulus information, and with no variation, there is no perception. So, surfaces are detected by variations in the texture conveyed by the stimulus information. Object perception also depends in part on a moving observer, so that it perceives a continually changing flow and movement in the perceptual array resulting from its movements. Objects and surfaces are "invariants," in contrast to the changing "perceptual flow" observed by the organism as it actively moves about within its environment. According to Mace's (1986) account of Gibson's theory, these invariants correspond to "stable features of the environment" (p. 144), which are simply objects. But of utmost importance to my account was Gibson's idea that perception is not simply about detecting invariant objects in the external environment but rather also about detecting the effects of its own actions and movements in the environment because of regular forms of change in the perceptual array that result from bodily movement, which vary independently of fixed objects. Turning one's head or moving to other positions changes the perceptual array in regular ways, such as objects moving unchanged to the side and even out of vision as one turns one's head, whereas looming out occurs with points on the surfaces of things moving along radiating lines from a point as the observer moves toward it or points on the surfaces of things converging to a point on the horizon as the observer moves away from it. Moving to a position that presents a different angle of view of an object changes the appearance of the object, but the appearance will nevertheless retain certain topologically invariant features, such as the order of spatially contiguous points on the surface, even if perceived distances between points change. Thus, perception could be subdivided into *exteroception,* the perception of invariant objects in the perceptual field, and *proprioception,* the perception of regular changes in the perceptual field produced by the observer's actions and movements, which provides the observer with information about itself.

Turner (1996) has a similar account:

> Suppose we see a baby shaking a rattle. Sequentially we can focus on the smile, the nose, the jerky movement of the shoulder, the frozen elbow, the hand, the rattle. Our focus changes, but we feel that, regardless, we continue to look at the *same story:* The child is playing with the rattle. We are able to unify all of these perceptions, all of these different foci. The mental spaces corresponding to the different foci will all have a child, a rattle, a rattling motion, and so on, and we connect these elements in each space to their counterparts in other spaces. We conceive of these various spaces as all attached to a single story. (p. 117)

Figure 23.9 Metaphoric Mapping of Subject-Object Schema From Perceptual to Scientific Domain



Next he says,

> Now imagine that we walk around to the other side of the baby. Our visual experience may change substantially. It is even possible that we will see none of what we saw before, strictly speaking. Yet our new view will not seem entirely new. The space of the new viewpoint will have a baby, a shoulder, a hand, a rattle, a rattling motion, and so on, and we will connect these elements to their counterparts in the spaces of other viewpoints and other foci, allowing us to think of the different small spatial stories we see as *one story,* viewed from different viewpoints and with different foci. (p. 117)

What we achieve is a *conceptual blend* of all these viewpoints and foci, with their linked corresponding elements providing invariant features of objects.

23.5.5.2. The Subject-Object Schema of Perception as Metaphor

The division of perception into exteroception and proprioception, which seems present from birth and is a constant feature of perception, provides the fundamental schema that I shall call the *subject-object* schema. It is this schema that we use metaphorically to judge when we have objective and subjective concepts in science. Now, it will be important to note that the use of the subject-object schema as metaphor is applied to conceptual integrations of information gathered at widely spaced points in time and space and recalled from memory or recording devices, whereas perception involves attentional integration of information gathered over very short intervals in time on the order of 50 to 200 milliseconds (Blumenthal, 1977). This is what makes this schema a metaphor when applied in the formation of abstract objective concepts because it is applied to a different domain than the attentional domain of perception.

The subject-object schema has the following elements: a subject, an object, an object as invariant in perception across different points of view, and a subject as the source of regular changes in perception independent of objects. The schema can be projected onto features of a scientific setting. The subject is the researcher, the object is an invariant property observed across many observations, and artifacts in the research setting are effects of the observational methods or apparatus on what is observed and are linked to the subject. In Figure 23.9, we show how the schema of subject/object is mapped to the scientific domain.

For example, consider the case of cold fusion. In 1989, Stanley Pons and Martin Fleischmann, two professors at the University of Utah with credible reputations in chemistry, claimed they had discovered an electrochemical process whereby nuclear fusion could take place at modest temperatures and pressures and yield energy in excess of that introduced into the system. In fact, one claimed he produced the effect in his kitchen using ordinary utensils as well as under more controlled conditions in the lab. The claims, if true, would have revolutionized the energy industry. No one suspected fraud, but the physics establishment suspected bad experimental technique and demanded immediate replication. Numerous experiments were conducted with usually failure to find the claimed results of fusion, and some laboratories that thought they had replicated the findings later discovered

experimental artifacts in their experiments that negated their findings. Later experiments in other laboratories were conducted to see if already known processes other than fusion could have produced the results of Pons and Fleischmann, and some believe they had. Other laboratories, however, continued to study the phenomenon, and some believed that some kind of new nuclear effect was generating energy although it might not be fusion. However, mainstream physics came to reject the idea of cold fusion (Goodstein, 1994; Platt, 1998).

In the cold fusion case, the requirement for replication and invariance in different laboratories and with different methods and instruments corresponded to the idea of an object as an invariant in the perceptual domain, independent of effects of the observer. Research artifacts due to a researcher's methods of instrumentation, observation, and execution of the experiment corresponded to effects of the observer's acts and motion on the perceptual array. The researcher corresponds to the subject, and the objective phenomenon in the scientific domain corresponds to the object in the perceptual domain.

The need to test a hypothesis with data other than that used in its formulation arises out of the need to establish that the invariant claimed by the hypothesis is independent of the theorist, just as objects are regarded as independent of the actions, perspective, and motions of the observer as the observer moves in the environment because they are invariants unaffected by the changes due to the actions of the observer. When one formulates a hypothesis, often one only has a bare framework for the hypothesis and uses observations of the phenomenon to be understood by the hypothesis to adjust the hypothesis to fit the phenomenon. But that ties the good fit to the phenomenon to the theorist who made the adjustments in his or her theory to get the fit. A test of independence for the invariance is not possible in the case of those adjustments because the final form of the hypothesis will necessarily fit those aspects of the phenomenon to which it was adjusted. To be a test, a test must have the logical possibility to fail the test. Thus, observations—especially under somewhat different conditions—that are not used in formulating the hypothesis must be used to test the hypothesis.

Closely associated with the need to test hypotheses against data not used in their formulation is the idea that theories and models are corrigible and defeasible with future experience. This draws upon experiences humans have of seeing something that looks like something they have encountered before and then seeing it from a different angle and discovering that what they see is not what they have seen before. Almost everyone has had the experience of coming across a person in a crowd from behind who looked just like a person he or she knew, only to discover as the person turned that the face was not the face of the person known. Our knowledge of objects is acquired piecemeal, from different perspectives, gradually over time, and expectations we have based on the past can be disconfirmed by current or future experience. So, objective knowledge, as experienced everyday, is corrigible, and that should guide us to expect scientific knowledge to be corrigible.

23.6. Objectivity, Degrees of Freedom, and Parsimony

In formulating hypothetical mathematical models to represent some phenomenon, often the researcher begins with a framework like that of a general structural equation model. Certain measured variables are regarded as indicators of certain latent exogenous variables, whereas other measured variables are regarded as indicators of certain latent endogenous variables. In providing a hypothetical causal structure between the latent variables and the observed indicator variables and between the various latent variables, the researcher then fixes certain path coefficients to some prespecified value. Fixing a path coefficient to zero means that one hypothesizes that a certain variable is not a cause of another variable. Fixing a path coefficient to a nonzero value means that one expects that a unit change in the causal variable will produce a change proportional to the value of the fixed coefficient in the affected variable. Freeing a path coefficient, however, contrary to what many believe, is not the same as asserting that there is a causal relation between the respective variables. The computer programs that estimate the free coefficient will adjust the free parameter and other free parameters until a best fit is found for the model to the data conditional on the fixed and constrained parameters of the model, which are carried along unchanged in the computations. A free parameter could turn out to be any value, including zero. So, freeing a parameter is not the same as asserting something about it in one's hypothesis but rather is an assertion of ignorance. One does not know a value to specify for the parameter. So, one's model is incomplete, and one needs to estimate unspecified parameters to get a model-based reproduction of the data to see if it fits the actual data. Freeing a parameter is adjusting one's hypothesis so that it fits the data as best as possible conditional on any further constraints

imposed on the parameters of the model. Thus, if there is to be any lack of fit, it will be due to the constrained parameters and not the free parameters. Thus, good fit for the model as a whole should be interpreted only in terms of reflecting a test of just the constraints on the model and not about the free parameters.

Now many models in science often have so many elements and connections between them that more parameters than elements in the data exist in the models. If theory does not already specify values for these parameters, then they must be estimated to get a model that is scaled to the data. Take, for example, a structural equation model that models the covariance matrix between some observed variables as functions of parameters relating the observed variables to latent variables and/or to functional or correlational relations between latent variables. Each element of the covariance matrix is thus a function of some of the parameters of the structural equation model. Thus, if the number of unknown parameters of the structural equation model exceeds the number of observed variances and covariances, it will not be possible to solve the equations relating observed variances and covariances to the model parameters for values of the unknown parameters. Thus, a necessary (but not sufficient) condition for being able to solve for these parameters is that the number of free parameters to estimate does not exceed the number of distinct observed parameters, which are the variances and covariances among the observed variables. There are $p(p+1)/2$ distinct variances and covariances for the observed variables. Corresponding covariances for the same pair of variables on either side of the principal diagonal of the covariance matrix are not distinct, so only one of each such pair is a distinct parameter. Determining that one has fixed appropriately a sufficient number of parameters in the model to allow the remaining free parameters to be determined by distinct elements of the variance-covariance matrix, regardless of their values, is known as the problem of identification. If a model is identified and the number of parameters to estimate equals the number of distinct elements, such a model is said to be *just-identified.* Just-identified models always fit their data perfectly. For example, in the problem of finding a curve to fit the five points in Figure 23.8, we found a curve for a sixth-degree polynomial that has seven coefficients:

$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5 + a_6x^6$$
$$= a_0 + a_1 + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5 + a_6x^6$$

Two of the parameters of the polynomial were fixed to certain values, leaving free only five parameters to be determined from values of the coordinates of the five points. With five knowns (the coordinates of the points) and five unknowns in the free parameters of the polynomials, the result was a just-identified equation that fit the points perfectly. Just-identified equations and just-identified models are not useful from the point of view of model testing because they always fit perfectly as a mathematical necessity. Furthermore, the constraints on parameters that achieve just-identification cannot be tested because the models fit perfectly. They are like a priori concepts that make the models possible. They are neither true nor false and are more a convention for how one will represent the data. For a test with the model to be possible, it will require introducing constraints on additional parameters.

If additional parameters of the polynomial equation are fixed to prespecified values, then it may not be possible for the remaining parameters to be solved for such that a curve passes through the points. Rather than exact fit, one may require only least squares fit. When additional parameters are specified beyond those making the model just-identified, the model becomes overidentified. The model tests the overidentifying constraints in the context of the just-identifying constraints. However, which are over-identifying and which are just-identifying constraints may no longer be uniquely defined. Different subsets of the constraints may be selected to serve as just-identifying constraints and the remaining constraints evaluated in terms of them. If each of two overidentified models cannot free up its constraints to find a common set of just-identifying constraints for both models, then the models may not be comparable because they are based on different conventions and/or untestable assumptions (with respect to the data).

Overidentified models are required to do scientific work because it is only with them that lack of fit becomes logically possible, allowing one to test models by means of assessing lack of fit to the data. But it is important to keep in mind what is tested: the overidentifying constraints in the context of some additional constraints that make identification possible. The whole model is not specified unless all of its parameters are specified; hence, the whole model is not tested. Lack of fit should only be addressed to the overidentified constraints.

Models can be compared in the degree to which they are testable. In Mulaik (2001), I proposed the degrees of freedom of a model as a measure of the disconfirmability of a model. I showed that for models that estimated parameters, the degrees of freedom of the model was the number of dimensions in which the reproduced data based on the model was free to differ from the observed data. The degrees of freedom of

a model are the differences in number between the number of distinct data points to fit and the number of free parameters. In other words, given p observed variables, for a structural equation model designed to fit $p(p + 1)/2$ distinct variances and covariances, estimating m parameters yields for the degrees of freedom $df = p(p + 1)/2 - m$. The fewer parameters estimated relative to the number of data points to fit, the more degrees of freedom. The fewness of parameters estimated is known as the *degree of parsimony* of the model, so degrees of freedom and parsimony are related concepts. As degrees of freedom increases, the number of free parameters decreases, and the model becomes more parsimonious. Parsimonious models have been advocated as the ideal for centuries. But it was Karl Popper (1934/1961) who argued that more parsimonious models are more falsifiable, thus giving a rationale for why parsimonious models are preferable. However, he was not able to work this out in detail to show why parsimonious models were more falsifiable. My account (Mulaik, 2001) shows that this is so because models that estimate fewer parameters are free to differ from the data in more dimensions.

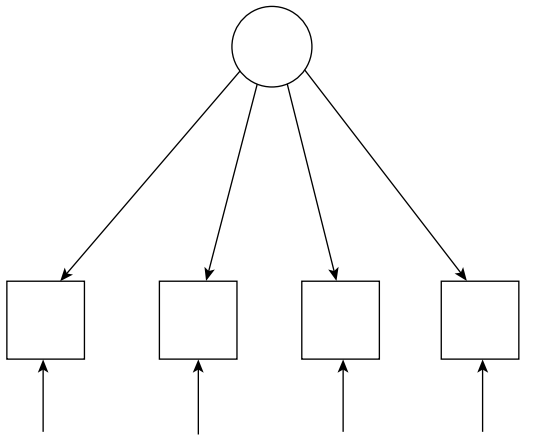
Because model testing is a part of establishing the objectivity of specified aspects of a hypothesis, degrees of freedom, which measure the degree to which a model can be tested, are also a part of assessing the objectivity of a model. Given two models of the same phenomena that fit the data equally well, the model with more degrees of freedom is to be preferred.

23.7. Objectivity and Multiple Indicators

If at least four observed variables are indicators of a latent variable, and one can show that a single common factor model is consistent with them, then that supports the objectivity of the latent common factor.

If a set of four or more variables has a covariance matrix that satisfies a single common factor model, then this establishes an invariant across them. A portion of each indicator's variation is proportional to the variation of a common latent variable. The variation of the common factor is the invariant across the indicators because each indicator displays it, although in muted or amplified form, depending on the factor loading. Each indicator's relationship to the latent common factor is analogous to seeing the same object from different points of view. There are features unique to each point of view as well as something invariant.

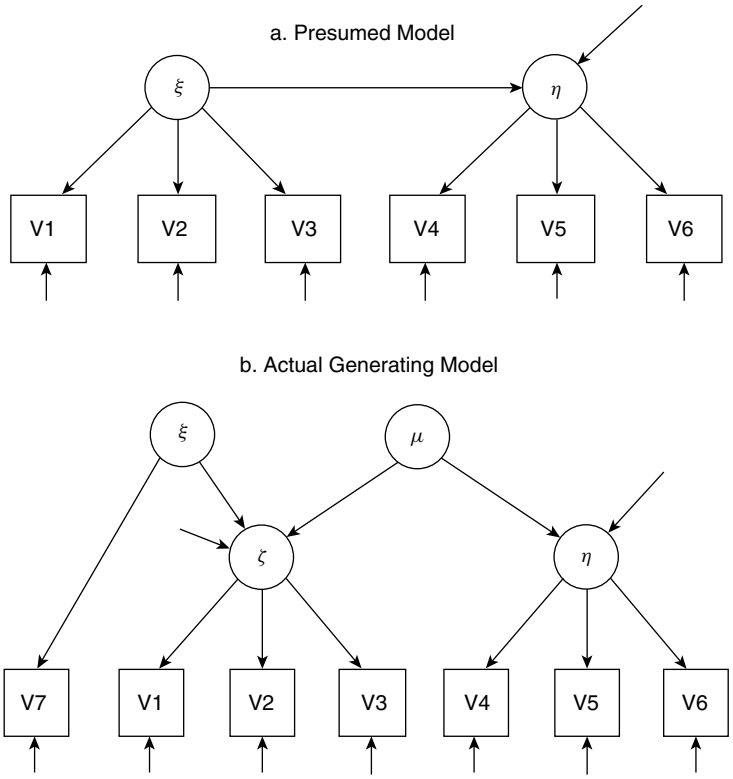
Figure 23.10 A Single Common Factor Can Gain Objective Support From Four or More Indicators



At least four indicators are required to establish the objectivity because then a test that the indicators have a common factor is possible. One can apply Spearman's tetrad difference test to the correlations among the indicators (Anderson & Gerbing, 1988; Glymour, Scheines, Spirtes, & Kelly, 1987; Hart & Spearman, 1913; Mulaik & Millsap, 2000). This test is as follows. One tests the following hypotheses: If the four variables, x_1, x_2, x_3, and x_4, have a common factor, then the correlations between the four variables should all be nonzero and satisfy the equations $\rho_{21}\rho_{34} - \rho_{23}\rho_{14} = 0$, $\rho_{24}\rho_{13} - \rho_{21}\rho_{34} = 0$, and $\rho_{24}\rho_{13} - \rho_{23}\rho_{14} = 0$, where ρ_{ij} denotes the correlation between variables i and j. Or one can evaluate the fit of a single common factor to them using a structural equations modeling program.

Some have thought that three indicators and a causal path from the latent variable to another latent variable that also has three indicators would be sufficient to establish that there is a common factor among the three indicators. It does, but it does not establish that the factor common to them is necessarily the factor you think it is. What it establishes is just that there is a factor common to the three indicators and the indicator of another latent variable. Mulaik and Millsap (2000) considered a case in which using only three indicators per latent variable would make it impossible to discover that an apparent causal effect between a hypothesized latent variable and another latent variable does not apply. Figure 23.11 is adapted from Figure 1 in Mulaik and Millsap (2000).

Figure 23.11 A Presumed Causal Model With Only Three Indicators and the Model That Generates the Data



In Figure 23.11a, one presumes that ξ is a cause of η One is able to test whether V1, V2, and V3 have a common factor by testing them jointly along with any one of the indicators of η using Hart and Spearman's (1913) tetrad difference test. However, if the data were actually generated, as in Figure 23.11b, then we could tell that something is wrong by introducing one other variable, V7, that is both an immediate and fourth indicator of ξ. In this case, we see that V1, V2, and V3 are immediate indicators of another common factor, ξ, which is itself a confluence of ξ and another methodological factor, μ which is also a cause of η. In fact, ξ is not a cause of η. Neither is ζ. To be sure, ξ is a common factor of V1, V2, and V3 as well as of V7. But V7 will be uncorrelated with any of the indicators of η, which would not be the case if ξ were a cause of η and V7. Whereas in the model in Figure 23.11a, any indicators including V1, V2, V3, and any one of the indicators of η would pass a tetrad difference test for a single common factor, that would not be the case for any three indicators of ξ and an indicator of η in the model in Figure 23.11b. Any four tests involving V7, two other indicators of ξ (say, V1 and V2), and an indicator of η (say, V5) would not pass the tetrad difference test. So, including a fourth indicator of ξ, chosen carefully to be sure that it does not use the same method of measurement as V1, V2, and V3, would increase the possibility of discovering something wrong with the model. In fact, including four or more indicators chosen under the belief that they measure the same latent variable with different methods improves the chances of rejecting the model, if unanticipated other causes are present, by increasing considerably the degrees of freedom or the number of tests that could be performed.

23.8. Conclusions

We have seen that objectivity concerns more than inter-subjective agreement between observers or maintaining an impersonal and unprejudiced attitude. It concerns various concepts of what constitutes an object, such as "an object is a thing bearing properties"

(inherence). Causality also involves objects because it concerns how attributes of objects are dependent on other attributes, often of other objects. We showed how this schema is used to develop concepts of a variable. Objects, we furthermore learned, can be conceived as reciprocally interacting systems in which the attributes of each object mutually affect those of other objects in the system. We also considered Kant's contribution to the development of the modern notions of objectivity—namely, his idea that an object is a concept regarded as universally valid that unites diverse intuitions according to a rule. This did not imply that empirical concepts are incorrigible but that one acts in reasoning with concepts as if the concept that applies to experiences in the present and the past will also necessarily apply to them in the future. But such expectations can be overturned with further experience. We considered how a modern school of linguists, philosophers, and literary theorists who also function as cognitive scientists argue that human thought routinely involves numerous forms of synthesis, the most common of which are metaphoric concepts that take metaphors from a source domain of embodied perception and coordinated action. Metaphors are mappings (in fact, most syntheses can be understood in terms of mappings) from one domain of experience to another so that inferences in the source domain can be transferred to the target domain. We noted that no one metaphor may allow us to fully understand a given phenomenon and that often several metaphors are required to achieve an accurate understanding. However, metaphors are only of intermediate complexity as conceptual syntheses go. More complex syntheses or conceptual integrations allow for the merger of elements from several input spaces into what are known as "conceptual blends."

We then considered that Kant's notion of an object is compelling as a concept that unites diverse intuitions according to a rule because it corresponds to fundamental features of embodied perception and action with respect to objects. We then adapted J. J. Gibson's notions of exteroception (perception of external objects as invariants in the perceptual flow) and proprioception (knowledge of an organism's actions and movements, taken from the regular changes it produces in the perceptual flow) to serve as a metaphor for scientific knowledge as the knowledge of objects. Science seeks knowledge of invariants across numerous observations taken at widely spaced points in time and space, and this corresponds to the way perception yields objects as invariants across sensory inputs spaced close together in space and time on the order of up to 200 milliseconds.

