Confirmatory Factor Analysis for Applied Research

Methodology in the Social Sciences
David A. Kenny, Series Editor

Confirmatory Factor Analysis for Applied Research

Timothy A. Brown

SERIES EDITOR'S NOTE by David A. Kenny

THE GUILFORD PRESS
New York   London

For my father, Kaye,
and Nick and Greg

About the Author

Timothy A. Brown, PsyD, is a professor in the Department of Psychology at Boston University, and Director of Research at Boston University's Center for Anxiety and Related Disorders. He has published extensively in the areas of the classification of anxiety and mood disorders, vulnerability to emotional disorders, psychometrics, and methodological advances in social sciences research. In addition to conducting his own grant-supported research, Dr. Brown serves as a statistical investigator or consultant on numerous federally funded research projects. He has been on the editorial boards of several scientific journals, including recent appointments as Associate Editor of the *Journal of Abnormal Psychology* and *Behavior Therapy.*

Series Editor's Note

For some reason, the topic of confirmatory factor analysis (CFA) has not received the attention that it deserves. Two closely related topics, exploratory factor analysis (EFA) and structural equation modeling (SEM), have dozens of textbooks written about them. Book-length treatments of CFA are rare and that is what makes this book distinctive.

One might think that there are so few books on CFA because it is so rarely used. However, this is not the case. Very often, those who conduct EFA follow up the analysis with CFA. Additionally, SEM always involves a measurement model and very often the best way to test that model is with CFA. Poor-fitting structural equation models are almost always due to CFA problems. Thus, to be proficient at SEM, the analyst must know CFA. This book very nicely explains the links between CFA and these two different methods, in particular the natural process of beginning with EFA, proceeding to CFA, and then SEM.

I think it is ironic that SEM has received so much more attention than CFA, because the social and behavioral sciences have learned much more from CFA than from SEM. In particular, through CFA we are able to understand the construct validity of attitudes and personality, and CFA provides important information about the relative stability of individual differences throughout the lifespan.

Unlike most books on factor analysis, this one spares us all the matrices with their transposes, Kronecker products, and inverses. Certainly matrix algebra is critical in the theory, proofs, and estimation of CFA, but for day-to-day practitioners, it just gets in the way. This is not to say that the author, Timothy A. Brown, doesn't discuss technical issues where necessary. The text is complicated where appropriate.

An example of one such complicated topic is the multitrait–multimethod matrix, first proposed by Donald Campbell and Donald Fiske. I am pleased that Brown decided to devote a full chapter to the topic. Interestingly, a generation of researchers tried to find EFA models for the matrix and never developed a completely satisfactory model. Another generation of researchers worked on several CFA models for the matrix, and Brown very nicely summarizes the models they produced.

Another useful feature of this book is that it contains an entire chapter devoted to issues of statistical power and sample sizes. Investigators need to make decisions, costly both in terms of time and money, about sample size. Very often they make those decisions using rather arbitrary procedures. The book outlines a formal and practical approach to that question.

For breadth of applications, the book provides examples from several different areas of the social and behavioral sciences. It also illustrates the analyses using several different software programs. Preferences for computer programs change as fast as preferences do for hair styles; thus, it is an advantage that the book is not tied to one computer program. Most readers would benefit from analyzing data of their own as they read the book.

Construct validity, instrument development and validation, reduction of the number of variables, and sources of bias in measurement, to name just a few, are subjects supported by high-quality CFA. Almost all research data include many variables; therefore, Brown's detailed and careful treatment of this important topic will be of benefit in almost all research situations. A gap in the field of multivariate data analysis that has existed for far too long has finally been filled. Researchers now have a readable, detailed, and practical discussion of CFA.

DAVID A. KENNY

Preface

This book was written for the simple reason that no other book of its kind had been published before. Although many books on structural equation modeling (SEM) exist, this is the first book devoted solely to the topic of confirmatory factor analysis (CFA). Accordingly, for the first time, many important topics are brought under one cover—for example, the similarities/differences between exploratory factor analysis (EFA) and CFA, the use of maximum likelihood EFA procedures as a precursor to CFA, diagnosing and rectifying the various sources for the ill-fit of a measurement model, analysis of mean structures, modeling with multiple groups (e.g., MIMIC), CFA scale reliability evaluation, formative indicator models, and higher-order factor analysis. After covering the fundamentals and various types of CFA in the earlier chapters, in later chapters I address issues like CFA with non-normal or categorical indicators, handling missing data, and power analysis/sample size determination, which are germane to SEM models of any type. Although it is equally important to CFA practice, another reason I included this material was because of the lack of adequate coverage in preexisting SEM sourcebooks. Thus, I hope the book will serve as a useful guide to researchers working with a latent variable model of any type. The book is not tied to specific latent variable software packages, and in fact the five most popular programs are featured throughout (Amos, EQS, LISREL, Mplus, SAS/CALIS). However, readers will note that this book is the first to provide an extensive treatment of Mplus, a program that is becoming increasingly popular with applied researchers for its ease of use with complex models and data (e.g., categorical outcomes, categorical latent variables, multilevel data).

The target readership of this book is applied researchers and graduate students working within any domain of social and behavioral sciences

(e.g., psychology, education, political science, management/marketing, sociology, public health). In the classroom, this book can serve as a primary or supplemental text in courses on advanced statistics, factor analysis, SEM, or psychometrics/scale development. For applied researchers, this book can be used either as a resource for learning the procedures of CFA or, for more experienced readers, as a reference guide for dealing with complex CFA models or data issues. What each chapter specifically covers is described in Chapter 1. The first five chapters deal with the fundamentals of CFA: what the researcher needs to know to conduct a CFA of any type. Thus, especially for readers new to CFA, it is recommended that the first five chapters be read in order, as this material is the foundation for the remainder of the book. Chapters 6 through 10 address specific types of CFA and other issues such as dealing with missing or categorical data and power analysis. The reading order of the second group of chapters is less important than for the first.

Advances in quantitative methodology are often slow to be picked up by applied researchers because such methods are usually disseminated in a manner inaccessible to many end users (e.g., formula-driven articles in mathematical/statistical journals). This is unfortunate, because multivariate statistics can be readily and properly employed by any researcher provided that the test's assumptions, steps, common pitfalls, and so on, are laid out clearly. Keeping with that philosophy, this book was written be a user-friendly guide to conducting CFA with real data sets, with an emphasis more on conceptual and practical aspects than on quantitative formulas. Several strategies were used to help meet this goal: (1) every key concept is accompanied by an applied data set and the syntax and output from the leading latent variable software packages; (2) tables are included that recap the procedures or steps of the methods being presented (e.g., how to conduct an EFA, how to write up the results of a CFA study); (3) numerous figures are provided that graphically illustrate some of the more complicated concepts or procedures (e.g., EFA factor rotation, forms of measurement invariance, types of nonpositive definite matrices, identification of formative indicator models), and (4) many chapters contain appendices with user-friendly illustrations of seemingly complex quantitative operations (e.g., data generation in Monte Carlo simulation research, calculation of matrix determinants and their role in model fit and improper solutions). I have also provided a website (http://people.bu.edu/tabrown/) with data and computer files for the book's examples and other materials (e.g., updates, links to other CFA resources). I hope that through the use of the aforementioned materials, even the most complicated CFA model or

data issue has been demystified and can now be readily tackled by the reader.

In closing, I would like to thank the people who were instrumental in the realization of this volume. First, thanks to Series Editor David A. Kenny, who, in addition to providing very helpful comments on specific sections, played an enormous role in helping me to shape the scope and coverage of this book. In addition, I would like to extend my appreciation to C. Deborah Laughton, Publisher, Methodology and Statistics, who provided many excellent suggestions and positive feedback throughout the process and who secured several terrific outside reviews. Indeed, I am grateful to the following reviewers, whose uniformly constructive and thoughtful feedback helped me strengthen the book considerably: Larry Price, Texas State University–San Marcos; Christopher Federico, University of Minnesota; and Ke-Hai Yuan, University of Notre Dame. I would also like to thank William Meyer, Production Editor at The Guilford Press, for his work in bringing a very technically complex manuscript to press. And finally, special thanks to my wife, Bonnie, for her continuous encouragement and support.

Contents

4 • Specification and Interpretation of CFA Models

Web address for the author's data and computer files
and other resources: http://people.bu.edu/tabrown/

1

Introduction

USES OF CONFIRMATORY FACTOR ANALYSIS

Confirmatory factor analysis (CFA) is a type of structural equation modeling (SEM) that deals specifically with measurement models, that is, the relationships between observed measures or *indicators* (e.g., test items, test scores, behavioral observation ratings) and latent variables or *factors*. A fundamental feature of CFA is its hypothesis-driven nature. It is unlike its counterpart, exploratory factor analysis (EFA), in that the researcher must prespecify all aspects of the CFA model. Thus, the researcher must have a firm a priori sense, based on past evidence and theory, of the number of factors that exist in the data, of which indicators are related to which factors, and so forth. In addition to its greater emphasis on theory and hypothesis testing, the CFA framework provides many other analytic possibilities that are not available in EFA. These possibilities include the evaluation of method effects and the examination of the stability or invariance of the factor model over time or informants. Moreover, for the reasons discussed below, CFA should be conducted prior to the specification of an SEM model.

CFA has become one of the most commonly used statistical procedures in applied research. This is because CFA is well equipped to address the types of questions that researchers often ask. Some of the most common uses of CFA are as follows.

Psychometric Evaluation of Test Instruments

CFA is almost always used during the process of scale development to examine the latent structure of a test instrument (e.g., a questionnaire). In

this context, CFA is used to verify the number of underlying dimensions of the instrument (factors) and the pattern of item–factor relationships (*factor loadings*). CFA also assists in the determination of how a test should be scored. When the latent structure is multifactorial (i.e., two or more factors), the pattern of factor loadings supported by CFA will designate how a test might be scored using subscales; that is, the number of factors is indicative of the number of subscales, and the pattern of item–factor relationships (which items load on which factors) indicates how the subscales should be scored. Depending on other results and extensions of the analysis, CFA may support the use of total scores (composite of all items) in addition to subscale scores (composites of subsets of items). For example, the viability of a single total score might be indicated when the relationships among the latent dimensions (subscales) of a test can be accounted for by one higher-order factor, and when the test items are meaningfully related to the higher-order factor (see higher-order CFA; Chapter 8). CFA is an important analytic tool for other aspects of psychometric evaluation. It can be used to estimate the *scale reliability* of test instruments in a manner that avoids the problems of traditional methods (e.g., Cronbach's alpha; see Chapter 8). Given recent advances in the analysis of categorical data (e.g., binary true/false test items), CFA now offers a comparable analytic framework to item response theory (IRT). In fact, in some ways, CFA provides more analytic flexibility than the traditional IRT model (see Chapter 9).

Construct Validation

Akin to a factor in CFA, a *construct* is a theoretical concept. In clinical psychology and psychiatry, for example, the mental disorders (e.g., major depression, schizophrenia) are constructs manifested by various clusters of symptoms that are reported by the patient or observed by others. In sociology, juvenile delinquency might be construed as a multidimensional construct defined by various forms of misconduct (e.g., property crimes, interpersonal violence, drug use, academic misconduct). CFA is an indispensable analytic tool for construct validation in the social and behavioral sciences. The results of CFA can provide compelling evidence of the *convergent* and *discriminant validity* of theoretical constructs. Convergent validity is indicated by evidence that different indicators of theoretically similar or overlapping constructs are strongly interrelated; for example, symptoms purported to be manifestations of a single mental disorder load

on the same factor. Discriminant validity is indicated by results showing that indicators of theoretically distinct constructs are not highly intercorrelated; for example, behaviors purported to be manifestations of different types of delinquency load on separate factors, and the factors are not so highly correlated as to indicate that a broader construct has been erroneously separated into two or more factors. One of the most elegant uses of CFA in construct validation is the analysis of multitrait–multimethod matrices (see Chapter 6). A fundamental strength of CFA approaches to construct validation is that the resulting estimates of convergent and discriminant validity are adjusted for measurement error and an error theory (see the "Method Effects" section, below). Thus, CFA provides a stronger analytic framework than traditional methods that do not account for measurement error (e.g., ordinary least squares approaches such as correlation/multiple regression assume variables in the analysis are free of measurement error).

Method Effects

Often, some of the covariation of observed measures is due to sources other than the substantive latent factors. For instance, consider the situation where four measures of employee morale have been collected; two indicators are employees' self-reports (e.g., questionnaires), the other two are obtained from supervisors (e.g., behavioral observations). It would be presumed that the four measures are intercorrelated because each is a manifest indicator of the underlying construct of morale. However, it is also likely that the employee self-report measures are more highly correlated with each other than with the supervisor measures, and vice versa. This additional covariation is not due to the underlying construct of morale, but reflects shared method variance. A *method effect* exists when additional covariation among indicators is introduced by the measurement approach. Method effects can also occur within a single assessment modality. For example, method effects are usually present in questionnaires that contain some combination of positively and negatively worded items (e.g., see Chapters 3 and 6). Unfortunately, EFA is incapable of estimating method effects. In fact, the use of EFA when method effects exist in the data can produce misleading results—that is, yield additional factors that are not substantively meaningful but instead stem from artifacts of measurement. In CFA, however, method effects can be specified as part of the error theory of the measurement model.

The advantages of estimating method effects within CFA include the ability to (1) specify measurement models that are more conceptually viable; (2) determine the amount of method variance in each indicator; and (3) obtain better estimates of the relationships of the indicators to the latent factors, and the relationships among latent variables (see Chapters 5 and 6).

Measurement Invariance Evaluation

Another key strength of CFA is its ability to determine how well measurement models generalize across groups of individuals or across time. *Measurement invariance* evaluation is an important aspect of test development. If a test is intended to be administered in a heterogeneous population, it should be established that its measurement properties are equivalent in subgroups of the population (e.g., gender, race). A test is said to be biased when some of its items do not measure the underlying construct comparably across groups. *Test bias* can be serious, such as in situations where a given score on a cognitive ability or job aptitude test does not represent the same true level of ability/aptitude in male and female respondents. Stated another way, the test would be biased against women if, for a given level of true intelligence, men tended to score several IQ units higher on the test than women. These questions can be addressed in CFA by multiple-groups solutions and MIMIC (multiple indicators, multiple causes) models (Chapter 7). For instance, in a multiple-groups CFA solution, the measurement model is estimated simultaneously in various subgroups (e.g., men and women). Other restrictions are placed on the multiple-groups solution to determine the equivalence of the measurement model across groups; for instance, if the factor loadings are equivalent, the magnitude of the relationships between the test items and the underlying construct (e.g., cognitive ability) are the same in men and women. Multiple-groups CFA solutions are also used to examine longitudinal measurement invariance. This is a very important aspect of latent variable analyses of repeated measures designs. In the absence of such evaluation, it cannot be determined whether temporal change in a construct is due to true change or to changes in the structure or measurement of the construct over time. Multiple-groups analysis can be applied to any type of CFA or SEM model. For example, these procedures can be incorporated into the analysis of multitrait–multimethod data to examine the generalizability of construct validity across groups.

WHY A BOOK ON CFA?

It also seems appropriate to begin this volume by addressing the question, "Is there really a need for a book devoted solely to the topic of CFA?" On the author's bookshelf sit 15 books on the subject of SEM. Why not go to one of these SEM books to learn about CFA? Given that CFA is a form of SEM, virtually all of these books provide some introduction to CFA. However, this coverage typically consists of a chapter at best. As this book will attest, CFA is a very broad and complex topic and extant SEM books only scratch the surface. This is unfortunate because, in applied SEM research, most of the work deals with measurement models (CFA). Indeed, many applied research questions are addressed using CFA as the primary analytic procedure (e.g., psychometric evaluation of test instruments, construct validation). Another large proportion of SEM studies focus on structural regression models, that is, the manner in which latent factors are interrelated. Although CFA is not the ultimate analysis in such studies, a viable measurement model (CFA) must be established prior to evaluating the structural (e.g., regressive) relationships among the latent variables of interest. When poor model fit is encountered in such studies, it is more likely that it stems from misspecifications in the measurement portion of the model (i.e., the manner in which observed variables are related to latent factors) than from the structural component that specifies the interrelationships of latent factors. This is because there are usually more things that can go wrong in the measurement model than in the structural model (e.g., problems in the selection of observed measures, misspecified factor loadings, additional sources of covariation among observed measures that cannot be accounted for by the latent factors). Existing SEM resources do not provide sufficient details on the sources of ill fit in CFA measurement models or how such models can be diagnosed and respecified. Moreover, advanced applications of CFA are rarely discussed in general SEM books (e.g., CFA with categorical indicators, scale reliability evaluation, MIMIC models, formative indicators).

Given the importance and widespread use of CFA, this book was written to provide an in-depth treatment of the concepts, procedures, pitfalls, and extensions of this methodology. Although the overriding objective of the book is to provide critical information on applied CFA that has not received adequate coverage in the past, it is important to note that the topics pertain to SEM in general (e.g., sample size/power analysis, missing data, non-normal or categorical data, formative indicators). Thus, it is

hoped that this book will also be a useful resource to researchers using any form of SEM.

COVERAGE OF THE BOOK

The first five chapters of this book present the fundamental concepts and procedures of CFA. Chapter 2 introduces the reader to the concepts and terminology of the common factor model. The common factor model is introduced in context of EFA. This book is not intended to be a comprehensive treatment of the principles and practice of EFA. However, an overview of the concepts and operations of EFA is provided in Chapter 2 for several reasons: (1) most of the concepts and terminology of EFA generalize to CFA; (2) it will foster the discussion of the similarities and differences of EFA and CFA in later chapters (e.g., Chapter 3); and (3) in programmatic research, an EFA study is typically conducted prior to a CFA study to develop and refine measurement models that are reasonable for CFA (thus, the applied CFA researcher must also be knowledgeable of EFA). An introduction to CFA is given in Chapter 3. After providing a detailed comparison of EFA and CFA, this chapter presents the various parameters, unique terminology, and fundamental equations of CFA models. Many other important concepts are introduced in this chapter that are essential to the practice of CFA and that must be understood in order to proceed to subsequent chapters—model identification, model estimation (e.g., maximum likelihood), and goodness of model fit. Chapter 4 illustrates and extends these concepts using a complete example of a CFA measurement model. In this chapter, the reader will learn how to program and interpret basic CFA models using several of the most popular latent variable software packages (LISREL, Mplus, Amos, EQS, CALIS). The procedures for evaluating the acceptability of the CFA model are discussed. In the context of this presentation, the reader is introduced to other important concepts such as model misspecification and Heywood cases. Chapter 4 concludes with a section on the material that should be included in the report of a CFA study. Chapter 5 covers the important topics of model respecification and model comparison. It deals with the problem of poor-fitting CFA models and the various ways a CFA model may be misspecified. This chapter also presents the technique of EFA within the CFA framework, an underutilized method of developing more viable CFA measurement models on the basis of EFA findings. The concepts of nested models, equivalent models, and method effects are also discussed.

The second half of the book focuses on more advanced or specialized topics and issues in CFA. Chapter 6 discusses how CFA can be conducted to analyze multitrait–multimethod (MTMM) data in the validation of social or behavioral constructs. Although the concepts of method effects, convergent validity, and discriminant validity are introduced in earlier chapters (e.g., Chapter 5), these issues are discussed extensively in context of MTMM models in Chapter 6. Chapter 7 discusses CFA models that contain various combinations of equality constraints (e.g., estimation of a CFA model with the constraint of holding two or more parameters to equal the same value), multiple groups (e.g., simultaneous CFA in separate groups of males and females), and mean structures (CFAs that entail the estimation of the intercepts of indicators and factors). These models are discussed and illustrated in context of the analysis of measurement invariance—that is, is the measurement model equivalent in different groups or within the same group across time? Two different approaches to evaluating CFA models in multiple groups are presented in detail: multiple-groups solutions and MIMIC models.

Chapter 8 presents three other types of CFA models: higher-order CFA, CFA approaches to scale reliability estimation, and CFA with formative indicators. Higher-order factor analysis is conducted in situations where the researcher can posit a more parsimonious conceptual account for the interrelationships of the factors in the initial CFA model. In the section on scale reliability estimation, it will be seen that the unstandardized parameter estimates of a CFA solution can be used to obtain point estimates and confidence intervals of the reliability of test instruments (i.e., reliability estimate = the proportion of the total observed variance in a test score that reflects true score variance). This approach has important advantages over traditional estimates of internal consistency (Cronbach's alpha). Models with formative indicators contain observed measures that "cause" the latent construct. In the typical CFA, indicators are defined as linear functions of the latent variable, plus error; that is, indicators are considered to be the effects of the underlying construct. In some situations, however, it may be more plausible to view the indicators as causing a latent variable; for example, socioeconomic status is a concept determined by one's income, education level, job status—not the other way around. Although formative indicators pose special modeling challenges, Chapter 8 shows how such models can be handled in CFA.

The last two chapters consider issues that must often be dealt with in applied CFA research, but that are rarely discussed in extant SEM sourcebooks. Chapter 9 addresses data set complications such as how to accom-

modate missing data and how to conduct CFA when the distributions of continuous indicators are non-normal. Various methods of handling each issue are discussed and illustrated (e.g., missing data: multiple imputation, direct maximum likelihood; non-normal data: alternative statistical estimators, bootstrapping, item parceling). Chapter 9 also includes a detailed treatment of CFA with categorical outcomes (e.g., tests with binary items such as true/false scales). In addition to illustrating the estimation and interpretation of such models, this section demonstrates the parallels and extensions of CFA to traditional IRT analysis. The final chapter of this book (Chapter 10) deals with the oft overlooked topic of determining the sample size necessary to achieve sufficient statistical power and precision of the parameter estimates in a CFA study. Two different approaches to this issue are presented (Satorra–Saris method, Monte Carlo method). The volume ends with an overview of two relatively new modeling possibilities involving CFA: multilevel factor models and factor mixture models.

OTHER CONSIDERATIONS

This book was written with the applied researcher and graduate student in mind. It is intended to be a user-friendly guide to conducting CFA with real data sets. To achieve this goal, conceptual and practical aspects of CFA are emphasized, and quantitative aspects are kept to a minimum (or separated from the main text; e.g., Chapter 3). Formulas are not avoided altogether, but are provided in instances where they are expected to foster the reader's conceptual understanding of the principles and operations of CFA. Although this book does not require a high level of statistical acumen, a basic understanding of correlation/multiple regression will facilitate the reader's passage through the occasional, more technically oriented section.

It is important that a book of this nature not be tied to a specific latent variable software program. For this reason, most of the examples provided in this book are accompanied with input syntax from each of the most widely used software programs (LISREL, Mplus, Amos, EQS, CALIS). Several comments about the examples are in order. First, readers will note that many of the syntax examples are first discussed in context of the LISREL program. This is not intended to imply a preference for LISREL over other software programs. Rather, it is more reflective of the historical fact that the etiological roots of SEM are strongly tied to LISREL. For instance, the widely accepted symbols of the parameters and computational formulas of a CFA model stem from LISREL notation (e.g., λ = factor loading). The

illustrations of LISREL matrix programming allow interested readers to more quickly understand computational aspects of CFA (e.g., using the provided formula, how the model-implied covariance of two indicators can be calculated on the basis of the CFA model's parameter estimates). Knowledge of this notation is also useful to readers who are interested in developing a deeper quantitative understanding of CFA and SEM in more technical sources (e.g., Bollen, 1989). However, the output from the Mplus program is relied on heavily in various examples in the book. Again, a preference for Mplus should not be inferred. This reflects the fact that the results of CFA are provided more succinctly by Mplus than by other programs (concise output = concise tables in this book).

Another potential pitfall of including computer syntax examples is the high likelihood that soon after a book goes into print, another version of the software will be released. When this book was written, the following versions of the software programs were current: LISREL 8.72, Mplus 3.11, EQS 5.7b, Amos 5.0.1, and SAS/CALIS 8.2. New releases typically introduce new features to the software, but do not alter the overall programming framework. In terms of this book, the most probable consequence of new software releases is that some claims about the (in)capabilities of the programs will become outdated. However, the syntax examples should be upwardly compatible (i.e., fully functional) with any subsequent software releases (e.g., although LISREL 7 does not contain many of the features of LISREL 8.72, syntax written in this version is fully operational in subsequent LISREL releases).

Especially in earlier chapters of this book, the computer syntax examples contain few, if any, programming shortcuts. Again, this is done to foster the reader's understanding of CFA model specification. This is another reason why LISREL is often used in the programming examples: that is, in LISREL matrix-based programming, the user must specify every aspect of the CFA model. Thus, CFA model specification is more clearly conveyed in LISREL as compared with some programs where these specifications occur "behind the scenes" (e.g., Mplus contains a series of defaults that automatically specify marker indicators, free and fixed factor loadings, factor variances and covariances, and so forth, in a standard CFA model). On a related note, many latent variable software programs (e.g., Amos, LISREL, EQS) now contain graphical interfaces that allow the user to specify the CFA model by constructing a path diagram with a set of drawing tools. Indeed, graphical interfaces are an increasingly popular method of model programming, particularly with researchers new to CFA and SEM. The primary reason why graphical input is not discussed in this book is

that it does not lend itself well to the written page. Yet there are other reasons why syntax programming can be more advantageous. For instance, it is often quicker to generate a CFA solution from a syntax program than by constructing a path diagram in a drawing editor. In addition, many of the advanced features of model specification are more easily invoked through syntax. Users who understand the logic of syntax programming (either matrix- or equation-based syntax operations) are able to move from one latent variable software package to another much more quickly and easily than users who are adept only in the use of a graphical interface of a given software program.

In an attempt to make the illustrations more provocative to the applied researcher, most of the examples in this book are loosely based on findings or test instruments in the extant literature. The examples are drawn from a variety of domains within the social and behavioral sciences—clinical psychology, personality psychology, social psychology, industrial/organizational psychology, and sociology. In some instances, the examples use actual research data, but in many cases the data were artificially generated strictly for the purposes of illustrating a given concept. Regardless of the origin of the data, the examples should not be used to draw substantive conclusions about the research domain or test instrument in question.

Many of the examples in this book use a variance–covariance matrix as input data (specifically, the correlations and standard deviations of the indicators are inputted, from which the program generates the sample variance–covariance matrix). This was done to allow interested readers to replicate examples directly from the information provided in the figures and tables of the book. Although matrix input is used as a convenience feature in this book, it is not necessarily the best method of reading data into an analysis. All leading latent variable programs are capable of reading raw data as text files, and many can read data saved in other software formats (e.g., SPSS .sav files, Microsoft Excel files). There are several advantages of using raw data as input. First, it is more convenient because the user does not need to compute the input matrices prior to conducting the latent variable analysis. Second, the input data are more precise when the software program computes the input matrices from raw data (user-generated matrices usually contain rounding error). Third, there are some situations where raw data must be analyzed; for instance, models that have missing, non-normal, or categorical data. Some sections of this book (e.g., Chapter 9) illustrate how raw data are read into the analysis. The inter-

ested reader can download the files used in these and other examples from the book's companion website (http://people.bu.edu/tabrown/).