Science tests assertions of invariants specified as hypotheses against data not used in the formulation of the hypotheses, so that if the hypothesis is upheld, it can be asserted that it was independent of the one who asserted the hypothesis and thus is an objective result. Science also demands replications of results to gain the different points of view of different laboratories, researchers, instruments, and indicators, which again corresponds to establishing invariants in perception as the organism moves to new positions. We also noted how methodological and conceptual artifacts in science correspond to subjective effects of the perceiver due to the perceiver's own actions. We then considered further how hypothesis testing yields conclusions that are provisionally independent of the researcher. Then, penultimately, we considered how degrees of freedom measure the disconfirmability of an incompletely specified hypothesis. And finally, we also considered how using multiple indicators strengthens the objectivity of a latent variable and also provides more ways and degrees of freedom for testing a hypothesis.

REFERENCES

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411–423.

Berkeley, G. (1901). *Principles of human knowledge* (A. C. Fraser, Ed.). Oxford, UK: Oxford University Press. (Original work published 1710)

Berkeley, G. (1901). *Three dialogues between Hylas and Philonous* (A. C. Fraser, Ed.). Oxford, UK: Oxford University Press. (Original work published 1713)

Blumenthal, A. L. (1977). *The process of cognition.* Englewood Cliffs, NJ: Prentice Hall.

Daston, L. (1994). Baconian facts, academic civility and the prehistory of objectivity. In A. Megill (Ed.), *Rethinking objectivity* (pp. 37–63). Durham, NC: Duke University Press.

Descartes, R. (1901). Discourse on the method of rightly conducting the reason and seeking truth in the sciences. In J. Veitch (Ed. & Trans.), *The method, meditations and philosophy of Descartes* (pp. 147–204). Washington, DC: M. Walter Dunne. (Original work published 1637)

Descartes, R. (1901). The meditations of Descartes. In J. Veitch (Ed. & Trans.), *The method, meditations and philosophy of Descartes* (pp. 205–280). Washington, DC: M. Walter Dunne. (Original work published 1641)

Fauconnier, G. (1994). *Mental spaces: Aspects of meaning construction in natural language.* Cambridge, UK: Cambridge University Press.

Fauconnier, G. (1997). *Mappings in thought and language.* Cambridge, UK: Cambridge University Press.

Fauconnier, G., & Turner, M. (2002). *The way we think: Conceptual blending and the mind's hidden complexities.* New York: Basic Books.

Gibson, J. J. (1966). *The senses considered as perceptual systems.* London: Allen & Unwin.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton-Mifflin.

Gibson, J. J. (1982). *Reasons for realism: Selected essays of James J. Gibson* (E. Reed & R. Jones, Eds.). Hillsdale, NJ: Lawrence Erlbaum.

Giere, R. (1988). *Explaining science: A cognitive approach.* Chicago: University of Chicago Press.

Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). *Discovering causal structure.* New York: Academic Press.

Goodstein, D. (1994). Whatever happened to cold fusion? Retrieved from the World Wide Web at www.its.caltech.edu/∼dg/fusion.html

Hart, B., & Spearman, C. (1913). General ability, its existence and nature. *British Journal of Psychology, 5,* 51–84.

Henrich, D. (1989). Kant's notion of a deduction and the methodological background of the first *Critique.* In E. Förster (Ed.), *Kant's transcendental deductions* (pp. 29–46). Stanford, CA: Stanford University Press.

Hume, D. (1968). *A treatise of human nature* (L. A. Selby Bigge, Ed.). Oxford, UK: Clarendon. (Original work published 1739)

Hume, D. (1977). *An enquiry concerning human understanding* (E. Steinberg, Ed.). Indianapolis, IN: Hackett. (Original work published 1748)

Johnson, M. (1993). *Moral imagination: Implications of cognitive science for ethics.* Chicago: University of Chicago Press.

Kant, I. (1996). *Critique of pure reason* (W. S. Pluhar, Trans.). Indianapolis, IN: Hackett. (Original work published 1787)

Lakoff, G. (1987). *Women, fire and dangerous things: What categories reveal about the mind.* Chicago: University of Chicago Press.

Lakoff, G. (1996). *Moral politics: What conservatives know that liberals don't.* Chicago: University of Chicago Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh.* New York: Basic Books.

Lakoff, G., & Nuñez, R. E. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being.* New York: Basic Books.

Lakoff, G., & Turner, M. (1989). *More than cool reason: A field guide to poetic metaphor.* Chicago: University of Chicago Press.

Locke, J. (1962). *Locke's essay concerning human understanding* (M. W. Calkins, Ed.). LaSalle, IL: Open Court. (Original work published 1694)

Mace, W. M. (1986). J. J. Gibson's ecological theory of information pickup: Cognition from the ground up. In T. J. Knapp & L. C. Robertson (Eds.), *Approaches to cognition: Contrasts and controversies* (pp. 137–157). Hillsdale, NJ: Lawrence Erlbaum.

Megill, A. (1994). The four senses of objectivity. In A. Megill (Ed.), *Rethinking objectivity* (pp. 1–20). Durham, NC: Duke University Press.

Merriam-Webster's collegiate dictionary (10th ed.). (1996). Springfield, MA: Merriam-Webster.

Michell, J. (1990). *An introduction to the logic of psychological measurement.* Hillsdale, NJ: Lawrence Erlbaum.

Mulaik, S. A. (1986). Toward a synthesis of deterministic and probabilistic formulations of causal relations by the functional relation concept. *Philosophy of Science, 53,* 313–332.

Mulaik, S. A. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Multivariate Behavioral Research, 22,* 267–305.

Mulaik, S. A. (1994a). The critique of pure statistics: Artifact and objectivity in multivariate statistics. In B. Thompson (Ed.), *Advances in social science and methodology* (pp. 247–296). Greenwich, CT: JAI.

Mulaik, S. A. (1994b). Kant, Wittgenstein, objectivity and structural equations modeling. In C. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 209–236). New York: Plenum.

Mulaik, S. A. (1995). The metaphoric origins of objectivity, subjectivity, and consciousness in the direct perception of reality. *Philosophy of Science, 62,* 283–303.

Mulaik, S. A. (2001). The curve-fitting problem: An objectivist view. *Philosophy of Science, 68,* 218–241.

Mulaik, S. A., & Millsap, R. (2000). Doing the four-step right. *Structural Equation Modeling, 7,* 36–73.

Pearson, K. (1911). *The grammar of science: Part I. Physical.* London: Adam & Charles Black. (Original work published 1892)

Peirce, C. S. (1931–1958). *Collected papers of Charles Sanders Peirce* (Vols. 1–8; C. Hartshorne & P. Weiss, Eds. [Vols. 1–6], & A. W. Burks, Ed. [Vols. 7–8]). Cambridge, MA: Harvard University Press.

Platt, C. (1998, November). What if cold fusion is real? *Wired, 6.* Available at Web archive: www.wired.com/wired/archive/6.11/coldfusion.html

Popper, K. R. (1961). *The logic of scientific discovery* (translated and revised by the author). New York: Science Editions. (Original work published 1934).

Schouls, P. A. (1980). *The imposition of method: A study of Descartes and Locke.* Oxford, UK: Clarendon.

Turner, M. (1996). *The literary mind: The origins of thought and language.* Oxford, UK: Oxford University Press.

Turner, M. (2001). *Cognitive dimensions of social science.* Oxford, UK: Oxford University Press.

Whewell, W. (1966). *The philosophy of the inductive sciences founded upon their history* (2nd ed.). New York: Johnson Reprint Corporation. (Original work published 1847)

Chapter 24

CAUSAL INFERENCE

PETER SPIRTES

RICHARD SCHEINES

CLARK GLYMOUR

THOMAS RICHARDSON

CHRISTOPHER MEEK

24.1. INTRODUCTION

A principal aim of many sciences is to model causal systems well enough to provide insight into their structures and mechanisms and to provide reliable predictions about the effects of policy interventions. To succeed in either of these aims, in general, one must specify a model at least approximately correctly. In the social and behavioral sciences, causal models are often described within any of a variety of statistical formalisms: categorical data models, logistic regression models, linear regression models, factor analysis models, principal components models, structural equation models, and so on. In practice, these models are obtained by a variety of methods: experimentation, investigator's convictions, uncontested background knowledge, automated search procedures such as stepwise regression and factor analysis, and any of a wide range of ad hoc model selection procedures. To understand the assumptions built into these and other classes of models, as well as their limitations, one must have a clear understanding of the connection between causal hypotheses and probabilistic or statistical hypotheses. And to understand the limitations of models produced by common methods, one should understand the theoretical limitations on all causal inference procedures.

In the past two decades, researchers from computer science, statistics, various social sciences, and philosophy have generalized methods, models, and concepts related to causal inference that were rooted in structural equation modeling. We will outline and illustrate the results of this research and address the following questions:

1. What is the difference between a causal model and a statistical model? What uses can be made of causal models that cannot be made of statistical models? (Section 24.2)
2. What assumptions relating causality to probability should we make? (Section 24.3)

AUTHORS' NOTE: We thank S. Mulaik for many helpful comments and references.

3. What are the theoretical limits on causal inference? We examine the answer to this question under a variety of different assumptions about background knowledge. (Section 24.4)
4. What are some reliable methods of causal inference? (We answer the question when it is assumed that there are no latent causes in Section 24.4.3 and when the possibility of latent causes is allowed in Section 24.5.)
5. Are the methods of causal inference commonly employed in various social sciences reliable? (Section 24.6)

We focus especially on two issues: (a) how, given possibly incomplete causal and statistical information, to predict the effects of interventions or policies and (b) how, and to what extent, causal information can be obtained from experimental and nonexperimental data under a variety of assumptions about the underlying processes.

24.2. Causal Models and Statistical Models

24.2.1. The Meaning of *Direct Cause*

Direct cause is one of a family of causal phrases (including *intervention, manipulation, direct effect*, etc.) that are easily inter-definable but not easily definable in noncausal terms. A variety of different definitions of causal concepts in terms of noncausal concepts have been proposed, but they are typically both complicated and controversial. We take a different approach here, using the concept of *direct cause* as an undefined primitive relationship between random variables and introducing generally (but often only implicitly) accepted axioms relating direct causes to probability distributions. Although in much of the philosophical literature, causation is taken to be a relation among events rather than variables, taking causation as a relation among variables fits much more closely with a variety of statistical models and methods of statistical inference. (Viewing causation as a relation among variables is also the approach taken in James, Mulaik, & Brett, 1982; Mulaik, 1986; Simon, 1953.) The advantage of the axiomatic approach is that the acceptability of these axioms does not necessarily depend on any particular definition of causality. This will allow us to discuss principles of causal inference that are acceptable to a variety of schools of thought about the meaning of *causality* (just as there are at least some principles of probabilistic inference that do not depend on the definition of *probability*). Philosophical discussions of the meaning and nature of causation are found in Cartwright (1989, 1999), Eells (1991), Hausman (1998), Shafer (1996), Sosa and Tooley (1993), and Pearl (2000).

24.2.2. Conditioning and Manipulating

Two fundamentally different operations map probability distributions[1] into other probability distributions. The first is *conditioning*, which corresponds roughly to mapping a probability distribution into a new distribution in response to finding out more information about the state of the world (or *seeing*). The second is *manipulating*, which corresponds roughly to mapping a probability distribution into a new probability distribution in response to changing the state of the world in a specified way (or *doing*, in the terminology of Pearl, 2000).[2] To illustrate the difference, we consider a simple example in which our pretheoretic intuitions about causation are uncontroversial; in subsequent sections, we will consider more interesting and realistic examples. Consider a population of flashlights, each of which has working batteries and light bulbs, and a switch that turns the light on when the switch is in the on position and turns the light off when the switch is in the off position. Each unit (flashlight) in the population has some properties (the switch position and whether or not the light is on). The properties are represented by the random variables *Switch* and *Light. Switch* can take on the values *on* or *off*, and *Light* can take on the values *on* or *off.* Although in this example, the random variables are binary to simplify the illustration, all of the concepts also apply to discrete variables with more than two categories and to continuous variables. The random variables have a joint distribution in the population. Suppose that in this case, the joint distribution is the following:

$$
\begin{aligned}
P(\mathit{Switch} = \mathit{on}, \mathit{Light} = \mathit{on}) &= 1/2 \\
P(\mathit{Switch} = \mathit{on}, \mathit{Light} = \mathit{off}) &= 0, \\
P(\mathit{Switch} = \mathit{off}, \mathit{Light} = \mathit{on}) &= 0 \\
P(\mathit{Switch} = \mathit{off}, \mathit{Light} = \mathit{off}) &= 1/2.
\end{aligned}
$$

1. We are deliberately ambiguous about the interpretation of *probability* here. The remarks here do not depend on whether a frequentist, propensity, or personalist interpretation of *probability* is assumed.

2. For more on the difference between conditioning and manipulating, see Rubin (1977); Spirtes, Glymour, and Scheines (2000); and Pearl (2000).

24.2.2.1. Conditioning

Much of statistics is devoted to finding efficient methods of estimating conditional distributions. Given a randomly chosen flashlight, the probability that the bulb is *on* is 1/2. However, if someone observes that a flashlight has a switch in the *off* position but does not directly observe whether the light is *off*, the probability of the light being *off*, conditional on the switch being *off*, is just the probability of the light being *off* in the subpopulation in which the switch is *off*; that is, $P(Light = off | Switch = off) = P(Light = off, Switch = off)/P(Switch = off) = 1$. So conditioning on an event maps the joint distribution of the variables into a new probability distribution. Similarly, the probability of the switch being *off*, conditional on the light being *off*, is just the probability of the switch being *off* in the subpopulation in which the light is *off*; that is, $P(Switch = off | Light = off) = P(Light = off, Switch = off)/P(Light = off) = 1$. An important feature of conditioning is that each conditional distribution is completely determined by the joint distribution (except when conditioning on an event that has probability 0).

24.2.2.2. Manipulating

Manipulating, like conditioning, maps a joint probability distribution into another joint distribution.[3] In contrast to conditioning, a manipulated probability distribution is not usually a distribution in a subpopulation of an existing population but is a distribution in a (possibly hypothetical) population formed by externally *forcing* a value on a variable in the system. Imagine that instead of seeing that a switch was *off*, we successfully manipulate the switch to *off*. It follows from the assumed structure and function of the flashlights that the probability of the light being *off* is 1. (Here, we are relying on pretheoretic intuitions about the example to derive the correct values for the manipulated probabilities [e.g., that working flashlights are on when the switch is on]. In later sections, we will describe formal methods by which the manipulated probabilities can be calculated.) We will adapt the notation of Lauritzen (2001) and denote the postmanipulation probability of the light being *off* as $P(Light = off \| Switch = off)$, using a double bar "‖" for manipulation, as distinguished from the single bar "|" of conditioning.[4] Note that in this case, $P(Light = off \| Switch = off) = P(Light = off | Switch = off)$. Analogously to the notation for conditioning, one can also put a set of variables $\mathbf{V}$ on the left side of the manipulation double bar, which represents the joint probability of $\mathbf{V}$ after manipulating the variables on the right side of the manipulating double bar.[5]

Suppose now that instead of manipulating *Switch* to *off*, we were to manipulate *Light* to *off*. Of course, the resulting probability distribution depends on how we manipulated *Light* to *off*. If we were to manipulate *Light* to *off* by unscrewing the light bulb, the probability that *Switch* is *off* is 1/2, the same as the probability that it was *off* prior to our manipulation. In that case, the manipulation is said to be an "ideal manipulation" of *Light* because an external cause was introduced (the unscrewing of the light bulb) that was a direct cause of *Light* and was not a direct cause of any other variable in the system. Under the assumptions described in Section 24.3, any ideal manipulation of a given variable to the same value will yield the same (probability distribution) over outcomes.

On the other hand, if we manipulated *Light* to *off* by pressing the *Switch* to *off*, then of course the probability that *Switch* is *off* after the manipulation is equal to 1. This, however, would not be an ideal manipulation of *Light* because the external cause was a direct cause of a variable in the system other than *Light* (namely, a direct cause of *Switch*). In general, the theory of predicting the effects of direct manipulations that we will describe assumes that the direct manipulations are successfully carried out and that they are ideal manipulations of variables in the system.

In the case where we perform an ideal manipulation of the light bulb to *off* (e.g., by unscrewing it), the manipulated probability does not equal the conditional probability; that is, $P(Switch = off \| Light = off) = 1/2 \neq P(Switch = off | Light = off) = 1$. This illustrates two key features of manipulations. The first is that in some cases, the manipulated probability is equal to the conditional probability (e.g., $P(Light =$

3. An early exposition of the view that the value of an effect is a probabilistic function of the value of a cause is presented in Mulaik (1986), who also discussed local independence in the context of chains of variables forming a Markov process.

4. What time period after the manipulation does the *postmanipulation distribution* refer to? In this case, long enough for the system to reach an equilibrium. In cases where there is no equilibrium or some other time period is referred to, the relevant variables should be indexed by time explicitly.

5. We use capitalized boldface to represent sets of variables, capitalized italics to represent variables, lowercase boldface to represent values of sets of variables, and lowercase italics to represent values of variables. If $\mathbf{V} = \{Switch, Light\}$ and $\mathbf{v} = \{Switch = off, Light = on\}$, then $P(\mathbf{v})$ represents $P(Switch = off, Light = on)$. There are a number of different alternative notations to the "‖" in Spirtes et al. (2000) and Pearl (2000).

off $\|Switch = off) = P(Light = off|Switch = off))$, and in other cases, the manipulated probability is not equal to the conditional probability (e.g., $P(Switch = off\|Light = off) \neq P(Switch = off|Light = off)$). In this example, conditioning on *Light* = *off* raised the probability of *Switch* = *off*, but manipulating *Light* to *off* did not change the probability of *Switch* = *off*. In general, if conditioning on the value of a variable X raises the probability of a given event, manipulating X to the same value may raise, lower, or leave the same the probability of a given event. Similarly, if conditioning on a given value of a variable lowers or leaves the probability of a given even the same, the corresponding manipulated probability may be higher, lower, or the same, depending on the domain.

The second key feature of manipulations is that even though *Light* = *on* if and only if *Switch* = *on* in the original population, the joint distributions that resulted from manipulating the values of *Switch* and *Light* were different. In contrast to conditioning, the results of manipulating depend on more than the joint probability distribution. The "more than the joint probability distribution," which the results of a manipulation of a specified variable depend on, is the causal relationships between variables. The reason that manipulating the switch position changed the status of the light is that the switch position is a cause of the status of the light; the reason that manipulating the light condition did not change the switch position is that the status of the light is not a cause of the switch position. Thus, discovering (at least implicitly) the causal relations between variables is a necessary step to correctly inferring the results of manipulations.

24.2.2.3. Other Kinds of Manipulations

Manipulating a variable to a particular value (e.g., *Switch* = *off*) is a special case of more general kinds of manipulations. For example, instead of assigning a value to a variable, a probability distribution can be assigned to a variable X. This is what occurs in randomized experiments. Suppose that we randomize the probability distribution of *Switch* to a distribution P', where $P'(Switch = on) = 1/4$, and $P'(Switch = off) = 3/4$. In that case, we denote the manipulated probability of *Light* = *on* as $P(Light = on \| P'(Switch))$; that is, a probability distribution appears on the right-hand side of the manipulation double bar. (The notation $P(Light = off\|Switch = off)$ is the special case where $P'(Switch = off) = 1$.)

More generally, given a set of variables $\mathbf{V}$ and manipulation of a set of variables $\mathbf{M} \subseteq \mathbf{V}$ to a distribution $P'(\mathbf{M})$, the joint distribution of $\mathbf{V}$ after the manipulation is denoted $P(\mathbf{V}\|P'(\mathbf{M}))$. From $P(\mathbf{V}\|P'(\mathbf{M}))$, it is possible to form marginal distributions and conditional distributions among the variables in $\mathbf{V}$ in the usual way. Thus, $P(X = x|Y = y\|P'(\mathbf{Z}))$ refers to the probability of X in the subpopulation where $Y = y$, after first manipulating the distribution of $\mathbf{Z}$ to $P'(\mathbf{Z})$.

To simplify the discussion, we will not consider manipulations that assign a conditional probability distribution to a variable (e.g., $P'(Light = off|Switch = on) = 1/2$ and $P'(Light = on|Switch = on) = 0)$), rather than assigning a marginal distribution to that variable. Also, when multiple manipulations are performed, we will assume for simplicity that in the joint manipulated distribution, the manipulated variables are independent.

24.2.3. Bayesian Networks: Causal and Statistical Interpretations

Bayesian networks are a kind of causal/statistical model that provides a convenient framework for representing and calculating the results of conditioning and manipulating. Bayesian networks also provide a convenient framework for discussing the relationship between causal relations and probability distributions. Bayesian networks are graphical models that generalize recursive structural equation models without correlated errors[6] (described in more detail in Section 24.2.4); they have both a statistical and a causal interpretation. We will describe the statistical interpretation first, then the causal interpretation, and finally the relationship between the two interpretations. Pearl (1988), Neapolitan (1990), Cowell (1999), and Jensen (2001) provide introductions to Bayesian networks. Lauritzen (2001), Pearl (2000), and Spirtes, Glymour, and Scheines (2000, chap. 3) describe the relation between the causal and statistical interpretations.

24.2.3.1. Statistical Interpretation

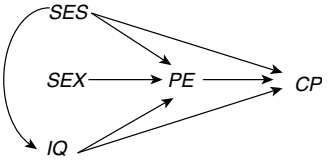
A Bayesian network consists of two parts: a directed acyclic graph (DAG) and a set of free parameters that map the graph onto a probability distribution via a rule that we will describe below. We will illustrate Bayesian networks using data from Sewell and Shah (1968), who studied five variables from a sample of

6. There are more general kinds of graphical models of which Bayesian networks are a special case that also have causal interpretations, but for the sake of simplicity, we postpone discussions of such models until later. See Whittaker (1990), Lauritzen (1996), Edwards (2000), and Spirtes et al. (2000).

10,318 Wisconsin high school seniors.[7] The variables and their values are as follows:

SEX	[male = 0, female = 1]
IQ = intelligence quotient	[lowest = 0, highest = 3]
CP = college plans	[yes = 0, no = 1]
PE = parental encouragement	[low = 0, high = 1]
SES = socioeconomic status	[lowest = 0, highest = 3]

Figure 24.1 Model of Causes of College Plans



The graph part of the Bayesian network that we will describe for the Sewell and Shah (1968) data is shown in Figure 24.1. We will explain the motivation behind hypothesizing the DAG in Section 24.4.3.

The following informal definitions describe various features of a directed graph.[8] A directed graph consists of a set of vertices and a set of directed edges, where each edge is an ordered pair of vertices. Let G be the directed graph in Figure 24.1. In G, the vertices are $\{IQ, SES, PE, CP, SEX\}$, and the edges are $\{SES \rightarrow IQ, IQ \rightarrow PE, SEX \rightarrow PE, IQ \rightarrow CP, SES \rightarrow PE, SES \rightarrow CP, PE \rightarrow CP\}$. In G, *SES* is a *parent* of *IQ*, *IQ* is a *child* of *SES*, and *SES* and *IQ* are *adjacent* because there is an edge $SES \rightarrow IQ$ in G. **Parents**(G, V) denotes the set of parents of a vertex V in directed graph G. A *path* in a directed graph is a sequence of adjacent edges (i.e., edges that share a common endpoint). A *directed path* in a directed graph is a sequence of adjacent edges all pointing in the same direction. For example, in G, $IQ \rightarrow PE \rightarrow CP$ is a directed path from *IQ* to *CP*. In contrast, $SES \rightarrow PE \leftarrow SEX$ is a path but not a directed path in G because the two edges do not point in the same direction. *CP* is a *descendant* of *SEX* (and *SEX* is an *ancestor* of *CP*) because there is a directed path from *SEX* to *CP*; in addition, by convention, each vertex is a descendant (and ancestor) of itself. A directed graph is *acyclic* when there is no directed path from any vertex to itself: In that case, the graph is a directed acyclic graph, or DAG for short. Table 24.1 shows these relationships for the DAG in Figure 24.1.

A DAG G over a set of variables $\mathbf{V}$ *represents* any joint distribution that can be factored according to the following rule:

$$P(\mathbf{v}) = \prod_{v \in \mathbf{V}} P(v|\mathbf{parents}(G, V)). \qquad (1)$$

In the case of continuous distributions, the probabilities in equation (1) can be replaced with density functions. For example, the DAG G of Figure 24.1 represents any joint probability distribution that can be factored according to the following formula:

$$\begin{aligned} P(iq, sex, ses, pe, cp) &= P(iq \mid ses) \times P(sex) \\ &\times P(ses) \times P(pe \mid ses, iq, sex) \\ &\times P(cp \mid pe, ses, iq), \end{aligned} \qquad (2)$$

where if iq is one of the values of *IQ*, $P(iq)$ is an abbreviation of $P(IQ = iq)$.

Equation (2) associates a set of probability distributions with the DAG in Figure 24.1; the ordered pair consisting of the DAG and the associated set of distributions is a statistical model. We would like to be able to refer to particular distributions in the statistical model by labeling each distribution in the model with a finite set of real numbers known as the values of the free parameters of the model. There are 128 different possible combinations of values of *SEX*, *IQ*, *SES*, *PE*, and *CP*. One way to refer to a particular joint distribution would be to list 128 real numbers, where each real number is the value of $P(iq, sex, ses, pe, cp)$ for some combination of values of *SEX*, *IQ*, *SES*, *PE*, and *CP*. (However, because the sum over all of the values of the variables of $P(IQ, SEX, SES, PE, CP)$ equals 1, only 127 numbers are really needed; the value of the 128th state is simply equal to 1 minus the sum of the first 127 numbers.) This would be one reasonable way to refer to any joint distribution $P(IQ, SEX, SES, PE, CP)$. However, not every joint distribution $P(IQ, SEX, SES, PE, CP)$ obeys the factorization of equation (2). If we want to refer only to joint distributions that factor according to equation (2), this is a poor method for two reasons. First, we are not guaranteed that the distribution referred to in this way can be factored according to equation (2). Second, we are using many more numbers than are actually needed to refer to just members of the set of probability distributions that can be factored according to equation (2).

A different method of referring to probability distributions that factor according to equation (2) uses equation (2) itself to map lists of numbers into probability distributions. For example, instead of specifying

7. Examples of the analysis of the Sewell and Shah (1968) data using Bayesian networks are given in Spirtes et al. (2000) and Heckerman (1998).

8. More formal definitions can be found in Spirtes et al. (2000) and Pearl (2000).

Table 24.1 Relationships Between Vertices in Figure 24.1

Vertex	*Children*	*Parents*	*Descendants*	*Ancestors*
SES	{*PE*, *CP*, *IQ*}	∅	{*PE*, *CP*, *IQ*, *SES*}	{*SES*}
SEX	{*PE*}	∅	{*PE*, *CP*, *SEX*}	{*SEX*}
IQ	{*PE*, *CP*}	{*SES*}	{*PE*, *CP*, *IQ*}	{*IQ*, *SES*}
PE	{*CP*}	{*SES*, *SEX*, *IQ*}	{*CP*, *PE*}	{*SES*, *SEX*, *IQ*, *PE*}
CP	∅	{*SES*, *PE*, *IQ*}	{*CP*}	{*SES*, *PE*, *IQ*, *SEX*, *CP*}

a particular value for $P(SEX = 0, IQ = 1, SES = 2, PE = 1, CP = 0)$ directly, we could simply specify values for $P(IQ = 1|SES = 2)$, $P(SEX = 0)$, $P(SES = 2)$, $P(PE = 1|SES = 2, IQ = 1, SEX = 0)$, and $P(CP = 0|PE = 1, SES = 2, IQ = 1)$. To choose some arbitrary values for the purposes of illustration, $P(IQ = 1|SES = 2) = .2$, $P(SEX = 0) = .5$, $P(SES = 2) = .1$, $P(PE = 1|SES = 2, IQ = 1, SEX = 0) = .3$, and $P(CP = 0|PE = 1, SES = 2, IQ = 1) = .4$. In that case, by equation (2), $P(SEX = 0, IQ = 1, SES = 2, PE = 1, CP = 0) = .2 \times .5 \times .1 \times .3 \times .4 = .0012$. By assigning numbers to $P(iq)$, $P(sex)$, $P(ses|iq)$, $P(pe|ses, iq, sex)$, and $P(cp|pe, ses, iq)$ for all values of *SEX*, *IQ*, *SES*, *PE*, and *CP*, the value of $P(IQ, SEX, SES, PE, CP)$ is determined for all values of *SEX*, *IQ*, *SES*, *PE,* and *CP,* and the joint distribution over $P(IQ, SEX, SES, PE, CP)$ is uniquely determined. Once again, however, this list would contain some redundant members. For example, once a value has been assigned to $P(SEX = 0)$, $P(SEX = 1)$, is determined to be $1 - P(SEX = 0)$. When all redundancies of this kind are removed, the resulting list contains 80 real numbers. If, for each vertex V, $P(V|\mathbf{parents}(G, V))$ is a probability distribution, the resulting joint distribution is guaranteed to factor according to equation (2). The fact that the statistical model with DAG *G* has fewer free parameters than the number of free parameters needed to refer to an arbitrary joint distribution over *SEX*, *IQ*, *SES*, *PE*, and *CP* entails that the distributions that *G* represents can be more efficiently estimated, stored in a smaller space, and used to more quickly calculate conditional probabilities. The quantities $P(iq)$, $P(sex)$, $P(ses|iq)$, $P(pe|ses, iq, sex)$, and $P(cp|pe, ses, iq)$ for all values of *IQ*, *SEX*, *SES*, *PE*, and *CP* (with the exception of the redundant quantities) are called the free parameters of the statistical model with DAG *G*.

By definition, a DAG *G* represents a probability distribution *P* if and only if *P* factors according to the DAG (equation (2)). Let $\mathrm{I}(\mathbf{X}, \mathbf{Y}|\mathbf{Z})_P$ mean that $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$ in distribution *P*—that is, $P(\mathbf{x}|\mathbf{y}, \mathbf{z}) = P(\mathbf{x}|\mathbf{z})$—for all values $\mathbf{y}$ and $\mathbf{z}$ such that $P(\mathbf{y}, \mathbf{z}) > 0$. By convention, $\mathrm{I}(\mathbf{X}, \varnothing|\mathbf{Z})_P$ is trivially true, and $\mathrm{I}(\mathbf{X},\mathbf{Y}|\varnothing)_P$ denotes unconditional independence of $\mathbf{X}$ and $\mathbf{Y}$. (The empty set here denotes an empty set of random variables, *not* a null event. Also, if a set contains a single variable, such as {*IQ*}, we will sometimes leave out the set brackets.)

The factorization of *P* according to *G* is equivalent to each variable *X* in the DAG being independent of all the variables that are neither parents nor descendants of *X* in *G*, conditional on all of the parents of *X* in *G*. Applying this rule to the example of the DAG in Figure 24.1 for any probability distribution that factors according to *G* (i.e., satisfies equation (2)), the following conditional independence relations hold in *P*:

$$I(\{IQ\}, \{SEX\}|\{SES\})_P\, I(\{SEX\}, \{IQ, SES\}|\varnothing)_P,$$
$$I(\{SES\}, \{SEX\}|\varnothing)_P\, I(\{PE\}, \varnothing|\{SES, IQ, SEX\})_P,$$
$$I(\{CP\}, \{SEX\}|\{PE, SES, IQ\})_P \qquad (3)$$

These conditional independence relations hold, regardless of what values are assigned to the free parameters associated with DAG *G*; we say that *G entails* the conditional independence relations. However, just because a conditional independence relation is not entailed by a DAG does not mean that it does not hold in *any* assignment of values to the free parameters: It just means that it does not hold in *every* assignment of values to the free parameters.

The conditional independence relations listed in (3) entail other conditional independence relations, for example, $I(\{SEX\}, \{SES\}|\{IQ\})_P$. There is an easily computable, purely graphical relationship, named *d-separation,* such that if a DAG *G* with vertex set $\mathbf{V}$ represents a probability distribution $P(\mathbf{V})$, $\mathbf{X}$ is *d*-separated from $\mathbf{Y}$ conditional on $\mathbf{Z}$ in *G* if and only if *G* entails that $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$ in $P(\mathbf{V})$. (See Pearl, 1988. The most complete reference with detailed proofs is Lauritzen, Dawid, Larsen, & Leimer, 1990. Scheines, 2003, is a Web site that contains a tutorial on *d*-separation.)

There are a number of equivalent formulations of the *d*-separation relation. The following definition is based on the intuition that only certain kinds of paths, which we shall call active paths, can pass information from

X to Y, conditional on $\mathbf{Z}$, and that a path is active only when every vertex on the path is active (i.e., capable of passing along information conditional on $\mathbf{Z}$). Then, two variables, X and Y, are d-separated conditional on $\mathbf{Z}$ when there are no active paths between them conditional on $\mathbf{Z}$; that is, there are no paths that can pass information from X to Y conditional on $\mathbf{Z}$. (The proofs of theorems about d-separation do not rely on any intuitions about passing information but rely solely on properties of conditional independence.) A vertex X_i is a *collider* on a path U in G if and only if there are edges $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$ on U in G; otherwise, it is a *noncollider.* Note that the endpoints of a path are always noncolliders on the path. A vertex on a path U is *active* conditional on a set of variables $\mathbf{Z}$ if and only if either it is not a collider on U and is not in $\mathbf{Z}$ or it is a collider on U and has a descendant in $\mathbf{Z}$. (Note that whether or not a vertex is active is relative to a particular path and a particular conditioning set.) A path U is *active* conditional on a set of variables $\mathbf{Z}$ if and only if every vertex on U is active conditional on $\mathbf{Z}$. If $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ are disjoint subsets of variables in G, $\mathbf{X}$ and $\mathbf{Y}$ are *d-connected* conditional on $\mathbf{Z}$ if and only if there is an active path conditional on $\mathbf{Z}$ between some $X \in \mathbf{X}$ and some $Y \in \mathbf{Y}$; otherwise, $\mathbf{X}$ and $\mathbf{Y}$ are *d-separated* conditional on $\mathbf{Z}$.

To illustrate these concepts, consider the DAG in Figure 24.1. It is easy to show that if a pair of variables are adjacent in a DAG, then they are d-connected conditional on any subset of the other variables. However, *SES* and *SEX* are not adjacent, and they are d-separated conditional on $\{IQ\}$ and are also d-separated conditional on $\varnothing$; *IQ* and *SEX* are not adjacent, and they are d-separated conditional on $\{SES\}$ and also d-separated conditional on $\varnothing$; and *SEX* and *CP* are not adjacent, and they are d-separated conditional on $\{PE, SES, IQ\}$.

Some of the consequences of the definition are fairly intuitive, but others are much less so. For example, it is intuitively obvious that the DAG of Figure 24.1 entails that *SEX* and *IQ* are unconditionally independent because there is no directed path between them, and there is no third variable that has directed paths to both of them. And the d-separation relation entails that *SEX* and *IQ* are unconditionally independent because they are d-separated conditional on the empty set; every path between *SEX* and *IQ* contains a collider, and no collider has a descendant in the empty set.

However, the condition that a vertex is active on a path conditional on $\mathbf{Z}$, if it is a collider on the path and has a descendant in $\mathbf{Z}$, is neither obvious nor intuitive in many instances. For example, *SEX* and *IQ* are not entailed to be independent conditional on $\{PE\}$ because on the path $SEX \rightarrow PE \leftarrow IQ$, *SEX* and *IQ* are noncolliders that are not in $\{PE\}$, and *PE* is a collider that has a descendant (itself) in $\{PE\}$; hence, *SEX* and *IQ* are d-connected conditional on $\{PE\}$.

24.2.3.2. Bayesian Networks: Causal Interpretation

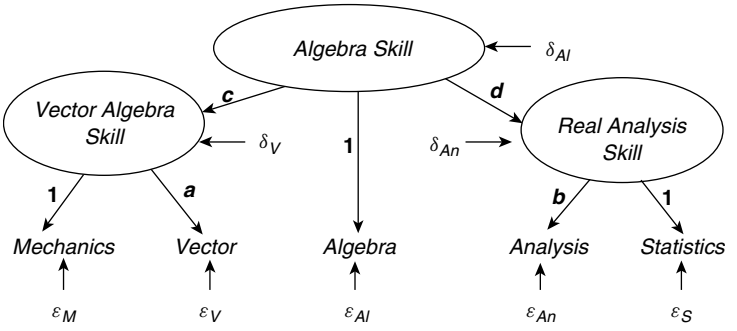
Note that the concept of "direct cause" is relative to a set of variables. Intuitively, if relative to $\{SEX, IQ, SES, PE, CP\}$, *SEX* is a direct cause of *PE*, and *PE* is a direct cause of *CP*, but *SEX* is not a direct cause of *PE*, then *SEX* is an indirect but not a direct cause of *CP*. Now consider a smaller set of variables such as $\{SEX, CP\}$, where the variables that record the details of the mechanism by which *SEX* is a cause of *CP* (i.e., by affecting *PE*) have been omitted. Relative to $\{SEX, CP\}$, *SEX* is a direct cause of *CP*.[9]

The graphical part of a Bayesian network can be given a causal interpretation. A set of random variables $\mathbf{S}$ is *causally sufficient* if $\mathbf{S}$ does not omit any variables that are direct causes (relative to $\mathbf{S}$) of any pair of variables in $\mathbf{S}$. Under the causal interpretation of a graph, a graph with a causally sufficient set of variables $\mathbf{S}$ *represents* the causal relations in a population N if there is a directed edge from A to B in the graph if and only if A is a direct cause of B relative to $\mathbf{S}$ for the population N. For example, under the causal interpretation of the DAG in Figure 24.1, there is a directed edge from *IQ* to *CP* if and only if *IQ* is a direct (relative to the set of variables in the DAG) cause of *CP* for the population. If the DAG in Figure 24.1 is a correct description of a causal system, then the set of variables $\{SEX, PE, CP\}$ is not causally sufficient because the set does not contain either *IQ* or *SES,* each of which is a direct cause of a pair of variables in the set. On the other hand, $\{SEX, CP\}$ is a causally sufficient set of variables because none of the other variables is a direct cause of both *SEX* and *CP*.

A *causal model* for a population N is a pair consisting of a causal graph over a causally sufficient set of variables $\mathbf{V}$ representing the causal relations in N and $P(\mathbf{V})$. The graphical part of a causal model M is denoted by $G(M)$. The kind of causation that we are describing in this chapter is causation between variables (or kinds of events, e.g., *Switch* and *Light*), not between individual events (e.g., the event of a particular flashlight having a *Switch* value *on* and the event of the same flashlight having the *Light* value *on*). Because the causal relation is between variables and not between events, it is possible that each of two

9. There is some controversy about whether *SEX* can be a cause of anything, in part because it is hard to imagine manipulating *SEX.* Very little of what follows depends on whether *SEX* should be considered a cause, and we will not consider any manipulations of *SEX.*

Figure 24.2 Model of Causes of Mathematical Marks



Mechanics: $=1\times$ Vector Algebra Skill $+\varepsilon_M$, Vector Algebra Skill $:= c\times$ Algebra Skill $+\delta_V$, Statistics $:= 1\times$ Real Analysis Skill $+\varepsilon_S$, Real Analysis Skill:$= d\times$ Algebra Skill $+\delta_{An}$

Vector $:= a\times$ Vector Algebra Skill $+\varepsilon_V$, Algebra $:= 1\times$ Algebra Skill $+\varepsilon_{Al}$, Analysis $:= b\times$ Real Analysis Skill $+\varepsilon_{An}$, Algebra Skill $:= \delta_{Al}$

variables can cause each other. For example, pedaling a bicycle can cause the wheel to spin, and (on some kinds of bicycles) spinning the wheel can cause the pedal to move. Thus, it is possible that a causal graph may be cyclic. The theory of cyclic causal graphs is important in econometrics, biology, and other subjects and is discussed in Section 24.5.2.4.2, but it is also considerably more difficult and less developed than the theory of acyclic causal graphs. For the rest of this chapter, we assume that all causal graphs are acyclic, unless we explicitly say otherwise.[10]

24.2.4. Structural Equation Models

Following Bentler (1985), the variables in a structural equation model (SEM) can be divided into two sets, the "error terms" and the substantive variables. The error terms are latent (which means only that their values are not recorded in the data), and some of the substantive variables may be latent as well. An SEM consists of a set of structural equations, one for each substantive variable, and the distributions of the error terms; together, these determine the joint distribution of the substantive variables. The structural equation for a substantive variable X_i is an equation with X_i on the left-hand side of the equation and the direct causes of X_i plus an error term ε_i on the right-hand side of the equation. The equations may take any mathematical form, although linear equations are most common. Kaplan (2000) and Bollen (1989) provide introductions to the theory of structural equation models. Many methodological issues relating to the construction and testing of structural equation models can be found at "SEMNET" at www.gsu.edu/∼mkteer/semnet.html.[11]

Figure 24.2 contains an example of a latent variable SEM. The original data set came from Mardia, Kent, and Bibby (1979).[12] The test scores for 88 students in five subjects (*Mechanics*, *Vector Algebra*, *Algebra*, *Analysis*, and *Statistics*) are the measured variables. The latent substantive variables are *Algebra Skill*, *Vector Algebra Skill*, and *Real Analysis Skill.* The distribution of the test scores is approximately multivariate Normal. In model M of Figure 24.2, the free parameters are the linear coefficients a, b, c, and d, as well as the variances and means of the error terms $\varepsilon_M, \varepsilon_V, \varepsilon_{Al}, \varepsilon_{An}, \varepsilon_S, \delta_{Al}, \delta_{An}$, and δ_V. (The coefficients in the structural equations for *Mechanics, Algebra*, and *Statistics* have been fixed at 1 to ensure identifiability, which is explained below.) Note that in the equations, we have used an assignment operator ":=" rather than the more traditional equals sign "=" to emphasize that the quantity on the right-hand side of the equation is not just equal to the random variable on

10. Glymour and Cooper (1999) provide a collection of articles that also covers many issues about causal inference with graphical models. The Web site www.ai.mit.edu/∼murphyk/Bayes/bnsoft.html describes the most popular software packages for graphical modeling. Robins (1986), and Van der Laan and Robins (2003) describe a nongraphical approach to causal inference based on Rubin's (1977) counterfactual approach to causal inference.

11. There are a number of statistical packages devoted to estimating and testing structural equation models. These include the commercial packages EQS (www.mvsoft.com), LISREL (www.ssicentral.com/lisrel/mainlis.htm), and CALIS, which is part of SAS (www.sas.com). EQS and LISREL also contain some search algorithms for modifying a given causal model. The statistical package R (www.r-project.org) also contains a "sem" package for estimating and testing structural equation models.

12. Analyses of these data are discussed in Whittaker (1990) and in Spirtes et al. (2000, chap. 6). Edwards (2000) points out some anomalous features of the data, indicating that they may have been preprocessed.

the left-hand side of the equation but also causes the random variable on the left-hand side of the equation. Thus, as Lauritzen (2001) suggests, it is more appropriate to call these *structural assignment models* rather than *structural equation models.*

Bayesian networks specify a joint distribution over variables with the aid of a DAG, whereas SEMs specify a value for each variable via equations. Superficially, they appear to be quite different, but they are not. An SEM contains information about both the joint probability distribution over the substantive variables and the causal relations between the substantive variables. The joint distribution of the error terms, together with the equations, determines the joint distribution of the substantive variables. In addition, each SEM is associated with a graph (called a *path diagram*) that represents the causal structure of the model and the form of the equations, where there is a directed edge from X to $Y (X \rightarrow Y)$ if X is a direct cause of Y, and there is a bi-directed edge between the error terms ε_X and ε_Y if and only if the covariance between the error terms is nonzero. In path diagrams, latent substantive variables are often enclosed in ovals. A DAG is a special case of a path diagram (without cycles or correlated errors). If the path diagram is a DAG, then an SEM is a special case of a Bayesian network, and it can be shown that the joint distribution factors according to equation (1), even when the equations are nonlinear. Any probability distribution represented by the DAG in Figure 24.2 satisfies the following factorization condition described by equation (1), where f is the density:

$$\begin{aligned} &f(\textit{mechanics, vector, algebra, analysis, statistics}) \\ &\quad = f\ (\textit{mechanics}|\textit{vector algebra skill}) \\ &\quad \times f(\textit{vector}|\textit{vector algebra skill}) \times f(\textit{algebra}| \\ &\quad \textit{algebra skill}) \times f(\textit{analysis}|\textit{real analysis skill}) \\ &\quad \times f(\textit{statistics}|\textit{real analysis skill}) \times f(\textit{vector} \\ &\quad \textit{algebra skill}|\textit{algebra skill}) \times f(\textit{real analysis} \\ &\quad \textit{skill}|\textit{algebra skill}) \times f(\textit{algebra skill}). \end{aligned}$$

Kiiveri and Speed (1982) first pointed out the connection between structural equation models and the factorization equation. If the path diagram is cyclic or contains correlated errors, the factorization condition does not in general hold, but other properties of graphical models do still hold in general of SEMs, as explained in Section 24.5.2.4.2.