SUMMARY

This chapter provided a general overview of the nature and purposes of CFA, including some of the fundamental differences between EFA and CFA. The ideas introduced in this chapter provide the background for a more detailed discussion of the nature of the common factor model and EFA, the subject of Chapter 2. It was noted that this book is intended to be a user-friendly guide to conducting CFA in real data sets, aimed at students and applied researchers who do not have an extensive background in quantitative methods. Accordingly, practical and conceptual aspects of CFA will be emphasized over mathematics and formulas. In addition, most of the chapters are centered on data-based examples drawn from various realms of the social and behavioral sciences. The overriding rationale of these examples was discussed (e.g., use of software programs, method of data input) to set the stage for their use in subsequent chapters.

2

The Common Factor Model and Exploratory Factor Analysis

This chapter introduces the reader to the concepts, terminology, and basic equations of the common factor model. Both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) are based on the common factor model. In this chapter, the common factor model is discussed primarily in the context of EFA. Nonetheless, most of the concepts and terminology (e.g., common and unique variances, factor loadings, communalities) of EFA are also used in CFA. This chapter discusses some of the fundamental similarities and differences of EFA and CFA. In applied research, EFA and CFA are often conducted in conjunction with one another. For instance, CFA is frequently used in the later stages of scale development after the factor structure of a testing instrument has been explored and refined by EFA. Thus, because the applied CFA researcher must have a working knowledge of EFA, the methods of conducting an EFA are reviewed in this chapter. This overview is also provided to allow more detailed comparisons of EFA and CFA in later chapters.

OVERVIEW OF THE COMMON FACTOR MODEL

Since its inception a century ago (Spearman, 1904, 1927), factor analysis has become one of the most widely used multivariate statistical procedures in applied research endeavors across a multitude of domains (e.g., psychology, education, sociology, management, public health). The fundamental intent of factor analysis is to determine the number and nature of

latent variables or *factors* that account for the variation and covariation among a set of observed measures, commonly referred to as *indicators*. Specifically, a factor is an unobservable variable that influences more than one observed measure and that accounts for the correlations among these observed measures. In other words, the observed measures are inter-correlated because they share a common cause (i.e., they are influenced by the same underlying construct); if the latent construct was partialed out, the intercorrelations among the observed measures would be zero.[1] Thus, factor analysis attempts a more parsimonious understanding of the co-variation among a set of indicators because the number of factors is less than the number of measured variables.

In applied research, factor analysis is most commonly used in psycho-metric evaluations of multiple-item testing instruments (e.g., question-naires; cf. Floyd & Widaman, 1995). For example, a researcher may have generated 20 questionnaire items that he or she believes are indicators of the unidimensional construct of self-esteem. In the early stages of scale development, the researcher might use factor analysis to examine the plau-sibility of this assumption (i.e., the ability of a single factor to account for the intercorrelations among the 20 indicators) and to determine if all 20 items are reasonable indicators of the underlying construct of self-esteem (i.e., how strongly is each item related to the factor?). In addition to psy-chometric evaluation, other common uses for factor analysis include con-struct validation (e.g., obtaining evidence of convergent and discriminant validity by demonstrating that indicators of selected constructs load onto separate factors in the expected manner; e.g., Brown, Chorpita, & Barlow, 1998) and data reduction (e.g., reducing a larger set of intercorrelated indicators to a smaller set of composite variables, and using these composites—i.e., factor scores—as the units of analysis in subsequent sta-tistical tests; e.g., Cox, Walker, Enns, & Karpinski, 2002).

These concepts emanate from the *common factor model* (Thurstone, 1947), which postulates that each indicator in a set of observed measures is a linear function of one or more common factors and one unique factor. Thus, factor analysis partitions the variance of each indicator (derived from the sample correlation/covariance matrix that is used as input for the analysis) into two parts: (1) *common variance*, or the variance accounted for by the latent factor, which is estimated on the basis of variance shared with other indicators in the analysis; and (2) *unique variance*, which is a combination of reliable variance that is specific to the indicator (i.e., sys-tematic latent factors that influence only one indicator) and random error variance (i.e., measurement error or unreliability in the indicator). There

are two main types of analyses based on the common factor model: *exploratory factor analysis* (EFA) and *confirmatory factor analysis* (CFA; Jöreskog, 1969, 1971a). Both EFA and CFA aim to reproduce the observed relationships among a group of indicators with a smaller set of latent variables, but they differ fundamentally by the number and nature of a priori specifications and restrictions made on the factor model. EFA is a data-driven approach such that no specifications are made in regard to the number of latent factors (initially) or to the pattern of relationships between the common factors and the indicators (i.e., the *factor loadings*). Rather, the researcher employs EFA as an exploratory or descriptive technique to determine the appropriate number of common factors and to uncover which measured variables are reasonable indicators of the various latent dimensions (e.g., by the size and differential magnitude of factor loadings). In CFA, the researcher specifies the number of factors and the pattern of indicator–factor loadings in advance, as well as other parameters such as those bearing on the independence or covariance of the factors and indicator unique variances. The prespecified factor solution is evaluated in terms of how well it reproduces the sample correlation (covariance) matrix of the measured variables. Thus, unlike EFA, CFA requires a strong empirical or conceptual foundation to guide the specification and evaluation of the factor model. Accordingly, EFA is typically used earlier in the process of scale development and construct validation, whereas CFA is used in later phases after the underlying structure has been established on prior empirical (EFA) and theoretical grounds. Other important differences between EFA and CFA are discussed in Chapter 3.

A brief applied example is used to illustrate some of the key concepts of the common factor model. In this basic example, four behavioral observation ratings (D1–D4) have been collected on 300 individuals admitted to an inpatient psychiatric facility. The four ratings are hopelessness (D1), feelings of worthlessness/guilt (D2), psychomotor retardation (D3), and sleep disturbance (D4). As shown in Table 2.1, these four clinical ratings (indicators) are moderately intercorrelated. It is conjectured that each of these ratings is a manifest indicator of the latent construct of depression; that is, each of the observed symptoms (e.g., hopelessness, worthlessness) has the shared influence of depression, the single latent variable (factor) that accounts for the intercorrelations among these observed measures. The only reason the indicators are correlated is that they share the common cause of depression; if this latent variable was partialed out, no relationship among these indicators would be seen.

TABLE 2.1. Intercorrelations among Four Behavioral Observation Ratings of Depression

	D1	D2	D3	D4
D1	1.00			
D2	0.70	1.00		
D3	0.65	0.66	1.00	
D4	0.62	0.63	0.60	1.00

Note. N = 300. D1 = hopelessness; D2 = feelings of worthlessness/guilt; D3 = psychomotor retardation; D4 = sleep disturbance.

Using the sample correlations presented in Table 2.1 as input, a factor analysis is conducted using the EFA routines provided in SPSS (FACTOR) and SAS (PROC FACTOR; see Table 2.2). For reasons noted later in this chapter, only a one-factor solution can be pursued. Because EFA uses correlations as the units of analysis, it can be run in SPSS and SAS by embedding the sample correlation matrix in the body of the syntax (as shown in Table 2.2), although both programs can generate this matrix by reading raw input data files. The procedures of EFA are discussed later in this chapter (e.g., methods of factor extraction and selection), but for purposes of this illustration, consider the selected results of the analysis presented in Table 2.2. Of particular interest is the output under the heading "Factor Matrix," which provides the factor loadings for the four clinical ratings. In EFA, the factor loadings are completely standardized estimates of the regression slopes for predicting the indicators from the latent factor, and thus are interpreted along the lines of standardized regression (β) or correlation (r) coefficients as in multiple regression/correlational analysis (cf. Cohen, Cohen, West, & Aiken, 2003).[2] For instance, the factor loading estimate for D1 (hopelessness) was .828, which would be interpreted as indicating that a standardized score increase in the latent factor (Depression) is associated with an .828 standardized score increase in hopelessness. Squaring the factor loadings provides the estimate of the amount of variance in the indicator accounted for by the latent variable (e.g., $.828^2$ = 68.5% variance explained). In factor analysis, the amount of variance in the indicator explained by the common factors is often referred to as the *communality* (shown in the SPSS output at the bottom of Table 2.1). Thus, for the D1 (hopelessness) indicator, the factor model estimates that 68.5%

TABLE 2.2. SPSS and SAS Syntax and Selected Output
for a Basic One-Factor Model

SPSS Syntax

```
MATRIX DATA VARIABLES=ROWTYPE_ D1 D2 D3 D4.
BEGIN DATA.
N 300 300 300 300
COR 1.0
COR .70 1.0
COR .65 .66 1.0
COR .62 .63 .60 1.0
END DATA.
FACTOR
  /MATRIX=IN(COR=*)
  /MISSING LISTWISE
  /ANALYSIS D1 D2 D3 D4
  /PRINT INITIAL EXTRACTION
  /CRITERIA FACTORS(1) ITERATE(25)
  /EXTRACTION ML
  /ROTATION NOROTATE.
```

SAS Syntax

```
data sm (type=CORR);
   input _type_ $ _name_ $ D1 D2 D3 D4;
cards;
n     . 300      .      .      .
corr  D1 1.00    .      .      .
corr  D2 0.70  1.00     .      .
corr  D3 0.65  0.66   1.00     .
corr  D4 0.62  0.63   0.60   1.00
;
proc factor data=sm method=ml nfactors=1 scree;
run;
```

Selected Output (SPSS)

```
Initial Statistics:


Variable   Communality * Factor  Eigenvalue  Pct of Var  Cum Pct
                       *
D1            .57811  *   1       2.93118        73.3        73.3
D2            .59175  *   2        .41039        10.3        83.5
D3            .53077  *   3        .35924         9.0        92.5
D4            .48795  *   4        .29919         7.5       100.0


Test of fit of the    1-factor model:

Chi-square statistic:     .2031, D.F.:   2, Significance:     .9035
```

TABLE 2.2. *(cont.)*

```
Factor Matrix:
            Factor  1
D1              .82822
D2              .84090
D3              .78766
D4              .75228

Final Statistics:

Variable  Communality *  Factor   SS Loadings   Pct of Var   Cum Pct
                        *
D1              .68595  *    1        2.57939        64.5         64.5
D2              .70712  *
D3              .62040  *
D4              .56592  *
```

of its total variance is *common variance* (variance explained by the latent variable of Depression), whereas the remaining 31.5% (i.e., 1 − .685 = .315) is *unique variance*. It was stated earlier that unique variance is some combination of specific factor and measurement error variance. It is important to note that EFA and CFA do not provide separate estimates of specific variance and error variance.

A path diagram of the one-factor measurement model is provided in Figure 2.1. The first part of the diagram presents the solution using common symbols for the various elements of factor models (using LISREL latent Y variable notation), and the second part of the diagram replaces these elements with the sample estimates obtained from the EFA presented in Table 2.1. Following the conventions of factor analysis and structural equation modeling (SEM), the latent factor of Depression is depicted by a circle or an oval, whereas the four clinical ratings (indicators) are represented by squares or rectangles. The unidirectional arrows (→) represent the factor loadings (λ, or *lambda*), which are the regression slopes (direct effects) for predicting the indicators from the latent factor (η, or *eta*). These arrows are also used to relate the unique variances (ε, or *epsilon*) to the indicators.[3]

A fundamental equation of the common factor model is

$$y_j = \lambda_{j1}\eta_1 + \lambda_{j2}\eta_2 + \ldots + \lambda_{jm}\eta_m + \varepsilon_j \tag{2.1}$$

where y_j represents the *j*th of *p* indicators (in the case *p* = 4; D1, D2, D3, D4) obtained from a sample of *n* independent subjects (in this case, *n* =
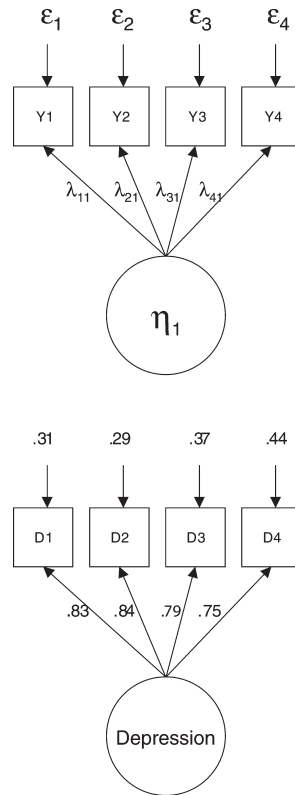
FIGURE 2.1. Path diagram of the one-factor model.

300), λ_{jm} represents the factor loading relating variable j to the mth factor η (in the case $m = 1$; the single factor of Depression), and ε_j represents the variance that is unique to indicator y_j and is independent of all η s and all other εs. As will be seen in subsequent chapters, similar notation is used to represent some of the equations of CFA. In this simple factor solution entailing a single latent factor (η_1) and four indicators, the regression functions depicted in Figure 2.1 can be summarized by four separate equations:

$$D1 = \lambda_{11}\eta_1 + \varepsilon_1 \qquad\qquad (2.2)$$
$$D2 = \lambda_{21}\eta_1 + \varepsilon_2$$
$$D3 = \lambda_{31}\eta_1 + \varepsilon_3$$
$$D4 = \lambda_{41}\eta_1 + \varepsilon_4$$

This set of equations can be summarized in a single equation that expresses the relationships among observed variables (y), latent factors (η), and unique variances (ε):

$$y = \Lambda_y\eta + \varepsilon \qquad (2.3)$$

or in expanded matrix form:

$$\Sigma = \Lambda_y\Psi\Lambda'_y + \Theta_\varepsilon \qquad (2.4)$$

where Σ is the $p \times p$ symmetric correlation matrix of p indicators, Λ_y is the $p \times m$ matrix of factor loadings λ (in this case, a 4×1 vector), Ψ is the $m \times m$ symmetric correlation matrix of the factor correlations (1×1), and Θ_ε is the $p \times p$ diagonal matrix of unique variances ε ($p = 4$). In accord with matrix algebra, matrices are represented in factor analysis and SEM by uppercase Greek letters (e.g., Λ, Ψ, and Θ) and specific elements of these matrices are denoted by lowercase Greek letters (e.g., λ, ψ, and ε). With minor variations, these fundamental equations can be used to calculate various aspects of the sample data from the factor analysis parameter estimates, such as the variances, covariances, and means of the input indicators (the last of these can be conducted in the context of CFA with mean and covariance structures; see Chapter 7). For example, the following equation reproduces the variance in the D1 indicator:

$$
\begin{aligned}
\text{VAR(D1)} = \sigma_{11} &= \lambda_{11}{}^2\psi_{11} + \varepsilon_1 \qquad (2.5) \\
&= .828^2(1) + .315 \\
&= 1.00
\end{aligned}
$$

where ψ_{11} = the variance of the factor η_1, and ε_1 = the unique variance of D1. Note that both ψ_{11} and σ_{11} equal 1.00 because the EFA model is completely standardized; that is, when variables are standardized, their variances equal 1.0. Similarly, the model estimate of the covariance (correlation) of D1 and D2 can be obtained from the following equation:

$$
\begin{aligned}
\text{COV(D1, D2)} = \sigma_{21} &= \lambda_{11}\psi_{11}\lambda_{21} \qquad (2.6) \\
&= (.828)(1)(.841) \\
&= .696
\end{aligned}
$$

Because the solution is completely standardized, this covariance would be interpreted as the factor model estimate of the sample correlation of D1

and D2. In other words, the model-implied correlation of the indicators is the product of their completely standardized factor loadings. Note that the sample correlation of D1 and D2 was .70, which is very close to the factor model-implied correlation of .696. As discussed in further detail in Chapter 3, the acceptability of factor analysis models is determined in large part by how well the parameter estimates of the factor solution (e.g., the factor loadings) are able to reproduce the observed relationships among the input variables. The current illustration should exemplify the point made earlier that common variance (i.e., variance explained by the latent factors as reflected by factor loadings and communalities) is estimated on the basis of the shared variance among the indicators used in the analysis. EFA generates a matrix of factor loadings (Λ) that best explain the correlations among the input indicators.

PROCEDURES OF EFA

Although a full description of EFA is beyond the scope of this book, an overview of its concepts and procedures is helpful to make later comparisons to CFA. The reader is referred to Fabrigar, Wegener, MacCallum, and Strahan (1999), Floyd and Widaman (1995), and Preacher and MacCallum (2003) for detailed guidelines on conducting EFA in applied data sets.

As stated earlier, the overriding objective of EFA is to evaluate the dimensionality of a set of multiple indicators (e.g., items from a questionnaire) by uncovering the smallest number of interpretable factors needed to explain the correlations among them. Whereas the researcher must ultimately specify the number of factors, EFA is an "exploratory" analysis because no a priori restrictions are placed on the pattern of relationships between the observed measures and the latent variables. This is a key difference between EFA and CFA. In CFA, the researcher must specify in advance several key aspects of the factor model such as the number of factors and patterns of indicator–factor loadings.

After determining that EFA is the most appropriate analytic technique for the empirical question at hand, the researcher must decide which indicators to include in the analysis and determine if the size and the nature of the sample are suitable for the analysis (for more details on these issues, see Chapters 9 and 10). Other procedural aspects of EFA include (1) selection of a specific method to estimate the factor model; (2) selection of the appropriate number of factors; (3) in the case of models that have more

than one factor, selection of a technique to rotate the initial factor matrix to foster the interpretability of the solution; and (4) if desired, selection of a method to compute factor scores.

Factor Extraction

There are many methods that can be used to estimate the common factor model such as maximum likelihood, principal factors, weighted least squares, unweighted least squares, generalized least squares, imaging analysis, minimum residual analysis, and alpha factoring, to name just some. For EFA with continuous indicators (i.e., observed measures that approximate an interval-level measurement scale), the most frequently used factor extraction methods are maximum likelihood (ML) and principal factors (PF). ML is also the most commonly used estimation method in CFA, and its fundamental properties are discussed in Chapter 3. A key advantage of the ML estimation method is that it allows for a statistical evaluation of how well the factor solution is able to reproduce the relationships among the indicators in the input data; that is, how closely do the correlations among the indicators predicted by the factor analysis parameters approximate the relationships seen in the input correlation matrix (see Eq. 2.6)? This feature is very helpful for determining the appropriate number of factors. However, as discussed in Chapter 9, ML estimation requires the assumption of multivariate normal distribution of the variables. If the input data depart substantially from a multivariate normal distribution, important aspects of the results of an ML-estimated EFA model can be distorted and not trustworthy (e.g., goodness of model fit, significance tests of model parameters). Another potential disadvantage of ML estimation is its occasional tendency to produce "improper solutions." An improper solution exists when a factor model does not converge on a final set of parameter estimates, or produces an "out of range" estimate such as an indicator with a communality above 1.0. However, PF has the strong advantages of being free of distributional assumptions and of being less prone to improper solutions than ML (Fabrigar et al., 1999). Unlike ML, PF does not provide goodness-of-fit indices useful in determining the suitability of the factor model and the number of latent variables. Thus, PF might be preferred in instances where marked non-normality is evident in the observed measures or perhaps when ML estimation produces an improper solution. However, as discussed later in this book, the presence of improper solutions may be a sign of more serious problems, such as a poorly specified factor model or a poorly behaved input data matrix. If dis-

tributional assumptions hold, ML might be favored because of its ability to produce a wide range of fit indices that guide other important aspects of the factor analytic procedure. As noted in Chapter 3, ML is a full information estimator that provides standard errors that can be used for statistical significance testing and confidence intervals of key parameters such as factor loadings and factor correlations. Strategies for dealing with non-normal, continuous outcomes and categorical indicators are discussed in Chapter 9.

Although related to EFA, principal components analysis (PCA) is frequently miscategorized as an estimation method of common factor analysis. Unlike the estimators discussed in the preceding paragraph (ML, PF), PCA relies on a different set of quantitative methods that are not based on the common factor model. PCA does not differentiate common and unique variance. Rather, PCA aims to account for the *variance* in the observed measures rather than explain the *correlations* among them. Thus, PCA is more appropriately used as a data reduction technique to reduce a larger set of measures to a smaller, more manageable number of composite variables to use in subsequent analyses. However, some methodologists have argued that PCA is a reasonable or perhaps superior alternative to EFA, in view of the fact that PCA possesses several desirable statistical properties (e.g., computationally simpler, not susceptible to improper solutions, often produces results similar to those of EFA, ability of PCA to calculate a participant's score on a principal component whereas the indeterminate nature of EFA complicates such computations). Although debate on this issue continues, Fabrigar et al. (1999) provide several reasons in opposition to the argument for the place of PCA in factor analysis. These authors underscore the situations where EFA and PCA produce dissimilar results; for instance, when communalities are low or when there are only a few indicators of a given factor (cf. Widaman, 1993). Regardless, if the overriding rationale and empirical objectives of an analysis are in accord with the common factor model, then it is conceptually and mathematically inconsistent to conduct PCA; that is, EFA is more appropriate if the stated objective is to reproduce the intercorrelations of a set of indicators with a smaller number of latent dimensions, recognizing the existence of measurement error in the observed measures. Floyd and Widaman (1995) make the related point that estimates based on EFA are more likely to generalize to CFA than are those obtained from PCA in that, unlike PCA, EFA and CFA are based on the common factor model. This is a noteworthy consideration in light of the fact that EFA is often used as a precursor to CFA in scale development and construct validation. A detailed demonstra-

tion of the computational differences between PCA and EFA can be found in multivariate and factor analytic textbooks (e.g., Tabachnick & Fidell, 2001).

Factor Selection

Next, the factor analysis is run using the selected estimation method (e.g., ML, PF). The results of the initial analysis are used to determine the appropriate number of factors to be extracted in subsequent analyses. This is often considered to be the most crucial decision in EFA because "underfactoring" (selecting too few factors) and "overfactoring" (selecting too many factors) can severely compromise the validity of the factor model and its resulting estimates (e.g., introduce considerable error in the factor loading estimates), although some research suggests that the consequences of overfactoring are less severe than those of underfactoring (cf. Fabrigar et al., 1999). Despite the fact that EFA is an exploratory or descriptive technique by nature, the decision about the appropriate number of factors should be guided by substantive considerations, in addition to the statistical guidelines discussed below. For instance, the validity of a given factor should be evaluated in part by its interpretability; for example, does a factor revealed by the EFA have substantive importance? A firm theoretical background and previous experience with the variables will strongly foster the interpretability of factors and the evaluation of the overall factor model. Moreover, factors in the solution should be well defined—that is, comprised of several indicators that strongly relate to it. Factors that are represented by two or three indicators may be underdetermined (have poor determinacy, see below) and highly unstable across replications. The solution should also be evaluated with regard to whether "trivial" factors exist in the data; for instance, factors based on differential relationships among indicators that stem from extraneous or methodological artifacts (e.g., method effects arising from subsets of very similarly worded or reverse-worded items; see Chapter 5).

It is also important to note that the number of factors (m) that can be extracted by EFA is limited by the number of observed measures (p) that are submitted to the analysis. The upper limit on the number of factors varies across estimation techniques. For instance, in EFA using PF, the maximum number of factors that can be extracted is $p - 1$.[4] In ML EFA, the number of parameters that are estimated in the factor solution (a) must be equal to or less than the number of elements (b) in the input correlation or covariance matrix (i.e., $a \leq b$). As the number of factors (m)

increases, so does the number of estimated parameters (a) in the solution. The fact that the maximum number of factors is mathematically limited by the input data can be problematic for ML analyses that use a small set of indicators; that is, the data may not support extraction of the number of factors that are posited to exist on conceptual grounds. For example, because only four observed measures ($p = 4$) were involved, it was possible to extract only one factor ($m = 1$) in the EFA presented in Table 2.2. Although a two-factor solution may be conceptually viable (e.g., Cognitive Depression: D1, D2; Somatic Depression: D3, D4), the number of parameters associated with a two-factor model (b) would exceed the number of pieces of information in the input correlation matrix (a); a and b can be readily calculated by the following equations:

$$a = (p * m) + [(m * (m + 1)] / 2) + p - m^2 \qquad (2.7)$$

$$b = [(p * (p + 1)] / 2 \qquad (2.8)$$

where p = number of observed variables (indicators), and m = number of factors.

Solving for b indicates that the input matrix contains 10 pieces of information (see Table 2.1), corresponding to the 6 correlations in the off-diagonal and the 4 standardized variances on the diagonal; that is, $b = (4 * 5) / 2 = 10$. Solving for a (when $m = 1$) indicates that there are 8 parameters that are estimated in a one-factor solution; that is, $a = (4 * 1) + [(1 * 2) / 2)] + 4 - 1 = 4 + 1 + 4 - 1 = 8$. Because the number of elements of the input matrix ($a = 10$) is greater than the number of parameters ($b = 8$), a single factor can be extracted from the data (as seen in Table 2.2, the degrees of freedom associated with the χ^2 fit statistic is 2, corresponding to the difference $a - b$, $10 - 8 = 2$; see Chapter 3). However, two factors cannot be extracted, because the number of parameters to be estimated in this model exceeds the number of elements of the input matrix by one, that is, $a = (4 * 2) + [(2 * 3) / 2)] + 4 - 4 = 8 + 3 + 4 - 4 = 11$.

Each aspect of the equation used to solve for a corresponds to specific parameters and mathematical restrictions in the EFA model (cf. Eq. 2.4). The first aspect, $(p * m)$, indicates the number of factor loadings (Λ_y). The second aspect, $([m * (m + 1)] / 2)$, indicates the number of factor variances and covariances (Ψ). The third aspect, p, corresponds to the number of residual variances (θ_ε). The final aspect, m^2, reflects the number of restrictions that are required to identify the EFA model (e.g., mathematically convenient restrictions, which include fixing factor variances to unity).

For example, as depicted in Figure 2.1, in the one-factor model there are four factor loadings ($p * m$), one factor variance ($[m * (m + 1)] / 2$), and four indicator residuals (p); however, for identification purposes, the factor variance is fixed to 1.0 ($m^2 = 1^2 = 1$) and thus the model contains eight estimated parameters. A two-factor solution would entail eight factor loadings ($4 * 2$), two factor variances and one factor covariance $[(2 * 3) / 2]$, and four residual variances (total number of parameters = 15). After subtracting the identifying restrictions ($m^2 = 2^2 = 4$; $15 - 4 = 11$), the number of parameters to be estimated in the two-factor model ($b = 11$) still exceeds the pieces in the input matrix ($a = 10$). Thus, two factors cannot be extracted from the data by ML when $p = 4$.

   Especially when an estimation procedure other than ML is used (e.g., PF), factor selection is often guided by the *eigenvalues* generated from either the *unreduced correlation matrix* (**R**; i.e., the input correlation matrix with unities—1.0s—in the diagonal) or the *reduced correlation matrix* (**R**$_r$; i.e., the correlation matrix with communality estimates in the diagonal). For example, the selected SPSS output in Table 2.2 provides eigenvalues from the unreduced correlation matrix under the heading "Initial Statistics."[5] Most multivariate procedures such as EFA rely on eigenvalues and their corresponding *eigenvectors* because they summarize variance in a given correlation or variance/covariance matrix. The calculation of eigenvalues and eigenvectors is beyond the scope of this chapter (for an informative illustration, see Tabachnick & Fidell, 2001), but for practical purposes, it is useful to view eigenvalues as representing the variance in the indicators explained by the successive factors. This is illustrated in the final two sections of Table 2.2; specifically, the eigenvalue corresponding to the single factor that was extracted to account for the interrelationships of the four ratings of clinical depression. On the SPSS printout, this eigenvalue is listed under the heading "SS Loadings" and equals 2.579. Calculating the sum of squares of the four factor loadings (i.e., $.82822^2 +$ $\ldots + .75228^2 = 2.579$) provides the eigenvalue for this factor. Dividing this eigenvalue by the total variance of the input matrix (because indicators are standardized, total variance is equal to the number of input measures, p) yields the proportion of variance in the indicators that is accounted for by the factor model (i.e., $2.579 / 4 = .645$) as also denoted under the heading "Pct of Var" (64.5%) in the "Final Statistics" section of the SPSS printout in Table 2.2.

   The previous paragraph discussed eigenvalues (e.g., 2.579) that were derived from the reduced correlation matrix (**R**$_r$) produced by the EFA solution. The SPSS printout (Table 2.2) also presents eigenvalues for **R**,

listed under the "Initial Statistics" heading (i.e., 2.93, .410, .359, .299). In line with the notion that eigenvalues communicate variance, note that the sum of the eigenvalues for **R** is 4 (i.e., total variance = number of input indicators, p). As was the case for eigenvalues associated with **R**$_r$, dividing the eigenvalue by 4 yields an estimate of explained variance (e.g., 2.93 / 4 = .733; see Table 2.2). Thus, eigenvalues guide the factor selection process by conveying whether a given factor explains a considerable portion of the total variance of the observed measures.

Three commonly used factor selection procedures are based on eigenvalues. They are (1) the Kaiser–Guttman rule; (2) the scree test; and (3) parallel analysis. The *Kaiser–Guttman rule* (also referred to as "the Kaiser criterion," or "the eigenvalues > 1.0 rule") is very straightforward: (1) obtain the eigenvalues derived from the input correlation matrix, **R** (as noted by Fabrigar et al., 1999, researchers occasionally make the mistake of using eigenvalues of the reduced correlation matrix, **R**$_r$); (2) determine how many eigenvalues are greater than 1.0; and (3) use that number to determine the number of nontrivial latent dimensions that exist in the input data. As seen in the "Initial Statistics" section of the selected SPSS output provided in Table 2.2, a single eigenvalue from the input correlation matrix (**R**) was above 1.0 (i.e., 2.93); thus, the Kaiser–Guttman rule would suggest a unidimensional latent structure.

The logic of the Kaiser–Guttman rule is that when an eigenvalue is less than 1.0, the variance explained by a factor is less than the variance of a single indicator. Recall that eigenvalues represent variance, and that EFA standardizes both the latent and observed variables (e.g., the variance that each standardized input variable contributes to the factor extraction is 1.0). Thus, because a goal of EFA is to reduce a set of input indicators (the number of latent factors should be smaller than the number of input indicators), if an eigenvalue is less than 1.0, then the corresponding factor accounts for less variance than the indicator (whose variance equals 1.0). The Kaiser–Guttman rule has wide appeal because of its simplicity and objectivity; in fact, it is the default in popular statistical software packages such as SPSS. Nevertheless, many methodologists have criticized this procedure because it can result in either overfactoring or underfactoring, and because of its somewhat arbitrary nature; for example, sampling error in the input correlation matrix may result in eigenvalues of .99 and 1.01, but nonetheless the Kaiser–Guttman rule would indicate the latter is an important factor whereas the former is not.

Another popular approach, called the *scree test* (Cattell, 1966), also uses the eigenvalues that can be taken from either the input or reduced

correlation matrix (although Fabrigar et al., 1999, note reasons why scree tests based $\mathbf{R}_r$ might be preferred). To provide a more realistic illustration of this procedure, a larger data set is used (p = 20). As shown in Figure 2.2, the scree test employs a graph whereby the eigenvalues form the vertical axis and the factors form the horizontal axis. The graph is inspected to determine the last substantial decline in the magnitude of the eigenvalues—or the point where lines drawn through the plotted eigenvalues change slope. A limitation of this approach is that the results of the scree test may be ambiguous (e.g., no clear shift in the slope) and open to subjective interpretation. This is evident in Figure 2.2 where the results could be interpreted as indicating either a four- or five-factor solution. However, as noted by Gorsuch (1983), the scree test performs reasonably well under conditions such as when the sample size is large and when well-defined factors are present in the data.

Another eigenvalue-based procedure for guiding factor selection is *parallel analysis* (Horn, 1965; Humphreys & Montanelli, 1975). The approach is based on a scree plot of the eigenvalues obtained from the sample data against eigenvalues that are estimated from a data set of random numbers (i.e., the means of eigenvalues produced by multiple sets of
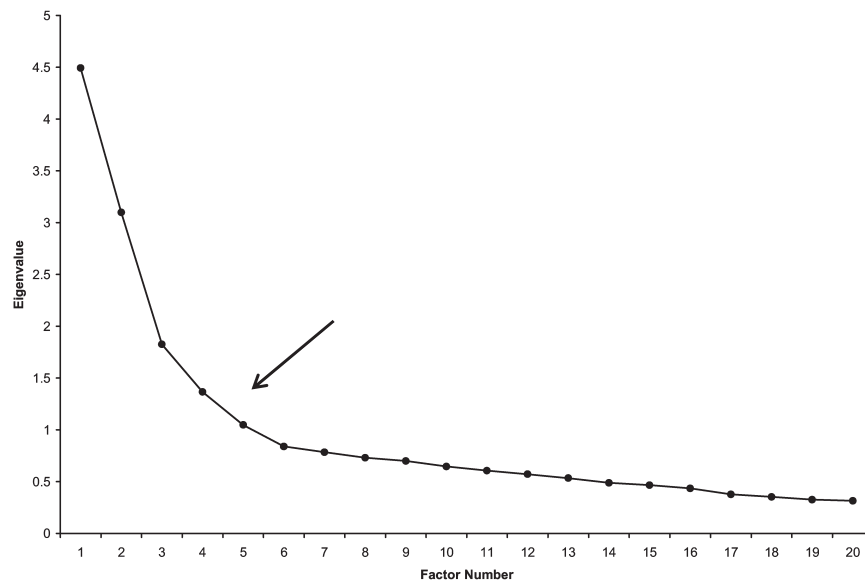
FIGURE 2.2. Scree test of eigenvalues from the unreduced correlation matrix. Arrow indicates region of curve where slope changes.

completely random data).[6] Both the observed sample and random data eigenvalues are plotted, and the appropriate number of factors is indicated by the point where the two lines cross. Thus, factor selection is guided by the number of real eigenvalues greater than the eigenvalues generated from the random data; that is, if the "real" factor explains less variance than the corresponding factor obtained from random numbers, it should not be included in the factor analysis. The term "parallel analysis" refers to the fact that the random data set(s) should parallel aspects of the actual research data (e.g., sample size, number of indicators). The rationale of parallel analysis is that the factor should account for more variance than is expected by chance (as opposed to more variance than is associated with a given indicator, per the logic of the Kaiser–Guttman rule). Using the 20-item data set, parallel analysis suggests four factors (see Figure 2.3). After the eigenvalue for the fourth factor, the eigenvalues from the randomly generated data (averages of 50 replications) exceed the eigenvalues of the research data. Although parallel analysis frequently performs well, like the
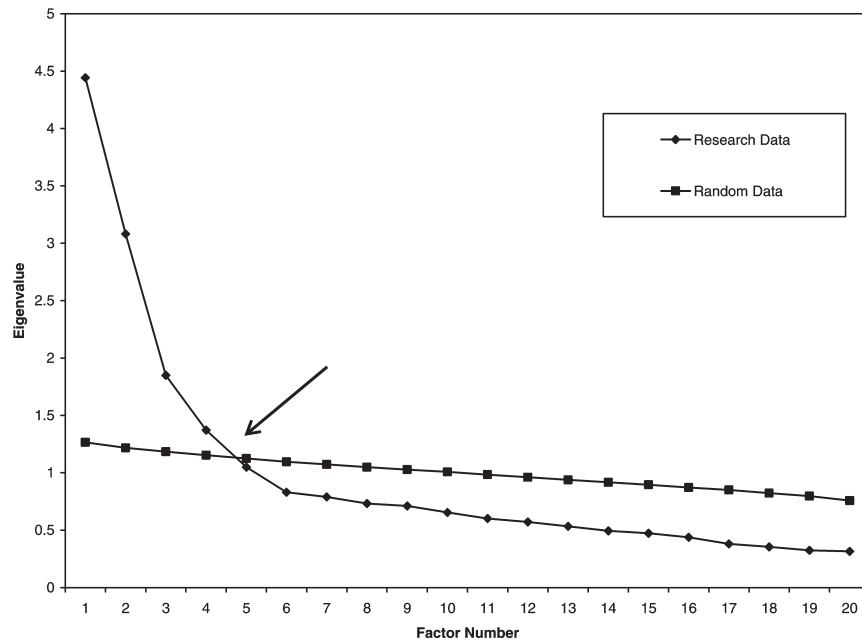
FIGURE 2.3. Parallel analysis using eigenvalues from research and random data (average of 50 replications). Arrow indicates that eigenvalues from random data exceed the eigenvalues from research data after the fourth factor.

scree test it is sometimes associated with somewhat arbitrary outcomes; for instance, chance variation in the input correlation matrix may result in eigenvalues falling just above or below the parallel analysis criterion. A practical drawback of the procedure is that it is not available in major statistical software packages such as SAS and SPSS, although parallel analysis is an option in the *Stata* software and various shareware programs found on the Internet (e.g., O'Connor, 2001). In addition, Hayton, Allen, and Scarpello (2004) have provided syntax for conducting parallel analysis in SPSS, although the user must save and summarize the eigenvalues generated from random data outside of SPSS.

As noted above, when a factor estimation procedure other than ML is employed, eigenvalue-based procedures such as application of the Kaiser–Guttman rule, the scree test, and parallel analysis can be used to assist in factor selection. Although these methods can also assist in determining the appropriate number of factors in ML factor analysis, ML has the advantage of being a full information estimator that allows for goodness-of-fit evaluation and statistical inference such as significance testing and confidence interval estimation. ML is covered extensively in later chapters, so only a brief overview relevant to EFA is provided here. It is helpful to consider ML EFA as a special case of SEM. For example, like CFA and SEM, ML EFA provides goodness-of-fit information that can be used to determine the appropriate number of factors. Various goodness-of-fit statistics (such as χ^2, and the root mean square of approximation, RMSEA; Steiger & Lind, 1980) provide different pieces of information about how well the parameters of the factor model are able to reproduce the sample correlations. As seen earlier in this chapter, the factor loadings of D1 and D2 yielded a predicted correlation of .696 (i.e., Eq. 2.6), which is very similar to the correlation of these indicators in the sample data (i.e., .70; see correlation between D1 and D2 in Table 2.1). If the remaining observed relationships in the input matrix are reproduced as well by the factor loading estimates in this solution, descriptive fit statistics such the χ^2 and RMSEA will indicate that the one-factor model provided a good fit to the data. As shown in Table 2.2, the SPSS output provides a χ^2 test of the fit of the one-factor solution. Because the χ^2 was statistically nonsignificant, $\chi^2(2) = .20$, $p = .90$, it could be concluded that the one-factor model provides a reasonable fit to the data. The nonsignificant χ^2 test suggests the correlation matrix predicted by the factor model parameters does not differ from the sample correlation matrix. However, it will be seen in Chapter 3 that χ^2 has serious limitations, and thus it should not be used as the sole index of overall model fit.

The goal of goodness-of-fit approaches is to identify the solution that reproduces the observed correlations considerably better than more parsimonious models (i.e., models involving fewer factors) but is able to reproduce these observed relationships equally or nearly as well as more complex solutions (i.e., models with more factors). Accordingly, the researcher conducting ML EFA is apt to estimate the factor model several times (specifying different numbers of factors) to compare the fit of the solutions. As in other approaches (e.g., eigenvalue-based methods), factor selection should not be determined by goodness of fit alone, but should be strongly assisted by substantive considerations (e.g., prior theory and research evidence) and other aspects of the resulting solution. Although a factor solution might provide a reasonable fit to the data, it may be unacceptable for other reasons such as the presence of factors that have no strong conceptual basis or utility (e.g., factors arising from methodological artifacts; see Chapter 5), poorly defined factors (e.g., factors in which only one or two indicators have strong primary loadings), indicators that do not have salient loadings on any factor, or indicators that have high loadings on multiple factors. Again, EFA is largely an exploratory procedure, but substantive and practical considerations should strongly guide the factor analytic process. Because of this and other issues (e.g., the role of sampling error), the results of an initial EFA should be interpreted cautiously and should be cross-validated (additional EFAs or CFAs should be conducted using independent data sets).

Factor Rotation

Once the appropriate number of factors has been determined, the extracted factors are rotated, to foster their interpretability. In instances when two or more factors are involved (rotation does not apply to one-factor solutions), rotation is possible because of the indeterminate nature of the common factor model—that is, for any given multiple-factor model, there exist an infinite number of equally good-fitting solutions, each represented by a different factor loading matrix. The term *simple structure* was coined by Thurstone (1947) to refer to the most readily interpretable solutions in which (1) each factor is defined by a subset of indicators that load highly on the factor; and (2) each indicator (ideally) has a high loading on one factor (often referred to as a *primary loading*) and has a trivial or close to zero loading on the remaining factors (referred to as a *cross-loading* or *secondary loading*). In applied research, factor loadings greater than or equal to .30 or .40 are often interpreted as *salient*; that is, the indicator is mean-

ingfully related to a primary or secondary factor. However, explicit or widely accepted guidelines do not exist and the criteria for salient and nonsalient loadings often depend on the empirical context. Thus, for models that contain two or more factors (where an infinite number of equally fitting solutions is possible), rotation is conducted to produce a solution with the best simple structure. It is important to emphasize that rotation does not alter the fit of the solution; for example, in ML EFA, model χ^2 is the same before and after factor rotation. Rather, factor rotation is a mathematical transformation (i.e., rotation in multidimensional space) that is undertaken to foster interpretability by maximizing factor loadings close to 1.0 and minimizing factor loadings close to 0.0. For a mathematical demonstration of this procedure, the reader is referred to Comrey and Lee (1992).

There are two types of rotation: *orthogonal* and *oblique*. In orthogonal rotation, the factors are constrained to be uncorrelated (i.e., factors are oriented at 90° angles in multidimensional space); in oblique rotation, the factors are allowed to intercorrelate (i.e., permit factor axis orientations of less than 90°). The correlation between two factors is equal to the cosine of the angle between the rotational axes. Because cos(90) = 0, the factors are uncorrelated in orthogonal rotation. In oblique rotations, the angle of the axis is allowed to be greater or less than 90°, and thus the cosine of the angle may yield a factor correlation between 0 and 1.0.

In applied social sciences research, orthogonal rotation is used most often, perhaps because it is the default in major statistical programs such as SPSS (*varimax rotation*), and the perception that orthogonally rotated solutions are more easily interpreted because the factor loadings represent correlations between the indicators and the latent factors (e.g., squaring the factor loading provides the proportion of variance in the indicator that the factor solution explains). In oblique solutions, factor loadings usually do not reflect simple correlations between the indicators and the factors unless the factors themselves have no overlap. Because oblique rotations allow the factors to intercorrelate, the correlations between indicators and factors may be inflated by the covariation of the factors; that is, an indicator may correlate with one factor in part through its correlation with another factor. However, orthogonal rotation may often produce misleading solutions in situations where the factors are expected to be intercorrelated; for example, a questionnaire whose latent structure entails several interrelated dimensions of a broader construct. In other words, although substantial correlations may exist among factors, orthogonal rotation constrains the solution to yield uncorrelated latent variables.

Thus, in most cases, oblique rotation is preferred because it provides a more realistic representation of how factors are interrelated. If the factors are in fact uncorrelated, oblique rotation will produce a solution that is virtually the same as one produced by orthogonal rotation. If the factors are interrelated, however, oblique rotation will yield a more accurate representation of the magnitude of these relationships. In addition, estimation of factor correlations provides important information such as the existence of redundant factors or a potential higher-order structure. Factor intercorrelations above .80 or .85 imply poor discriminant validity and suggest that a more parsimonious solution could be obtained (see Chapter 5). If all factors in the solution are moderately intercorrelated at roughly the same magnitude, a single higher-order factor may account for these relationships (see Chapter 8). Moreover, when EFA is used as a precursor to CFA (see Chapter 5), oblique solutions are more likely to generalize to CFA than orthogonal solutions (i.e., constraining factors to be uncorrelated in CFA will typically result in poor model fit).

Several forms of oblique rotation have been developed (e.g., promax, quartamin, orthooblique). When oblique rotation is requested, most software programs (such as SPSS) output both a *pattern matrix* and a *structure matrix*. The loadings in the pattern matrix convey the *unique* relationship between a factor and an indicator. They are interpreted in the same fashion as partial regression coefficients in standard multiple regression; that is, the coefficient represents the relationship between the predictor (latent factor) and outcome (indicator), while controlling for the influence of all other predictors (other latent factors). Thus, indicator variance that is explained by more than one factor is omitted from the loadings in the pattern matrix. The structure matrix is calculated by multiplying the pattern matrix by the factor correlation matrix (oblique rotation produces a factor correlation matrix, orthogonal rotation does not). Hence, loadings in the structure matrix reflect both the unique relationship between the indicator and factor (as in the pattern matrix) *and* the relationship between the indicator and the shared variance among the factors. In other words, the loadings in the structure matrix reflect a zero-order relationship between the indicator and a given factor without holding the other factors in the solution constant. Unless the correlations among factors are minimal, loadings in the structure matrix will typically be larger than those in the pattern matrix because they are inflated by the overlap in the factors (akin to zero-order correlations vs. partial regression coefficients in standard multiple regression). Although there is some debate about whether the pattern matrix or structure matrix should be used, by far the pattern matrix is

most often interpreted and reported in applied research. In fact, some popular latent variable software programs provide only the pattern matrix (e.g., Mplus). As noted above, the mathematical operations for generating the structure matrix are quite straightforward when the pattern matrix and factor intercorrelation are available. Thus, with the aid of software (e.g., SAS PROC IML), either matrix can be readily computed on the basis of the other (i.e., structure matrix = pattern matrix multiplied by the factor correlation matrix; pattern matrix = structure matrix multiplied by the inverse of factor correlation matrix).

The factor intercorrelations produced by an oblique rotation of a PCA solution are often lower than those of obliquely rotated solutions based on the common factor model (e.g., PF, ML; Fabrigar et al., 1999; Widaman, 1993). This is because in common factor analysis, random error is removed from the factors. In PCA, random error is included in the components because PCA does not differentiate common and unique variance. Thus, another potential adverse consequence of PCA is to mistakenly conclude that components share modest variance when in fact the intercorrelations have been attenuated by random error or to conclude that components are distinct when in fact the error-disattenuated correlations would be above .80. Because factor correlations arising from common factor analysis are more likely to be closer to population values, this is another reason why methodologists usually advocate EFA over PCA.

Factor rotation is illustrated in Figure 2.4 using a real data set of eight indicators collected from a sample of 500 participants. A scree test and parallel analysis suggest a two-factor solution. Results indicate that the first four indicators (Y1–Y4) load on Factor 1 and the remaining four indicators (Y5–Y8) load on Factor 2. Figure 2.4 displays geometric representations of unrotated, orthogonally rotated (varimax), and obliquely rotated (promax) factor matrices. ML estimation produced the unrotated factor loadings presented in Figure 2.4A. Figure 2.4B shows the results of the varimax rotation. The factor axes remain at 90° angles, but are rotated in the most optimal fashion to maximize high factor loadings and minimize low loadings. Rotation produces a transformation matrix. Using matrix algebra, the unrotated factor loading matrix is multiplied by the transformation matrix to produce the rotated factor loading matrix. In this data set, the varimax transformation matrix is as follows:

	Factor 1	Factor 2
Factor 1	.93347	.35867
Factor 2	−.35867	.93347

A. Unrotated Factor Matrix
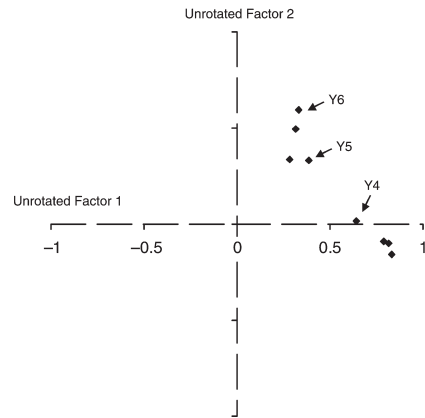
	Factor	
	1	2
Y1	.834	−.160
Y2	.813	−.099
Y3	.788	−.088
Y4	.642	.015
Y5	.386	.329
Y6	.333	.593
Y7	.313	.497
Y8	.284	.336

B. Orthogonally Rotated Factor Matrix (Varimax)
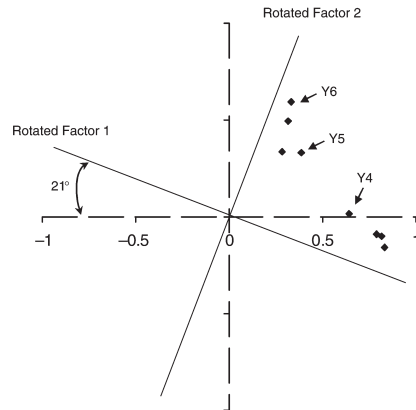
	Factor	
	1	2
Y1	.836	.150
Y2	.794	.199
Y3	.767	.201
Y4	.594	.244
Y5	.242	.445
Y6	.098	.673
Y7	.114	.576
Y8	.145	.416

C. Obliquely Rotated Factor Matrix (Promax)
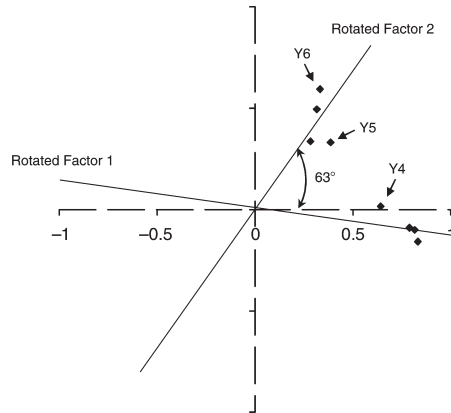
	Factor	
	1	2
Y1	.875	−.062
Y2	.817	.003
Y3	.788	.012
Y4	.588	.106
Y5	.154	.418
Y6	−.059	.704
Y7	−.018	.595
Y8	.055	.413

FIGURE 2.4. Geometric representations of unrotated, orthogonally rotated, and obliquely rotated factor matrices.

34

These values convey how much the axes were rotated to foster simple structure. Specifically, the values on the diagonal (.93347) are cosines and the values on the off-diagonal (.35867, −.35867) are sines and –sines. As shown in Figure 2.4B, the axes were rotated 21° to better transect the clusters of indicators. Within rounding error, the cos(21) equals .933 and the sin(19) equals .359, the same as the transformation coefficients shown above. Because orthogonal rotation was used, the axes of Factor 1 and Factor 2 remain at right angles, and thus the factors are constrained to be uncorrelated, that is, cos(90) = 0.

To witness the effects of rotation on maximizing and minimizing factor loadings, consider the fifth indicator, Y5. Before rotation, the loadings of Y5 on Factor 1 and Factor 2 were very similar (.386 and .329, respectively; Figure 2.4A). A 21° rotation of the factor axes raised Y5's position on the Factor 2 axis (.445), and decreased this indicator's position on the Factor 1 axis (.242) (Figure 2.4B). Although this transformation fosters the interpretability of the solution, it does not alter the communality of Y5 or any other indicator. In a solution entailing more than one latent variable, communalities in an orthogonal EFA are calculated by taking the sum of squared loadings for a given indicator across all factors.[7] Before and after rotation, the proportion of variance explained in Y5 is .257; unrotated solution: $.386^2 + .329^2 = .257$, rotated solution: $.242^2 + .445^2 = .257$. Thus, rotation does not alter the fit of the factor solution.

Figure 2.4B also suggests that oblique rotation may be more appropriate. A factor solution is best defined when the indicators are clustered around the upper end of their respective factor axes. The higher up the axis, the higher the factor loading; if an indicator is in close proximity to one factor axis, it does not load highly on another factor. As shown in Figure 2.4B, orthogonal rotation moved the Factor 2 axis closer to the Y5–Y8 indicators. As a result, the rotation had an overall effect of increasing the primary loadings of Y5–Y8 on Factor 2, and decreasing their cross-loadings on Factor 1 (compared to the unrotated solution, Figure 2.4A). However, orthogonal rotation moved the Factor 1 axis away from the Y1–Y4 indicators, which had the general effect of increasing the magnitude of the cross-loadings of Y1–Y4 on Factor 2; for example, Y4: .015 versus .244 for the unrotated and rotated solutions, respectively. Indeed, in instances where all the indicators fall in between the factor axes after orthogonal rotation (as seen in Figure 2.4B), the restriction of maintaining a 90° orientation of the factor axes may not be tenable. Oblique rotations such as promax begin with an orthogonal rotation, but then "break" the 90° angle to allow the factor axes to pass through the clusters of indicators better.

The angle of the factor axes reflects the factor correlation. If factors are uncorrelated, the angle of factor axes will remain close to 90°. If the factors are correlated, the angle of factor axes will deviate from 90°.

Figure 2.4C provides a geometric depiction of oblique rotation (promax) of the two-factor solution. Note that the axes of Factor 1 and Factor 2 were both turned inward somewhat to better transect the two clusters of indicators. As compared with orthogonal rotation, the oblique rotation increased the values of most primary loadings further. A more notable impact of oblique rotation was its success at moving the cross-loadings closer to zero (but as before, the overall fit of the solution is the same). To accomplish this, the angle of the factor axes was shifted from 90° to 63° (see Figure 2.4C). The results of the analysis indicate that the correlation of Factor 1 and Factor 2 is .45. This corresponds to the cosine of the factor angle; that is, $\cos(63) = .45$.