24.3. Causality and Probability

To reliably draw causal conclusions from the frequencies of the values of random variables in a sample, we will need to employ some assumptions that relate causal relations to probability distributions. In this section, we will describe and discuss several such assumptions.

24.3.1. The Causal Markov Assumption

We have described both a causal and statistical interpretation of graphs. What is the relationship between these two interpretations? We make the following assumption (equivalent to an assumption stated informally in Kiiveri & Speed, 1982):

Causal Markov assumption (factorization). For a causally sufficient set of variables $\mathbf{V}$ in a population N, if an acyclic causal graph G represents the causal relations among $\mathbf{V}$ in N, then G also represents $P(\mathbf{V})$; that is,

$$P(\mathbf{v}) = \prod_{v \in \mathbf{V}} P(v|\mathbf{parents}(G, V)). \tag{4}$$

In the example of the causal DAG in Figure 24.1, the causal Markov assumption implies

$$\begin{aligned} &P(\textit{sex, iq, ses, pe, cp}) \\ &\quad = P(\textit{iq}|\textit{ses}) \times P(\textit{sex}) \times P(\textit{ses}) \\ &\qquad \times P(\textit{pe}|\textit{ses, iq, sex}) \times P(\textit{cp}|\textit{pe, ses, iq}). \end{aligned} \tag{5}$$

An equivalent way of stating the causal Markov assumption in terms of conditional independence relations is the following.

Causal Markov assumption (independence). For a causally sufficient set of variables $\mathbf{V}$ in a population N, if an acyclic causal graph G represents the causal relations among $\mathbf{V}$ in N, each vertex X in $\mathbf{V}$ is independent of the set of vertices that are neither parents nor descendants of X in G, conditional on the parents of X in G.

In the example of the causal DAG in Figure 24.1, the independence version of the causal Markov assumption implies that the conditional independence relations listed in equation (3) of Section 24.2.3.1 hold in the probability distribution P in population N.

The causal Markov assumption is implicit in much of the practice of structural equation modeling (without cycles or correlated errors). In an SEM with Gaussian error terms, $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$ if and only if for each $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$, the partial correlation of $\mathbf{X}$ and $\mathbf{Y}$ given $\mathbf{Z}$ (denoted $\rho(X, Y|\mathbf{Z})$) is equal to zero. Thus, causal analysis of linear Gaussian SEMs depends on the analysis of vanishing partial correlations and their consequences. Simon's (1954)

famous analysis of "spurious correlation" is precisely an application of the causal Markov assumption to explain correlated errors. The examples that Bollen (1989) gives of why a disturbance term for a variable X might be correlated with one of the causes of X other than sampling problems are all due to causal relations between the disturbance term and other causes of X. In the context of linear or nonlinear structural equation models, the assumption that causally unconnected error terms are independent entails the full causal Markov assumption. Spirtes et al. (2000, chap. 3) discuss the causal Markov assumption, as well as conditions under which it should not be assumed (e.g., if the correct causal graph is cyclic, then a different version of the assumption should be made).

The causal Markov assumption implies that there is a procedure for calculating the effects of manipulations of variables in Bayesian networks and SEMs. The ideas are explained in the next two subsections.

24.3.2. Calculating the Effects of Manipulations in Bayesian Networks

In a Bayesian network with causal DAG G, the effect of an ideal manipulation can be calculated according to the following rule. If the distribution prior to the manipulation is $P(\mathbf{v})$, and the distribution after the manipulation is $P(\mathbf{v}\|P'(\mathbf{S}))$, then

$$P(\mathbf{v}\|P'(\mathbf{s})) = P'(\mathbf{s}) \times \prod_{v\in\mathbf{v}\backslash\mathbf{s}} P(v|\mathbf{parents}(G, V)),$$

where $\mathbf{v}$ is a set of values of variables in $\mathbf{V}$, $\mathbf{s}$ is a set of values of variables in $\mathbf{S}$, $\mathbf{parents}(G, V)$ is a set of values of variables in $\mathbf{Parents}(G, V)$, and $\mathbf{v}\backslash\mathbf{s}$ is a set of values for variables that are in $\mathbf{V}$ but not in $\mathbf{S}$.[13] (A proof is given in Spirtes et al., 2000, chap. 3.) That is, in the original factorization of $P(\mathbf{V})$, one simply replaces

$$\prod_{s\in\mathbf{S}} P(s|\mathbf{parents}(G, S)),$$

with $P'(\mathbf{s})$, where $\mathbf{S}$ is the set of manipulated variables. The manipulation operation depends on what the correct causal graph is because for each $S \in \mathbf{S}$, G appears in the term $P(s|\mathbf{parents}(G, S))$. Also, because the value of $\mathbf{S}$ in the manipulation does not causally depend on the values of the parents of $\mathbf{S}$, the postmanipulation DAG that represents the causal structure does not contain any edges into $\mathbf{S}$. (More general kinds of manipulations do not have this latter property.)

To return to the flashlight example, the premanipulation causal DAG is *Switch* → *Light*, and the premanipulation distribution is

$$\begin{aligned}
&P(Switch = on, Light = on) = P(Switch = on)\\
&\quad \times P(Light = on|Switch = on) = 1/2 \times 1 = 1/2,\\
&P(Switch = off, Light = on) = P(Switch = off)\\
&\quad \times P(Light = on|Switch = off) = 1/2 \times 0 = 0,\\
&P(Switch = on, Light = off) = P(Switch = on)\\
&\quad \times P(Light = off|Switch = on) = 1/2 \times 0 = 0,\\
&P(Switch = off, Light = off) = P(Switch = off)\\
&\quad \times P(Light = off|Switch = off) = 1/2 \times 1 = 1/2.
\end{aligned}$$

Suppose that *Light* is manipulated to the distribution $P'(Light = off) = 1$. Then the postmanipulation distribution $P(switch, light\,\|\,P'(Light))$ is found by substituting $P'(light)$ for $P(light|switch)$ for each value *light* of *Light* and each value *switch* of *Switch*:

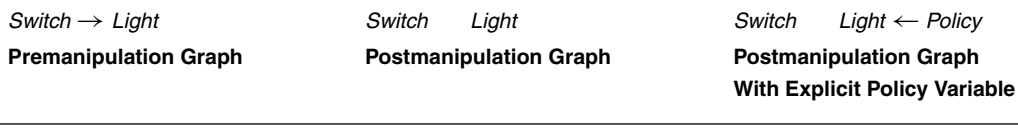
$$\begin{aligned}
&P(Switch = on, Light = on\|P'(Light)) = P(Switch\\
&\quad = on) \times P'(Light = on) = 1/2 \times 0 = 0,\\
&P(Switch = off, Light = on\|P'(Light)) = P(Switch\\
&\quad = off) \times P'(Light = on) = 1/2 \times 0 = 0,\\
&P(Switch = on, Light = off\|P'(Light)) = P(Switch\\
&\quad = on) \times P'(Light = off) = 1/2 \times 1 = 1/2,\\
&P(Switch = off, Light = off\|P'(Light)) = P(Switch\\
&\quad = off) \times P'(Light = off) = 1/2 \times 1 = 1/2.
\end{aligned}$$

In the postmanipulation distribution, *Switch* does not cause *Light,* and *Light* and *Switch* are independent. Hence, the postmanipulation graph that represents the postmanipulation distribution is formed by breaking all of the edges into *Light* and has no edge from *Switch* to *Light.* Although *Switch* and *Light* are symmetric in the premanipulation distribution $P(Light = light, Switch = switch)$, the effects of manipulating them are asymmetric because *Light* and *Switch* are not symmetric in the causal DAG. Manipulations in Bayesian networks are described in Spirtes et al. (2000, chaps. 3, 7), Pearl (2000), and Lauritzen (2001).

Spirtes et al. (2000, chap. 3) describe a representation of manipulations that explicitly includes a new cause of *Light* in the postmanipulation causal graph. The new cause is the exogenous *Policy* variable that has the value *off* in the premanipulation population and *on* in the postmanipulation population. This alternative representation shows that one of the assumptions that makes a manipulation "ideal" is that the cause of *Light* in the postmanipulation distribution (the *Policy*

13. For example, if $\mathbf{v} = \{Switch = on, Light = off\}$, and $\mathbf{s} = \{Light = off\}$, then $\mathbf{v}\backslash\mathbf{s} = \{Switch = on\}$.

Figure 24.3 Three Types of Causal Graphs



variable) is an exogenous variable that is a direct cause only of *Light* and hence is independent of all of the nondescendants of *Light* (by the causal Markov assumption).

The theory of manipulations presented here answers questions only about the effects of ideal manipulations. In some cases, someone implementing a policy may intend that an action be an ideal manipulation when in reality it is not. However, whether any particular action that is taken to manipulate a variable is ideal is not part of the theory but has to be answered outside of the theory.

24.3.3. Manipulations in SEMs

In SEMs, there is a different but equivalent representation of a manipulation. Suppose we were to manipulate the scores of all of the students by giving them the answers to the questions on the *Analysis* test before they take it. Applying the analysis of manipulations given in Strotz and Wold (1960), the effect of an ideal manipulation of the *Analysis* test score on the joint distribution can be calculated by replacing the structural equation for *Analysis* with a new structural equation that represents its manipulated value (or, more generally, the manipulated distribution of *Analysis*). In this example, the structural equation $Analysis := b \times Real\ Analysis\ Skill + \varepsilon_{An}$ would be replaced by the equation $Analysis := 100$. The postmanipulation distribution is just the distribution entailed by the distribution of the error terms together with the new set of structural equations. The model that results from the manipulation of *Analysis* has a path diagram of the manipulated population formed by breaking all of the edges into the manipulated variable. In this example, the edge from *Real Analysis Skill* to *Analysis* would be removed.

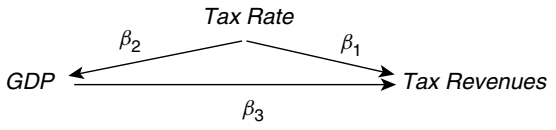
In both SEMs and causal Bayesian networks, a distinction can be drawn between the direct effect of one variable on another and the total effect of one variable on another. The total effect of A on B measures the change in B given a manipulation that makes a unit change in A. In the example of Figure 24.2, the total effect of *Algebra Skill* on *Vector* is given by $a \times c$, the product of the coefficient a associated with the edge from *Vector Algebra Skill* and the coefficient c associated with the edge from *Algebra Skill* to *Vector Algebra Skill.* In linear SEMs, the direct effect of A on B is a measure of how much B changes given a manipulation that makes a unit change in A, whereas all variables other than A and B are manipulated to hold their current values fixed. The direct effect of A on B is given by the coefficient associated with the edge from A to B, or zero if there is no such edge. For example, the direct effect of *Vector Algebra Skill* on *Vector* is a, and the direct effect of *Algebra Skill* on *Vector* is zero. In nonlinear systems, such as Bayesian networks, there is no single number that summarizes the effects of manipulations: The difference between $P(B)$ and $P(B\|A)$ can depend on both the value of B and the value of A, and even if it does not, the effect cannot be summarized by a single number. Manipulations in SEMs are described in Strotz and Wold (1960), Spirtes et al. (2000, chap. 3), Pearl (2000), and Lauritzen (2001).

24.3.4. Causal Faithfulness Assumptions

The causal Markov assumption states that causal graphs entail conditional independence relations, but it says nothing about what conditional independence relations entail about causal graphs. Observing independence between *Switch* and *Light* does not entail, by the causal Markov assumption alone, that *Switch* does not cause *Light* or *Light* does not cause *Switch.* It has been shown that in an SEM, given just the causal Markov assumption and allowing for the possibility of unmeasured common causes of measured variables, any direct effect of A on B is compatible with any covariance matrix among the measured variables (Robins, Scheines, Spirtes, & Wasserman, 2003). Hence, to draw conclusions about direct effects from observational data, some additional assumptions must be made. We will examine three such assumptions of increasing strength in this section.

If a probability distribution $P(\mathbf{V})$ is represented by a DAG G, then P is *faithful* to G if and only if every conditional independence relation that holds in $P(\mathbf{V})$ is entailed (by d-separation) by G—that is, holds for all values of the free parameters and not just some

Figure 24.4 Distribution is Unfaithful to DAG When $\beta_1 = -\beta_2 \times \beta_3$



values of the free parameters. (In Pearl, 2000, this is called *stability.*) Otherwise, $P(\mathbf{V})$ is *unfaithful* to G. (In linear Gaussian SEMs, this is equivalent to $P(\mathbf{V})$ being *faithful* to the path diagram of G if and only if every zero partial correlation that holds in $P(\mathbf{V})$ is entailed by d-separation by G.) In the rest of the section, we will discuss faithfulness in SEMs because the application to SEMs is somewhat simpler than the application to Bayesian networks. Some form of the assumption of faithfulness is used in every science and amounts to no more than the belief that an improbable and unstable cancellation of parameters does not hide real causal influences. When a theory cannot explain an empirical regularity, except by invoking a special parameterization, most scientists are uneasy with the theory and look for an alternative. Figure 24.4 shows an example of how an unfaithful distribution can arise. For example, suppose the DAG in Figure 24.4 represents the causal relations among standardized variables *Tax Rate, GDP*, and *Tax Revenues.* In this case, there are no vanishing partial correlation constraints entailed for all values of the free parameters. But $\rho(\textit{Tax Rate, Tax Revenues}) = \beta_1 + (\beta_2 \times \beta_3)$, so if $\beta_1 = -(\beta_2 \times \beta_3)$, then *Tax Rate* and *Tax Revenues* are uncorrelated, even though the DAG does not entail that they are uncorrelated (i.e., there is a path that d-connects *Tax Rate* and *Tax Revenues* conditional on the empty set, namely, the edge from *Tax Rate* to *Tax Revenues*). The SEM postulates a direct effect of *Tax Rate* on *Tax Revenues* (β_1) and a canceling indirect effect through the $GDP(\beta_2 \times \beta_3)$. The parameter constraint indicates that these effects *exactly* offset each other, leaving no total effect whatsoever.

It is clear from this example that unfaithful distributions greatly complicate causal inference. Because *Tax Rate* and *Tax Revenues* are completely uncorrelated, the alternative incorrect model Tax Rate → GDP ← Tax Revenue would tend to have a better goodness of fit statistic (because it is simpler and fits the sample correlation matrix almost as well). The violation of faithfulness described in the example only occurs for very special values of the parameters, that is, $\beta_1 = -(\beta_2 \times \beta_3)$. In general, the probability of the set of free parameter values for any DAG that lead to unfaithful distributions is zero, for any "smooth" prior probability distribution[14] (e.g., Normal, exponential, etc.) over the free parameters. This motivates the following Bayesian assumption. (The methods for and consequences of assigning prior probabilities to causal graphs and parameters to perform Bayesian inferences are described in more detail in Section 24.4. Although we state these assumptions for SEMs for convenience, there are more general versions of these assumptions that apply to Bayesian networks more generally.)

Causal faithfulness prior assumption. Suppose that there is a population N with distribution $P(\mathbf{V})$ and a DAG G that represents the causal relations in N. If X and Y are d-connected conditional on $\mathbf{Z}$ in G (i.e., G does not entail that $\rho(X, Y|\mathbf{Z}) = 0$ for all values of the free parameters), then the set of free parameter values for which $\rho(X, Y|\mathbf{Z}) = 0$ has prior probability zero.

This assumption is implicitly made by any Bayesian who has a prior over the parameters taken from the usual families of distributions. Of course, this argument is not relevant to those who reject Bayesian arguments or to Bayesians who place a prior over the parameters that are not "smooth" and assign a nonzero probability to violations of faithfulness.

A stronger version of the causal faithfulness prior assumption that does not require acceptance of the existence of prior probability distributions is the following.[15]

Causal faithfulness assumption (SEMs). Suppose that there is a population N with distribution $P(\mathbf{V})$ and a DAG G that represents the causal relations in N. If X and Y are d-connected conditional on $\mathbf{Z}$ in G (i.e., G does not entail that $\rho(X, Y|\mathbf{Z}) = 0$ for all values of the free parameters), then $\rho(X, Y|\mathbf{Z}) \neq 0$.

The causal faithfulness assumption is a kind of simplicity assumption. If a distribution P is faithful to a SEM M_1 without latent variables or correlated errors, and P also results from assigning values to the free parameters of another SEM M_2 to which P is not faithful, then M_1 has fewer free parameters than M_2.

The faithfulness assumption limits the SEMs considered to those SEMs in which population constraints are entailed by graphical structure, rather than by particular values of the parameters. Causal faithfulness

14. That is, it is absolutely continuous with respect to the Lebesgue measure.

15. This is a stronger assumption because it eliminates all parameters that lead to violations of faithfulness from the sample space, instead of simply leaving them in the sample space and assigning them prior probability zero.

should not be assumed when there are deterministic relationships among the substantive variables or equality constraints on free parameters because either of these can lead to violations of the assumption.

An equivalent formulation of the causal faithfulness assumption states that if $\rho(X, Y|\mathbf{Z}) = 0$, then the true causal graph contains no d-connecting path between X and Y conditional on $\mathbf{Z}$. The causal faithfulness assumption, as stated, has implications only for cases in which a partial correlation is exactly zero. It is compatible with a partial correlation being arbitrarily small, whereas an edge coefficient (which is the strength of a d-connecting path consisting of a single edge) is arbitrarily large. The following stronger version of the causal faithfulness assumption eliminates this latter possibility. Let the strength of a d-connecting path be the product of the edge coefficients in the path times the product of the edges in paths from colliders to members of the conditioning set.

Strong causal faithfulness assumption (SEMs). Suppose that there is a population N with distribution $P(\mathbf{V})$ and a DAG G that represents the causal relations in N. If $\rho(X, Y|\mathbf{Z})$ is small, then there is no strong d-connecting path between X and Y conditional on $\mathbf{Z}$.

This statement could be made precise in several ways. One way in which it could be made precise is to assume that the strength of a d-connecting path between X and Y conditional on $\mathbf{Z}$ is no more than some constant k times $\rho(X, Y|\mathbf{Z})$. So the strong causal faithfulness assumption is really a family of assumptions, indexed by k.

Unlike the causal faithfulness assumption, violations of the strong causal faithfulness assumption are not probability zero for every "smooth" prior over the parameters. However, common modeling practices suggest that modelers often implicitly assume some version of a strong causal faithfulness assumption. For example, it is often the case that in causal modeling in various domains, a large number of measured variables $\mathbf{V}$ are reduced by regressing some variable of interest Y on the other variables and eliminating from consideration those variables that have small regression coefficients. Because (for standardized variables) a small regression coefficient of Y when X is regressed on all variables in $\mathbf{V}$ (except for X itself) entails that $\rho(X, Y|\mathbf{V}\backslash\{X, Y\})$ is small, this amounts to assuming that a small partial correlation is evidence for a small linear coefficient of X in the structural equation for Y.

The various forms of the causal faithfulness assumption are described and discussed in Spirtes et al. (2000). We will not further discuss the plausibility of the assumptions here, but we will trace out the consequences of each of these assumptions.

24.4. Model Estimation, Causal Inference and Consistency

One goal of causal inference is to infer the correct causal structure, that is, the correct causal graph or some set of graphs containing the correct causal graph. We will refer to this as *graphical model estimation.* A second goal is to infer the effect of a manipulation, which typically is a function of the graphical model and the values of its free parameters. This is an estimation of (functions of) the free parameters. The estimation of graphical models and the estimation of parameters in graphical models are customarily treated as entirely different problems, but formally, they are essentially the same problem: to use data to gain approximate information from among a vast space of possibilities consistent with prior knowledge. Techniques of graphical model estimation closely parallel approaches to parameter estimation. The first virtue of a "point" estimation procedure of any kind is that, in the long run, it certainly converges to the true value of whatever feature—parameter value or graphical model—is to be estimated. We distinguish three such "consistency" properties of estimators.

24.4.1. The Classical Framework

In the classical framework, an estimator $\hat{\theta}_n$ is a function that maps samples of size n into real numbers. An estimator is a *pointwise consistent* estimator of a quantity θ (e.g., the average effect of a manipulation of X on Y) if, for each possible value of θ, in the limit as the sample size approaches infinity, the probability of the distance between the estimator and the true value of θ being greater than any fixed finite value approaches zero. More formally, let O^n be a sample of size n of the observed variables $\mathbf{O}$, $\Omega(G)$ be the set of probability distributions that arise from assigning legal values to the free parameters of DAG G, Γ be some set of DAGs, and $\theta(P, G)$ be some causal parameter of interest (that is a function of the distribution P and the DAG G). Let $\Omega\Gamma = \{(P, G): G \in \Gamma, P \in \Omega(G)\}$ (i.e., the set of all DAG-legal parameter pairs) and $d[\hat{\theta}_n(O^n), \theta(P, G)]$ be the distance between $\hat{\theta}_n(O^n)$ and $\theta(P, G)$. An estimator $\hat{\theta}$ is *pointwise consistent* if, for all $(P, G) \in \Omega\Gamma$, for every $\varepsilon > 0$, $P^n(d[\hat{\theta}_n(O^n), \theta(P, G)] > \varepsilon) \rightarrow 0$; that is, the probability of the distance between the estimate and the true parameter being greater than any

fixed size ε greater than 0 approaches 0 as the sample size increases.