Factor Scores

After an appropriate factor solution has been established, the researcher may wish to calculate factor scores using the factor loadings and factor correlations. Factor scores are used for various purposes such as to serve as proxies for latent variables, and to determine a participant's relative standing on the latent dimension. Conceptually, a factor score is the score that would have been observed for a person if it had been possible to measure the latent factor directly. In applied research, factor scores are often computed by creating *coarse factor scores*, which are simple unweighted composites of the raw scores of indicators (e.g., averaging or summing) found to have salient loadings on the factor. However, there are many reasons why coarse factor scores may poorly represent latent factors (e.g., they may be highly intercorrelated even when the factors are truly orthogonal; Grice, 2001). Alternatively, factor scores can be estimated by multivariate methods that use various aspects of the reduced or unreduced correlation matrix and factor analysis coefficients. The resulting values are called *refined factor scores*. A frequently used method of estimating refined factor scores is Thurstone's (1935) least squares regression approach, although several other strategies have been developed (e.g., Bartlett, 1937; Harman, 1976; McDonald, 1981). Most statistical software packages provide options to compute refined factor scores by one or more of these methods. In the majority of instances, refined factor scores have less bias than coarse factor scores and thus are favored over coarse factor scores as proxies for latent factors (Grice, 2001). However, a complicating issue in factor score estimation is the indeterminate nature of the common factor

model. With respect to factor scores, this indeterminacy means that there is an infinite number of sets of factor scores that could be computed from any given factor analysis that would be equally consistent with the same factor loadings (Grice, 2001). The degree of indeterminacy depends on several aspects, such as the ratio of items to factors and the size of the item communalities (e.g., factors defined by several items with strong communalities have better determinacy). If a high degree of indeterminacy is present, the sets of factor scores can vary so widely such that an individual ranked high on the dimension in one set may receive a low ranking on the basis of another set. In such scenarios, the researcher has no way of discerning which set of scores or rankings is most accurate.

Thus, although typically neglected in applied factor analytic research, the degree of factor score indeterminacy should be examined as part of EFA, especially in instances when factor scores are to be computed for use in subsequent statistical analyses. Grice (2001) has specified three criteria for evaluating the quality of factor scores: (1) *validity coefficients*—correlations between the factor score estimates and their respective factor scores; (2) *univocality*—the extent to which the factor scores are excessively or insufficiently correlated with other factors in the same analysis; and (3) *correlational accuracy*—how closely the correlations among factor scores correspond to the correlations among the factors. For instance, Gorsuch (1983) has recommended that validity coefficients should be at least .80, although higher values (e.g., > .90) may be required in some situations (e.g., when factor scores are used as dependent variables). Unfortunately, procedures for evaluating factor scores are not standard options in most software packages. An exception is the Mplus program, where validity coefficients can be requested as part of CFA or EFA within CFA by using the FSDETERMINACY option of the OUTPUT command (see Chapter 5). To fill this void, Grice (2001) has developed SAS PROC IML computer code for assessing the degree of factor score indeterminacy (validity coefficients, univocality, correlational accuracy) in the context of EFA (these programs can be downloaded from psychology.okstate.edu/faculty/jgrice/factorscores/).

SUMMARY

Procedural recommendations for conducting applied EFA are summarized in Table 2.3. In addition to providing a practical overview of exploratory factor analysis (e.g., procedural considerations for factor estimation, selection, rotation, and interpretation), the goal of this chapter was to introduce

TABLE 2.3. Fundamental Steps and Procedural Recommendations for EFA

Factor Extraction

- Use an estimator based on the common factor model, such as:
  - —*Principal factors*: no distributional assumptions, less prone to improper solutions than maximum likelihood
  - —*Maximum likelihood*: assumes multivariate normality, but provides goodness of fit evaluation and, in some cases, significance tests and confidence intervals of parameter estimates

Factor Selection

- Determine the appropriate number of factors by:
  - —*Scree plot* of eigenvalues from the reduced correlation matrix,
  - —*Parallel analysis*, and/or
  - —*Goodness of model fit* (e.g., χ^2, RMSEA; see Chapter 3)

Factor Rotation

- In multifactorial models, rotate the solution to obtain simple structure by:
  - —Using an *oblique rotation* method (e.g., promax, quartimin)

Interpret the Factors and Evaluate the Quality of the Solution

- Consider the meaningfulness and interpretability of the factors:
  - —Factors should have substantive meaning and conceptual/empirical relevance
  - —Rule out nonsubstantive explanations such as method effects (e.g., factors comprised of reverse- and non-reverse-worded items; see Chapters 3 and 5)
- Eliminate poorly defined factors, such as:
  - —Factors on which only two or three items have salient loadings
  - —Factors defined by items that have small loadings (i.e., *low communalities*)
  - —Factors with low *factor determinacy* (poor correspondence between the factors and their factor scores; see Grice, 2001)
- Eliminate poorly behaved items (indicators), such as:
  - —Items with high loadings on more than one factor (i.e., *cross-loadings*)
  - —Items with small loadings on all factors (i.e., *low communalities*)

Re-Run and (Ideally) Replicate the Factor Analysis

- If items or factors are dropped in preceding step, re-run the EFA in the same sample
- Replicate the final EFA solution in an independent sample
- Consider further replications/extensions of the factor solution by:
  - —Developing tentative CFA models (e.g., EFA in the CFA framework; see Chapter 5)
  - —Larger-scale CFA investigations
  - —*Measurement invariance* evaluation in population subgroups (e.g., equivalence of solution between sexes; see Chapter 7)

Note. EFA, exploratory factor analysis; RMSEA, root mean square error of approximation; CFA, confirmatory factor analysis.

key concepts that will be carried forward in the subsequent chapters on CFA (e.g., observed vs. latent variables, factor loadings, factor correlations, common and unique variance, basic equations and notation). Some fundamental differences of EFA and CFA were described. Unlike EFA, in CFA the number of factors and the pattern of indicator–factor loadings are specified in advance on the basis of strong empirical knowledge or theory. The acceptability of the CFA model is evaluated in part by descriptive fit statistics that convey the ability of the solution to reproduce the observed relationships among the input indicators, although similar testing can be applied in EFA when the ML estimator is used. As will be seen in Chapter 3, EFA and CFA differ in several other important manners.

NOTES

1. Exceptions to this rule are discussed in Chapters 3 and 4 (e.g., when indicator measurement errors are correlated).

2. For instance, in the current example, which entails a single factor, the factor loadings can be interpreted as zero-order correlation coefficients between the latent factor and the observed measures (i.e., factor loading = the standardized regression slope = zero-order correlation). In solutions involving multiple, correlated factors (oblique rotation), factor loadings from the factor pattern matrix are interpreted as partial regression coefficients.

3. The reader will encounter many variations in this notational system across factor analysis and SEM texts. For instance, because indicator unique variances (ε) are not observed, it is common to see these parameters depicted as circles in path diagrams. In Chapter 3, this notation is expanded by differentiating latent X (exogenous) and latent Y (endogenous) solutions.

4. In PCA, the limit on the number of components is equal to p.

5. Because eigenvalues are drawn from the unreduced correlation matrix ($\mathbf{R}$), PCA is always conducted initially regardless of the type of factor analysis requested (e.g., PF).

6. Some researchers (e.g., Glorfeld, 1995) have recommended that the 95th percentile of eigenvalues from random data be used in place of average eigenvalues, in part to adjust for parallel analysis's slight tendency to overfactor (regardless of method used, research has shown that parallel analysis is accurate in the vast majority of cases; e.g., Humphreys & Montanelli, 1975; Zwick & Velicer, 1986).

7. Although communalities can also be hand calculated from the estimates of an obliquely rotated EFA solution, this computation is less straightforward because the factors are permitted to be intercorrelated and thus the factor loadings are partial regression coefficients. Later chapters (e.g., Chapter 3) discuss the tracing rules necessary to compute these estimates.

3

Introduction to CFA

The purpose of this chapter is to introduce the reader to the purposes, parameters, and fundamental equations of CFA. Now that the common factor model and EFA have been described in Chapter 2, CFA and EFA are compared more thoroughly. On the basis of these comparisons, the advantages and purposes of CFA will become apparent. The notation and computation of the parameters of the CFA model are presented. This chapter also deals with the important concepts of model identification, maximum likelihood estimation, and goodness-of-fit evaluation. The concepts introduced in this chapter should be studied carefully, as they are germane to all examples of CFA presented in subsequent chapters of this book.

SIMILARITIES AND DIFFERENCES OF EFA AND CFA

Common Factor Model

Like EFA, the purpose of CFA is to identify latent factors that account for the variation and covariation among a set of indicators. Both EFA and CFA are based on the common factor model, and thus many of the concepts and terms that were discussed in Chapter 2 apply to CFA, such as factor loadings, unique variances, communalities, and residuals. However, while EFA is generally a descriptive or exploratory procedure, in CFA the researcher must prespecify all aspects of the factor model: the number of factors, the pattern of indicator–factor loadings, and so forth. As noted in Chapter 2, CFA requires a strong empirical or conceptual foundation to

guide the specification and evaluation of the factor model. Accordingly, CFA is typically used in later phases of scale development or construct validation after the underlying structure has been tentatively established by prior empirical analyses using EFA, as well as on theoretical grounds.

EFA and CFA often rely on the same estimation methods (e.g., *maximum likelihood*, or ML). When a full information estimator such as ML is used, the factor models arising from EFA and CFA can be evaluated in terms of how well the solution reproduces the observed variances and covariances among the input indicators (i.e., goodness-of-fit evaluation). In addition, the quality of EFA and CFA models is determined in part by the size of resulting parameter estimates (e.g., magnitude of factor loadings and factor intercorrelations) and how well each factor is represented by observed measures (e.g., number of indicators per factor, size of indicator communalities, factor determinacy).

Standardized and Unstandardized Solutions

The tradition in EFA is to *completely standardize* all variables in the analysis. Specifically, a correlation matrix is used as input in EFA and both the latent factors and indicators are completely standardized: factor variances equal 1.0; factor loadings are interpreted as correlations or standardized regression coefficients.[1] Although CFA also produces a completely standardized solution, much of the analysis does not standardize the latent or observed variables. Instead of using a correlation matrix (i.e., a correlation matrix is a completely standardized variance–covariance matrix), CFA typically analyzes a variance–covariance matrix (needed to produce an unstandardized CFA solution) or raw data that are used by the software program to produce an input variance–covariance matrix. Thus, the CFA input matrix is comprised of indicator variances on the diagonal (a variance equals the indicator's standard deviation squared; i.e., $V = SD^2$), and indicator covariances in the off-diagonal (a covariance can be calculated by multiplying the correlation of two indicators by their *SD*s; i.e., $COV_{xy} = r_{xy}SD_xSD_y$). In addition to a completely standardized solution, the results of CFA include an *unstandardized solution* (parameter estimates expressed in the original metrics of the indicators) and possibly a *standardized solution* (relationships involving unstandardized indicators and standardized latent variables). Unstandardized, standardized, and completely standardized solutions are discussed in more detail in Chapter 4. Of particular note here is the fact that many key aspects of CFA are based on unstandardized estimates, such as the standard errors and significance testing of model

parameters. As will be seen in Chapter 7, the various forms of measurement invariance evaluation (e.g., equivalence of parameter estimates within and across groups) are also based on the unstandardized solution.

The unstandardized means of the indicators can also be included in the CFA. Thus, in contrast to EFA, which focuses on completely standardized values, CFA may entail the analysis of both unstandardized variance–covariance structures and mean structures (as the result of standardization in EFA, indicator means are presumed to be zero). As discussed in Chapter 7, when indicator means are included as input in CFA, the analysis can estimate the means of the latent factors and the intercepts of the indicators. Akin to multiple regression, an indicator intercept is interpreted as the predicted value of the indicator when the latent factor—or predictor—is zero. The analysis of mean structures is particularly relevant to multiple-groups CFA models where the researcher may be interested in comparing groups on the latent means (the SEM parallel to analysis of variance) or determining the equivalence of a testing instrument's measurement properties across groups (e.g., inequality of item intercepts is indicative of test bias/differential item functioning; see Chapter 7).

The outcome of EFA is reported as a completely standardized solution. In applied CFA research, completely standardized solutions are most commonly reported. However, SEM methodologists often express a strong preference for reporting unstandardized solutions because the analysis itself is based on unstandardized variables, and completely standardized values are potentially misleading. For instance, the true nature of the variance and relationships among indicators and factors can be masked when these variables have been standardized, and when the original metric of variables is expressed in meaningful units, unstandardized estimates more clearly convey the importance or substantive significance of the effects (cf. Willett, Singer, & Martin, 1998). Chapter 4 demonstrates that, through basic equations, one can readily calculate a (completely) standardized solution from the unstandardized solution, and vice versa.

Indicator Cross-Loadings/Model Parsimony

In addition, EFA and CFA differ markedly in the manner by which indicator cross-loadings are handled in solutions entailing multiple factors (in unidimensional models, the issues of cross-loadings and factor rotation are irrelevant). As noted in Chapter 2, all indicators in EFA freely load on all factors and the solution is rotated to maximize the magnitude of primary

loadings and minimize the magnitude of cross-loadings. Factor rotation does not apply to CFA. This is because the identification restrictions associated with CFA are achieved in part by fixing most or all indicator cross-loadings to zero (see Figure 3.1). In other words, rotation is not necessary in CFA because simple structure is obtained by specifying indicators to load on just one factor. CFA models are typically more parsimonious than EFA solutions because while primary loadings and factor correlations are freely estimated, no other relationships are specified between the indicators and factors (e.g., no loadings that relate indicators Y1–Y4 to Factor 2, and indicators Y5–Y8 to Factor 1; see Figure 3.1A). Thus, with few exceptions (cf. EFA in the CFA framework, Chapter 5), CFA attempts to reproduce the observed relationships among input indicators with fewer parameter estimates than EFA.

Table 3.1 presents the factor loading matrices of three analyses of the same data set ($N = 1,050$ adolescents): (1) a CFA (Model A); (2) an EFA with oblique rotation (Model B); and (3) an EFA with orthogonal rotation (Model C). Eight antisocial behaviors were used as indicators in the analyses. Each analysis entailed two factors: Property Crimes (e.g., shoplifting, vandalism) and Violent Crimes (e.g., fighting, aggravated assault). The path diagrams of Models A and B in Figure 3.1 correspond to Models A and B in Table 3.1. The Model B path diagram could be edited to conform to an orthogonal EFA by removing the double-headed curved arrow reflecting the factor correlation. As can be seen in Table 3.1, in EFA each indicator loads on all factors. Rotation (either orthogonal or oblique) is used to foster the interpretability of the factor loadings (i.e., maximize high loadings, minimize low loadings). For the reasons stated in Chapter 2, rotation does not affect the fit of the EFA solution; that is, the indicator communalities are identical in orthogonal and oblique EFA (see Table 3.1 and Appendix 3.1). The CFA model is more parsimonious than the EFA models because all indicator cross-loadings are prespecified to equal zero; that is, Y1–Y4 on Violent Crimes = 0, Y5–Y8 on Property Crimes = 0 (see Model A in Table 3.1 and Figure 3.1). Thus, there are only 8 factor loading estimates in the CFA, as compared with 16 factor loading estimates in the EFAs. Accordingly, rotation in CFA is not required.

Another consequence of fixing cross-loadings to zero in CFA is that factor correlation estimates in CFA tend to be of higher magnitude than analogous EFA solutions. This can be seen in the current data set where the factor correlations between Property Crimes and Violent Crimes were .57 and .62 for the oblique EFA and CFA, respectively (Table 3.1). The rea-

Model A: Confirmatory Factor Model (all measurement error is random)
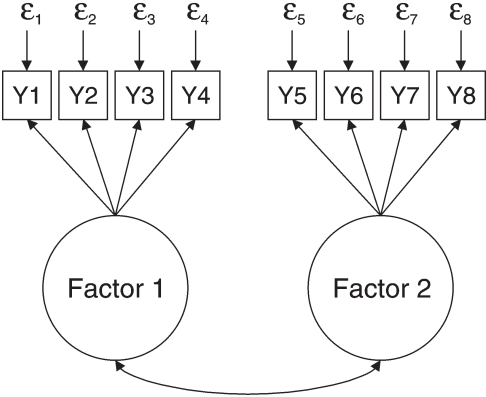
Model B: Exploratory Factor Model (oblique rotation)
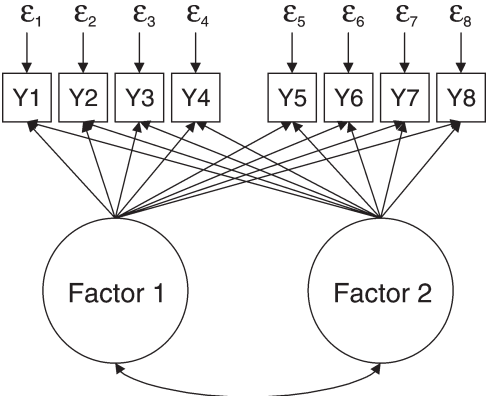
Model C: Confirmatory Factor Model (with a correlated measurement error)
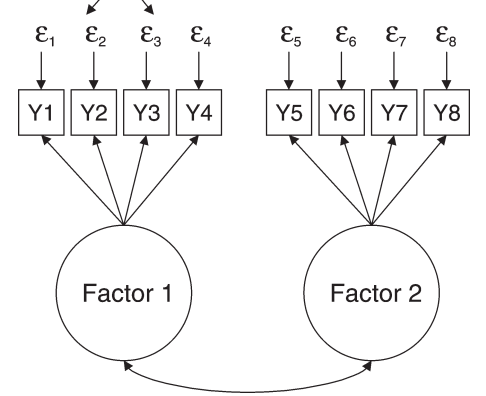
FIGURE 3.1. Path diagrams of confirmatory and exploratory factor models.

TABLE 3.1. Factor Loadings Matrices from EFA and CFA
of Adolescent Antisocial Behaviors

A. CFA (factor correlation = .6224)

| | Factor | | Communality |
	Property Crimes	Violent Crimes	
Y1	.7996	.0000	.64
Y2	.6451	.0000	.42
Y3	.5699	.0000	.32
Y4	.4753	.0000	.23
Y5	.0000	.7315	.53
Y6	.0000	.5891	.35
Y7	.0000	.7446	.55
Y8	.0000	.5803	.34

B. EFA (oblique rotation, factor correlation = .5722)

| | Factor | | Communality |
	Property Crimes	Violent Crimes	
Y1	.9187	−.0958	.75
Y2	.5422	.1045	.37
Y3	.5300	.0372	.30
Y4	.4494	.0103	.21
Y5	.0434	.7043	.53
Y6	−.1178	.6999	.41
Y7	.1727	.6106	.52
Y8	.0264	.5756	.35

C. EFA (orthogonal rotation, factor correlation = 0)

| | Factor | | Communality |
	Property Crimes	Violent Crimes	
Y1	.8493	.1765	.75
Y2	.5509	.2574	.37
Y3	.5185	.1898	.30
Y4	.4331	.1408	.21
Y5	.2587	.6826	.53
Y6	.1032	.6314	.41
Y7	.3535	.6312	.52
Y8	.2028	.5552	.35

Note. N = 1,050. Y1 = shoplifting, Y2 = vandalism, Y3 = theft, Y4 = broke into building/vehicle, Y5 = fighting, Y6 = aggravated assault, Y7 = hit family/teachers, Y8 = threatened others.

son for this outcome, along with an explanation of the communality and model-implied estimates in EFA and CFA, is provided in Appendix 3.1.

Unique Variances

Unlike EFA, the CFA framework offers the researcher the ability to specify the nature of relationships among the measurement errors (unique variances) of the indicators. Although both EFA and CFA differentiate common and unique variances, within EFA the specification of relationships among unique variances is not made. Because CFA typically entails a more parsimonious solution (i.e., CFA usually attempts to reproduce the observed relationships among indicators with fewer parameter estimates than EFA), it is possible to estimate such relationships when this specification is substantively justified and other identification requirements are met (see Chapter 5). Consequently, because of EFA's identification restrictions, factor models must be specified under the assumption that measurement error is random. In contrast, correlated measurement error can be modeled in a CFA solution. The CFA model presented in Model A of Figure 3.1 (i.e., Figure 3.1A) depicts a two-factor measurement model where all measurement error is presumed to be random. The underlying assumption of this specification is that the observed relationship between any two indicators loading on the same factor (e.g., Y7 and Y8) is due entirely to the shared influence of the latent dimension; that is, if Factor 2 was partialed out, the intercorrelations between these indicators would be zero. The model presented in Figure 3.1C depicts the same CFA measurement model, with the exception that a correlated error has been specified between Y2 and Y3. This specification assumes that, whereas indicators Y2 and Y3 are related in part because of the shared influence of the latent dimension (Factor 1), some of their covariation is due to sources other than the common factor. In measurement models, the specification of correlated errors may be justified on the basis of source or method effects that reflect additional indicator covariation that resulted from common assessment methods (e.g., observer ratings, questionnaires), reversed or similarly worded test items, or differential susceptibility to other influences such as response set, demand characteristics, acquiescence, reading difficulty, or social desirability (cf. Brown, 2003; Marsh, 1996). The specification in Figure 3.1C depicts a *correlated uniqueness* approach to modeling error covariances (i.e., zero-order relationships are freely estimated between pairs of indicators). As noted in the discussion of multitrait–multimethod CFA solutions (Chapter 6), the relationships among indica-

tor errors can also be modeled using a *correlated methods* approach that entails the specification of methods factors in addition to the latent factors of substantive interest.

The inability to specify correlated errors (i.e., the nature of the relationships among unique variances) is a very significant limitation of EFA. For instance, in applied factor analytic research of questionnaires comprised of a combination of positively and reversed worded items, a common consequence of this EFA limitation is the tendency to extract and interpret methods factors that have little substantive basis (cf. Brown, 2003; Marsh, 1996). For example, a very extensive psychometric literature exists on the Rosenberg (1965) Self-Esteem Scale (SES), a questionnaire that consists of four positively worded items (e.g., "I feel good about myself") and three negatively worded items (e.g., "At times I think I am no good at all"). Early EFA research routinely produced two SES factors comprised of negatively and positively worded items that were interpreted as substantively meaningful (e.g., "positive self-evaluation" vs. "negative self-evaluation"). However, as compellingly argued by Marsh (1996), a strong conceptual basis did not exist in support for distinct dimensions of positive and negative self-esteem. Instead, Marsh (1996) noted that these two-factor solutions were an artifact of response styles associated with the wording of the items (e.g., response biases such as acquiescence). Using CFA, Marsh (1996) evaluated various SES measurement models corresponding to previously reported solutions (e.g., one-factor model without error covariances, two-factor models) and correlated uniqueness (residual) models. Results indicated the superiority of a unidimensional solution ("Global Self-Esteem") with method effects (correlated residuals) associated with the negatively worded items. Although having a compelling substantive basis (i.e., the existence of a single dimension of self-esteem, but need for an error theory to account for the additional covariation among similarly worded items), this model could not be estimated in EFA because EFA does not allow for the specification of correlated indicator errors.

Model Comparison

The preceding sections of this chapter have documented some of the key ways that CFA offers greater modeling flexibility than EFA (i.e., prespecification of the number of factors, patterns of item–factor relationships, presence or absence of error covariances). In addition to these aspects, the CFA framework allows the researcher to impose other restrictions on the

factor solution, such as constraining all the factor loadings or all the unique variances to be equal (e.g., as in the evaluation of the conditions of tau equivalence or parallel tests; see Chapter 7). The viability of these constraints can be evaluated by statistically comparing whether the fit of the more restricted solution is worse than a comparable solution without these constraints. Direct statistical comparison of alternative solutions is possible when the models are nested. As discussed further in Chapter 5, a *nested model* contains a subset of the free parameters of another model, which is often referred to as the *parent model*. For example, consider the following two models: (1) *Model P*: a one-factor model comprised of six indicators allowed to freely load onto the factor; and (2) *Model N*: a one-factor model identical to Model P, except that the factor loadings are constrained to load equally onto the factor. Although the models are structurally the same (i.e., they consist of one factor and the same six indicators), they differ in their number of freely estimated versus constrained parameters. When parameters are *freely estimated*, the researcher allows the analysis to find the values for the parameters in the CFA solution (e.g., factor loadings, factor correlations, unique variances) that optimally reproduce the variances and covariances of the input matrix. In the case of *fixed parameters*, the researcher assigns specific values (e.g., fixes cross-loadings to zero to indicate no relationship between an indicator and a factor; cf. lack of an arrow between Y1 and Factor 2 in Figure 3.1A). When parameters are *constrained*, the researcher does not specify the parameters' exact values, but places other restrictions on the magnitude these values can take on (e.g., in the case of Model N, the researcher instructs the analysis to optimally reproduce the input matrix under the condition that all factor loadings are the same). Thus, Model N is nested under Model P (the parent model) because it contains a subset of Model P's free parameters. Accordingly, the fit of Model N can be statistically compared to the fit of Model P (through methods such as the χ^2 difference test; see Chapter 4) to directly evaluate the viability of the condition of equal factor loadings (i.e., tau equivalence = do the six indicators relate equally to the latent factor?). Because EFA entails only freely estimated parameters (fixed parameters cannot be specified), comparative model evaluation of this nature is not possible.

These procedures (e.g., χ^2 difference testing) can be used to statistically compare other forms of nested models in CFA. For instance, CFA can be used to statistically determine whether the various measurement parameters of a factor model (e.g., factor loadings) are the same in two or more groups (e.g., males and females; see Chapter 7).

PURPOSES AND ADVANTAGES OF CFA

From the preceding sections of this chapter, the objectives and advantages of CFA in relation to EFA may now be apparent. Although both EFA and CFA are based on the common factor model and often use the same estimation method (e.g., maximum likelihood), the specification of CFA is strongly driven by theory or prior research evidence. Thus, unlike the approach in EFA, in which the researcher can only prespecify the number of factors, the CFA researcher usually tests a much more parsimonious solution by indicating the number of factors, the pattern of factor loadings (and cross-loadings, which are usually fixed to zero), and an appropriate error theory (e.g., random or correlated indicator error). In contrast to EFA, CFA allows for the specification of relationships among the indicator uniquenesses (error variances), which may have substantive importance (e.g., correlated errors due to method effects). Thus, every aspect of the CFA model is specified in advance. The acceptability of the specified model is evaluated by goodness of fit and the interpretability and strength of the resulting parameter estimates (overall goodness of fit also applies to EFA when the ML estimator is used). As noted previously, CFA is more appropriate than EFA in the later stages of construct validation and test construction when prior evidence and theory support "more risky" a priori predictions regarding latent structure. For example, the modeling flexibility and capabilities of CFA (e.g., specification of an error theory) afford sophisticated analyses of construct validity, such as in the multitrait–multimethod approach (see Chapter 6) where the convergent and discriminant validity of dimensions are evaluated in context of (partialing out the influence of) varying assessment methods.