However, pointwise consistency is only a guarantee about what happens in the large sample limit, not at any finite sample size. Pointwise consistency is compatible with there being, at each sample size, some value of the causal parameter such that the probability of the estimator being far from the true value is high. Suppose that one were interested in answering questions of the following kind: What sample size is needed to guarantee that, regardless of the true value of the causal quantity, it is "improbable" that the estimator is "far" from the truth? *Improbable* and *far* are vague terms, but they can be made precise. *Improbable* can be made precise by choosing a positive real ε, such that any probability less than ε is improbable. *Far* can be made precise by choosing a positive real δ such that any distance greater than δ is "far." Then, the question can be rephrased as follows: What sample size is needed to guarantee that, regardless of the true value of the causal quantity, the highest probability that an estimator is more than δ away from the truth is less than ε? Given only pointwise consistency, the answer may be "infinite." However, a stronger form of consistency, *uniform consistency*, guarantees that answers to questions of the form given above are always finite for any given ε and δ greater than zero. More formally, an estimator $\hat{\theta}$ is *uniform consistent* if, for every $\varepsilon, \delta > 0$, there exists a sample size N, such that for all sample sizes $n > N$, $\sup_{(P,G)\in\Omega G} P^n(d[\hat{\theta}_n(O^n), \theta(P, G)] > \delta) < \varepsilon$.[16] There is no difficulty in extending these concepts to cover vectors of real numbers as well, as long as the distance between vectors is well defined.

24.4.2. The Bayesian Framework

In the Bayesian framework, one method of point estimation of a quantity θ proceeds by

1. assigning a prior probability to each causal graph,
2. assigning joint prior probabilities to the parameters conditional on a given causal graph,
3. calculating the posterior probability of θ (which we assume to be a function of the posterior probabilities of the graphs and the graph parameter values),
4. turning the posterior probability over the average effect of the manipulation into a point estimate by returning the value of θ that has the highest posterior probability.

16. See Bickel and Doksum (2001).

Note that such an estimator is a function not only of the data but also of the prior probabilities and can have a weaker sense of consistency than pointwise consistency. If the set of causal models (graph-probability distribution pairs) for which the estimator converges in probability to the correct value has a prior probability of 1, then we will say that it is *Bayes consistent* (with respect to the given set of priors). Because a pointwise consistent estimator converges in probability to the correct value for all causal models in the sample space, pointwise consistency entails Bayes consistency.

We will explain under what conditions there are—and are not—estimation procedures with these consistency properties and also describe some open issues. We are concerned with three kinds of estimation: values of parameters in a model given the model, graphical models given background knowledge, and effects of manipulations on specific variables, that is, $P(x \| P'(\mathbf{Y}))$. In all the examples we discuss, we will make the causal Markov assumption. We also assume that there is a causally sufficient set of variables $\mathbf{V}$ that is jointly Normal or contains all discrete variables. In both the multivariate Normal and discrete cases, the quantity to be estimated, $P(X \| P'(\mathbf{Y}))$, is parameterized by a finite vector of real numbers. We do not always assume that all variables in $\mathbf{V}$ are observed. We will assume that the causal graph is acyclic unless explicitly stated otherwise. We also assume that there are no correlated errors, unless explicitly stated otherwise (this case will be discussed further in the section on latent variables). Unless otherwise noted, we assume the samples are independent and identically distributed. Some weakening of these data assumptions is possible without changing the basic results that follow.

24.4.3. Causal Inference Assuming the Measured Variables Are Causally Sufficient

First, we consider the case in which there is a causally sufficient set of variables $\mathbf{V}$ that are all measured.

24.4.3.1. Known Causal Graph

There are uniform consistent estimators of the free parameters of a causal model for multivariate Normal or discrete DAG models. In the case of multivariate Normal distributions, a uniform consistent maximum likelihood estimate of the edge coefficients can be obtained by regressing each variable on its parents in the causal DAG. In the case of discrete DAG models,

a uniform consistent maximum likelihood estimate of the parameters $P(v|\mathbf{parents}(G, V))$ can be obtained by using the relative frequency of v conditional on $\mathbf{parents}(G, V)$.

As we saw in Sections 24.3.2 and 24.3.3, $P(x \| P'(\mathbf{Y}))$ is a function of the parameters. It follows that there are uniform consistent estimates of $P(x \| P'(\mathbf{Y}))$. (To avoid some complications, we assume in the case of discrete variables that $P'(\mathbf{Y})$ does not assign a nonzero probability to any value of $\mathbf{Y}$ that has probability 0 in the unmanipulated population.)

24.4.3.2. Unknown Causal Graph

Given the causal Markov assumption, but none of the causal faithfulness assumptions, there are no Bayes, pointwise, or uniform consistent estimators of the direction of edges for any true, unknown causal graph, as long as the variables are dependent. This is because any (multivariate Normal or discrete) distribution can be represented by some submodel of a DAG in which every pair of vertices are adjacent, regardless of the orientation of the edges.

However, given any of the causal faithfulness assumptions, in many cases, some orientations of edges are incompatible with the distribution $P(\mathbf{V})$, and considerably more information about the causal structure, and hence about effects of ideal manipulations, can be reliably derived from samples from $P(\mathbf{V})$. This is explained in the next several subsections.

24.4.3.3. Distribution Equivalence

Consider the college plans example. There are a variety of ways of scoring how well such a discrete model fits a sample, which include $p(\chi^2)$, and the BIC or Bayesian information criterion[17] (Bollen & Long, 1993). The BIC assigns a score that rewards a model for assigning a high likelihood to the data (under the maximum likelihood estimate of the values of the free parameters) and penalizes a model for being complex (which, for causal DAG models without latent variables, can be measured in terms of the number of free parameters in the model). The BIC is also a good approximation to the posterior probability in the large sample limit.

However, to evaluate how well the data support this *causal* model, we need to know whether there are other *causal* models compatible with background knowledge that fit the data equally well. In this case, for each of the DAGs in Figure 24.5 and for *any* data set D, the two models fit the data equally well and receive the same score (e.g., $p(\chi^2)$ or BIC scores). Informally, G_1 and G_2 are **O**-*distribution equivalent* if any marginal probability distribution over the observed variables $\mathbf{O}$ generated by an assignment of values to the free parameters of graph G_1 can also be generated by an assignment of values to the free parameters of graph G_2 and vice versa.[18] If G_1 and G_2 have no latent variables, then we will simply say that G_1 and G_2 are *distribution equivalent.* If two distribution equivalent models are equally compatible with background knowledge and have the same degrees of freedom, the data do not help choose between them, so it is important to be able to find the complete set of path diagrams that are distribution equivalent to a given path diagram. (The two models in Figure 24.5 have the same degrees of freedom.)

As we will illustrate below, it is often far from obvious what constitutes the complete set of DAGs or path diagrams that are distribution equivalent to a given DAG or path diagram, particularly when there are latent variables, cycles, or correlated errors. We will call such a complete set an **O**-*distribution equivalence class.* (Again, if we consider only models without latent variables, we will call such a complete set a *distribution equivalence class.*) If it is the complete set of graphs without correlated errors or directed cycles (i.e., DAGs that are **O**-distribution equivalent), we will call it a *simple* **O**-*distribution equivalence class.*

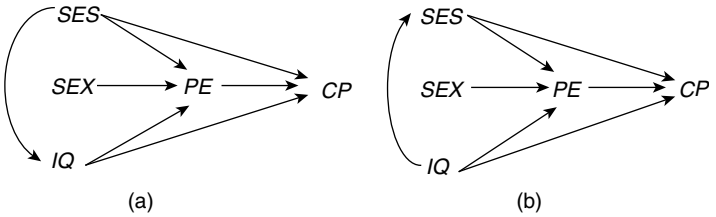
24.4.3.4. Features Common to a Simple Distribution Equivalence Class

An important question that arises with respect to simple distribution equivalence classes is whether it is possible to extract the features that the set of simple distribution-equivalent path diagrams has in common.

17. The Bayesian information criterion (BIC) for a directed acyclic graph (DAG) is defined as $\log P(D|\hat{\theta}_G, G) - (d/2)\log N$, where D is the sample data, G is a DAG, $\hat{\theta}_G$ is the vector of maximum likelihood estimates of the parameters for DAG G, N is the sample size, and d is the dimensionality of the model, which in DAGs without latent variables is simply the number of free parameters in the model. Because $P(G|D) \propto P(D|G)P(G)$, if $P(G)$ is the same for each DAG, the BIC score approximation for $P(D|G)$ can be used as a score for approximating $P(G|D)$.

18. For technical reasons, a more formal definition requires a slight complication. G is a *subgraph* of G' when G and G' have the same vertices, and G has a (not necessarily proper) subset of the edges in G'. G_1 and G_2 are **O**-*distribution equivalent* if, for every model M such that $G(M) = G_1$, there is a model M' with $G(M')$ that is a subgraph of G_2, and the marginal over $\mathbf{O}$ of $P(M')$ equals the marginal over $\mathbf{O}$ of $P(M)$, and for every model M' such that $G(M') = G_2$, there is a model M with $G(M)$ that is a subgraph of G_1, and the marginal over $\mathbf{O}$ of $P(M)$ equals the marginal over $\mathbf{O}$ of $P(M')$.

Figure 24.5 An Example of a Simple Distribution Equivalence Class



For example, each of the graphs in Figure 24.5 has the same adjacencies. The edge between *IQ* and *SES* points in different directions in the two graphs in Figure 24.5. However, $PE \rightarrow CP$ is the same in both members of the simple distribution equivalence class. This is informative because even though the data do not help choose between members of the simple distribution equivalence class, insofar as the data are evidence for the disjunction of the members in the simple distribution equivalence class, it is evidence for the orientation $PE \rightarrow CP$. In Section 24.4.3.6, we describe how to extract all of the features common to a simple distribution equivalence class of path diagrams.

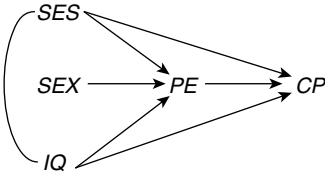
24.4.3.5. Distribution Equivalence for Path Diagrams Without Correlated Errors or Directed Cycles

Recall from Section 24.2.3.2 that a causal model is an ordered pair consisting of a causal graph and a probability distribution. If, for causal model M, there is another causal model M' with a different causal graph but the same number of degrees of freedom and the same marginal distribution over the measured variables in M, then $p(\chi^2)$ for M' equals $p(\chi^2)$ for M, and they have the same BIC scores. Such models are guaranteed to exist if there are models that have the same number of degrees of freedom and contain graphs that are distribution equivalent to each other. Theorem 1 (Spirtes et al., 2000, chap. 4; Verma & Pearl, 1990) shows how distribution equivalence can be calculated quickly. X is an *unshielded collider* in a DAG G if and only if G contains edges $A \rightarrow X \leftarrow B$, and A is not adjacent to B in G.

Theorem 1. For multivariate Normal distributions or discrete distributions, two causal models with directed acyclic causal graphs but no correlated errors are distribution equivalent if and only if they contain the same vertices, the same adjacencies, and the same unshielded colliders.

See also Stetzl (1986), Lee and Hershberger (1990), and MacCallum, Wegener, Uchino, and Fabrigar (1993) for discussions of model equivalence in structural equation models.

Figure 24.6 Pattern



24.4.3.6. Extracting Features Common to a Simple Distribution Equivalence Class

Theorem 1 is also the basis of a representation (called a pattern in Verma & Pearl, 1990) of an entire simple distribution equivalence class. The pattern that represents the set of DAGs in Figure 24.5 is shown in Figure 24.6.

A pattern has the same adjacencies as the DAGs in the simple distribution equivalence class that it represents. In addition, an edge is oriented as $X \rightarrow Z$ in the pattern if and only if it is oriented as $X \rightarrow Z$ in every DAG in the simple distribution equivalence class and as $X - 2$ otherwise. Meek (1995), Chickering (1995), and Andersson, Madigan, and Perlman (1995) show how to quickly generate a pattern that represents the simple equivalence class of a DAG from the DAG. Section 24.4.3.9.4 discusses the problem of constructing a causal pattern from sample data.

24.4.3.7. Calculating the Effects of Manipulations From a Pattern

The rules that specify which effects of manipulations can be calculated from a pattern and how to calculate them, as well as which effects of manipulations cannot be calculated from a pattern, are described in Spirtes et al. (2000, chap. 7). Here we give some examples without proof.

Suppose that it is known that the pattern in Figure 24.6 is the true causal pattern (i.e., the true causal DAG is a member of the simple distribution equivalence class represented by that pattern). The pattern represents the set of DAGs in Figure 24.5. The DAG in Figure 24.5b predicts that $P(iq \| P'(SES)) = P(iq)$ because IQ is not an effect of SES in that DAG. However, the DAG in Figure 24.5a predicts that $P(iq \| P'(SES)) \neq P(iq)$ because IQ is an effect of SES in Figure 24.5a. Hence, knowing only that the true causal DAG is represented by the pattern in Figure 24.6 does not determine a unique answer for the value of $P(iq \| P'(SES))$, and there are no consistent estimators (of any kind) of $P(iq \| P'(SES))$.

In contrast, both of the DAGs in Figure 24.5 predict that $P(pe|ses, iq \| P'(IQ)) = P(pe|ses, iq)$, where $P(pe|ses, iq \| P'(IQ))$ denotes the probability of pe conditional on ses and iq, after IQ has been manipulated to $P'(IQ)$. It follows that if it is known that the true causal pattern is the pattern in Figure 24.6, there are uniform consistent estimators of $P(pe|ses, iq \| P'(IQ))$.

Finally, there are conditional distributions that do change under manipulation but that can be calculated from quantities that do not change under manipulation. If it is known that the true causal pattern is the pattern in Figure 24.6, there are uniform consistent estimators of $P(cp|pe \| P'(PE))$.

$$P(cp|pe \| P'(PE)) = \sum_{IQ,\ SES} P(cp|pe,\ ses,\ iq) \times P(iq|ses) \times P(ses). \quad (6)$$

Given the Sewell and Shah (1968) data, and assuming that the pattern in Figure 24.6 is the correct pattern, the following are estimates of $P(cp|pe \| P'(PE))$:

$$\begin{aligned} P(CP = 0|PE = 0 \| P'(PE)) &= .095 \\ P(CP = 1|PE = 0 \| P'(PE)) &= .905, \\ P(CP = 0|PE = 1 \| P'(PE)) &= .484 \\ P(CP = 1PE = 1 \| P'(PE)) &= .516. \end{aligned}$$

24.4.3.8. Consistent Estimators of the Effects of Manipulations

Suppose that neither the true causal pattern nor the true causal DAG is given, that the only given data are samples from a jointly Normal probability distribution or discrete variables, and that the set of variables is known to be causally sufficient. Under what assumptions and conditions are there Bayes, pointwise, or uniform consistent estimators of the effects of manipulations?

If E_n is an estimator of some quantity Q, then under their standard definitions, Bayes, pointwise, and uniform consistency of E_n require that as the sample size n increases, E_n approaches Q, regardless of the true value of Q. Under this definition, there are no consistent estimators of any kind of effects of any manipulation, even given the strong causal faithfulness assumption. However, given the causal faithfulness prior assumption, the causal faithfulness assumption, or the strong causal faithfulness assumption, there are slightly weakened senses of Bayes, pointwise, and uniform consistency, respectively, under which there are consistent estimators of the effects of some manipulations. In the weakened sense, an estimator can return "don't know" as well as a numerical estimate, and a "don't know" estimate is considered to be zero distance from the truth. For an estimator to be nontrivial, there must be some values of Q for which, with probability 1, in the large sample limit the estimator does not return "don't know." From now on, we will use *Bayes consistent estimator, pointwise consistent estimator*, and *uniform consistent estimator* in this weakened sense.

Suppose that we are given a causally sufficient set of multivariate Normally distributed or discrete variables $\mathbf{V}$ and the causal Markov assumption but not any version of the causal faithfulness assumption. If the time order is known, and there are no deterministic relations among the variables, then there are uniform consistent estimators of any manipulation. If the time order is not known, then for any X and Y that are dependent, regardless of what the true probability distribution $P(\mathbf{V})$ is, there are no Bayes, pointwise, or uniform consistent estimators of $P(y \| P'(X))$. This is because there is always a DAG compatible with the causal Markov assumption in which X is a cause of Y and another DAG in which X is not a cause of Y.

Table 24.2 summarizes the results reviewed above. In all cases, it is assumed that the causal Markov assumption is true, that there are no deterministic relations among variables, and that all distributions are multivariate Normal or all variables are discrete. Some combinations of conditions are missing because the strong causal faithfulness assumption entails the causal faithfulness assumption, which entails the causal faithfulness prior assumption. The first four columns are combinations of assumptions that are possible, and the last three columns give the consequences of those assumptions. The "$\Leftarrow$" symbol marks entailment relations among the assumptions and the results. Not surprisingly, the stronger the version of causal faithfulness that is assumed, the stronger the sense of consistency that can be achieved.

Table 24.2 Existence of Estimator Under Different Assumptions: Nonlatent Case

	Assumptions			*Existence Results*		
Time Order	*Causal Faithfulness Prior* $\Leftarrow$	*Causal Faithfulness* $\Leftarrow$	*Strong Causal Faithfulness*	*Existence of Bayes Consistent* $\Leftarrow$	*Existence of Pointwise Consistent* $\Leftarrow$	*Existence of Uniform Consistent*
No	No	No	No	No	No	No
No	Yes	No	No	Yes	No	No
No	Yes	Yes	No	Yes	Yes	No
No	Yes	Yes	Yes	Yes	Yes	Yes
Yes	No	No	No	Yes	Yes	Yes

We will describe the construction of consistent estimators of manipulations in Section 24.4.3.9.5. Even given the strong causal faithfulness assumption, because all of the DAGs represented by a given pattern are distribution equivalent, only the correct causal pattern can be pointwise consistently estimated in the large sample limit. So any consistent estimator of the effects of manipulations is sometimes going to return "don't know." In general, consistent estimators return numerical estimates (as opposed to "don't know") whenever the value of the manipulation is a function of the true causal *pattern* (as opposed to the true causal DAG) and the true distribution (as described in Section 24.4.3.7). The results in Table 24.2 about the causal faithfulness assumption are proved in Robins et al. (2003), and the results about a version of the strong causal faithfulness assumption are proved in Spirtes et al. (2000, chap. 12) and Zhang and Spirtes (2003).

In general, the consistency of the estimators described in this chapter applies only to a limited class of models (represented by directed acyclic graphs) and a limited class of distributional families (multivariate Normal or discrete). In addition, they do not always use all available background knowledge (e.g., parameter equality constraints). How well an estimator performs on actual data depends on at least five factors:

1. the correctness of the background knowledge input to the algorithm,
2. whether the causal Markov assumption holds,
3. which of the strong causal faithfulness assumptions (indexed by k) holds,
4. whether the distributional assumptions made by the statistical tests of conditional independence hold,
5. the power of the conditional independence tests used by the estimators.

Each of these assumptions may be incorrect in particular cases. Hence, the output of the estimators described in this chapter should be subjected to further tests wherever possible. However, the problem is made even more difficult because even under the strong causal faithfulness assumption, for computational reasons, it is not known how to probabilistically bound the size of errors. It is possible to perform a "bootstrap" test of the stability of the output of an estimation algorithm by running it multiple times on samples drawn with replacement from the original sample. However, although this can show that the output is stable, it does not show that the output is close to the truth because the probability distribution might be unfaithful, or very close to unfaithful, to the true causal graph. We recommend, as well, running search procedures on simulated data of the same size as the actual data, generated from a variety of initially plausible models. The results can give an indication of the probable accuracy of the search procedure and its sensitivity to search parameters and to the complexity of the data-generating process. Of course, if the actual data are generated by a radically different structure, or if the actual underlying probability distribution or sampling characteristics do not agree with those in the simulations, these indications may be misleading. Also, it should be kept in mind that even when a model suggested by an estimator fits the data very well, it is possible that there are other models that will also fit the data well and are equally compatible with background knowledge, particularly when the sample size is small.

24.4.3.9. Consistent Estimation of Causal Models

In this section, we discuss some of the methodological implications of the results presented in the previous sections for estimation of models (or, in more common terminology, model selection or model search). The proper methodology depends on whether one is interested in constructing statistical models (used for calculating conditional probabilities) or in constructing

causal models (used for calculating manipulations of the true causal pattern). Throughout we will use the college plans data as an example. Analogous methodological conclusions can be drawn for SEMs. At this point, we will not consider the issues of how the variables in the college plans data set were constructed or impose any constraints on the models drawn from background knowledge. Such further considerations could be incorporated into the search algorithms discussed below and could alter their output.

The estimation of statistical models and of causal DAGs or patterns calls for very different methodologies because a good statistical model could be a very bad causal model. (For example, two DAGs represented by the same pattern may be equally good statistical models, but only one of them is a good causal model.)

The problem is that constructing estimators of causal DAGs or patterns is very difficult for several reasons. Even if latent variables are excluded, the space of DAGs (or patterns) is enormous: The number of different models grows super-exponentially with the number of variables. Of course, background knowledge, such as time order, can vastly reduce the space. Nevertheless, even given background knowledge, the number of a priori plausible alternatives is often orders of magnitude too large to search by hand.

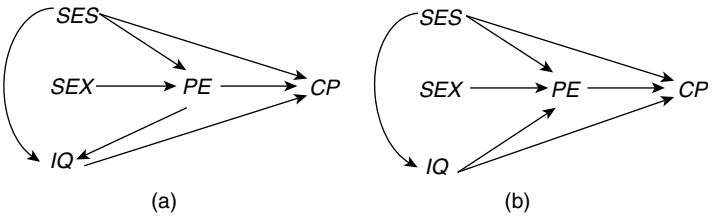
24.4.3.9.1. Estimation of statistical models. Suppose that a model of the college plans data is to be used to predict the value of *CP* from the other observed variables. One way to do this is to estimate $P(cp|sex, iq, ses, pe)$ and choose the value of *CP* with the highest probability. The relative frequency of *CP* conditional on *sex, iq, ses*, and *pe* in a random sample is a uniform consistent estimator of $P(cp|sex, iq, ses, pe)$. If the sample size is large, then the relative frequency will be a good estimator of $P(cp|sex, iq, ses, pe)$; however, if the sample size is small, then it typically will not be a good estimator because the number of sample points with fixed given values for *SEX, IQ, SES*, and *PE* will be small or possibly zero, and the estimator will have very high variance and a high mean squared error. (If the variables were continuous, the analogous operation would be to regress *CP* on *SEX, IQ, SES*, and *PE*.) A number of machine learning techniques, including variable selection algorithms, neural networks, support vector machines, decision trees, and nonlinear multiple regression, can be applied to obtain a prediction rule (see Mitchell, 1997). Once the statistical model is constructed, it can be evaluated in several different ways. For example, the sample can be initially divided into a training set and a test set. Then the model can be constructed on a training set, and the mean squared error of predictions can be calculated on the test set. There are also a variety of other cross-validation techniques that can be used to evaluate models. If several different models are constructed, the one with the smallest mean squared error on the test set can be chosen. Note that it does not matter if there are several different statistical models that predict *CP* equally well: In that case, any of them can be used because the goal is not to identify causes of *CP* but only to predict its value. If the goal is to predict *CP* from *PE* alone, then the sample size is large enough that the relative frequency of *CP* conditional on *PE* is a good estimator of $P(cp|pe)$.

24.4.3.9.2. Bayesian estimation of causal DAGs. In the ideal Bayesian framework, a prior probability is assigned over the space of causal DAGs and over the values of the free parameters of each DAG, and then the posterior probability of each DAG is calculated from the data. To turn this into a point estimate of a causal DAG, we can output the causal DAG with the highest posterior probability. In practice, it requires too much computation to calculate posterior probabilities, and we settle for calculating ratios of posterior probabilities of alternative DAGs. (There is no reason in principle why a theory of Bayesian estimation of causal patterns could not also be developed.)

Under the family of priors ("BDe priors") described in Heckerman (1998) (that satisfy the causal faithfulness prior assumption), asymptotically with probability 1, the posterior of the true causal DAG will not be smaller than the posterior of any other DAG. If every DAG has a nonzero prior probability, to be Bayes consistent, a point estimator based on choosing the DAG with the highest probability has to output "don't know" unless the DAG with the highest probability is the sole member of a simple distributional equivalence class. That is because under the family of BDe priors, different DAGs represented by the same pattern will typically all have nonzero posterior probabilities, even in the limit.

There are a number of computational difficulties associated with calculating posterior probabilities over either the space of causal DAGs or the space of causal patterns. Because there are a huge number of possible DAGs, it is a nontrivial problem to assign priors to each causal DAG and to the parameters for each causal DAG. Heckerman (1998) discusses techniques by which this can be accomplished. A collection of articles about learning graphical models, including the Bayesian approach, is given in Jordan (1998).

Figure 24.7 Bayesian Search Output (Assuming No Latent Common Causes)



It is computationally impossible in practice to calculate the posterior probability for a single causal DAG, let alone all causal DAGs. However, techniques have been developed to quickly calculate the ratio of posterior probabilities of any two DAGs. As an approximation of a Bayesian solution, then, it is possible to search among the space of DAGs (or the space of patterns) and output the DAGs (or patterns) with the highest posterior probabilities. (A variation of this is performing a search over the space of DAGs but turning each of the DAGs output into the pattern that represents the DAG as a final step, to determine whether the point estimate of the effect of a manipulation is Bayes consistent.) A wide variety of searches from the machine learning literature have been proposed as search algorithms for locating the DAGs with the highest posterior probabilities. These include simple hill climbing (at each stage choosing the DAG with the highest posterior probability from among all of the DAGs that can be obtained from the current best candidate DAG by a single modification), genetic algorithms, simulated annealing, and so on (for a summary, see Spirtes et al., 2000, chap. 12).

As an example, consider again the college plans data. Under the assumption of no latent common causes, with *SEX* and *SES* having no parents and *CP* having no children, and under a variety of different priors, the two DAGs with the highest prior probability (which differ in the direction of the edge between *PE* and *IQ*) that were found are shown in Figure 24.7a, b. The DAG in Figure 24.7b is the same as the DAG in Figure 24.5a. The DAG in Figure 24.7a, however, has a posterior probability that is on the order of 10^{10} times more probable than the DAG in Figure 24.7b. This is because although the DAG in Figure 24.7b fits the data better, the DAG in Figure 24.7a is much simpler, having only 68 free parameters. (The large number of free parameters is due to the fact that the variables are discrete, and hence the free parameters are not the covariance matrix and the means, as in a multivariate Normal distribution, but the probability of each variable conditional on its parents. See Section 24.2.3.1.)

An interesting unresolved question is what the results of a score-based search would be if further constraints were imposed on the parameters (e.g., if the probability of *CP* conditional on *PE*, *SES*, and *IQ* were obtained from a logistic regression).

24.4.3.9.3. Score-based estimation of causal DAGs or patterns. For computational reasons, the full Bayesian solution of calculating the posterior probability of each DAG or the posterior probability of the effect of a manipulation cannot be carried out. The approximate Bayesian solution in effect uses the posterior probability as a way of assigning scores to DAGs, which can then be incorporated into a procedure that searches for the DAGs (or patterns) with the highest score. There are a variety of other scores that (assuming the causal faithfulness assumption) have the property that, in the large sample limit with probability 1, the true DAG will have a score that is not exceeded by any other DAG. See Heckerman (1998) and Bollen and Long (1993), who describe a number of different approaches to scoring models. As in the case of Bayesian inference, a variety of searches using these scores can be performed. Instead of outputting a DAG as the result of a score-based search, a pattern could be output by turning the DAG with the highest score into the pattern that represents it.[19] Chickering and Meek (2002) describe a pointwise consistent score-based search over the space of patterns for the correct causal pattern. In the worst case, it is too computationally intensive to carry out, but if the true graph is sparse, it can be carried out for at least dozens of variables.

19. Buntine (1996) provides an overview of different approaches to search over Bayesian networks. There are also many articles on this subject in the *Proceeding of the Conference on Uncertainty in Artificial Intelligence* (www.auai.org) and the *Proceedings of the International Workshop on Artificial Intelligence and Statistics.*

24.4.3.9.4. Constraint-based estimation of causal patterns. The PC algorithm is another example of a pointwise consistent estimator of causal patterns. It takes as input a covariance matrix or discrete data counts, distributional assumptions, optional background knowledge (e.g., time order), and a significance level, and outputs a pattern. The significance level cannot be interpreted as the probability of Type I error for the pattern output but merely as a parameter of the search. From simulation studies, it appears that it is best to set the significance level quite high for small sample sizes (e.g., .15 or .2 for sample size 100) and quite low for large sample sizes (e.g., .01 or .001 for sample size 10,000), with larger samples required for discrete models. The search proceeds by performing a sequence of conditional independence tests. (The name *constraint based* comes from the testing of constraints entailed by a pattern—in this case, conditional independence constraints.) The length of time that the algorithm takes to run depends on how many parents each variable has. In the worst case (where some variable has all the other variables as parents), the time it takes to perform the search grows exponentially as the number of variables grows. However, in some cases, where each variable has relatively few parents, it can perform searches on 100 measured variables or more. How large a set of causal models is represented by the output asymptotically depends on what the true causal DAG is. The output of the search is a pointwise consistent estimate of the true causal pattern under the causal Markov and causal faithfulness assumptions (if the significance level of the tests performed approaches zero as the sample size approaches infinity). However, it has been shown that there are no uniform consistent estimators of causal patterns under any of the causal faithfulness assumptions described in Section 24.3.4 (although there are uniform consistent estimates of the effects of manipulations under the strong causal faithfulness assumption).

One advantage of a constraint-based search algorithm is that it does not require any estimation of the free parameters of a model.[20] One disadvantage of a constraint-based algorithm is that it outputs only a single pattern and gives no indication of whether other patterns explain the data almost as well as the output pattern but represent very different casual graphs. A partial answer to this problem is to run the algorithm with different significance levels or to perform a bootstrap test of the output. In addition, the fit of one of the models represented by the pattern can be tested in the usual way, using a chi-square test. Such a test can be done either on the same data used in the search or preferably on data that were not input to the search algorithm.

The output of the PC algorithm on the college plans data (on significance levels ranging from .001 to .05) is the pattern in Figure 24.6. A bootstrap test of the PC algorithm (with significance level .001) produced the same model as in Figure 24.6 on 8 out of 10 samples. On the other 2 samples, the edge between *PE* and *CP* was not oriented.

The pattern output by the PC algorithm represents the second most probable DAG found in Heckerman (1998), and given the restrictions assumed by Heckerman, this DAG is the only DAG represented by the pattern.

Although the set of causal models represented by the pattern in Figure 24.6 were the best models without latent variables found by the PC algorithm, it can be shown that the set of conditional independence relations judged to hold in the population by performing conditional independence tests are not faithful to any causal model without latent variables. We will discuss relaxing the "no latent variable assumption" imposed by the PC algorithm in Section 24.5.2.[21]

24.4.3.9.5. From estimators of causal models to estimators of the effects of manipulations. The overall strategy for consistently estimating the effects of a manipulation are illustrated in the following example of how to estimate $P(cp|pe\|P'(PE))$:

1. Use the data and some search algorithm to estimate the true causal pattern. For example, under the causal faithfulness assumption, the PC algorithm is a pointwise consistent estimator of causal patterns and outputs the pattern in Figure 24.6.
2. If the effect of the manipulation is not uniquely determined by the pattern (as described in

20. In some methodologies, in which the goal is to infer the correct causal graph or some set containing the correct causal graphs, the parameters are considered a kind of nuisance parameter that are needed to test the fit of a model but not of interest in themselves. See Mulaik and Millsap (2000).

21. Other applications of constraint-based causal inference algorithms are described in Glymour and Cooper (1999) and Spirtes et al. (2000). Biological applications are described in Shipley (2000). Some applications to econometrics are described in Swanson and Granger (1997), and the concept of causality in economics is described in Hoover (2001). HUGIN (www.hugin.com) is a commercial program that contains an implementation of a modification of the PC algorithm. TETRAD IV (www.phil.cmu.edu/projects/tetrad) is a free program that contains a number of search algorithms, including the PC and FCI algorithms (described in Section 24.5.2.1).

Section 24.4.3.7), output "don't know." In this case, $P(cp|pe\|P'(PE))$ is uniquely determined from the pattern.
3. Find a DAG that the pattern represents. In this case, for example, the pattern in Figure 24.6 represents the DAG in Figure 24.5a.
4. Consistently estimate the free parameters of the DAG. In this example, there are maximum likelihood estimates of the values of the free parameters that are uniformly consistent.
5. Use the estimate of the pattern and the estimate of the values of the free parameters to estimate $P(cp|pe\|P'(PE))$ using equation (6) of Section 24.4.3.7.

Under the strong causal faithfulness assumption, this procedure is a uniform consistent estimator of the effects of a manipulation despite the fact that the estimator of the true causal pattern is not uniformly consistent. (Informally, there are no uniform consistent estimators of causal patterns because of the difficulty of distinguishing between a causal model G and a causal model G' that is formed by adding arbitrarily weak edges to the DAG in G. The DAGs in G and G' are far apart in the sense that they contain DAGs that are represented by patterns that contain different edges; however, in terms of predicting the effects of manipulations, they are quite close to each other, as long as the extra edges are weak.)

One analogous method of estimation uses a Bayesian search algorithm to estimate the true causal DAG, rather than the PC algorithm to estimate the true causal pattern. In Heckerman (1998), the DAG with the highest posterior probability that was found is the one in Figure 24.7a (and there are no other DAGs in the simple distributional equivalence class that are compatible with the background assumptions made by Heckerman). It follows from the DAG in Figure 24.7a that

$$P(cp|pe\|P'(PE)) = \sum_{SES} P(cp|pe, ses) \times P(ses).$$

The following are the estimates for $P(cp|pe\|P'(PE))$ given the Sewell and Shah (1968) data, using maximum likelihood estimates of the free parameters of the DAG in Figure 24.7a:

$$P(CP = 0|PE = 0\|P'(PE)) = .080$$
$$P(CP = 1|PE = 0\|P'(PE)) = .920,$$
$$P(CP = 0|PE = 1\|P'(PE)) = .516$$
$$P(CP = 1PE = 1\|P'(PE)) = .484.$$

These estimates are close to the ones derived from using the PC algorithm.

24.5. Latent Variable Models

For the parametric families of distributions that we have considered, it is not necessary to introduce latent variables into a model to be able to construct uniform consistent estimators of conditional probabilities. Introducing a latent variable into a model may aid in the construction of consistent estimators that have smaller mean squared error on small samples. This is particularly true of discrete variable models, in which models such as one that has a DAG with one latent variable that is a parent of every measured variable (sometimes called a *latent class model*) has often proved useful in making predictions.

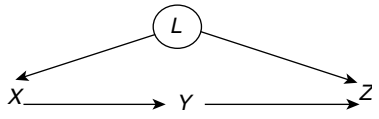
However, when a model is to be used to predict the effects of manipulations, then the introduction of latent variables into a graph is not merely useful for the sake of constructing low-variance estimators but can also be essential for constructing consistent estimators. Unfortunately, as described in this section, latent variables causal models, as opposed to causal models in which the measured variables are causally sufficient, face a number of extra problems that complicate estimation of the effects of manipulations.

24.5.1. Known Causal Graph

In some cases, the parameters of a DAG model with latent variables can still be consistently estimated despite the presence of latent variables. There are a number of algorithms for such estimations, including instrumental variables estimators, and iterative algorithms that attempt to maximize the likelihood. If the model parameters can be estimated, then because the effects of manipulations are functions of the model parameters, the effects of manipulations can also be consistently estimated. However, consistently estimating the model parameters of a latent variable model presents a number of significant difficulties.

1. It is not always the case that the model parameters are functions of the distribution over the measured variables. This is true of most factor-analytic models, for example. In that case, the model parameters are said to be "underidentified." For parametric families of distributions, whether or not a causal parameter is underidentified is essentially an algebraic problem. Unfortunately, known algorithms for determining whether a causal parameter is underidentified are too computationally expensive to be run on more than a few variables. There are a number of computationally feasible known necessary conditions for

Figure 24.8 The Backdoor Criterion



underidentification and a number of computationally feasible known sufficient conditions for underidentification (see Becker, Merckens, & Wansbeek, 1994; Bollen, 1989; Geiger & Meek, 1999).

2. Even when the model parameters are identifiable, the family of marginal distributions (over the observed variables) associated with a DAG with latent variables lacks many desirable statistical properties possessed by the family of distributions associated with a DAG without measured variables. For example, for SEMs with Normally distributed variables, there are no known general proofs of the asymptotic existence of maximum likelihood estimators of the model parameters (see Geiger, Heckerman, King, & Meek, 1999).

3. The estimation of model parameters is often done by iterative algorithms, which are computationally expensive, suffer from convergence problems, and can get stuck in local maxima (see Bollen, 1989).

There are also cases in which not all of the model parameters are identifiable, but the effects of some manipulations are identifiable (see Pearl, 2000; Pearl & Robins, 1995). A simple example is given by application of "the backdoor criterion" (Pearl, 2000) to the model in Figure 24.8, where X, Y, and Z are binary and measured, and L is ternary and unmeasured. In that case, the model parameters are unidentifiable. However, it can be shown that if $P'(y = 0) = 1$,

$$P(z|y\|P'(Y)) = \sum_X P(z|x, y)P(x).$$

Whether the effect of a given manipulation is identifiable for some parametric families of distributions is an algebraic question. However, the known general algorithms for calculating the solutions are too computationally expensive to be applied to models with more than a few variables. Special cases, which are computationally feasible, are given in Pearl (2000).

24.5.2. Unknown Causal Graph

We consider under what conditions and assumptions there are consistent estimators of the effects of manipulations when the measured set of variables may not be causally sufficient.

24.5.2.1. Distribution and Conditional Independence Equivalence

It is possible that two directed graphs entail the same set of conditional independence relations over a set of measured variables but are not **O**-distributionally equivalent, as long as at least one of them contains a latent variable, a correlated error, or a cycle. For example, the DAG in Figure 24.2 entails no conditional independence relations among only the measured variables $\mathbf{O} = \{$*Mechanics, Vector, Algebra, Analysis, Statistics*$\}$; all of the conditional independence relations that it entails involve conditioning on some latent variable. Any DAG G' with the same measured variables, but no latent variables, and in which every pair of measured variables are adjacent also entails no conditional independence relations among the measured variables. Hence, the DAG in Figure 24.2 and G' entail the same set of conditional independence relations among the measured variables (i.e., the empty set). However they are not **O**-distribution equivalent because the DAG in Figure 24.2 entails the nonconditional independence constraint $\rho($*Mechanics, Analysis*$) \times \rho($*Vector, Statistics*$) - \rho/($*Mechanics, Statistics*$) \times \rho($*Vector, Analysis*$) = 0$ for all values of its free parameters, whereas G' does not entail the constraint for all values of its free parameters. (Spearman, 1904, described these "vanishing tetrad constraints," and Glymour, Scheines, Spirtes, & Kelly, 1987, describe an algorithm that shows how to deduce such constraints from graphical structure.)

Although it is theoretically possible to determine when two SEMs or two Bayesian networks with latent variables are **O**-distributionally equivalent or to find features common to an **O**-distributional equivalence class, in practice, algorithms are not computationally feasible (Geiger & Meek, 1999) for models with more than a few variables. In addition, if an unlimited number of latent variables are allowed, the number of DAGs that are **O**-distributionally equivalent may be infinite. This implies that the strategy that was used to estimate the effects of manipulations when there were no latent variables cannot be carried forward unchanged to the case in which there may be latent variables.

We will describe two strategies to deal with the difficulty in identifying **O**-distribution equivalence classes.

The first strategy is to perform searches over a special class of graphical models, *multiple indicator models*, which simplifies the search process. This is described in Section 24.5.2.4.1.

The second strategy, described in this section, is not as informative as the computationally infeasible strategy of searching for **O**-distribution equivalence classes but is nevertheless correct.

If **O** represents the set of measured variables in graphs G_1 and G_2, then G_1 and G_2 are **O**-*conditional independence equivalent* if and only if they entail the same set of conditional independence relations among the variables in **O** (i.e., they have the same d-separation relations). It is often far from obvious what constitutes a complete set of graphs **O**-conditional independence equivalent to a given graph (Spirtes & Richardson, 1997; Spirtes, Richardson, Meek, Scheines, & Glymour, 1998). We will call such a complete set an **O**-*conditional independence equivalence class.* If it is the complete set of graphs without correlated errors or directed cycles (i.e., DAGs that are **O**-conditional independence equivalent), we will call it a *simple* **O**-*conditional independence equivalence class.*

A simple **O**-conditional independence equivalence class contains an infinite number of DAGs because there is no limit to the number of latent variables that may appear in a DAG.

24.5.2.2. Constraint-Based Search Algorithms

There are algorithms (e.g., PC) that give a pointwise consistent estimate of the simple conditional independence (and distributional) equivalence class of a DAG without latent variables by outputting a pattern that represents all of the features that the DAGs in the equivalence class have in common. Similarly, there is an algorithm (the FCI algorithm) that outputs a pointwise consistent estimate of the simple **O**-conditional independence equivalence class of the true causal DAG (assuming the causal Markov and causal faithfulness principles), in the form of a graphical structure called a partial ancestral graph that represents some of the features that the DAGs in the equivalence class have in common. The FCI algorithm takes as input a sample, distributional assumptions, optional background knowledge (e.g., time order), and a significance level and outputs a partial ancestral graph. Because the algorithm uses only tests of conditional independence among sets of observed variables, it avoids the computational problems involved in calculating posterior probabilities or scores for latent variable models.

Just as the pattern can be used to predict the effects of some manipulations, a partial ancestral graph can also be used to predict the effects of some distributions. Instead of calculating the effects of manipulations for which every member of the simple **O**-distribution equivalence class agrees, we can calculate the effects only of those manipulations for which every member of the simple **O**-conditional independence equivalence agrees. This will typically predict the effects of fewer manipulations than could be predicted given the simple **O**-distributional equivalence class (because a larger set of graphs has to make the same prediction), but the predictions made will still be correct.

Applying the FCI algorithm to the Sewell and Shah (1968) data yields output that predicts that $P(CP = 0|PE = 0\|P'(PE)) = P(CP = 0|PE = 0)$ and the following estimates:

$$P(CP = 0|PE = 0\|P'(PE)) = .063$$
$$P(CP = 1|PE = 0\|P'(PE)) = .937,$$
$$P(CP = 0|PE = 1\|P'(PE)) = .572$$
$$P(CP = 1PE = 1\|P'(PE)) = .428.$$

Again, these estimates are close to the estimates given by the output of the PC algorithm and the output of the Bayesian search algorithm. A bootstrap test of the output run at significance level .001 yielded the same results on 8 out of 10 samples. In the other 2 samples, the algorithm could not calculate the effect of the manipulation.