In addition, CFA offers a very strong analytic framework for evaluating the equivalence of measurement models across distinct groups (e.g., demographic groups such as sexes, races, or cultures). This is accomplished by either multiple-group solutions (i.e., simultaneous CFAs in two or more groups) or MIMIC models (i.e., the factors and indicators are regressed onto observed covariates representing group membership; see Chapter 7). Although cursory methods of examining the concordance of factor structures within EFA are available (e.g., Ahmavaara, 1954), the CFA framework is superior in terms of its modeling flexibility (e.g., ability to specify partial invariance models; cf. Byrne, Shavelson, & Muthén, 1989) and its ability to examine every potential source of invariance in the factor solution, including latent means and indicator intercepts. These

capabilities permit a variety of important analytic opportunities in applied research, such as the evaluation of whether a scale's measurement properties are invariant across population subgroups; for example, are the number of factors, factor loadings, item intercepts, and so forth, that define the latent structure of a questionnaire equivalent in males and females? Measurement invariance is an important aspect of scale development, as this endeavor determines whether a testing instrument is appropriate for use in various groups (Chapter 7). Indeed, multiple-groups CFA can be used to evaluate the generalizability of a variety of important constructs (e.g., are the diagnostic criteria sets used to define mental disorders equivalent across demographic subgroups such as race and gender?). Moreover, this approach can be used to examine group differences in the means of the latent dimensions. While analogous to analysis of variance (ANOVA), the CFA approach is superior to ANOVA because group comparisons are made in the context of measurement invariance (unlike ANOVA, which simply assumes that a given observed score reflects the same level of the latent construct in all groups).

Similarly, another advantage of CFA and SEM is the ability to estimate the relationships among variables adjusting for measurement error. A key limitation of ordinary least squares (OLS) approaches such as correlational and multiple regression analysis is the assumption that variables have been measured without error; that is, they are perfectly reliable, meaning that all of an observed measure's variance is true score variance. However, this assumption rarely holds in the social and behavioral sciences, which rely heavily on variables that have been assessed by questionnaires, independent observer ratings, and the like. Consequently, estimates derived from OLS methods (e.g., correlations, regression coefficients) are usually attenuated to an unknown degree by measurement error in the variables used in the analysis.[2] However, CFA and SEM allow for such relationships to be estimated after adjusting for measurement error and an error theory (extent of random and correlated measurement error). For example, in the CFA model presented in Figure 3.1A, the relationship between the two constructs would be reflected by their factor intercorrelation (r between Factor 1 and Factor 2), as opposed to the observed relationships among the indicators that load on these factors. Indeed, this factor correlation is a better estimate of the population value of this relationship than any two indicator pairings (e.g., r between Y1 and Y4) because it has been adjusted for measurement error; that is, shared variance among the factor's indicators is operationalized as true score variance, which is passed onto the latent factor.

In Chapter 2, the problems with the computation and use of factor scores in EFA were briefly noted. On occasion, a researcher will wish to relate the factors revealed by EFA to other variables. Typically (but see Gorsuch, 1997), this requires the researcher to compute factor scores to serve as proxies for the latent factors in subsequent analyses. However, this practice is limited by the issue of factor score indeterminacy: for any given EFA, there are an infinite number of sets of factor scores that could be computed that are equally consistent with the factor loadings (see Chapter 2). In CFA and SEM, indeterminacy of factor scores is not a problem because this analytic framework eliminates the need to compute factor scores; that is, the latent factors themselves are used in the analysis. Unlike EFA, CFA/SEM offers the researcher considerable modeling flexibility, such that additional variables can be readily brought into the analysis to serve as correlates, predictors, or outcomes of the latent variables (e.g., see MIMIC models, Chapter 7).

Frequently, CFA is used as a precursor to SEM models that specify structural relationships (e.g., regressions) among the latent variables. SEM models can be broken down into two major components: (1) the *measurement model*, which specifies the number of factors, how the various indicators are related to the latent factors, and the relationships among indicator errors (i.e., a CFA model); and (2) the *structural model*, which specifies how the various latent factors are related to one another (e.g., direct or indirect effects, no relationship, spurious relationship). Consider the two basic path diagrams in Figure 3.2. Whereas the two diagrams depict models entailing the same set of indicators and the same latent factors, the first diagram (A) represents a measurement model (a CFA model entailing three intercorrelated factors) and the second diagram (B) reflects a structural model to indicate that the relationship between Factor X and Factor Y is fully mediated by Factor Z (as with factor loadings, direct effects among latent variables are depicted by unidirectional arrows in Figure 3.2B). Thus, whereas the relationships among the latent variables are allowed to freely intercorrelate in the CFA model (analogous to an oblique EFA solution), the exact nature of the relationships is specified in the structural model; that is, Factor X has a direct effect on Factor Z, Factor Z has a direct effect on Factor Y, and Factor X has an indirect effect on Factor Y. Note that in the measurement (CFA) model, there are three parameters relating the factors to one another: factor correlations between X and Y, X and Z, and Y and Z (depicted by double-headed, curved arrows in Figure 3.2A). In the structural model, there are only two structural parameters, $X \rightarrow Y$ and $Y \rightarrow Z$. As discussed later in this chapter, the structural portion

Model A: Measurement Model
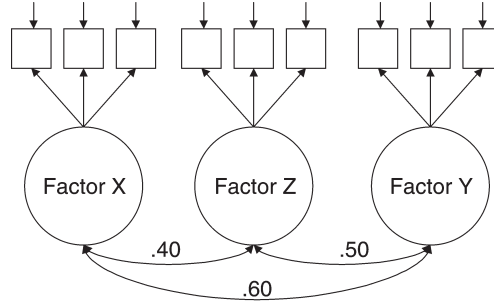
Model B: Structural Model
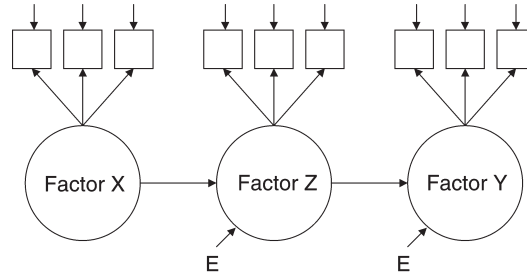
FIGURE 3.2. Path diagrams of measurement and structural models.

of this solution is overidentified, meaning that there exist fewer structural parameters (i.e., 2: X → Y and Y → Z) in the model than the number of possible relationships among the latent factors (i.e., 3: correlations between X and Y, X and Z, and Y and Z). Thus, the structural model is more parsimonious than the measurement model because it attempts to reproduce the relationships among the latent variables with one less freely estimated parameter. Because of the overidentified nature of the structural portion of this model, its goodness of fit may be poorer than that of the measurement model. As illustrated by a tracing rule presented later in this chapter (e.g., Eq. 3.16), the structural portion of this model will result in poor fit if the product of the Factor X → Factor Z path and Factor Z → Factor Y path does not closely approximate the correlation between Factors X and Y estimated in the measurement model. Indeed, the indirect effects structural model in Figure 3.2B would be poor fitting because the product of the X → Z and Z → Y direct effects [(.40)(.50) = .20] does not approximate the correlation between Factors X and Y (.60; see Figure 3.2A).

The purpose of this discussion is to illustrate that goodness of model fit is determined by how adequately both the measurement and structural portions of a model are specified. A key aspect of CFA evaluation is the ability of the parameters from the measurement model (e.g., factor loadings and factor correlations) to reproduce the observed relationships among the indicators. If the CFA model is misspecified (e.g., failure to specify the correct number of factors, pattern of factor loadings, etc.), a poor-fitting solution will result. However, poor fit may also arise from a misspecified structural model, which, like the model depicted in Figure 3.2B, often possesses fewer freely estimated parameters than its corresponding measurement model. Because there are a variety of potential sources of poor fit in CFA models involving multiple indicators, the researcher should establish a viable measurement model prior to pursuing a structural solution. If model testing is initiated with a structural solution, it is usually difficult to determine the extent to which poor fit is attributable to the measurement and structural aspects of the solution. For instance, consider the scenario where the measurement model in Figure 3.2A is well specified (i.e., good-fitting, strong and interpretable factor loadings and factor correlations), but the researcher begins by testing the structural model shown in Figure 3.2B. Although poor fit would be due to the misspecified structural model (i.e., inability to reproduce the Factor X and Factor Y relationship), the researcher may falsely suspect the measurement aspect of the model. In most cases, poor fit cannot arise from the structural portion of a CFA measurement model because the factors are usually specified as freely intercorrelated. Thus, CFA solutions are a useful prelude to SEM models that aim to reproduce the relationships among latent variables with a more parsimonious set of structural parameters.

PARAMETERS OF A CFA MODEL

All CFA models contain factor loadings, unique variances, and factor variances. Factor loadings are the regression slopes for predicting the indicators from the latent factor. Unique variance is variance in the indicator that is not accounted for by the latent factors. Unique variance is typically presumed to be measurement error and is thus often referred to as such (other synonymous terms include "error variance" and "indicator unreliability"). In an unstandardized solution, a factor variance expresses the sample variability or dispersion of the factor; that is, the extent to which sample participants' relative standing on the latent dimension is similar or different. If

substantively justified, a CFA may include error covariances (referred to as "correlated uniquenesses," "correlated residuals," or "correlated errors"), which suggest that two indicators covary for reasons other than the shared influence of the latent factor (e.g., see Figure 3.1C). In CFA solutions, error covariances are often specified on the basis of method effects (e.g., the indicators were measured by a common method), although other sources of these relationships are possible. When the CFA solution consists of two or more factors, a factor covariance (a "factor correlation" being the completely standardized counterpart) is usually specified to estimate the relationship between the latent dimensions. However, one may fix factor covariances to zero, akin to an orthogonal EFA solution.

CFA is often confined to the analysis of variance–covariance structures. In this instance, the aforementioned parameters (factor loadings, error variances and covariances, factor variances and covariances) are estimated to reproduce the input variance–covariance matrix. The analysis of covariance structures is based on the implicit assumption that indicators are measured as deviations from their means (i.e., all indicator means equal zero). However, the CFA model can be expanded to include the analysis of mean structures, in which case the CFA parameters also strive to reproduce the observed sample means of the indicators (which are included along with the sample variances and covariances as input data). Accordingly, such CFA models also include parameter estimates of the indicator intercepts (predicted value of the indicator when the factor is zero) and the latent factor means, which are often used in multiple-groups CFA to test whether distinct groups differ in their relative standing on latent dimensions (see Chapter 7).

Latent variables in CFA may be either exogenous or endogenous. An *exogenous variable* is a variable that is not caused by other variables in the solution (such as Factor X in Figure 3.2B). Conversely, an *endogenous variable* is caused by one or more variables in the model (i.e., other variables in the solution exert direct effects on the variable, as in Factor Y in Figure 3.2B). Thus, exogenous variables can be viewed as synonymous to X, independent, or predictor (causal) variables. Similarly, endogenous variables are equivalent to Y, dependent, or criterion (outcome) variables. However, in the case of structural models, an endogenous variable may be the cause of another endogenous variable, as is the case of Factor Z in Figure 3.2B.

CFAs are typically considered to be exogenous (latent X) variable solutions (e.g., Figure 3.2A). However, when the CFA analysis includes covariates (i.e., predictors of the latent factors or indicators as in MIMIC models; Chapter 7) or higher-order factors (see Chapter 8), the latent factors are endogenous (latent Y). Even when the CFA is a pure measurement

model, some researchers (e.g., methodologists using LISREL software) choose to specify the analysis as a latent Y solution. There are various reasons for this, including the ability to accomplish useful specification tricks in LISREL (e.g., LISREL specifications for scale reliability evaluation; Chapter 8), greater simplicity, and the fact that many statistical papers use latent Y specifications to present information. The issue of "latent X" versus "latent Y" CFA specification is not relevant to many latent variable software programs (e.g., Mplus, EQS, Amos), which on the surface do not appear to rely on matrix operations. In LISREL, specifying a pure CFA measurement model as a latent X or latent Y solution has no impact on the fit and parameter estimates of the solution.

Figures 3.3 and 3.4 present the LISREL notation for the parameters and matrices of a CFA solution for latent X and latent Y specifications,
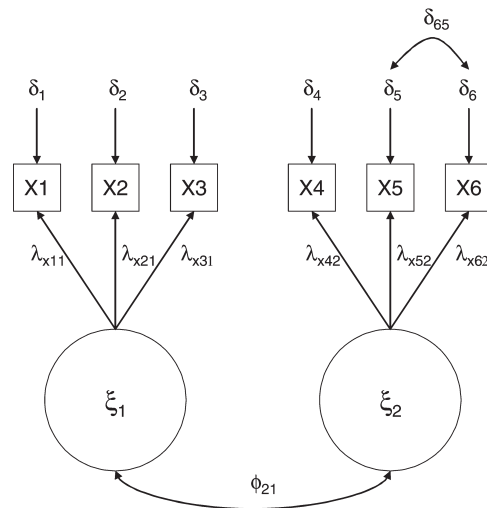
Name	Parameter	Matrix	Type	Description
Lambda-X	λ_x	Λ_x	Regression	Factor loadings
Theta delta	δ	Θ_δ	Variance–covariance	Error variances and covariances
Phi	ϕ	Φ	Variance–covariance	Factor variances and covariances
Tau-X	τ_x		Mean vector	Indicator intercepts
Kappa	κ		Mean vector	Latent means
Xi (Ksi)	ξ		Vector	Names of exogenous variables

FIGURE 3.3. Latent X notation for a two-factor CFA model with one error covariance. Factor variances, factor means, and indicator intercepts are not depicted in the path diagram.
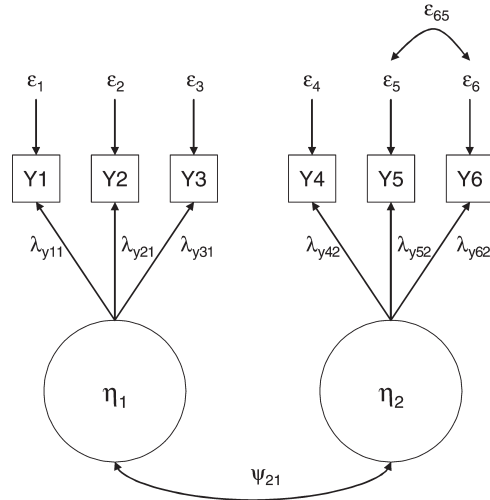
Name	Parameter	Matrix	Type	Description
Lambda-Y	λ_y	Λ_y	Regression	Factor loadings
Theta epsilon	ε	Θ_ε	Variance–covariance	Error variances and covariances
Psi	ψ	Ψ	Variance–covariance	Factor variances and covariances
Tau-Y	τ_y		Mean vector	Indicator intercepts
Alpha	α		Mean vector	Latent means
Eta	η		Vector	Names of endogenous variables

FIGURE 3.4. Latent Y notation for a two-factor CFA model with one error covariance. Factor variances, factor means, and indicator intercepts are not depicted in the path diagram.

respectively. As noted in the preceding paragraph, it is not necessary to understand this notation in order to specify CFA models in most software packages. However, knowledge of this notational system is useful because most sourcebooks and quantitative papers rely on it to describe the parameters and equations of CFA and SEM analyses. Consistent with material presented in context of EFA in Chapter 2, lowercase Greek symbols correspond to specific parameters (i.e., elements of a matrix such as λ), whereas capital Greek letters reflect an entire matrix (e.g., the full matrix of factor loadings, Λ). As in EFA, factor loadings are symbolized by lambdas (λ) with x and y subscripts in the case of exogenous and endogenous latent variables, respectively. The unidirectional arrows ($\rightarrow$) from the factors (e.g., ξ_1, η_1) to the indicators (e.g., X1, Y1) depict direct

effects (regressions) of the latent dimensions onto the observed measures; the specific regression coefficients are the lambdas (λ). Thetas (Θ) represent matrices of indicator error variances and covariances; theta-delta (Θ_δ) in the case of indicators of latent X variables, theta-epsilon (Θ_ε) for indicators of latent Y variables. For notational ease, the symbols δ and ε are often used in place of θ_δ and θ_ε, respectively, in reference to elements of Θ_δ and Θ_ε (as is done throughout this book). Although unidirectional arrows connect the thetas to the observed measures (e.g., X1–X6), these arrows do not depict regressive paths; that is, Θ_δ and Θ_ε are symmetric variance–covariance matrices consisting of error variances on the diagonal, and error covariances, if any, in the off-diagonal. Although less common, some notational systems do not use directional arrows in the depiction of error variances in order to avoid this potential source of confusion. For instance, one notational variation is to symbolize error variances with ovals because, like latent factors, measurement errors are not observed.

Factor variances and covariances are notated by phi (ϕ) and psi (ψ) in latent X and latent Y models, respectively. Curved, bidirectional arrows are used to symbolize covariances (correlations); in Figures 3.3 and 3.4, curved arrows indicate the covariance between the factors (ϕ_{21}, ψ_{21}) and the error covariance of the X5 and X6 indicators (δ_{65}, ε_{65}). When relationships are specified as covariances, the researcher is asserting that the variables are related (e.g., ξ_1 and ξ_2). However, this specification makes no claims about the nature of the relationship due to either the lack of knowledge regarding the directionality of the association (e.g., $\xi_1 \rightarrow \xi_2$) or the unavailability to the analysis of variables purported to account for this overlap (e.g., ξ_1 and ξ_2 are related because they share a common cause that is not represented by observed measures or latent variables in the analysis). Nonetheless, as discussed in Chapter 8, higher-order factor analysis is a useful approach for explaining the covariances among factors when a strong theory exists in regard to the patterning of the factor interrelationships.

The parameters in Figures 3.3 and 3.4 also possess numerical subscripts to indicate the specific elements of the relevant matrices. For example, λ_{x11} (Figure 3.3) indicates that the X1 measure loads on the first exogenous factor (ξ_1), and λ_{x21} indicates that X2 also loads on ξ_1. This numeric notation assumes that the indicators were ordered X1, X2, X3, X4, X5, and X6 in the input variance–covariance matrix. If the input matrix was arranged in this fashion, the lambda X matrix (Λ_x) in Figure 3.3 would be as follows:

$$
\begin{array}{ccc}
 & \xi_1 & \xi_2 \\
X1 & \lambda_{x11} & 0 \\
X2 & \lambda_{x21} & 0 \\
X3 & \lambda_{x31} & 0 \\
X4 & 0 & \lambda_{x42} \\
X5 & 0 & \lambda_{x52} \\
X6 & 0 & \lambda_{x62}
\end{array}
\tag{3.1}
$$

where the first numerical subscript refers to the row of Λ_x (i.e., the positional order of the X indicator) and the second numerical subscript refers to the column of Λ_x (i.e., the positional order of the exogenous factors, ξ); e.g., λ_{x52} conveys that the fifth indicator in the input matrix (X5) loads on the second latent X factor (ξ_2). Thus, Λ_x and Λ_y are full matrices whose dimensions are defined by p rows (number of indicators) and m columns (number of factors). The zero elements of Λ_x (e.g., λ_{x12}, λ_{x41}) indicate the absence of cross-loadings (e.g., the relationship between X1 and ξ_2 is fixed to zero). This is also depicted in Figures 3.3 and 3.4 by the absence of directional arrows between certain indicators and factors (e.g., no arrow connecting ξ_2 to X1 in Figure 3.3).

A similar system is used for variances and covariances among factors (ϕ in Figure 3.3; ψ in Figure 3.4) and indicator errors (δ and ε in Figures 3.3 and 3.4, respectively). However, because these aspects of the CFA solution reflect variances and covariances, they are represented by $m \times m$ symmetric matrices with variances on the diagonal and covariances in the off-diagonal. For example, the phi matrix (Φ) in Figure 3.3 would look as follows:

$$
\begin{array}{ccc}
 & \xi_1 & \xi_2 \\
\xi_1 & \phi_{11} & \\
\xi_2 & \phi_{21} & \phi_{22}
\end{array}
\tag{3.2}
$$

where ϕ_{11} and ϕ_{22} are the factor variances, and ϕ_{21} is the factor covariance. Similarly, the theta-delta matrix (Θ_δ) in Figure 3.3 is the following $p \times p$ symmetric matrix:

$$
\begin{array}{ccccccc}
 & X1 & X2 & X3 & X4 & X5 & X6 \\
X1 & \delta_{11} & & & & & \\
X2 & 0 & \delta_{22} & & & & \\
X3 & 0 & 0 & \delta_{33} & & & \\
X4 & 0 & 0 & 0 & \delta_{44} & & \\
X5 & 0 & 0 & 0 & 0 & \delta_{55} & \\
X6 & 0 & 0 & 0 & 0 & \delta_{65} & \delta_{66}
\end{array}
\tag{3.3}
$$

where δ_{11} through δ_{66} are the indicator errors and δ_{65} is the covariance of the measurement errors of indicators X5 and X6. For notational ease, the diagonal elements are indexed by single digits in Figures 3.3 and 3.4 (e.g., δ_6 is the same as δ_{66}). The zero elements of Θ_δ (e.g., δ_{21}) indicate the absence of error covariances (i.e., these relationships are fixed to zero).

In CFA with mean structures (see Chapter 7), indicator intercepts are symbolized by tau (τ) and latent exogenous and endogenous means are symbolized by kappa (κ) and alpha (α), respectively. Because the focus has been on CFA, only parameters germane to measurement models have been discussed thus far. LISREL notation also applies to structural components of models that entail directional relationships among exogenous and endogenous variables. For instance, gamma (γ, matrix: Γ) denotes regressions between latent X and latent Y variables, and beta (β, matrix: $\mathbf{B}$) symbolizes directional effects among endogenous variables. Most of the CFA illustrations provided in this book do not require gamma or beta parameters. Exceptions include CFA with covariates (e.g., MIMIC models, Chapter 7) where the measurement model is regressed on observed background measures (e.g., a dummy code indicating male vs. female), higher-order CFA (Chapter 8), and models with formative indicators (Chapter 8).

FUNDAMENTAL EQUATIONS OF A CFA MODEL

CFA aims to reproduce the sample variance–covariance matrix by the parameter estimates of the measurement solution (e.g., factor loadings, factor covariances, etc.). To illustrate, Figure 3.3 has been revised such that parameter estimates have been inserted for all factor loadings, factor correlation, and indicator errors (see now Figure 3.5). For ease of illustration, completely standardized values are presented, although the same concepts and formulas apply to unstandardized solutions. The first set of measures (X1, X2, X3) are indicators of one latent construct (ξ_1), whereas the second set of measures (X4, X5, X6) are indicators of another latent construct (ξ_2). It could be said, for example, that indicators X4, X5, and X6 are *congeneric* (cf. Jöreskog, 1971a) because they share a common factor (ξ_2).[3] An indicator would not be considered congeneric if it loaded on more than one factor.

In the case of congeneric factor loadings, the variance of an indicator is reproduced by multiplying its squared factor loading by the variance of the factor, and then summing this product with the indicator's error variance. The predicted covariance of two indicators that load on the same factor is computed as the product of their factor loadings times the variance

Predicted Variances–Covariances (Correlations):

```
      X1         X2         X3         X4         X5         X6

X1  1.00000
X2  0.72000  1.00000
X3  0.76500  0.68000  1.00000
X4  0.36000  0.32000  0.34000  1.00000
X5  0.33750  0.30000  0.31875  0.60000  1.00000
X6  0.31500  0.28000  0.29750  0.56000  0.72500  1.00000
```
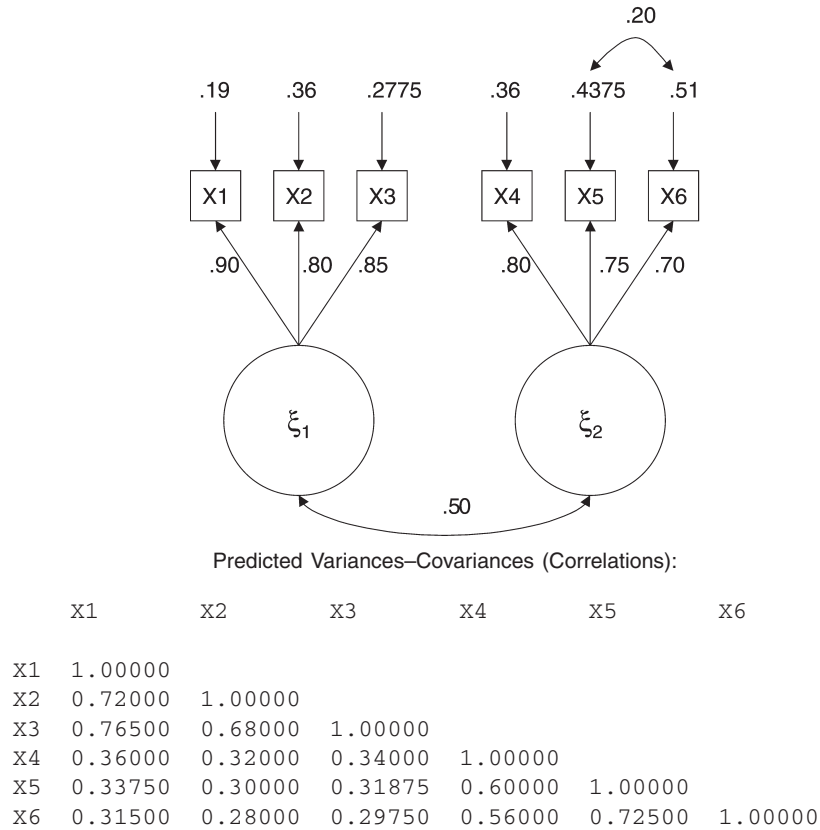
FIGURE 3.5. Reproduction of the input matrix from the parameter estimates of a two-factor measurement model (completely standardized solution).

of the factor. The model-implied covariance of two indicators that load on separate factors is estimated as the product of their factor loadings times the factor covariance. For example, based on the parameter estimates in the solution presented in Figure 3.5, the variance of X2 would be reproduced by the following equation (using latent X notation):

$$\text{VAR(X2)} = \sigma_{22} = \lambda_{x21}^2\phi_{11} + \delta_2 \tag{3.4}$$
$$= .80^2(1) + .36$$
$$= 1.00$$

In the case of completely standardized solutions (such as the current illustration), one can reproduce the variance of an indicator by simply squaring its factor loading ($.80^2$) and adding its error ($.36$), because the factor variance

will always equal 1.00 (however, the factor variance must be included in this calculation when dealing with unstandardized solutions). Note that the variance of ξ_2 also equals 1.00 because of the completely standardized model (e.g., variance of X6 = $\lambda_{x62}^2\phi_{22} + \delta_6$ = .70^2 + .51 = 1.00).

The squared factor loading represents the proportion of variance in the indicator that is explained by the latent factor (often referred to as a *communality*; see Chapter 2). For example, the communality of X2 is:

$$\eta_{22}^2 = \lambda_{x21}^2 \qquad\qquad (3.5)$$
$$= .80^2$$
$$= .64$$

indicating that ξ_1 accounts for 64% of the variance in X2. Similarly, in the completely standardized solution presented in Figure 3.5, the errors represent the proportion of variance in the indicators that is not explained by the latent factor; for example, δ_2 = .36, indicating that 36% of the variance in X2 is unique variance (e.g., measurement error). These errors (residual variances) can be readily calculated as 1 minus the squared factor loading. Using the X2 indicator, the computation would be

$$\delta_2 = 1 - \lambda_{x21}^2 \qquad\qquad (3.6)$$
$$= 1 - .80^2$$
$$= .36$$

The predicted covariance (correlation) between X2 and X3 would be estimated as follows:

$$COV(X2, X3) = \sigma_{3,2} = \lambda_{x21}\phi_{11}\lambda_{x31} \qquad\qquad (3.7)$$
$$= (.80)(1)(.85)$$
$$= .68$$

As before, in the case of completely standardized solutions the factor variance will always equal 1.00, so the predicted correlation between two congeneric indicators can be calculated by the product of their factor loadings; for example, model-implied correlation of X4, X5 = .80(.75) = .60.

The predicted covariance (correlation) between X3 and X4 (indicators that load on separate factors) would be estimated as follows:

$$COV(X3, X4) = \sigma_{4,3} = \lambda_{x31}\phi_{21}\lambda_{x42} \qquad\qquad (3.8)$$
$$= (.85)(.50)(.80)$$
$$= .34$$

Note that the factor correlation (ϕ_{21}) rather than the factor variance is used in this calculation.

Figure 3.5 presents the 6 variances and 15 covariances (completely standardized) that are estimated by the two-factor measurement model. This model also contains a correlation between the errors of the X5 and X6 indicators (δ_{65} = .20). In this instance, the covariation between the indicators is not accounted for fully by the latent factor (ξ_2); that is, X5 and X6 share additional variance owing to influences other than the latent construct (e.g., method effects). Thus, the equation to calculate the predicted correlation of X5 and X6 includes the correlated error:

$$\text{COV(X5, X6)} = \sigma_{6,5} = (\lambda_{x52}\phi_{22}\lambda_{x62}) + \delta_{65} \qquad (3.9)$$
$$= (.75)(1)(.70) + .20$$
$$= .725$$

CFA MODEL IDENTIFICATION

In order to estimate the parameters in CFA, the measurement model must be *identified.* A model is identified if, on the basis of known information (i.e., the variances and covariances in the sample input matrix), it is possible to obtain a unique set of parameter estimates for each parameter in the model whose values are unknown (e.g., factor loadings, factor correlations, etc.). Model identification pertains in part to the difference between the number of freely estimated model parameters and the number of pieces of information in the input variance–covariance matrix. Before addressing this issue, an aspect of identification specific to the analysis of latent variables is discussed—scaling the latent variable.

Scaling the Latent Variable

In order to conduct a CFA, every latent variable must have its scale identified. By nature, latent variables are unobserved and thus have no defined metrics (units of measurement). Thus, these units of measurement must be set by the researcher. In CFA, this is typically accomplished in one of two ways.

In the first and by far the most popular method, the researcher fixes the metric of the latent variable to be the same as one of its indicators. The indicator selected to pass its metric onto the latent factor is often referred to as a *marker* or *reference indicator.* The guidelines for selecting and speci-

fying marker indicators are discussed in Chapter 4. When a marker indicator is specified, a portion of its sample variance is passed on to the latent variable. Using Figure 3.5, suppose X1 was selected as the marker indicator for ξ_1 and had a sample variance (σ_{11}) of 16. Because X1 had a completely standardized factor loading on ξ_1 of .90, 81% of its variance was explained by ξ_1; $.90^2 = .81$ (cf. Eq. 3.5). Accordingly, 81% of the sample variance in X1 is passed on to ξ_1 to represent the factor variance of ξ_1:

$$\phi_{11} = \lambda_{x11}^2 \sigma_{11} \qquad\qquad (3.10)$$
$$= (.81)16$$
$$= 12.96$$

As will be shown in Chapter 4, these estimates are part of the unstandardized CFA solution.

   In the second method, the variance of the latent variable is fixed to a specific value, usually 1.00. Consequently, a standardized and a completely standardized solution are produced. Although the latent variables have been standardized (i.e., their variances are fixed to 1.0), the fit of this model is identical to the unstandardized model (i.e., models estimated using marker indicators). While useful in some circumstances (e.g., as a parallel to the traditional EFA model), this method is used less often than the marker indicator approach. The former strategy produces an unstandardized solution (in addition to a completely standardized solution), which is useful for several purposes, such as tests of measurement invariance across groups (Chapter 7) and evaluations of scale reliability (Chapter 8).

Statistical Identification

Besides scaling the latent variable, the parameters of a CFA model can be estimated only if the number of freely estimated parameters does not exceed the number of pieces of information in the input variance–covariance matrix. The concept was introduced in Chapter 2 in context of the identification of EFA models estimated by ML. A model is *underidentified* when the number of unknown (freely estimated) parameters exceeds the number of pieces of known information (i.e., elements of the input variance–covariance matrix). An underidentified model cannot be solved because there are an infinite number of parameter estimates that result in perfect model fit. To illustrate, consider the following basic equation:

$$x + y = 7 \qquad\qquad (3.11)$$

In this instance, there are two unknowns (x and y), and one known ($x + y$ = 7). This equation is underidentified because the number of unknown parameters (x and y) exceeds the known information; consequently, there are an infinite number of pairs of values that x and y could take on to solve for $x + y = 7$ ($x = 1$, $y = 6$; $x = 2$, $y = 5$, etc.). In the context of CFA, consider the model depicted in Figure 3.6A. In CFA, the knowns are usually the sample variances and covariances of the input indicators. When the CFA involves the analysis of mean structures, the sample means of the indicators are also included in the count of pieces of known information, and the indicator intercepts and latent means are included in the count of parameter estimates (see Chapter 7). For the model depicted in Figure 3.6A, the input matrix would be comprised of three knowns (pieces of information): the variances of X1 and X2, and the covariance of X1 and X2. The unknowns of the CFA solution are the freely estimated model parameters. In the Figure 3.6A model, there are 4 freely estimated parameters: 2 factor loadings (λ_{x11}, λ_{x21}), and 2 indicator errors (δ_1, δ_2). In this example, the metric of ξ_1 was set by fixing its variance to 1.0. Thus, because the factor variance (ϕ_{11}) was fixed, it is not included in the count of unknowns. Alternatively, we may have opted to define the metric of ξ_1 by choosing either X1 or X2 to serve as a marker indicator. In this case, the factor variance (ϕ_{11}) would contribute to the count of freely estimated parameters, but the factor loading of the marker indicator would not be included in this tally because it was fixed to pass its metric on to ξ_1 (see Chapter 4 for more details). Regardless of which method is used to define the units of measurement of ξ_1, the count of freely estimated parameters in Figure 3.6A equals 4.

Thus, the CFA model in Figure 3.6A is underidentified because the number of unknowns (4 freely estimated parameters) exceeds the number of knowns (3 elements of the input matrix = 2 variances, 1 covariance). This model would aim to reproduce the sample covariance of X1 and X2. Suppose, for example, this sample covariance corresponds to a correlation of .64. There are an infinite number of set of values that λ_{x11}, λ_{x21}, δ_1, and δ_2 could take on to reproduce an X1–X2 correlation of .64. Recall that in a completely standardized solution, the predicted correlation between two indicators that load on the same factor is the product of their factor loadings. Thus, there are endless pairs of values that could be estimated for λ_{x11} and λ_{x21} that would produce a perfectly fitting model (e.g., λ_{x11} = .80, λ_{x21} = .80; λ_{x11} = .90, λ_{x21} = .711; λ_{x11} = .75, λ_{x21} = .853, etc.).

Model A: Under-Identified ($df = -1$)
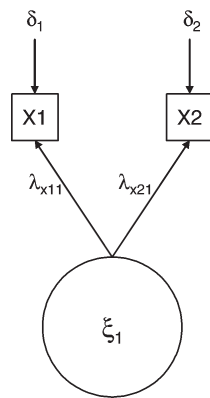
Input Matrix (3 elements)

	X1	X2
X1	σ_{11}	
X2	σ_{21}	σ_{22}

Freely Estimated Model Parameters = 4
(e.g., 2 factor loadings, 2 error variances)

Model B: Just-Identified ($df = 0$)
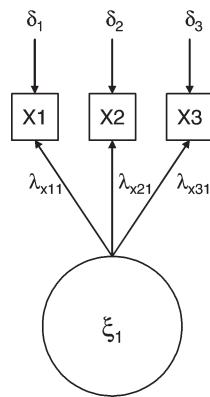
Input Matrix (6 elements)

	X1	X2	X3
X1	σ_{11}		
X2	σ_{21}	σ_{22}	
X3	σ_{31}	σ_{32}	σ_{33}

Freely Estimated Model Parameters = 6
(e.g., 3 factor loadings, 3 error variances)

Model C: Under-Identified ($df = -1$)
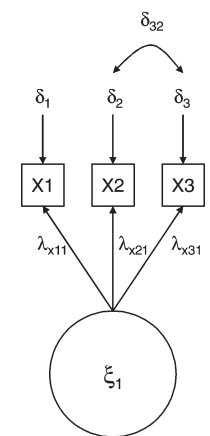
Input Matrix (6 elements)

	X1	X2	X3
X1	σ_{11}		
X2	σ_{21}	σ_{22}	
X3	σ_{31}	σ_{32}	σ_{33}

Freely Estimated Model Parameters = 7
(e.g., 3 factor loadings, 3 error variances, 1
error covariance)

FIGURE 3.6. Examples of underidentified and just-identified CFA models.

Incidentally, it would be possible to identify the Figure 3.6A model if additional constraints are imposed on the solution. For instance, the researcher could add the restriction of constraining the factor loadings to equality. In this case, the number of knowns (3) would equal the number of unknowns (3) and the model would be *just-identified*. As discussed in further detail in the next paragraph, in just-identified models there exists one unique set of parameter estimates that perfectly fit the data. In the current illustration, the only factor loading parameter estimate that would reproduce the observed X1–X2 correlation (.64) is .80; λ_{x11} = .80 and λ_{x21} = .80, solved by imposing the equality constraint. Although these constraints might assist in model identification by reducing the number of freely estimated parameters (such as in the current example), such restrictions are often unreasonable on the basis of evidence or theory.

Unlike underidentified models, just-identified models can be "solved." In fact, because the number of knowns equals the number of unknowns, in just-identified models, there exists a single set of parameter estimates that *perfectly* reproduce the input matrix. Before further applying this concept to CFA, consider this example from simultaneous equations algebra:

$$x + y = 7 \qquad\qquad (3.12)$$
$$3x - y = 1$$

In this example, the number of unknowns (x, y) equals the number of knowns $(x + y = 7, 3x - y = 1)$. Through basic algebraic manipulation, it can be readily determined that $x = 2$ and $y = 5$; that is, there is only one possible pair of values for x and y.

Now consider the CFA model in Figure 3.6B. In this example, the input matrix consists of 6 knowns (3 variances, 3 covariances), and the model consists of 6 freely estimated parameters: three factor loadings and three indicator errors (again assume that the variance of ξ_1 was fixed to 1.0). This CFA model is just-identified and would produce a unique set of parameter estimates $(\lambda_{x11}, \lambda_{x21}, \lambda_{x31}, \delta_1, \delta_2, \delta_3)$ that perfectly reproduce the correlations among X1, X2, and X3 (see Appendix 3.2). Thus, although just-identified CFA models can be fit to the sample input matrix, goodness-of-model-fit evaluation does not apply because, by nature, such solutions always have perfect fit. This is also why goodness of fit does not apply to traditional statistical analyses such as multiple regression; that is, these models are inherently just-identified. For instance, in OLS multiple regression, all observed variables are connected to one another either by direct effects, $X_1, X_2 \rightarrow Y$, or freely estimated correlations, $X_1 \leftrightarrow X_2$.

It is important to note that while a CFA model of a construct consisting of 3 observed measures may meet the conditions of identification (as in Figure 3.6B), this is true only if the errors of the indicators are not correlated with each other. For instance, the model depicted in Figure 3.6C is identical to Figure 3.6B, with the exception of a correlated residual between indicators X2 and X3. This additional parameter (δ_{32}) now brings the count of freely estimated parameters to 7, which exceeds the number of elements of the input variance–covariance matrix (6). Thus, the Figure 3.6C model is underidentified and cannot be fit to the sample data.

A model is *overidentified* when the number of knowns (i.e., number of variances and covariances in the input matrix) exceeds the number of freely estimated model parameters. For example, the one-factor model depicted in Figure 3.7 (Model A) is structurally overidentified because there are 10 elements of the input matrix (4 variances for X1–X4, 6 covariances), but only 8 freely estimated parameters (4 factor loadings, 4 error variances; the variance of ξ_1 is fixed to 1.0). The difference in the number of knowns (*b*) and the number of unknowns (*a*; i.e., freely estimated parameters) constitutes the model's *degrees of freedom* (*df*). Overidentified solutions have positive *df*, just-identified models have 0 *df* (because the number of knowns equals the number of unknowns), and underidentified models have negative *df* (cannot be solved or fit to the data). Thus, the Figure 3.7A model is overidentified with *df* = 2:

$$df = b - a \hspace{3cm} (3.13)$$
$$= 10 - 8$$
$$= 2$$

The second model in Figure 3.7 (Model B) is also overidentified with *df* = 1 (assuming that ξ_1 and ξ_2 have a non-zero correlation; see discussion of "empirical underidentification" below). As in the Figure 3.7A solution, there are 10 elements of the input matrix. However, the Figure 3.7B model consists of 9 freely estimated parameters (4 factor loadings, 4 error variances, 1 factor covariance), thus resulting in 1 *df*.

As a final example of an overidentified solution, consider the measurement model that was presented in Figure 3.5. In this case, there are 21 pieces of information in the input matrix (6 variances, 15 covariances). The reader may note that it becomes cumbersome to count the elements of the input matrix as the number of variables increases. Fortunately, the following formula, initially presented in Chapter 2 (Eq. 2.8), readily provides this count:

Model A: Over-Identified ($df = 2$)

Input Matrix (10 elements)

	X1	X2	X3	X4
X1	σ_{11}			
X2	σ_{21}	σ_{22}		
X3	σ_{31}	σ_{32}	σ_{33}	
X4	σ_{41}	σ_{42}	σ_{43}	σ_{44}

Freely Estimated Model Parameters = 8
(e.g., 4 factor loadings, 4 error variances)

Model B: Over-Identified ($df = 1$)

Input Matrix (10 elements)

	X1	X2	X3	X4
X1	σ_{11}			
X2	σ_{21}	σ_{22}		
X3	σ_{31}	σ_{32}	σ_{33}	
X4	σ_{41}	σ_{42}	σ_{43}	σ_{44}

Freely Estimated Model Parameters
= 9 (e.g., 4 factor loadings, 4 error
variances, 1 factor covariance)

(cont.)

**FIGURE 3.7.** Examples of overidentified and empirically underidentified CFA models. In all examples, the metric of the latent factor is defined by fixing its variance to 1.0.

$$b = p(p + 1) / 2 \qquad (3.14)$$

where b is the number of elements of the input matrix, and p is the number of indicators included in the input matrix. Thus, in the Figure 3.5 model involving 6 indicators (X1–X6):

$$b = 6(6 + 1) / 2 = 21$$

Model C: Empirically Under-Identified ($df = 0$)
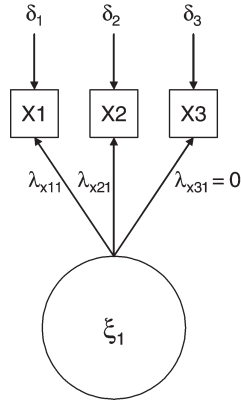
Input Matrix (6 elements)

	X1	X2	X3
X1	σ_{11}		
X2	σ_{21}	σ_{22}	
X3	$\sigma_{31}{}^*$	$\sigma_{32}{}^*$	σ_{33}

$^*\sigma_{31} = 0,\ ^*\sigma_{32} = 0$

Freely Estimated Model Parameters = 6
(e.g., 3 factor loadings, 3 error
variances)

Model D: Empirically Under-Identified ($df = 1$)
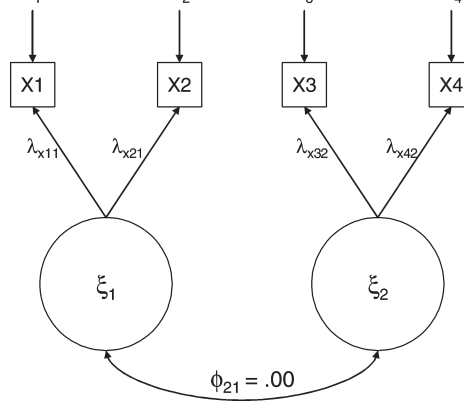
Input Matrix (10 elements)

	X1	X2	X3	X4
X1	σ_{11}			
X2	σ_{21}	σ_{22}		
X3	$\sigma_{31}{}^*$	$\sigma_{32}{}^*$	σ_{33}	
X4	$\sigma_{41}{}^*$	$\sigma_{42}{}^*$	σ_{43}	σ_{44}

$^*\sigma_{31} = 0,\ ^*\sigma_{32} = 0,\ ^*\sigma_{41} = 0,$
$^*\sigma_{42} = 0$

Freely Estimated Model Parameters
= 7 (e.g., 3 factor loadings, 3 error
variances, 1 error covariance)

FIGURE 3.7. (*cont.*)

reflecting the 6 variances (p) and the 15 covariances $[p(p - 1) / 2]$.

The specification of the Figure 3.5 model entails 14 freely estimated parameters (in this example, assume that the metric of ξ_1 and ξ_2 was set by selecting X1 and X4 as marker indicators): 4 factor loadings, 6 error variances, 1 error covariance, 2 factor variances, 1 factor covariance. The loadings of X1 and X4 are not included in this count because they were fixed in order to pass their metrics onto ξ_1 and ξ_2. Therefore, this model is overidentified with $df = 7$ (21 knowns minus 14 unknowns).

Degrees of freedom are used in many descriptive indices of *goodness of model fit* (such as model χ^2; see the "Descriptive Goodness of Fit Indices" section of this chapter). Indeed, an important aspect of overidentified solutions is that goodness-of-fit evaluation applies—specifically, how well the model is able to reproduce the input variances and covariances (i.e., the input matrix) using a lesser number of unknowns (i.e., freely estimated model parameters). Thus, as in just-identified models, the available known information indicates that there is one best value for each freely estimated parameter in the overidentified solution. Unlike just-identified models, overidentified models rarely fit the data perfectly (i.e., a perfectly fitting model is one whose parameter estimates recreate the input variance–covariance matrix exactly). The principles of goodness of fit are discussed in fuller detail in the last section of this chapter.

Specification of a model to have at least zero *df*s is a necessary but not sufficient condition for identification. Figure 3.7 also illustrates examples of *empirically underidentified* solutions (Kenny, 1979). In an empirically underidentified solution, the model is statistically just- or overidentified, but aspects of the input matrix prevent the analysis from obtaining a unique and valid set of parameter estimates. The most obvious example of empirical underidentification would be the case where all covariances in the input matrix equal 0. However, empirical underidentification can result from other patterns of (non)relationships in the input data. For example, Figure 3.7C depicts a model that is identical to the just-identified model presented in Figure 3.6B, yet its input matrix reveals an absence of a relationship between X3 and X1 and X2. This aspect of the input matrix would render the Figure 3.7C model functionally equivalent to the underidentified solution presented in Figure 3.6A. Similarly, Figure 3.7D represents a model that is identical to the overidentified model in Figure 3.7B. However, because the factor correlation is 0 (due to a lack of relationship of X1 and X2 with X3 and X4), this solution would be analogous to simultaneously attempting to estimate two Figure 3.6A solutions (underidentified). For these and other reasons (e.g., increased power and precision of parameter estimates; cf. Marsh, Hau, Balla, & Grayson, 1998), methodologists recommend that latent factors be defined by a minimum of three indicators to avoid this possible source of underidentification; for example, even if $\phi_{21} = 0$, a solution for the Figure 3.7D model could be obtained if ξ_1 and ξ_2 were measured by three indicators each.

Empirical underidentification will also occur if the marker indicator selected to define the metric of the latent variable is not correlated with the

other indicators. For instance, suppose that in the Figure 3.7A model, X1 was used as the marker indicator for ξ_1 but was not correlated with X2, X3, and X4 (i.e., σ_{21}, σ_{31}, and σ_{41} were not significantly different from 0). Consequently, λ_{x11} would equal 0 and thus no variance would be passed onto ξ_1 ($\phi_{11} = \lambda_{x11}^2\sigma_{11} = 0\sigma_{11} = 0$; cf. Eq. 3.10), producing an empirically underidentified solution.

   If an attempt is made to fit an empirically underidentified model, the computer software will fail to yield a solution or will provide an improper solution (usually accompanied by error messages or warnings). For example, one possible consequence is so-called *Heywood cases*, which refer to parameter estimates that have out-of-range values (e.g., negative indicator error variances). Improper solutions may arise from a variety of causes, such as poorly specified models (e.g., in Figure 3.7C, X3 is not a reasonable indicator of ξ_1) and problems with the input matrix (e.g., pairwise deletion, collinearities, small sample size; cf. Wothke, 1993). These problems are discussed in detail in Chapter 5.

Guidelines for Model Identification

On the basis of the preceding discussion, some basic guidelines for model identification can be summarized:

   1. Regardless of the complexity of the model (e.g., one-factor vs. multiple factors, size of indicator set), latent variables must be scaled by either specifying marker indicators or fixing the variance of the factor (usually to a value of 1.0);
   2. Regardless of the complexity of the model, the number of pieces of information in the input matrix (e.g., indicator variances and covariances) must equal or exceed the number of freely estimated model parameters (e.g., factor loadings, factor variances/covariances, indicator error variances/covariances);
   3. In the case of one-factor models, a minimum of three indicators is required. When three indicators are used (and no correlated errors are specified; e.g., Figure 3.6B), the one-factor solution is just-identified and goodness-of-fit evaluation does not apply, although this model can still be evaluated in terms of the interpretability and strength of its parameter estimates (e.g., magnitude of factor loadings). When four or more indicators are used (and no correlated errors are specified; e.g., Figure 3.7A), the model is overidentified (i.e., there are more elements of the input matrix

than freely estimated model parameters) and goodness of fit can be used in the evaluation of the acceptability of the solution;

4. In the case of models that entail two or more factors and two indicators per latent construct, the solution will be overidentified, provided that every latent variable is correlated with at least one other latent variable and the errors between indicators are uncorrelated (e.g., Figure 3.7B). However, because such solutions are susceptible to empirical underidentification (e.g., Figure 3.7D), a minimum of three indicators per latent variable is recommended.

These guidelines assume models where each indicator loads on only one latent factor and are free of correlated measurement error (i.e., Θ is a diagonal matrix and thus all off-diagonal elements equal 0). As discussed in Chapter 5, the issue of identification is more complicated in models that involve double-loading indicators (i.e., *noncongeneric* indicators) or correlated indicator errors (e.g., Figure 3.6C).

ESTIMATION OF CFA MODEL PARAMETERS

The objective of CFA is to obtain estimates for each parameter of the measurement model (i.e., factor loadings, factor variances and covariances, indicator error variances and possibly error covariances) that produce a predicted variance–covariance matrix (symbolized as Σ) that resembles the sample variance–covariance matrix (symbolized as S) as closely as possible. For instance, in overidentified models (such as Figure 3.7A), perfect fit is rarely achieved (i.e., $\Sigma \neq S$). Thus, in the case of a CFA model such as Figure 3.7A, the goal of the analysis is to find a set of factor loadings (λ_{x11}, λ_{x21}, λ_{x31}, λ_{x41}) that yield a predicted covariance matrix (Σ) that best reproduces the input matrix (S); for example, find parameter estimates for λ_{x11} and λ_{x21} such that the predicted correlation between X1 and X2 ($\lambda_{x11}\phi_{11}\lambda_{x21}$) closely approximates the sample correlation of these indicators (σ_{21}) (although in the actual estimation process, this occurs simultaneously for all parameters and implied covariances). This process entails a *fitting function*, a mathematical operation to minimize the difference between Σ and S. By far, the fitting function most widely used in applied CFA research (and SEM, in general) is *maximum likelihood* (ML). The fitting function that is minimized in ML is:

$$F_{ML} = \ln|S| - \ln|\Sigma| + \text{trace}[(S)(\Sigma^{-1})] - p \qquad (3.15)$$

where $|S|$ is the determinant of the input variance–covariance matrix, $|\Sigma|$ is the determinant of the predicted variance–covariance matrix, p is the order of the input matrix (i.e., the number of input indicators), and ln is the natural logarithm. Although a full explication of this function is beyond the scope of this chapter (cf. Bollen, 1989; Eliason, 1993), a few observations are made in an effort to foster its conceptual understanding (see also Appendix 3.3). The determinant and trace summarize important information about matrices such as S and Σ. The determinant is a single number (i.e., a scalar) that reflects a generalized measure of variance for the entire set of variables contained in the matrix. The trace of a matrix is the sum of values on the diagonal (e.g., in a variance–covariance matrix, the trace is the sum of variances). The objective of ML is to minimize the differences between these matrix summaries (i.e., the determinant and trace) for S and Σ. The operations of this formula are most clearly illustrated in the context of a perfectly fitting model, as quoted from Jaccard and Wan (1996):

> In this case, the determinant of S will equal the determinant of Σ, and the difference between the logs of these determinants will equal 0. Similarly, $(S)(\Sigma^{-1})$ will equal an identity matrix with all ones in the diagonal. When the diagonal elements are summed (via the trace function), the result will be the value of p. Subtracting p from this value yields 0. Thus, when there is perfect model fit, F_{ML} equals 0. (pp. 85–86)

The calculation and use of F_{ML} is discussed in greater detail later in this chapter and in Appendix 3.3.

The underlying principle of ML estimation in CFA is to find the model parameter estimates that maximize the probability of observing the available data if the data were collected from the same population again. In other words, ML aims to find the parameter values that make the observed data most likely (or conversely, maximize the likelihood of the parameters given the data). Finding the parameter estimates for an overidentified CFA model is an *iterative* procedure. That is, the computer program (such as LISREL, Mplus, EQS, or Amos) begins with an initial set of parameter estimates (referred to as *starting values* or *initial estimates*, which can be automatically generated by the software or specified by the user) and repeatedly refines these estimates in an effort to reduce the value of F_{ML} (i.e., minimize the difference between Σ and S). Each refinement of the parameter estimates to minimize F_{ML} is an *iteration*. The program conducts internal checks to evaluate its progress in obtaining parameter estimates that best reproduce S (i.e., that result in the lowest F_{ML} value). *Convergence* of

the model is reached when the program arrives at a set of parameter estimates that cannot be improved upon to further reduce the difference between Σ and S. In fact, for a given model and S, the researcher may encounter minor differences across programs in F_{ML}, goodness-of-fit indices, and so forth, as the result of variations across software packages in minimization procedures and stopping criteria.