However, when the FCI algorithm is applied to the mathematical marks data set, the output, although a pointwise consistent estimate of the simple **O**-conditional equivalence class containing the true causal DAG, is not informative because it is not possible to predict the effects of any manipulation from the output; also the running time of the algorithm that constructs it is exponential in the number of variables. (The sample size for the mathematics marks data is quite small [88], and the actual output is from a set of conditional independence relations that would be entailed by the DAG in Figure 24.2 if *Algebra* were a very good measure of *Algebra Skill.*) We will discuss modifications of the FCI algorithm that make it useful for inferences about latent variable models such as those in Figure 24.2 in Section 24.5.2.4.1.

24.5.2.3. Bayesian and Score-Based Searches for Latent Variable Models

Score-based searches of latent variable models and Bayesian searches of latent variable models face similar difficulties. The search space is infinite, and a

Table 24.3 Existence of Estimator Under a Variety of Assumptions: Latent Case

	Assumptions			*Existence Results*		
Time Order	*Causal Faithfulness Prior* ⇐	*Causal Faithfulness* ⇐	*Strong Causal Faithfulness*	*Existence of Bayes Consistent* ⇐	*Existence of Pointwise Consistent* ⇐	*Existence of Uniform Consistent*
No	No	No	No	No	No	No
No	Yes	No	No	Yes	No	No
No	Yes	Yes	No	Yes	Yes	No
No	Yes	Yes	Yes	Yes	Yes	No
Yes	No	No	No	No	No	No

good strategy for deciding which parts of the space to search is not known. In principle, there is no problem in calculating the posterior probability of a latent variable model or its BIC score, but it is typically computationally infeasible. However, Heckerman (1998) describes some methods of approximation that are computationally feasible and applies them to several different latent variable models of the college plans data set (finding a latent variable model that is much more probable than any nonlatent variable model) (see also Rusakov & Geiger, 2003). Even for multivariate Normal or discrete latent variable models, the existence of maximum likelihood estimates in the large sample limit has not been demonstrated.

24.5.2.4. Bayes, Pointwise, and Uniform Consistent Estimators

Suppose that the only given data are samples from discrete variable or a multivariate Normal probability distribution and that the set of variables is not known to be causally sufficient. Under what assumptions and conditions are there pointwise or uniform consistent estimators of manipulations that are functions of the sample when the causal DAG is not given? The answers are provided in Table 24.3. Note that the only two lines that have changed from Table 24.2 are the last two lines, in which neither a known time order nor the strong causal faithfulness assumption entails the existence of uniform consistent estimators.

In each case, when the possibility of latent common causes is allowed, there are more cases in which the consistent estimators return "don't know" than if it is assumed that there are no latent common causes (see Robins et al., 2003; Spirtes et al., 2000, chap. 12; Zhang & Spirtes, 2003).

Whether there are other reasonable assumptions under which there exist uniform consistent estimators of the effects of manipulations when latent variables are not ruled out by background knowledge, as well as whether there are analogous results for other classes of distributions and various assumptions about background knowledge, are open questions.

24.5.2.4.1. Multiple-indicator models. The model shown in Figure 24.2 is an example of a multiple-indicator model. Multiple-indicator models can be divided into two parts. The causal relationships between the latent variables are called the structural model. The rest of the model is called the measurement model. The structural model is *Vector Algebra Skill* ← *Algebra Skill* → *Real Analysis Skill*, and the measurement model consists of the graph with all of the other edges. Typically, the structural model is the part of the model that is of interest.[22]

There are several strategies for using multiple-indicator models to make causal inferences that first attempt to find the correct measurement model, through a combination of background knowledge and search, and then use the measurement model to search for the correct structural model. Mulaik and Millsap (2000) describe a four-step process for testing multiple-indicator models. Anderson and Gerbing (1982) and Spirtes et al. (2000, chap. 10) describe methods for constructing measurement models that assume that it is known which measured variables (indicators) measure which latent variable and then detect those indicators that affect each other, or are affected by more than one latent variable. If the number of latents is not known, or it is not known which measured variables are indicators of which latents, it might be hoped that factor analysis could be used to create a correct measurement model. However, Glymour (1997) describes some simulation experiments in which factor analysis does not do well at

22. Sullivan and Feldman (1979) provide an introduction to multiple-indicator models. Lawley and Maxwell (1971) describe factor analysis in detail. Bartholomew and Knott (1999) provide an introduction to a variety of different kinds of latent variable models.

constructing measurement models or even at identifying the number of latent variables. See also the discussion in Mulaik and Millsap (2000) about problems with using factor analysis for choosing the number of latent variables, as well as Mulaik (1972) for the foundations of factor analysis.

If the measurement model is known, then it can be used to perform searches for the structural model in several different ways. For example, the measurement model can be used to perform tests of conditional independence among the latent variables. For example, to test whether $\rho(\textit{Vector Algebra Skill}, \textit{Real Analysis Skill} \mid \textit{Algebra Skill}) = 0$, a chi-square test comparing the model in Figure 24.2 with the model that differs only by the addition of an edge between *Vector Algebra Skill* and *Real Analysis Skill* can be performed. If the difference between the two models is not significant, then the partial correlation is judged to be zero. Thus, given the measurement models, the FCI or PC algorithm can be applied directly to estimate the structural model. At a significance level of .2, the PC algorithm produces the pattern *Vector Algebra Skill—Algebra Skill—Real Analysis Skill*, which is the pattern that represents the structure model part of the model shown in Figure 24.2.

24.5.2.4.2. Distribution and conditional independence equivalence for path diagrams with correlated errors or directed cycles. The representation of feedback using cyclic graphs, as well as the theory of inference to cyclic graphs from data, is not as well developed as for DAGs, except in special cases. There are general algorithms for testing distribution equivalence for multivariate Normal graphical models with correlated errors or directed cycles, but the known algorithms are generally computationally infeasible for more than a few variables (Geiger & Meek, 1999). For multivariate Normal variables, Spirtes (1995) and Koster (1996) proved that all of the conditional independence relations entailed by a graph with correlated errors and cycles are captured by the (natural extension of) the d-separation relation to cyclic graphs, and Pearl and Dechter (1996) and Neal (2000) proved an analogous result for discrete variables. However, Spirtes proved that given nonlinear relationships among continuous variables, it is possible for $\mathbf{X}$ to be d-separated from $\mathbf{Y}$ conditional on $\mathbf{Z}$ but for $\mathbf{X}$ and $\mathbf{Y}$ to be dependent conditional on $\mathbf{Z}$. There are computationally feasible algorithms for testing conditional independence equivalence for multivariate Normal graphical models with (a) correlated errors or (b) directed cycles but no latent variables. There are extensions of the PC algorithm to multivariate Normal graphs with cycles (Richardson, 1996), but there is no known algorithm for inferring graphs with both cycles and latent variables. Lauritzen and Richardson (2002) discuss the representation of feedback using not cyclic graphs but an extension of DAGs called *chain graphs*.

24.6. Some Common Errors in Model Specification

In this section, we examine the soundness of several practices in the social sciences sometimes used to draw causal inferences.

24.6.1. The Formation of Scales

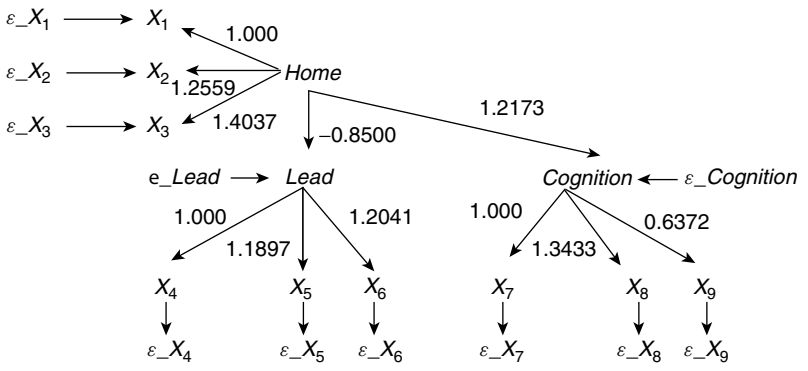
It is a common practice when attempting to discover the causal relations between latent variables to take all of the indicators of a given latent variable and average them together to form a "scale" (although this practice is also sometimes warned against; see, e.g., the discussion on SEMNET). This practice is often called *parceling* in the structural equation modeling literature. The scale variable is then substituted for the latent variable in the analysis. The practice is codified in the formal theory of measurement as *conjoint additive measurement,* and the following simulated example shows why this practice does not yield reliable information about the causal relationships between latent variables.

For the hypothetical model in Figure 24.9, hereafter the "true model," 2,000 pseudo-random data points were generated. (The numbers next to the edges are the linear coefficients associated with the edge.) All exogenous variable error terms are independent standard Normal variables.

Suppose the question under investigation is the effect of *Lead* on *Cognition*, controlling for the *Home* environment. Given the true model, the correct answer is 0; that is, *Lead* has no direct effect on *Cognition* according to this model. Consider first the ideal case in which we suppose that we can directly and perfectly measure *Home*, *Lead,* and *Cognition.* To test the effect of *Lead* on *Cognition,* we might regress *Cognition* on *Lead* and *Home.* Finding that the linear coefficient on *Lead* is $-.00575$, which is insignificant ($t = -.26$, $p = .797$), we correctly conclude that *Lead's* effect is insignificant.

Second, consider the case in which *Lead* and *Cognition* were directly measured but *Home* was measured with a scale that averaged X_1, X_2, and X_3, the indicators of *Home* in the true model: *Homescale* =

Figure 24.9 Simulated Lead Study

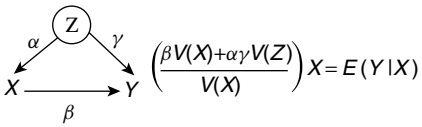


$(X_1 + X_2 + X_3)/3$. Further suppose we estimate the effect of *Lead* on *Cognition* by regressing *Cognition* on *Lead* and controlling for *Home* with the *Homescale* variable. We find that the coefficient on *Lead* is now $-.178$, which is significant at $p = .000$, and we incorrectly conclude that *Lead's* effect on *Cognition* is deleterious.

Third, consider the case in which *Lead, Cognition*, and *Home* were all measured with scales: *Homescale* $= (X_1 + X_2 + X_3)/3$, *Leadscale* $= (X_4 + X_5 + X_6)/3$, and *Cogscale* $= (X_7 + X_8 + X_9)/3$. Suppose we estimate the effect of *Lead* on *Cognition* by regressing *Cogscale* on *Homescale* and *Leadscale.* This gives a coefficient on *Leadscale* of $-.109$, which is still highly significant at $p = .000$, so we would again incorrectly conclude that *Lead's* effect is deleterious.

Finally, consider a strategy in which we build a scale for *Home* as we did above; that is, *Homescale* $= (X_1 + X_2 + X_3)/3$. Then, in the DAG in Figure 24.9, we remove the variables X_1, X_2, and X_3 and replace *Home* with *Homescale.* In one important respect, the result is worse. In this case, the regression coefficient of *Lead* on *Cognition*, controlling for the home environment (*Homescale*), is -0.137, which is highly significant at $t = -5.271$ and thus substantively in error as an estimate of the structural coefficient. However, the model that replaces *Home* with *Homescale* as a whole fits quite well ($\chi^2 = 14.57, df = 12, p = .265$), and all the distributional assumptions are satisfied, so nothing in the statistical treatment of this case would indicate that we have misspecified the model, even though the estimate of the influence of *Lead* is quite incorrect. Note, however, that the correct partial ancestral graph for the latent variables models (which contains undirected edges between each pair of *Lead, Homescale,* and *Cognition*) would correctly indicate that no causal conclusion about the effect of *Lead* on *Cognition* could be drawn.

Figure 24.10 The Problem of Confounding

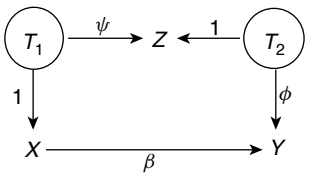


24.6.2. Regression

It is common knowledge among practicing social scientists that for the coefficient of X in the regression of Y on X to be interpretable as the direct effect of X on Y, there should be no "confounding" variable Z that is a cause of both X and Y (see Figure 24.10).

The coefficient from the regression of Y on X alone will be a consistent estimator only if either α or γ is equal to zero. Furthermore, observe that the bias term $\alpha\gamma V(Z)/V(X)$ (where $V(Z)$ is the variance of Z) may be either positive or negative and of arbitrary magnitude. However, the coefficient of X in the regression of Y on X and Z is a consistent estimator of β because $\text{Cov}(X, Y|Z)/V(X|Z) = \beta$. The danger presented by failing to include confounding variables is well understood by social scientists. Indeed, it is often used as the justification for considering a long "laundry list" of "potential confounders" for inclusion in a given regression equation. What is perhaps less well understood is that including a variable that is not a confounder can also lead to biased estimates of the structural coefficient. In the following example, Z may temporally precede both X and Y.

Figure 24.11 Estimates Biased by Including More Variables



In the DAG depicted in Figure 24.11, note that X and Y are unconfounded. Two unmeasured confounders, T_1 and T_2 (of X and Z, and Y and Z, respectively), are uncorrelated with one another. It can be shown that the coefficient of X in the regression of Y on X and Z is not a consistent estimate of β (unless $\rho(X, Z) = 0$ or $\rho(Y, Z) = 0$) and may even have a completely different sign. In the case where $\beta = 0$, the coefficient of X in the regression of Y on X will be zero in the population but will become nonzero once Z is included.

Statistical folklore often appears to suggest that it is better to include rather than exclude a variable from a regression (barring statistical problems such as small sample size). If the goal of a model is to predict the value of an unmeasured variable, rather than the result of a manipulation, this is sound advice (ignoring for the moment such statistical problems as small sample size or collinearity). However, if the purpose of a model is to describe causal relations or to predict the effects of a manipulation, this is not a theoretically sound practice. The notion that adding more variables is always advisable is perhaps given support by reference to "controlling for Z," with the implication being that controlling for Z eliminates a source of bias; in fact, though, it can add to the bias. The conclusion to be drawn from these examples is that there is no sense in which one is "playing safe" by including rather than excluding "potential confounders"; if they turn out not to be potential confounders, then this could change a consistent estimate into an inconsistent estimate. Of course, this does not mean that, on average, one is not better off regressing on more variables than fewer: Whether or not this is the case depends on the distribution of the parameters in the domain. Greenland (2003) argues on the basis of simulations of simple epidemiological models that in the domain of epidemiology, on average (apart from sampling problems), the bias reduction caused by conditioning on more variables is generally greater than the bias introduced by conditioning on more variables.

The situation is also made somewhat worse by the use of misleading definitions of *confounder*: Sometimes, a confounder is said to be a variable that is strongly correlated with both X and Y or even a variable whose inclusion changes the coefficient of X in the regression. Because, for sufficiently large $\rho(X, Z)$ or $\rho(Y, Z)$, Z in Figure 24.11 would qualify as a confounder under either of these definitions, it follows that under either definition, including "confounding variables" in a regression may make a hitherto consistent estimator inconsistent.

If Y is regressed on a set of variables $\mathbf{W}$, including X, we can ask the following: In which SEMs will the partial regression coefficient of X be a consistent estimate of the structural coefficient β associated with the $X \rightarrow Y$ edge? The coefficient of X is a consistent estimator of β if $\mathbf{W}$ does not contain any descendant of Y in G, and X is d-separated from Y given $\mathbf{W}$ in the DAG formed by deleting the $X \rightarrow Y$ edge from G.[23] If this condition does not hold, then for almost all instantiations of the parameters in the SEM, the coefficient of X will fail to be a consistent estimator of β. It follows directly from this that (almost surely) β cannot be estimated consistently via any regression equation if either there is an edge $X \leftrightarrow Y$ (i.e., ε_X and ε_Y are correlated) or if X is a descendant of Y (so that the path diagram is cyclic).

24.6.3. LISREL and Related Beam Search Procedures

Many editions of the LISREL program and similar programs (e.g., EQS) contain automated procedures for modifying an initial model to find an alternative that provides a better fit to the data. These procedures have several difficulties. If the causal graph of the initial model is not a subgraph of the true causal graph, they cannot give correct results. Because they rely on computationally demanding iterative fits of successive models, the procedure uses implicit heuristics (e.g., freeing at each stage only the single parameter that results in most improved fit)—a procedure known in computer science as one-step look ahead or beam search—and is often unsound. Not surprisingly, therefore, no proofs of their asymptotic correctness are available. Extensive simulation studies (Spirtes et al., 2000, chap. 11) have shown the procedures to be unreliable on large, finite samples obtained from known

23. This criterion is similar to Pearl's (1993) back-door criterion, except that the back-door criterion was proposed as a means of estimating the *total* effect of X on Y.

Gaussian SEMs. (The program manuals also suggest that the output of such searches may be unreliable and should be treated only as suggestions.)

24.6.4. Aggregation

In many subjects, including the social sciences, causal models are developed for variables that are aggregated over many units. Income averages, IQ averages, and so on are examples. The models developed for such variables may sometimes be used as models of causal relations at the individual level (e.g., that higher IQs cause higher incomes). Except in special cases, such inferences are fallacious. Conditional independence relations among aggregated variables may not indicate conditional independence relations among variables at the unit level or vice versa. Proofs of sufficient and necessary conditions for such inferences to be valid are given in Chu, Glymour, Scheines, and Spirtes (2003). (This is related to the "ecological fallacy," under which it is assumed that correlations at the aggregate level are the same as correlations among individuals.)

REFERENCES

Anderson, J., & Gerbing, D. (1982). Some methods for respecifying measurement models to obtain unidimensional construct measurement. *Journal of Marketing Research, 19*, 453–460.

Andersson, S., Madigan, D., & Perlman, M. (1995). *A characterization of Markov equivalence classes for acyclic digraphs* (Tech. Rep. No. 287). Seattle: Department of Statistics, University of Washington.

Bartholomew, D., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London: Edward Arnold.

Becker, P., Merckens, A., & Wansbeek, T. (1994). *Identification, equivalent models, and computer algebra.* San Diego: Academic Press.

Bentler, P. (1985). *Theory and implementation of EQS: A structural equations program.* Los Angeles: BMDP Statistical Software, Inc.

Bickel, P., & Doksum, K. (2001). *Mathematical statistics: Basic ideas and selected topics.* Hillsdale, NJ: Prentice Hall.

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

Bollen, K., & Long, J. (1993). *Testing structural equation models.* Newbury Park, CA: Sage.

Buntine, W. (1996). A guide to the literature on learning graphical models. *IEEE Transactions on Knowledge and Data Engineering, 8*, 195–210.

Cartwright, N. (1989). *Nature's capacities and their measurement.* New York: Oxford University Press.

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science.* New York: Cambridge University Press.

Chickering, D. (1995). A transformational characterization of equivalent Bayesian network structures. In P. Besnard & S. Hanks (Eds.), *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 87–98). San Mateo, CA: Morgan Kaufmann.

Chickering, D., & Meek, C. (2002). Finding optimal Bayesian networks. In A. Darwiche & N. Friedman (Eds.), *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002)* (pp. 94–102). San Francisco: Morgan Kaufmann.

Chu, T., Glymour, C., Scheines, R., & Spirtes, P. (2003). A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. *Bioinformatics, 19,* 1147–1152.

Cowell, R. (Ed.). (1999). *Probabilistic networks and expert systems.* New York: Springer-Verlag.

Edwards, D. (2000). *Introduction to graphical modelling* (2nd ed.). New York: Springer-Verlag.

Eells, E. (1991). *Probabilistic causality.* New York: Cambridge University Press.

Geiger, D., Heckerman, D., King, H., & Meek, C. (1999). On the geometry of DAG models with hidden variables. In D. Heckerman & J. Whittaker (Eds.), *Artificial Intelligence and Statistics 99* (pp. 76–85). San Francisco: Morgan Kauffman.

Geiger, D., & Meek, C. (1999). On solving statistical problems with quantifier elimination. In H. Prade & K. Laskey (Eds.), *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI—99)* (pp. 216–225). San Francisco: Morgan Kaufmann.

Glymour, C. (1997). Social statistics and genuine inquiry: Reflections on the bell curve. In B. Devlin, S. Fienberg, D. Resnick, & K. Roeder (Eds.), *Intelligence, genes and success* (pp. 257–280). New York: Springer-Verlag.

Glymour, C., & Cooper, G. (Eds.). (1999). *Computation, causation and discovery.* Cambridge: MIT Press.

Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). *Discovering causal structure.* San Diego: Academic Press.

Greenland, S. (2003). Quantifying biases in causal models: Classical confounding versus collider-stratification bias. *Epidemiology, 14,* 300–306.

Hausman, D. (1998). *Causal asymmetries.* New York: Cambridge University Press.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M. Jordan (Ed.), *Learning in graphical models* (pp. 301–354). Cambridge: MIT Press.

Hoover, K. (2001). *Causality in macroeconomics.* New York: Cambridge University Press.

James, R., Mulaik, S., & Brett, J. (1982). *Causal analysis: Assumptions, models, and data.* Beverly Hills, CA: Sage.

Jensen, F. (2001). *Bayesian networks and decision graphs.* New York: Springer-Verlag.

Jordan, M. (1998). *Learning in graphical models.* Cambridge: MIT Press.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions.* Thousand Oaks, CA: Sage.

Kiiveri, H., & Speed, T. (1982). Structural analysis of multivariate data: A review. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 209–289). San Francisco: Jossey-Bass.

Koster, J. (1996). Markov properties of non-recursive causal models. *Annals of Statistics, 24*, 2148–2177.

Lauritzen, S. (1996). *Graphical models.* Oxford, UK: Oxford University Press.

Lauritzen, S. (2001). Causal inference from graphical models. In O. Barnsdorff-Nielsen, D. Cox, & C. Kluppenlberg (Eds.), *Complex stochastic systems* (pp. 63–107). London: Chapman & Hall.

Lauritzen, S., Dawid, A., Larsen, B., & Leimer, H. (1990). Independence properties of directed Markov fields. *Networks, 20*, 491–505.

Lauritzen, S., & Richardson, T. (2002). Chain graph models and their causal interpretations (with discussion). *Journal of the Royal Statistical Society, Series B, 64*, 321–361.

Lawley, D., & Maxwell, A. (1971). *Factor analysis as a statistical method.* London: Butterworth.

Lee, S., & Hershberger, S. (1990). A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behavioral Research, 25*, 313–334.

MacCallum, R., Wegener, D., Uchino, B., & Fabrigar, L. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114*, 185–199.

Mardia, K., Kent, J., & Bibby, J. (1979). *Multivariate analysis.* New York: Academic Press.

Meek, C. (1995). Causal inference and causal explanation with background knowledge. In P. Besnard & S. Hanks (Eds.), *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 403–410). San Mateo, CA: Morgan Kaufmann.

Mitchell, T. (1997) *Machine learning.* Cambridge, MA: WCB/McGraw-Hill.

Mulaik, S. (1972). *The foundations of factor analysis.* New York: McGraw-Hill.

Mulaik, S. (1986). Toward a synthesis of deterministic and probabilistic formulations of causal relations by the functional relation concept. *Philosophy of Science, 53*, 313–332.

Mulaik, S., & Millsap, R. (2000). Doing the 4-step right. *Structural Equation Modelling, 7*, 36–73.

Neal, R. (2000). On deducing conditional independence from d-separation in causal graphs with feedback: The Uniqueness Condition is not sufficient. *Journal of Artificial Intelligence Research, 12*, 87–91.

Neapolitan, R. (1990). *Probabilistic reasoning in expert systems.* New York: John Wiley.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* San Mateo, CA: Morgan Kaufman.

Pearl, J. (2000). *Causality: Models, reasoning and inference.* New York: Cambridge University Press.

Pearl, J., & Dechter, R. (1996). Identifying independencies in causal graphs with feedback. In F. Jensen & E. Horvitz (Eds.), *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence* (pp. 240–246). San Francisco: Morgan Kaufmann.

Pearl, J., & Robins, J. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In P. Besnard & S. Hanks (Eds.), *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence* (pp. 444–453). San Francisco: Morgan Kaufmann.

Richardson, T. (1996). A discovery algorithm for directed cyclic graphs. In F. Jensen & E. Horvitz (Eds.), *Uncertainty in artificial intelligence: Proceedings of the Twelfth Conference* (pp. 462–469). San Francisco: Morgan Kaufmann.

Robins, J. (1986). A new approach to causal inference in mortality studies with sustained exposure periods: Application to control of the healthy worker survivor effect. *Mathematical Modeling, 7*, 1393–1512.

Robins, J., Scheines, R., Spirtes, P., & Wasserman, L. (2003). Uniform consistency in causal inference. *Biometrika, 90*, 491–515.

Rubin, D. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics, 2*, 1–26.

Rusakov, D., & Geiger, D. (2003). Automated analytic asymptotic evaluation of the marginal likelihood of latent models. In C. Meek & U. Kjorulff (Eds.), *Proceedings of the 19th Conference on Uncertainty and Artificial Intelligence* (pp. 501–508). San Francisco: Morgan Kaufmann.

Scheines, R. (2003). *Causal and statistical reasoning.* Retrieved from www.phil.cmu.edu/projects/csr/

Sewell, W., & Shah, V. (1968). Social class, parental encouragement, and educational aspirations. *American Journal of Sociology, 73*, 559–572.

Shafer, S. (1996). *The art of causal conjecture.* Cambridge: MIT Press.

Shipley, W. (2000). *Cause and correlation in biology.* Cambridge, UK: Cambridge University Press.

Simon, H. (1953). Causal ordering and identifiability. In W. Hood & T. Koopmans (Eds.), *Studies in econometric methods* (pp. 49–74). New York: John Wiley.

Simon, H. (1954). Spurious correlation: a causal interpretation. *JASA, 49,* 467–479.

Sosa, E., & Tooley, M. (Eds.). (1993). *Causation.* New York: Oxford University Press.

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology, 15*, 201–293.

Spirtes, P. (1995). Directed cyclic graphical representation of feedback models. In P. Besnard & S. Hanks (Eds.), *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 491–498). San Mateo, CA: Morgan Kaufmann.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge: MIT Press.

Spirtes, P., & Richardson, T. (1997). A polynomial time algorithm for determining DAG equivalence in the presence of latent variables and selection bias. In *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics* (pp. 489–500).

Spirtes, P., Richardson, T., Meek, C., Scheines, R., & Glymour, C. (1998). Using path diagrams as a structural equation modeling tool. *Sociological Methods and Research, 27*, 148–181.

Stetzl, I. (1986). Changing causal relationships without changing the fit: Some rules for generating equivalent LISREL models. *Multivariate Behavior Research, 21*, 309–331.

Strotz, R., & Wold, H. (1960). Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica, 28*, 417–427.

Sullivan, J., & Feldman, S. (1979). *Multiple indicators: An introduction.* Beverly Hills, CA: Sage.

Swanson, N., & Granger, C. (1997). Impulse response function based on a causal approach to residual orthogonalization in vector autoregression. *Journal of the American Statistical Association, 92*(437), 357–367.

Van der Laan, M., & Robins, J. (2003). *Unified methods for censored longitudinal data and causality.* New York: Springer-Verlag.

Verma, T., & Pearl, J. (1990). Equivalence and synthesis of causal models. In J. Lemmer, M. Henrion, P. Bonissone, & L. Kanal (Eds.), *Proceedings of the Sixth Conference on Uncertainty in AI* (pp. 220–227). Mountain View, CA: Association for Uncertainty in AI, Inc.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics.* New York: John Wiley.

Zhang, J., & Spirtes, P. (2003). Strong faithfulness and uniform consistency in causal inference. In C. Meek & U. Kjorulff (Eds.), *Proceedings of the 19th Conference on Uncertainty and Artificial Intelligence* (pp. 632–639). San Francisco: Morgan Kaufmann.

Name Index

SUBJECT INDEX

About the Editor

David Kaplan received his PhD in Education from UCLA in 1987. He is now Professor of Education and (by courtesy) Psychology at the University of Delaware. His research interests are in the development and application of statistical models to problems in educational evaluation and policy analysis. His current program of research concerns the development of dynamic latent continuous and categorical variable models for studying the diffusion of educational innovations. He can be reached at dkaplan@udel.edu, and his Web site is at www.udel.edu/dkaplan.

About the Contributors

Terry Ackerman is a professor in the department of Educational Research Methodology at the University of North Carolina at Greensboro. His research interests include both unidimensional and multidimensional IRT modeling, differential item/test functioning, and diagnostic assessment testing. He can be reached at taackerm@uncg.edu.

James E. Albert received his PhD in statistics from Purdue University in 1979 and joined Bowling Green State University as a professor in mathematics and statistics at that time. He is editor of *The American Statistician*. His academic interests include Bayesian inference, the analysis of sports data, and teaching statistics.

Frank M.T.A. Busing, MA, is a researcher at the Department of Psychology, Leiden University, The Netherlands. His research interests are multidimensional scaling, multidimensional unfolding, and software development. He can be reached at busing @fsw.leidenuniv.nl.

Hua-Hua Chang received his PhD in statistics from the University of Illinois at Urbana-Champaign in 1992. He is now Associate Professor in the Department of Educational Psychology at the University of Texas at Austin. His research interests include large-scale assessment, computerized adaptive testing, differential item functioning, and cognitive skills diagnosis.

Jamieson L. Duvall is a PhD candidate in the Department of Psychology at the University of Kentucky. His research interests include the influence of self-regulatory processes on individuals' decisions to consume alcohol and the use of latent variable modeling techniques in the analysis of multiwave data.

Gerd Gigerenzer is Director of the Max Planck Institute for Human Development, Berlin. His research interests include fast and frugal heuristics, decision making, risk, bounded rationality, and social rationality. His books include *Reckoning With Risk; Bounded Rationality: The Adaptive Toolbox* (with R. Selten); *Adaptive Thinking;* and *Simple Heuristics That Make Us Smart* (with P. Todd and the ABC Research Group.)

Clark Glymour is Alumni University Professor at Carnegie Mellon University and Senior Research Scientist at the Institute for Human and Machine Cognition in Pensacola, Florida.

Gregory R. Hancock earned his PhD in education from the University of Washington in 1991 and is currently Professor in the Department of Measurement, Statistics and Evaluation at the University of Maryland, College Park. He is past chair of the SEM special interest group of the American Educational Research Association. He also serves on the editorial board of a number of journals, including *Structural Equation Modeling: A Multidisciplinary Journal,* and teaches workshops all over the United States. He may be reached at ghancock@umd.edu.

Donald Hedeker is Professor of Biostatistics in the School of Public Health at the University of Illinois at Chicago. He received his PhD in quantitative psychology from the University of Chicago in 1989. His research interests focus on the development and dissemination of statistical methods for clustered and longitudinal data, with particular emphasis on mixed-effects models for categorical outcomes.

Larry V. Hedges is Professor of Sociology, Psychology, and Public Policy Studies at the University of Chicago. His research interests include the development of statistical methods for social sciences (particularly methods for meta-analysis), models for cognitive processes, the demography of academic

achievement, and education policy. He can be contacted at lhedges@uchicago.edu.

Willem J. Heiser, PhD, is Professor of Psychology, including Statistical Methods and Data Theory, at Leiden University, The Netherlands. His research interests are multidimensional scaling and unfolding, nonlinear multivariate analysis, and classification methods. He can be reached at heiser@fsw.leidenuniv.nl.

Rick H. Hoyle is Senior Research Scientist in the Terry Sanford Institute of Public Policy and the Department of Psychology: Social and Health Sciences at Duke University. He is Associate Director for Data Services in the Center for Child and Family Policy and Director of the Data Core in the Trans-Disciplinary Prevention Research Center, funded by the National Institute on Drug Abuse. He is a Fellow of the Society for the Psychological Study of Social Issues and Editor of the *Journal of Social Issues*. He is editor of *Structural Equation Modeling: Concepts, Issues, and Applications* and *Statistical Strategies for Small Sample Research* and author (with Harris and Judd) of *Research Methods in Social Relations* (7th ed.).

Valen E. Johnson received his PhD in statistics from the University of Chicago and was Professor at Duke University until 2002, when he became Professor of Biostatistics at University of Michigan. He is a fellow of the American Statistical Association. He is coauthor of *Ordinal Data Modeling* (with James Albert) and author of *Grade Inflation: A Crisis in College Education.* His research interests include ordinal and rank data modeling, Bayesian image analysis, Bayesian reliability modeling, convergence diagnostics for Markov chain Monte Carlo algorithms, Bayesian goodness-of-fit diagnostics, and educational assessment.

Spyros Konstantopoulos is an Assistant Professor of Education and Social Policy at Northwestern University. His research interests include the extension and application of statistical methods to issues in education, social science, and policy studies. His methodological work involves statistical methods for quantitative research synthesis (meta-analysis) and mixed effects models with nested structure (hierarchical linear models). His substantive work encompasses research on class size effects, technology (computer use) effects, teacher and school effects, program evaluation, labor market performance of young adults, and the social distribution of academic achievement.

Stefan Krauss is a Research Scientist at the Max Planck Institute for Human Development in Berlin. His research interests include Bayesian reasoning, statistical thinking, and educational psychology.

Jay Magidson is founder and president of Statistical Innovations, a Boston-based software and consulting company. He is widely published in various professional journals and was awarded a patent for an innovative graphical display. He is developer of the SI-CHAID and GOLDMineR programs and codeveloper (with Jeroen Vermunt) of the Latent GOLD and Latent GOLD Choice programs. His research interests include applications of advanced statistical modeling in the social sciences, especially latent class, discrete choice, and segmentation modeling.

Christopher Meek is a Senior Researcher in the Machine Learning and Applied Statistics Group at Microsoft Research. His main research interest is in statistical approaches to learning from data. His work has focused on methods for learning and applying probabilistic models to a variety of different domains including handwriting recognition, data mining, recommendation systems, and text classification. Since joining Microsoft Research, he has worked on many applications including data mining tools in SQL Server and Commerce Server, and handwriting recognition in the Tablet PC. Christopher received his PhD from Carnegie-Mellon University in 1997.

Jacqueline J. Meulman is Professor of Applied Data Theory in the Department of Educational Sciences, Data Theory Group, at Leiden University. She also has an adjunct position as Professor of Psychology in the Department of Psychology at the University of Champaign-Urbana. Her research interests include nonlinear methods for multivariate data analysis, clustering methods, statistical learning, and methods for data analysis in systems biology (genomics, proteomics, metabolomics). She can be reached at meulman@fsw.leidenuniv.nl, and her website is at www.datatheory.nl/pages/meulman.

Peter C. M. Molenaar is senior Professor, head of the Department of Psychological Methodology, and previous head of the Cognitive Developmental Psychology Group at the University of Amsterdam. His areas of statistical experience include dynamic factor analysis, applied nonlinear dynamics, adaptive filtering techniques, spectrum analysis, psychophysiological signal analysis, artificial neural network modeling, covariance structure modeling, and behavior genetic modeling. He has published widely in the above-mentioned areas, emphasizing applications to

cognitive development, brain-behavior relationships, brain maturation and cognition, genetic influences on EEG during the life span, and optimal control of psychotherapeutic processes.

Scott Morris is Associate Professor of Psychology at Illinois Institute of Technology, where he teaches courses in multivariate statistics and personnel selection. His primary research interests are decision making in employment contexts and meta-analysis. He received his PhD in Industrial/Organizational Psychology from the University of Akron in 1994.

Stanley A. Mulaik is Professor Emeritus of Psychology at the Georgia Institute of Technology, where he teaches courses in multivariate statistics, factor analysis, structural equation modeling, psychometric theory, and theories of personality. His research interests are in the philosophy of science, especially in connection with the philosophy of causality and objectivity. He received his PhD in clinical psychology from the University of Utah in 1963, was at the Psychometric Laboratory of the University of North Carolina at Chapel Hill from 1966-1970, and has been a member of the faculty at the Georgia Institute of Technology in Atlanta, Georgia, since 1970. He was semi-retired in 2000.

Bengt Muthén, PhD, is Professor at the Graduate School of Education and Information Studies at UCLA. He was the 1988–1989 President of the Psychometric Society. He currently has an Independent Scientist Award from the National Institutes of Health for methodology development in the alcohol field. He is one of the developers of the Mplus computer program, which implements many of his statistical procedures. His research interests focus on the development of applied statistical methodology in areas of education and public health.

Ratna Nandakumar is Professor in the School of Education at the University of Delaware, where she teaches courses in applied statistics and educational measurement. Her research areas are in the development, refinement, and application of statistical methodologies for educational and psychological test data. She has worked in the area of dimensionality and DIF assessments. She received her PhD from the University of Illinois at Urbana-Champaign in 1987.

Richard E. Neapolitan has been a researcher in the area of uncertainty in artificial intelligence, particularly Bayesian networks, since the mid-1980s. In 1990, he wrote the seminal text *Probabilistic Reasoning in Expert Systems,* which helped to unify the field of Bayesian networks. He established the field further with his book *Learning Bayesian Networks,* which appeared in 2003. Besides authoring books, he has published numerous cross-disciplinary articles spanning the fields of computer science, mathematics, philosophy of science, and psychology. Presently, he is Professor and Chair of Computer Science at Northeastern Illinois University, while also serving as a visiting scholar at Monash University in Australia.

John R. Nesselroade has served on the faculties of West Virginia University and the Pennsylvania State University and currently is the Hugh Scott Hamilton Professor of Psychology at the University of Virginia. He received his PhD in psychology from the University of Illinois in 1967 under the primary supervision of Raymond B. Cattell. He is a frequent Visiting Senior Scientist at the Max Planck Institute for Human Development in Berlin, Germany.

Shizuhiko Nishisato received his PhD from the University of North Carolina in Chapel Hill under the supervision of R. Darrell Bock. His entire career was devoted to the development of dual scaling. A former editor of *Psychometrika*, a former President of the Psychometric Society, and a Fellow of the American Statistical Association, Dr. Nishisato is Professor Emeritus at the University of Toronto, Canada.

Gregory J. Palardy is Assistant Professor in the Research, Evaluation, Measurement, and Statistics program in the Department of Educational Psychology at the University of Georgia. His research interests include multilevel models, structural equation models, longitudinal models, and school effectiveness evaluation.

Thomas Richardson received his BA in Mathematics and Philosophy from the University of Oxford in 1992, an MSc in Logic and Computation, and a PhD in Logic, Computation, and Methodology of Science from Carnegie Mellon University in 1995 and 1996 respectively. A Fellow of the Center for Advanced Studies in the Behavioral Sciences at Stanford, he was a Rosenbaum Fellow at the Isaac Newton Institute, Cambridge, UK in 1997. He co-chaired the 2001 Workshop on Artificial Intelligence and Statistics and was an associate editor for the *Journal of the Royal Statistical Society Series B* (Methodological) from 1999-2003. His research work focuses on algorithms for uncovering causal structure from data, particularly in contexts where hidden variables may be present. He joined the Statistics Department at the University of Washington in 1996, where he is Associate Professor.

David Rindskopf is Professor of Educational Psychology and Psychology at the City University of New York (CUNY) Graduate Center. His research and teaching focus on missing data, categorical data, factor analysis, structural equation models, multilevel models, research methods, and item response theory. He is a Fellow of the American Statistical Association, and President of the Society of Multivariate Experimental Psychology (2003–2004). He can be reached at drindskopf@gc.cuny.edu.

Louis A. Roussos is Assistant Professor in the Department of Educational Psychology at the University of Illinois at Urbana-Champaign. His research interests include computerized testing, DIF, multidimensional IRT, and skills diagnosis.

Russell W. Rumberger is Professor in the Gervirtz Graduate School of Education at the University of California, Santa Barbara, and Director of the University of California Linguistic Minority Research Institute. He received a PhD in education and an MA in economics from Stanford University in 1978 and a BS in electrical engineering from Carnegie-Mellon University in 1971. His research on education and work has focused on the economic payoffs to schooling and on educational requirements of work. His research on at-risk students has focused on the causes, consequences, and solutions to the problem of school dropouts; the causes and consequences of student mobility; the schooling of English-language learners; and the impact of school segregation on student achievement.

André A. Rupp is Assistant Professor at the University of Ottawa in Ottawa, Ontario, Canada. His research interests include inferential Limits of Assessments, validity and reliability theory, parameter invariance, item response modeling, cognitively diagnostic assessment, applied statistics, research methodology, quantitative register analysis, and language assessment.

Richard Scheines received his PhD in Philosophy of Science from the University of Pittsburgh in 1987. He is now Professor of Philosophy, Automated Learning and Discovery, and Human-Computer Interaction at Carnegie Mellon University. Besides research on causal discovery, he has developed an online course on causation and statistics: the Carnegie Mellon Curriculum on Causal and Statistical Reasoning (www. phil.cmu.edu/projects/csr).

Michael Seltzer is an Associate Professor in the Graduate School of Education and Information Studies at UCLA. His research focuses on the development and application of statistical methods for multilevel modeling and longitudinal analysis.

Judith D. Singer is the James Bryant Conant Professor at the Harvard University Graduate School of Education. She holds a doctorate in statistics from Harvard University and teaches courses in applied statistics and specializes in quantitative methods for measuring change over time and analyzing the occurrence, timing, and duration of events. She is the coauthor (with John B. Willett) of *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence* (2003).

Peter Spirtes is a Research Scientist at the Institute for Human and Machine Cognition, a Professor of Philosophy at Carnegie Mellon University, and on the board of the Center for Automated Learning and Discovery at Carnegie Mellon University. He received a PhD in eistory and philosophy of Science from the University of Pittsburgh in 1981 and an MS in computer science from the University of Pittsburgh in 1983. His research centers on learning causal relationships from nonexperimental data and is interdisciplinary in nature, involving philosophy, statistics, graph theory, and computer science. His current research centers on extending the application of the results about causal inference to a wider class of phenomena and investigating the extent to which these search procedures can be made more reliable on small samples.

William Stout is Professor Emeritus in the department of statistics at the University of Illinois, Urbana-Champaign, and former director of the University of Illinois Statistical Laboratory for Educational and Psychological Measurement. He received his PhD in mathematics from Purdue University and is a Fellow of the Institute of Mathematical Statistics and past president of the Psychometric Society. He has held past associate editorships of *Psychometrika* and *JEBS*. His research interests include test equity, latent multidimensionality, and skills diagnosis of standardized test data.

Anita J. van der Kooij, MA, is a researcher at the Department of Educational Sciences, Data Theory Group, Leiden University, The Netherlands. Her research interests are optimal scaling techniques for multivariate analysis, and software development. She can be reached at kooij@fsw.leidenuniv.nl.

Jeroen K. Vermunt is Professor in the Department of Methodology and Statistics at Tilburg University, The Netherlands. He holds a PhD in social sciences from the same university. He teaches and publishes on methodological topics such as categorical data techniques, methods for the analysis of longitudinal

and event history data, latent class and finite mixture models, latent trait models, and multilevel and random-effects models. He is developer of the LEM program for categorical data analysis and codeveloper (with Jay Magidson) of the Latent GOLD and Latent GOLD Choice software packages for latent class and finite mixture modeling.

Oliver Vitouch is Professor of Psychology and head of the Cognitive Psychology Unit (CPU) at the University of Klagenfurt, Austria. He has published in various subfields of cognitive psychology and the cognitive neurosciences and entertains a long-term love affair with methodological and epistemological issues.

John B. Willett is the Charles William Elliot Professor at the Harvard University Graduate School of Education. He holds a doctorate in quantitative methods from Stanford University and masters' degrees in statistics and psychometrics from Stanford and Hong Kong Universities, respectively. He teaches courses in applied statistics and specializes in quantitative methods for measuring change over time and analyzing the occurrence, timing, and duration of events. He is the coauthor (with Judith D. Singer) of *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence* (2003).

Bruno D. Zumbo, PhD, is Professor at the University of British Columbia, Vancouver, British Columbia, Canada. His professional interests center upon developing statistical theory and quantitative methods for conducting measurement, research, and evaluation.