Occasionally, a latent variable solution will fail to converge. Convergence is often related to the quality and complexity of the specified model (e.g., the number of restrictions imposed on the solution) and the adequacy of the starting values. In the case of complex (yet properly specified) models, convergence may not be reached because the program has stopped at the maximum number of iterations, which is set by either the program's default or a number specified by the user. This problem may be rectified by simply increasing the maximum number of iterations or possibly using the preliminary parameter estimates as starting values. However, a program may also cease before the maximum number of iterations has been reached because its internal checks indicate that progress is not being made in obtaining a solution that minimizes F_{ML}. Although this can be an indication of more serious problems such as a grossly misspecified model, this outcome may stem from more innocuous issues such as the scaling of the indicators and the adequacy of the starting values.[4]

Starting values affect the minimization process in a number of ways. If these initial values are similar to the final model parameter estimates, fewer iterations are required to reach convergence. Yet if the starting values are quite different from the final parameter estimates, there is a greater likelihood of nonconvergence or that the solution will contain Heywood cases (e.g., communalities > 1.0, negative error variances). Fortunately, the CFA researcher usually does not need to be concerned about starting values because most latent variable software programs have incorporated sophisticated methods for automatically generating these initial estimates (e.g., Version 8 of LISREL uses the instrumental variables and two-stage least squares methods to compute starting values). However, in more complex models (such as those involving nonlinear constraints; cf. CFA approach to scale reliability evaluation in Chapter 8), it is sometimes necessary for the user to provide these values. The strategy for selecting starting values varies somewhat, depending on the type of model (e.g., multiple-groups solution, a solution with nonlinear constraints), and thus only a few broad recommendations are provided here.

It is useful to ensure that starting values are consistent with the metric of the observed data; for example, starting values for error variances

should not be greater than the variance of the indicator; initial estimates of factor variances should not be greater than the observed variance of the marker indicator. Often, the most important parameters to provide starting values are the error variances. Initial values can also be generated by noniterative statistical procedures conducted on the data set (e.g., OLS estimates) or basic algebraic calculations (cf. Appendix 3.2). If such estimates are not available, these initial estimates may be gleaned from prior research involving similar models or indicator sets. In addition, some SEM sourcebooks contain guidelines for calculating starting values when the researcher can make a reasonable guess about whether the parameter estimates in question constitute a weak, moderate, or strong effect (e.g., Bollen, 1989). In the case of complex models, it can be fruitful to begin by fitting only portions of the model initially, and then use the resulting parameter estimates as starting values in the larger solution.

One reason why ML is widely used in CFA model estimation is that it possesses desirable statistical properties, such as the ability to provide standard errors (*SE*s) for each of the model's parameter estimates. These *SE*s are used for conducting statistical significance tests of the parameter estimates (i.e., z = unstandardized parameter estimate divided by its *SE*) and for determining the precision of these estimates (e.g., 95% confidence interval = parameter $\pm[SE * 1.96]$). Moreover, F_{ML} is used in the calculation of many goodness-of-fit indices (see below).

However, it is important to note that ML is only one of many methods that can be used to estimate CFA and other types of structural equation models. Indeed, ML has several requirements that render it an unsuitable estimator in some circumstances. As compared with some estimators, ML is more prone to Heywood cases. In addition, ML is more likely to produce markedly distorted solutions if minor misspecifications have been made to the model. Some key assumptions of ML are that (1) the sample size is large (asymptotic); (2) the indicators have been measured on continuous scales (i.e., approximate interval-level data); and (3) the distribution of the indicators is multivariate normal. The latter two assumptions apply to indicators of latent factors, not to other observed measures that may exist in the analysis, such as nominal variables that serve as covariates (e.g., see MIMIC models, Chapter 7). Although the actual parameter estimates (e.g., factor loadings) may not be affected, non-normality in ML analysis can result in biased standard errors (and hence faulty significance tests) and a poorly behaved χ^2 test of overall model fit. If non-normality is extreme (e.g., marked floor effects, as would occur if most of the sample responded to items using the lowest response choice—e.g., 0 on a 0–12 scale), then

ML will produce incorrect parameter estimates; that is, the assumption of a linear model is invalid. Thus, in the case of non-normal, continuous indicators, it is better to use a different estimator, such as ML with robust standard errors and χ^2 (MLM; Bentler, 1995). MLM provides the same parameter estimates as ML, but both the model χ^2 and standard errors of the parameter estimates are corrected for non-normality in large samples [e.g., MLM produces the Satorra–Bentler scaled (mean adjusted) χ^2 in which the typical normal theory χ^2 is divided by a scaling correction to better approximate χ^2 under non-normality]. If one or more of the factor indicators is categorical (or non-normality is extreme), normal theory ML should not be used. In this instance, estimators such as weighted least squares (WLS; also known as asymptotic distribution free, ADF), robust weighted least squares (e.g., WLSMV), and unweighted least squares (ULS) are more appropriate. WLS can also be used for non-normal, continuous data, although MLM is often preferred, given its ability to outperform WLS in small and medium-sized samples (Curran, West, & Finch, 1996; Hu, Bentler, & Kano, 1992). Other limitations of WLS are discussed in Chapter 9, in the context of a detailed presentation of the most widely used estimators of non-normal and categorical data.

Because of its widespread popularity, ML is used in most of the examples in this book. However, the reader should be aware that the vast majority of the principles and procedures discussed in the remaining chapters will apply, regardless of the type of fitting function that is employed. Exceptions are noted when important procedural differences exist across estimators (e.g., χ^2 difference testing using the Satorra–Bentler scaled χ^2, Chapter 9).

Illustration

To illustrate the concepts of parameter estimation and F_{ML} minimization, consider this simple example. As shown in Figure 3.8, a basic path model is tested using single indicators of behavioral inhibition (x), school refusal (y), and social anxiety (z), and an $N = 200$ school-age children. Of particular interest in this example is whether the relationship between behavioral inhibition (x) and school refusal (y) is fully mediated by social anxiety (z). Although the model is somewhat unrealistic (e.g., it assumes no measurement error in x, y, and z and does not conform to the typical strategy for evaluating mediated effects; cf. Baron & Kenny, 1986), its simplified nature will foster the illustration of the concepts and calculations introduced in the preceding and subsequent sections. Note that two of the
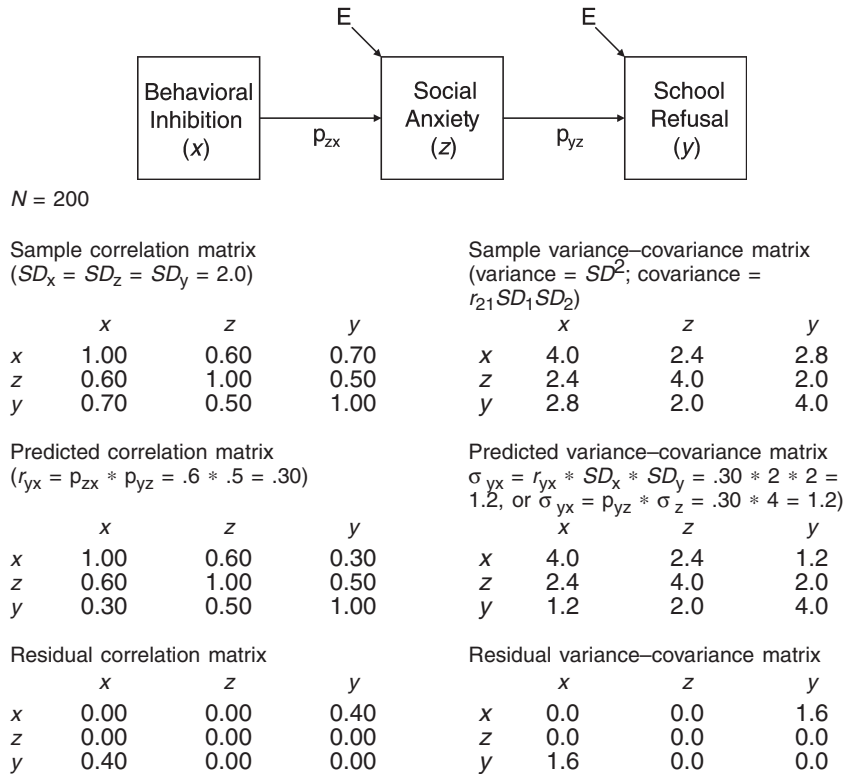
effects in this model are *tautological*, meaning that the paths between *x* and *z* and *z* and *y* must equal their observed relationships; in other words, given the way that the model is specified (e.g., *x* and *z*, and *z* and *y* are linked by direct effects), full reproduction of their observed covariances (correlations) is guaranteed (for algebraic proof of this fact, see Jaccard & Wan, 1996). However, the model also possesses one *nontautological* (i.e., overidentified) relationship involving *x* and *y*. In other words, although the model will generate a unique set of parameter estimates (i.e., paths between *x* and *z* and *z* and *y*, variance of *x*, residual variances of *z* and *y*), complete reproduction of the observed relationship between *x* and *y* is not ensured. Specifically, using a simple tracing rule (Bollen, 1989; Loehlin, 2004; Wright, 1934), the predicted correlation (and covariance) between *x* and *y* will be the product of the paths between *x* and *z* and *z* and *y*; that is:

$$r_{xy} = (\mathrm{p}_{xz})(\mathrm{p}_{zy}) \qquad (3.16)$$

(although ML uses a different approach to estimate the model and Σ, the results are the same). The model-implied relationship between *x* and *y* will not necessarily equal the observed relationship between these variables. Thus, the proximity of Σ to S depends entirely on the ability of the path model to reproduce the observed zero-order relationship between *x* and *y*. The model is thus overidentified with 1 *df* corresponding to the nontautological relationship between *x* and *y*. As discussed earlier, another way to determine whether the model has a single *df* is to take the difference between the number of elements of the input matrix ($b = 6 = 3$ variances, 3 covariances) and the number of freely estimated parameters ($a = 5$ = 2 regressive paths, the variance of *x*, and the two residual variances of *y* and *z*). Because the model is overidentified, goodness-of-fit evaluation will apply. A good-fitting model will be obtained if the model's parameter estimates reasonably approximate the observed relationship between *x* and *y*. If this model-implied relationship differs considerably from the observed *x*/*y* relationship, a poor-fitting model will result.

As shown in Figure 3.8, the model-implied correlation between *x* and *y* is .30; $r_{yx} = (\mathrm{p}_{zx})(\mathrm{p}_{yx}) = .6(.5) = .30$ (the covariance of *x* and *y* is predicted to be 1.2). This predicted relationship differs from the observed correlation between *x* and *y* (.70), thus suggesting a poor-fitting model. Table 3.2 presents the use of SAS PROC IML to calculate the residual matrix (sample matrix minus the predicted matrix) and F_{ML} (see Appendix 3.3 for hand calculation and a conceptual overview of F_{ML}). Because the relationship between *x* and *y* is the only nontautological effect in this model, this is the

only element of the residual matrix that can take on a value other than zero. As seen in Figure 3.8, the residual correlation and covariance for x and y are .40 and 1.6, respectively. Table 3.2 illustrates the calculation of F_{ML} on the basis of variance–covariance matrices, although the same F_{ML} value would be obtained if correlation matrices were used (yet, as noted earlier, variance–covariance matrices are often preferred in order to obtain unstandardized solutions, valid standard errors, and to permit other options such as multiple-groups evaluation). As shown in Table 3.2, the fitted model resulted in an F_{ML} value of 0.4054651, reflecting the discrepancy between S and Σ. If perfect fit had been attained, all elements of the residual matrix and F_{ML} would be 0. Although this value in and of itself is not readily interpretable (unless $F_{ML} = 0$), it is used in the calculation of a variety of goodness-of-fit indices.

$N = 200$

Sample correlation matrix
($SD_x = SD_z = SD_y = 2.0$)

	x	z	y
x	1.00	0.60	0.70
z	0.60	1.00	0.50
y	0.70	0.50	1.00

Sample variance–covariance matrix
(variance = SD^2; covariance = $r_{21}SD_1SD_2$)

	x	z	y
x	4.0	2.4	2.8
z	2.4	4.0	2.0
y	2.8	2.0	4.0

Predicted correlation matrix
($r_{yx} = p_{zx} * p_{yz} = .6 * .5 = .30$)

	x	z	y
x	1.00	0.60	0.30
z	0.60	1.00	0.50
y	0.30	0.50	1.00

Predicted variance–covariance matrix
$\sigma_{yx} = r_{yx} * SD_x * SD_y = .30 * 2 * 2 = 1.2$, or $\sigma_{yx} = p_{yz} * \sigma_z = .30 * 4 = 1.2$)

	x	z	y
x	4.0	2.4	1.2
z	2.4	4.0	2.0
y	1.2	2.0	4.0

Residual correlation matrix

	x	z	y
x	0.00	0.00	0.40
z	0.00	0.00	0.00
y	0.40	0.00	0.00

Residual variance–covariance matrix

	x	z	y
x	0.0	0.0	1.6
z	0.0	0.0	0.0
y	1.6	0.0	0.0

FIGURE 3.8. Illustration of model estimation of a simple path model.

TABLE 3.2. SAS PROC IML Syntax and Output for Computing a Residual Matrix and F_{ML}

SAS PROC IML Syntax

```
PROC IML;
OPTIONS NOCENTER;
*SAMPLE COVARIANCE MATRIX;
S =     {4.0 2.4 2.8,
         2.4 4.0 2.0,
         2.8 2.0 4.0};
*PREDICTED COVARIANCE MATRIX;
SIGM = {4.0 2.4 1.2,
         2.4 4.0 2.0,
         1.2 2.0 4.0};
RES = S - SIGM;         *RESIDUAL COVARIANCE MATRIX;
SDET = DET(S);          *DETERMINANT OF THE SAMPLE COVARIANCE MATRIX;
SIGMDET = DET(SIGM);    *DETERMINANT OF THE PREDICTED COVARIANCE MATRIX;
LOGS = LOG(SDET);       *NATURAL LOG OF SAMPLE MATRIX DETERMINANT;
LOGSIGM = LOG(SIGMDET); *NATURAL LOG OF PREDICTED MATRIX DETERMINANT;
SIGMINV = INV(SIGM);    *INVERSE OF PREDICTED MATRIX;
SDIV = S*SIGMINV;       *MULTIPLICATION OF SAMPLE MATRIX AND PREDICTED
INVERSE;
STRACE = TRACE(SDIV);   *TRACE OF THE RESULTING SDIV MATRIX;
SORDER = NROW(S);       *ORDER OF SAMPLE MATRIX = NUMBER OF INDICATORS;
*CALCULATION OF FML;
FML = ABS((LOGS - LOGSIGM) + STRACE - SORDER);
PRINT S;
PRINT SIGM;
PRINT RES;
PRINT SDET;
PRINT SIGMDET;
PRINT LOGS LOGSIGM;
PRINT SDIV;
PRINT STRACE;
PRINT SORDER;
PRINT FML;
```

Annotated SAS PROC IML Output

```
The SAS System

          S    (the sample variance-covariance matrix)

      4        2.4        2.8
    2.4          4          2
    2.8          2          4
```

(cont.)

TABLE 3.2. *(cont.)*

```
         SIGM (the predicted variance-covariance matrix)

      4        2.4       1.2
    2.4         4         2
    1.2         2         4
   RES       (the residual variance-covariance matrix)

      0         0        1.6
      0         0         0
    1.6         0         0


   SDET      (the determinant of the sample variance-covariance matrix)

   20.48


 SIGMDET    (the determinant of the predicted variance-covariance
             matrix)

   30.72


    LOGS   LOGSIGM (the natural logs of the sample and predicted
                   matrices)

3.0194488 3.4249139


    SDIV   (matrix produced by multiplying the sample matrix and the
             inverse of the predicted matrix)

      1 -0.266667 0.5333333
-3.47E-17         1          0
  0.625     -0.375          1


  STRACE    (trace of SDIV; sum of diagonal elements = 3)

       3


  SORDER    (order of the sample matrix = number of input indicators)

       3


   FML  (F value reflecting minimization of the maximum likelihood
        criterion)

0.4054651
```

DESCRIPTIVE GOODNESS-OF-FIT INDICES

The classic goodness-of-fit index is χ^2. Under typical ML model estimation, χ^2 is calculated as:

$$\chi^2 = F_{ML}(N - 1) \qquad\qquad (3.17)$$

(except in Mplus, where χ^2 is calculated by multiplying F_{ML} by N instead of $N - 1$). For instance, using $N - 1$, the Figure 3.8 model χ^2 is 80.69 (0.4054651 ∗ 199). Because this model is associated with 1 df, the critical χ^2 value (α = .05) is 3.84 (i.e., $\chi^2 = z^2 = 1.96^2 = 3.8416$). The model χ^2 of 80.69 exceeds the critical value of 3.84, and thus the null hypothesis that $S = \Sigma$ is rejected. Thus, a statistically significant χ^2 (latent variable software programs provide the exact probability value of the model χ^2) supports the alternate hypothesis that $S \neq \Sigma$, meaning that the model estimates do not sufficiently reproduce the sample variances and covariances (i.e., the model does not fit the data well).

Although χ^2 is steeped in the traditions of ML and SEM (e.g., it was the first fit index to be developed), it is rarely used in applied research as a sole index of model fit. Indeed, important criticisms of χ^2 include the following: (1) in many instances (e.g., small N, non-normal data) its underlying distribution is not χ^2 distributed (compromising the statistical significance tests of the model χ^2); (2) it is inflated by sample size (e.g., if N were to equal 100 in the Figure 3.8 model, $\chi^2 = 40.14$), and thus large N solutions are routinely rejected on the basis of χ^2 even when differences between S and Σ are negligible; and (3) it is based on the very stringent hypothesis that $S = \Sigma$. As discussed below, many alternative fit indices are based on less stringent standards such as "reasonable" fit and fit relative to an independence model.[5] Nevertheless, χ^2 is used for other purposes, such as nested model comparisons (see Chapters 4, 5, and 7) and the calculation of other fit indices (e.g., the Tucker–Lewis index; see below). While χ^2 is routinely reported in CFA research, other fit indices are usually relied on more heavily in the evaluation of model fit.

Although a host of fit indices are available, only a handful are described and recommended here. These fit indices were selected on the basis of their popularity in the applied literature and, more important, their favorable performance in Monte Carlo research. Other widely used indices such as the goodness-of-fit index and adjusted goodness-of-fit index are not included because of evidence of their poor behavior in simu-

lation studies (e.g., Hu & Bentler, 1998; Marsh, Balla, & McDonald, 1988). However, it should be noted that this topic is surrounded by considerable controversy (e.g., what indices should be used in what contexts? and what cutoff values should be used to indicate acceptable fit?). Interested readers are referred to Hu and Bentler (1995, 1998, 1999) for more information.

Fit indices can be broadly characterized as falling under three categories: absolute fit, fit adjusting for model parsimony, and comparative or incremental fit. This typology is not perfect, as some fit indices, such as the Tucker–Lewis index, have features of more than one category. Most latent variable software packages (e.g., LISREL, Mplus, Amos, EQS) provide each of the fit indices described below. Because each type of index provides different information about model fit, researchers are advised to consider and report at least one index from each category when evaluating the fit of their models.

Absolute Fit

Absolute fit indices assess model fit at an absolute level; in various ways, they evaluate the reasonability of the hypothesis that $S = \Sigma$ without taking into account other aspects such as fit in relation to more restricted solutions. Thus, χ^2 is an example of an absolute fit index. Another index that falls in this category is the standardized root mean square residual (SRMR). Conceptually, the SRMR can be viewed as the average discrepancy between the *correlations* observed in the input matrix and the correlations predicted by the model (though in actuality, the SRMR is a positive square root average; see Eq. 3.18). Accordingly, it is derived from a residual correlation matrix (e.g., see Figure 3.8). A similarly named index, the root mean square residual (RMR), reflects the average discrepancy between observed and predicted *covariances*. However, the RMR can be difficult to interpret because its value is affected by the metric of the input variables; thus, the SRMR is generally preferred. In most instances (e.g., models involving a single input matrix), the SRMR can be calculated by (1) summing the squared elements of the residual correlation matrix and dividing this sum by the number of elements in this matrix (on and below the diagonal), that is, $b = p(p + 1) / 2$ (Eq. 3.14), and (2) taking the square root (SQRT) of this result. For example, the SRMR of the Figure 3.8 solution would be computed as follows:

$$\text{SRMR} = \text{SQRT}[(0^2+0^2+0^2+.4^2+0^2+0^2)/ 6] = .163 \qquad (3.18)$$

The SRMR can take a range of values between 0.0 and 1.0, with 0.0 indicating a perfect fit (i.e., the smaller the SRMR, the better the model fit).

Parsimony Correction

Although sometimes grouped under the category of absolute fit (e.g., Hu & Bentler, 1999), these indices differ from χ^2, SRMR, and so forth, by incorporating a penalty function for poor model parsimony (i.e., number of freely estimated parameters as expressed by model df). For example, consider a scenario where two different models, Model A and Model B, fit a sample matrix (S) equally well at the absolute level; yet the specification of Model B entails more freely estimated parameters than Model A (i.e., Model A has more dfs than Model B). Indices from the parsimony class would thus favor Model A over Model B because the Model A solution fit the sample data with fewer freely estimated parameters.

A widely used and recommended index from this category is the root mean square error of approximation (RMSEA; Steiger & Lind, 1980). The RMSEA is a population-based index that relies on the *noncentral χ^2 distribution*, which is the distribution of the fitting function (e.g., F_{ML}) when the fit of the model is not perfect. The noncentral χ^2 distribution includes a *noncentrality parameter* (NCP), which expresses the degree of model misspecification. The NCP is estimated as $\chi^2 - df$ (if the result is a negative number, NCP = 0). When the fit of a model is perfect, NCP = 0 and a central χ^2 distribution holds. When the fit of the model is not perfect, the NCP is greater than 0 and shifts the expected value of the distribution to the right of that of the corresponding central χ^2 (cf. Figure 1 in MacCallum, Browne, & Sugawara, 1996). The RMSEA is an "error of approximation" index because it assesses the extent to which a model fits *reasonably* well in the population (as opposed to testing whether the model holds exactly in the population; cf. χ^2). To foster the conceptual basis of the calculation of RMSEA, the NCP is rescaled to the quantity d: $d = \chi^2 - df / (N - 1)$. The RMSEA is then computed:

$$\text{RMSEA} = \text{SQRT}[d / df] \qquad (3.19)$$

where df is the model df (although slight variations exist in some programs; e.g., Mplus uses N instead of $N - 1$). As can be seen in Eq. 3.19, the RMSEA compensates for the effect of model complexity by conveying discrepancy in fit (d) per each df in the model. Thus, it is sensitive to the

number of model parameters; being a population-based index, the RMSEA is relatively insensitive to sample size. The RMSEA from the Figure 3.8 solution would be:

$$RMSEA = SQRT[.40 / 1)] = .63$$

where $d = (80.69 - 1) / 199 = .40$.

Although its upper range is unbounded, it is rare to see the RMSEA exceed 1.0. As with the SRMR, RMSEA values of 0 indicate perfect fit (and values very close to 0 suggest good model fit).

The noncentral χ^2 distribution can be used to obtain confidence intervals for RMSEA (a 90% interval is typically used). The confidence interval indicates the precision of the RMSEA point estimate. Methodologists recommend including this confidence interval when reporting the RMSEA (e.g., MacCallum et al., 1996). However, researchers should be aware that the width of this interval is affected by sample size and the number of freely estimated parameters in the model (e.g., unless N is very large, complex models are usually associated with wide RMSEA confidence intervals).[6]

Moreover, to address the overstringent nature of χ^2 (i.e., it tests for "perfect" fit), Browne and Cudek (1993) have developed a statistical test of closeness of model fit using the RMSEA. Specifically, "close" fit (CFit) is operationalized as RMSEA values less than or equal to .05. This test appears in the output of most software packages as the probability value that RMSEA is ≤ .05. Nonsignificant probability values (i.e., $p > .05$) may be viewed in accord with acceptable model fit, although some methodologists have argued for stricter guidelines (e.g., $p > .50$; Jöreskog & Sörbom, 1996a).

Comparative Fit

Comparative fit indices (also referred to as *incremental fit indices*; e.g., Hu & Bentler, 1998) evaluate the fit of a user-specified solution in relation to a more restricted, nested baseline model. Typically, this baseline model is a "null" or "independence" model in which the covariances among all input indicators are fixed to zero, although no such constraints are placed on the indicator variances. As you might expect, given the relatively liberal criterion of evaluating model fit against a solution positing no relationships among the variables, comparative fit indices often look more favorable (i.e., more suggestive of acceptable model fit) than indices from the pre-

ceding categories. Nevertheless, some indices from this category have been found to be among the best behaved of the host of indices that have been introduced in the literature.

One of these indices, the comparative fit index (CFI; Bentler, 1990), is computed as follows:

$$\text{CFI} = 1 - \max[(\chi^2_T - df_T), 0]/\max[(\chi^2_T - df_T), (\chi^2_B - df_B), 0] \quad (3.20)$$

where χ^2_T is the χ^2 value of the target model (i.e., the model under evaluation), df_T is the df of the target model, χ^2_B is the χ^2 value of the baseline model (i.e., the "null" model), and df_B is the df of the baseline model; max indicates to use the largest value—for example, for the numerator, use $(\chi^2_T - df_T)$ or 0, whichever is larger. The χ^2_B and df_B of the null model are included as default output in most software programs. If the user wishes to obtain these values in programs that do provide this information, χ^2_B and df_B can be calculated by fixing all relationships to 0 (but freely estimating the indicator variances). The CFI has a range of possible values of 0.0 to 1.0, with values closer to 1.0 implying good model fit. Like the RMSEA, the CFI is based on the noncentrality parameter (i.e., $\lambda = \chi^2_T - df_T$, included in standard output of some programs such as LISREL), meaning that it uses information from expected values of χ^2_T or χ^2_B (or both, in the case of the CFI) under the noncentral χ^2 distribution associated with $S \neq \Sigma$ (e.g., central χ^2 is a special case of the noncentral χ^2 distribution when $\lambda = 0$). Using the results of the Figure 3.8 model, the CFI would be

$$\text{CFI} = 1 - [(80.69 - 1) / (226.75 - 3)] = .645$$

Another popular and generally well-behaved index falling under this category is the Tucker–Lewis index (TLI; Tucker & Lewis, 1973; referred to as the non-normed fit index in some programs). In addition, the TLI has features that compensate for the effect of model complexity; that is, as does the RMSEA, the TLI includes a penalty function for adding freely estimated parameters that do not markedly improve the fit of the model. The TLI is calculated by the following formula:

$$\text{TLI} = [(\chi^2_B / df_B) - (\chi^2_T / df_T)] / [(\chi^2_B / df_B) - 1] \quad (3.21)$$

where, as with the CFI, χ^2_T is the χ^2 value of the target model (i.e., the model under evaluation), df_T is the df of the target model, χ^2_B is the χ^2

value of the baseline model (i.e., the "null" model), and df_B is the df of the baseline model. Unlike the CFI, the TLI is non-normed, which means that its values can fall outside the range of 0.0 to 1.0. However, it is interpreted in a fashion similar to the CFI in that values approaching 1.0 are interpreted in accord with good model fit. The TLI for the Figure 3.8 solution is

$$TLI = [(226.75 / 3) - (80.69 / 1)] / [(226.75 / 3) - 1] = -.068$$

The collective goodness-of-fit indices from each category point to the poor fit of the Figure 3.8 solution, $\chi^2(1) = 80.69$, $p < .05$, SRMR = .163, RMSEA = .633, CFI = .645, TLI = −.068. Although straightforward in the Figure 3.8 example, the issues and guidelines for using these descriptive indices of overall model fit are considered more fully in the next section of this chapter.

Guidelines for Interpreting Goodness-of-Fit Indices

As noted earlier, issues surrounding goodness-of-fit indices are hotly debated (e.g., which indices should be used? what cutoff criteria should be applied to indicate good and poor model fit?). If the reader were to peruse published books and journal articles on this topic, he or she would note few areas of consensus in regard to recommended fit index cutoffs. Thus, it would be inappropriate to unequivocally recommend cutoffs in this book. Indeed, such endeavor is complicated by the fact that fit indices are often differentially affected by various aspects of the analytic situation such as sample size, model complexity, estimation method (e.g., ML, WLS), amount and type of misspecification, normality of data, and type of data [e.g., TLI and RMSEA tend to falsely reject models when N is small (Hu & Bentler, 1999); SRMR does not appear to perform well in CFA models based on categorical indicators (Yu, 2002)]. The reader is referred to Hu and Bentler (1998, 1999) for evidence and a detailed discussion of how such aspects may impact the performance of these fit indices (although it should be again noted that the RMSEA, SRMR, CFI, and TLI were selected in this book partly on the basis of their overall satisfactory performance in the Hu and Bentler simulations). It is also important to emphasize that goodness-of-fit indices are only one aspect of model evaluation. As discussed in Chapter 4, although model evaluation usually begins with the examination of these fit indices, it is equally important to examine a solution in terms of potential areas of localized strain (e.g., are there specific relationships the model does not adequately reproduce?) and the inter-

pretability and strength of the resulting parameter estimates (e.g., absence of Heywood cases; statistical significance, direction, and size of parameter estimates are in accord with prediction). With these caveats fully in mind, a few prominent guidelines for fit indices are reviewed.

In one of the more recent and comprehensive evaluations of cutoff criteria, the findings of simulation studies conducted by Hu and Bentler (1999) suggest the following guidelines. Support for contentions of reasonably good fit between the target model and the observed data (assuming ML estimation) is obtained in instances where (1) SRMR values are close to .08 or below; (2) RMSEA values are close to .06 or below; and (3) CFI and TLI values are close to .95 or greater. Hu and Bentler's (1999) use of the phrase "close to" is not accidental, because the recommended cutoff values were found to fluctuate as a function of modeling conditions (e.g., type of misspecified model) and whether or not an index was used in combination with other fit indices (e.g., the acceptability of Type I and Type II error rates often improved when a combination of indices was employed). Other methodologists have handled these complexities by providing descriptive anchors for various ranges of fit index values rather than specifying explicit cutoffs. For instance, Browne and Cudeck (1993) propose, as a rule of thumb, that RMSEA values less than 0.08 suggest adequate model fit (i.e., a "reasonable error of approximation," p. 144), RMSEA values less than 0.05 suggest good model fit, and that models with RMSEA ≥ 0.1 should be rejected (MacCallum et al., 1996, further elaborated on these guidelines by asserting that the RMSEAs in the range of .08–0.10 suggest "mediocre" fit). Additional support for the fit of the solution would be evidenced by a 90% confidence interval of the RMSEA whose upper limit is below these cutoff values (e.g., 0.08). As noted earlier, a nonsignificant CFit (RMSEA < .05) could also be interpreted in accord with acceptable model fit. Similarly, methodologists have noted that while CFI and TLI values below .90 should lead the researcher to strongly suspect (reject) the solution, CFI and TLI values in the range of .90–.95 may be indicative of acceptable model fit (e.g., Bentler, 1990). However, when fit indices fall in these "marginal" ranges, it is especially important to consider the consistency of model fit as expressed by the various types of fit indices in tandem with the particular aspects of the analytic situation; for example, when N is somewhat small, an RMSEA = .08 may be of less concern if all other indices are strongly in a range suggesting "good" model fit. Again, this underscores the importance of considering fit indices from multiple fit categories (absolute fit, parsimony correction, comparative fit) in tandem with examining other relevant aspects of the solution (e.g., localized areas of ill

fit; interpretability and size of parameter estimates). These aspects of model evaluation are discussed in Chapter 4.

SUMMARY

This chapter provided a detailed comparison of EFA and CFA. Although EFA and CFA are based on the common factor model, CFA has the fundamental advantage over EFA in that it allows the researcher to control every aspect of the model specification (e.g., generate an unstandardized solution, specify correlated errors, place various constraints on the solution, such as fixing cross-loadings to zero or holding model parameters to equality). As shown in later chapters, these features allow CFA to address important research questions for which EFA is not well suited (or incapable), such as the comparison of factor models across groups (e.g., measurement invariance of tests; Chapter 7), the analysis of multitrait–multimethod data in construct validation (Chapter 6), the analysis of mean structures (Chapter 7), scale reliability evaluation (Chapter 8), and the inclusion of covariates in the factor model (e.g., MIMIC, Chapter 7).

In addition, this chapter introduced the fundamental concepts of CFA such as model notation, basic equations, model identification, model estimation (e.g., ML), and goodness-of-fit evaluation. The appendices to this chapter are provided to foster the reader's understanding of each of these core principles using data-based illustrations. In Chapter 4, these concepts will be exemplified and extended in the context of a fully worked-through CFA of an applied data set.

NOTES

1. Although rarely seen in applied research, it is possible to conduct EFA using a (unstandardized) variance–covariance matrix. In fact, use of a covariance matrix in ML EFA would be preferred when the data are not scale-free (i.e., EFAs based on a correlation and covariance matrix produce markedly different model χ^2 values).

2. When measurement error is systematic (e.g., method effects), it can also lead to positively biased parameter estimates.

3. Historically, the term "congeneric" has been used more strictly to refer to one-factor measurement models (i.e., a set of indicators that load on one factor). In this book, the term is used more generally in reference to sets of indicators that load on just one factor (e.g., multifactorial measurement models that contain no double-loading indicators).

4. Some programs (e.g., Mplus) prefer that the indicators submitted to the latent variable analysis be kept on a similar scale. For instance, problems with convergence may occur if the variances of the indicators are markedly heterogeneous. This can be resolved by rescaling the indicator(s) by a linear transformation. Say, for example, that X1, X2, and X3 are measured on 0–8 scales, but X4 has scores ranging from 1 to 70 (Figure 3.7A). To avoid potential problems with convergence, X4 can be rescaled simply by dividing its observed scores by 10.

5. An initial attempt to address the shortcomings of χ^2 dates back to the seminal structural equation modeling study by Wheaton, Muthén, Alwin, and Summers (1977). In this study, the authors introduced the χ^2/df ratio in attempt to use χ^2 in a manner that would foster more realistic model evaluation. Although the χ^2/df ratio has become very popular in applied research, its use is strongly discouraged (in fact, even Wheaton, 1987, subsequently recommended that the χ^2/df ratio not be used). For instance, researchers frequently use this index under the false understanding that it adjusts for χ^2's excessive power to reject $S = \Sigma$ when in fact the χ^2/df ratio is equally sensitive to sample size (i.e., $N - 1$ is used in the calculation of both χ^2 and χ^2/df). Other reasons contraindicating the use of the χ^2/df ratio include the fact that it has not undergone the scrutiny of simulation research (cf. Hu & Bentler, 1998, 1999) and the lack of consistent standards of what χ^2/df value represents good or bad model fit.

6. The assumptions of the noncentral distribution and RMSEA do not hold if the model is grossly misspecified. For other cautions on the use of RMSEA (e.g., impact of non-normality and sample size), the reader is referred to Yuan (2005).

Appendix 3.1

Communalities, Model-Implied Correlations, and Factor Correlations in EFA and CFA

I. COMMUNALITY ESTIMATES

As noted in Chapter 2, the calculation of communalities in orthogonal EFA is straightforward. In these models, the communality is the sum of squared factor loadings for a given indicator across all factors. For instance, in the two-factor model of adolescent antisocial behavior presented in Table 3.1, the communality of the Y1 indicator is equal to its squared loading on Property Crimes plus its squared loading on Violent Crimes; that is, $.8493^2 + .1765^2 = .75$.

Communalities can also be easily computed in CFA. In orthogonal EFA, this computation is straightforward because the factors are not correlated (i.e., the factor loadings reflect the zero-order correlations between the indicator and the factors). In CFA, this computation is clear-cut because indicators are typically specified to load on one factor only (i.e., all cross-loadings are fixed to zero). Thus, the communality is simply the factor loading squared. For instance, in the CFA, the communality of the Y1 indicator is $.7996^2 = .64$ (cf. Eq. 3.5).

In oblique EFA (or in a CFA where an indicator is specified to load on two or more factors), the calculation of a communality requires an understanding of the path diagram tracing rules and equations (e.g., in this chapter, see the "Fundamental Equations of a CFA Model" section). Because factors are intercorrelated in oblique EFA, the factor loadings are partial regression coefficients (pattern matrix). Thus, computation of a communality in oblique EFA is a function of both the factor loadings and the factor correlations. Below is a path diagram of the oblique EFA solution from Table 3.1 for the Y1 and Y5 indicators only.
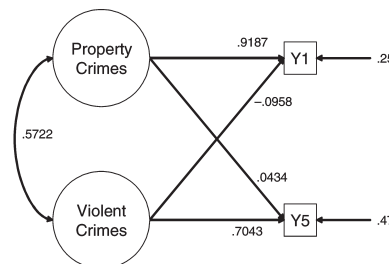
In this example, the communality of Y1 is

$$\eta^2_{y1} = \lambda_{11}^2 + \lambda_{12}^2 + 2\lambda_{11}\phi_{21}\lambda_{12} \qquad (3.22)$$
$$= .9187^2 + -.0958^2 + 2(.9187)(.5722)(-.0958)$$
$$= .75$$

The communality of Y5 is

$$\eta^2_{y5} = \lambda_{52}^2 + \lambda_{51}^2 + 2\lambda_{52}\phi_{21}\lambda_{51} \qquad (3.23)$$
$$= .7043^2 + .0434^2 + 2(.7043)(.5722)(.0434)$$
$$= .53$$

Thus, for example, the communality of Y1 is the sum of the unique direct effects of Property Crimes and Violent Crimes on Y1 (λ_{11}^2 and λ_{12}^2, respectively), and the variance in Y1 that Property Crimes and Violent Crimes jointly explain [$2(\lambda_{11})(\phi_{21})(\lambda_{12})$]. In other words, Property Crimes and Violent Crimes are correlated (ϕ_{21} = .5722) and some portion of the variance in Y1 that is explained by Property Crimes is also explained by Violent Crimes.

II. MODEL-IMPLIED CORRELATIONS

The quality of the factor solutions is determined in part by how well the parameter estimates of the solution are able to reproduce the observed correlations of the indicators. The correlations predicted by the parameters of the factor solution are often referred to as *model-implied estimates*. The model-implied correlations from oblique EFA, orthogonal EFA, and CFA are illustrated below, again using the Y1 and Y5 indicators as an example.

In the oblique EFA, the model-implied correlation between Y1 and Y5 is

$$\text{Corr}(y5, y1) = \lambda_{11}\lambda_{51} + \lambda_{52}\lambda_{12} + \lambda_{11}\phi_{21}\lambda_{52} + \lambda_{51}\phi_{21}\lambda_{12} \qquad (3.24)$$
$$= (.9187)(.0434) + (.7043)(-.0958)$$
$$+ (.9187)(.5722)(.7043) + (.0434)(.5722)(-.0958)$$
$$= .34$$

The calculation of the model-implied correlation of Y1 and Y5 is a bit easier in orthogonal EFA because the factors are not correlated. Thus, the predicted correlation is calculated solely from the summed cross-products of the factor loadings:

$$\text{Corr}(y5, y1) = \lambda_{11}\lambda_{51} + \lambda_{52}\lambda_{12} \qquad (3.25)$$
$$= (.8493)(.2587) + (.6826)(.1765)$$
$$= .34$$

Again note that the model-implied correlation of Y1 and Y5 is the same in orthogonal and oblique EFA: the choice of rotation method does not affect the fit of the factor solution.

In CFA models where each indicator loads on only one factor (cf. Table 3.1, Figure 3.1A), computation of the model-implied correlation is very straightforward (i.e., although factors are intercorrelated, each indicator loads on a single factor). The correlation of Y1 and Y5 that is predicted from the CFA solution is (cf. Eq. 3.8)

$$
\begin{aligned}
\text{Corr}(y5, y1) &= \lambda_{11}\phi_{21}\lambda_{52} \\
&= (.7996)(.6224)(.7315) \\
&= .36
\end{aligned}
\tag{3.26}
$$

III. FACTOR CORRELATIONS

In Chapter 3, it was stated that CFA factor correlations are often higher than factor correlations emanating from oblique EFA of the same data set. This occurred in the analyses presented in Table 3.1, where the factor correlations estimated by oblique EFA and CFA were .57 and .62, respectively. In an oblique EFA, the model-implied correlation of indicators with primary loadings on separate factors can be estimated in part by the indicator cross-loadings (cf. Eq. 3.24). In contrast, the model-implied correlation of indicators loading on separate factors in CFA is estimated solely by the primary loadings and the factor correlation (cf. Eq. 3.26). For example, as compared with oblique EFA, in the CFA more burden is on the factor correlation to reproduce the correlation between Y1 and Y5 because there are no cross-loadings to assist in this model-implied estimate (cf. Figure 3.1A). Therefore, in the iterative process of establishing CFA parameter estimates that best reproduce the sample correlation matrix, the magnitude of the factor correlation estimate may be increased somewhat (relative to oblique EFA) to better account for the relationships of indicators that load on separate factors (assuming that these indicators are correlated to some degree).

Obtaining a Solution
for a Just-Identified Factor Model

Factor analysis typically entails matrix algebra and other mathematical operations that are very cumbersome to demonstrate and conduct by hand calculation. However, the solution for a basic, just-identified factor model such as the one depicted in Figure 3.6B can be calculated on the basis of principles discussed in this chapter and with the help of the algebra of simultaneous equations. Consider the following input matrix and just-identified, one-factor model:

Input matrix:

```
      X1       X2       X3
X1   1.000
X2   0.595    1.000
X3   0.448    0.544    1.000
```
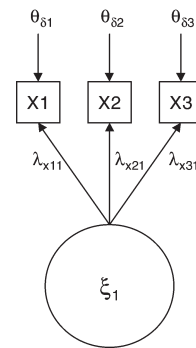
For ease of notation, let $a = \lambda_{x11}$, $b = \lambda_{x21}$, and $c = \lambda_{x31}$.

As discussed is this chapter, the number of knowns (6 elements of the input matrix) equals the number of unknowns (6 parameters = 3 factor loadings, 3 errors; the factor variance is fixed to 1.0). Thus, the model is just-identified, and its parameter estimates will perfectly reproduce the input matrix. It was also discussed that, for two indicators loading on the same factor, multiplying their factor loadings provides the model estimate of their zero-order correlation. Because the current model is just-identified, the products of the loadings will perfectly reproduce the zero-order relationships among X1, X2, and X3. The systems of equations are as follows:

Equation 1. $ab = .595$
Equation 2. $ac = .448$
Equation 3. $bc = .544$

This problem is also just-identified, as there are 3 unknowns (a, b, c) and 3 knowns (the 3 equations). Systems of equations can be solved in various ways (e.g., substitution, elimination, matrices). In the example, substitution is used by first solving for a:

Step 1. Rearrange Equations 1 and 2, so that b and c are the outputs.

Equation 1. $ab = .595$ $b = .595/a$
Equation 2. $ac = .448$ $c = .448/a$
Equation 3. $bc = .544$

Step 2. Substitute Equations 1 and 2 in Equation 3.

Equation 3. b c $= .544$
                     $(.595/a)(.448/a) = .544$

Step 3. Solve Equation 3.

Equation 3. $(.595/a)(.448/a) = .544$
                     $.26656/a^2$ $= .544$
                     $.26656/.544$ $= a^2$
                     $.49$ $= a^2$
                     $.70$ $= a$

Step 4. Now that we know that $a = .70$, it is straightforward to solve for b and c using the original equations.

Equation 1. $.70b = .595$ $.595/.70 = b$ $b = .85$
Equation 2. $.70c = .448$ $.448/.70 = c$ $c = .64$

Thus, the factor loadings are .70, .85, and .64, for λ_{x11}, λ_{x21}, and λ_{x31}, respectively. As an accuracy check, note that multiplying these loadings together perfectly reproduces the input correlations; that is, $.70(.85) = .595$, $.70(.64) = .448$, $.85(.64) = .544$.

Step 5. Because this is a completely standardized solution, the errors for X1, X2, and X3 can be obtained by squaring the loadings and subtracting the result from 1.0.

$$\delta_1 = 1 - .70^2 = .51$$
$$\delta_2 = 1 - .85^2 = .2775$$
$$\delta_3 = 1 - .64^2 = .5904$$

Now every unknown parameter in the factor model has been solved, and the solution is as follows:
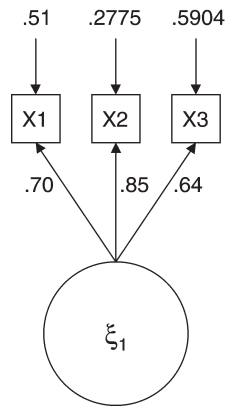
Hand Calculation of F_{ML} for the Figure 3.8 Path Model

Calculation of F_{ML} is best left to the computer because the computations needed to produce this value are quite tedious and complex in most data sets. However, given the simplicity of the Figure 3.8 model, hand calculation of F_{ML} is presented in effort to foster the reader's conceptual understanding of the ML fit function:

$$F_{ML} = \ln|S| - \ln|\Sigma| + \text{trace}[(S)(\Sigma^{-1})] - p$$

The presentation will focus primarily on the first half of this equation, $\ln|S| - \ln|\Sigma|$, given its relationship to other widely known statistics (e.g., F ratio, likelihood ratio), and the fact that, in this example, the remaining part of the equation will equal zero. The remaining portion of the F_{ML} formula, $\text{trace}[(S)(\Sigma^{-1})] - p$, pertains to the distances of the indicator variances in S and Σ. In many instances (as in the Figure 3.8 solution), this difference will be zero because the CFA model perfectly reproduces the observed variance of the input indicators; cf. $\sigma = \lambda_x^2\phi_1 + \delta$, and thus $\text{trace}[(S)(\Sigma^{-1})] - p = 3 - 3 = 0$ (see Table 3.2). However, there are some instances where differences in the diagonal elements (indicator variances) of S and Σ do not equal zero and thus contribute to F_{ML} (e.g., in the CFA evaluation of tau equivalence, see Chapter 7). But in the Figure 3.8 solution, the value of F_{ML} is determined solely by the difference in the natural logs of the determinants of S and Σ.
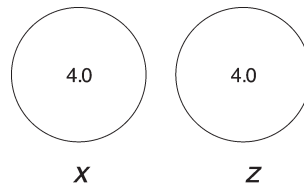
DETERMINANTS ($|S|$, $|\Sigma|$)

As noted in Chapter 3, a determinant is a matrix algebra term for a single number (i.e., a scalar) that reflects a generalized measure of variance for the entire set of variables contained in a matrix. Stated another way, it is an index of the amount of nonredundant variance that reflects the extent to which variables in the matrix are free to vary. This concept is illustrated in the following zero-order (2×2) and multivariate (3×3) examples, drawing from the sample variance–covariance matrix in Figure 3.8.

Consider the scenario where x and z share none of their variance (i.e., $r_{xz} = .00$). If x and z had SDs = 2.0, the 2 × 2 variance–covariance matrix would be as follows:

	x	z
x	4.0	0.0
z	0.0	4.0

The variance and (lack of) covariance of x and z are represented in the following Venn diagram:

Using a basic principle of matrix algebra, the determinant of a 2 × 2 matrix,

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is $|\mathbf{A}| = ad - bc$

In the present example, $\mathbf{A}$ is a variance–covariance matrix and thus the equation $ad - bc$ represents subtracting the covariance (bc) from the variance (ad). Solving this equation, $4(4) - 0(0)$ results in a determinant of 16. This value indicates that all the variance in the matrix of x and z (4 * 4 = 16) is nonredundant; that is, x and z are entirely free to vary. Because the Venn diagram indicates no overlapping variance between x and z, the determinant could be calculated simply by multiplying the variance of x and z.

Consider the situation where x and z share all of their variance (i.e., $r_{xz} = 1.0$). The variance–covariance matrix and Venn diagram would look as follows:

|	x	z
x	4.0	4.0
z	4.0	4.0
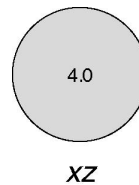
In this case, the determinant equals 0, $ad - bc = 4(4) - 4(4) = 0$, which reflects that all the variance in x and z (ad) is redundant (bc); in other words, there is no generalized variability in this matrix (no freedom to vary). Alternatively, using the Venn diagram, the determinant could be solved by multiplying the variance in x (4) by the unique variance in z (0): 4 * 0 = 0.

When a matrix has a determinant of 0, it is said to be *singular* (conversely, matrices with non-zero determinants are *nonsingular*), meaning that one or more rows or columns are linearly dependent on other rows or columns (as would occur if a variance–covariance was constructed using the subscales and total score of a questionnaire if the total score is formed by summing the subscales). Because determinants are used in calculating the *inverse* of a matrix (e.g., Σ^{-1}, see F_{ML} equation), a singular matrix is problematic because it has no inverse; the matrix algebra equivalent to division is multiplying a matrix by the inverse of another matrix; for example, $(S)(\Sigma^{-1})$, see F_{ML} equation. Thus, many statistics for the matrix cannot be computed. Users of latent variable software programs may occasionally encounter error messages stating that a matrix is singular or nonpositive definite. These errors often stem from linear dependency or some other problems in the input matrix (e.g., multicollinearity/one or more eigenvalues = 0; use of pairwise deletion as a missing data strategy; N exceeds the number of input indicators), although the matrix generated by the specified model must also be positive definite (Wothke, 1993).

Now consider the zero-order relationship of x and z that was observed in the Figure 3.8 model (r_{xz} = .60). The variance–covariance and Venn diagram are as follows:

	x	z
x	4.0	2.4
z	2.4	4.0
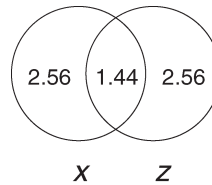
The determinant of this matrix is 10.24; $ad - bc$ = 4(4) − 2.4(2.4) = 16 − 5.76 = 10.24. Some of the total variance in x and z (ad = 16) is redundant (bc = 5.76); subtracting bc from ad provides the measure of generalized variability in this matrix. Also note that dividing bc by ad yields the proportion of shared variance in x and z: 5.76 / 16 = .36 = r^2. Alternatively, using the Venn diagram, the determinant could be computed by multiplying the variance of x (4) by the unique variance in z (2.56): 4 ∗ 2.56 = 10.24. The unique variance in z is calculated by multiplying its variance (4) by its proportion of unique variance ($1 - r^2$ = .64): 4 ∗ .64 = 2.56.

Computing the determinant for matrices of the order 3 × 3 and higher is more complicated. In order to begin with a computationally simpler example, consider the scenario where the correlation between x and z is .60, z and y are correlated .50, but x and y share none of their variance (i.e., r_{xy} = 0.0). This variance–covariance matrix and Venn diagram would follow:

	x	z	y
x	4.0	2.4	0.0
z	2.4	4.0	2.0
y	0.0	2.0	4.0
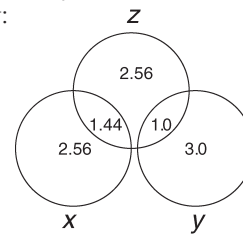
Using another method from matrix algebra, the determinant of a 3×3 matrix,

$$\mathbf{A} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

is: $|\mathbf{A}| = (aei - bdi) + (bfg - ceg) + (cdh - afh)$

Although there are different approaches to solving for the determinant of a 3×3 matrix (e.g., "butterfly method," use of *minors* and their matrices of *cofactors*), the above method is computationally similar to the strategy used to solve for 2×2 matrices (although alternate approaches, such as the method of minors and cofactors, generalize better to matrices of the order of 4×4 and higher; cf. Bollen, 1989). Here, we begin with the first element of the matrix (*a*) and multiply it by all other elements on its diagonal (*aei*). Then, the element in the column adjacent to *a* is selected (*b*) and is multiplied by the remaining elements in the diagonal going in the opposite direction of *aei* (*bdi*; compare with method used for 2×2 matrices). Note that the remaining element in *bdi* (i.e., *i*) is obtained by bringing the column [*c f i*] in front of the column [*a d g*]. Then *bdi* is subtracted from *aei*, and the process is repeated two more times for the remaining elements in the first row of **A** (*bfg – ceg*, *cdh – afh*). The results are then summed.

Applying this method to the above variance–covariance matrix yields a determinant of 24.96:

aei − bdi:	$(4.0 * 4.0 * 4.0) - (2.4 * 2.4 * 4.0) =$	40.96
bfg − ceg:	$(2.4 * 2.0 * 0.0) - (0.0 * 4.0 * 0.0) =$	0.00
cdh − afh:	$(0.0 * 2.4 * 2.0) - (4.0 * 2.0 * 2.0) =$	−16.00
	Total:	24.96

Thus, of the total variance in the matrix of *x*, *z*, and *y* ($4 * 4 * 4 = 64$), a considerable portion is overlapping (39.04) and not included in the determinant. Note that dividing 39.04 by 64 yields .61, which would reflect R^2 if *z* was regressed onto *x* and *y* (it works out cleanly in this example because *x* and *y* do not correlate; $r_{xz}^2 = .36$, $r_{zy}^2 = .25$, $.36 + .25 = .61$). Indeed, using the Venn diagram method, we could readily solve for the determinant by multiplying the variance of *x* (4) by the variance of *y* (4) by the unique variance in *z* (1.56): $4 * 4 * 1.56 = 24.96$; the unique variance in *z* is computed by subtracting its variance (4) with the variance explained by *x* ($4 * .36 = 1.44$) and *y* ($4 * .25 = 1.0$); $4 - 1.44 - 1 = 1.56$ (again simplified by having no overlap between *x* and *y*).

It is hoped that this provides the reader with a better conceptual understanding of a determinant. Let's now turn to the sample (*S*) and predicted (Σ) matrices from Figure 3.8:

	S				Σ		
	x	z	y		x	z	y
x	4.0	2.4	2.8	x	4.0	2.4	1.2
z	2.4	4.0	2.0	z	2.4	4.0	2.0
y	2.8	2.0	4.0	y	1.2	2.0	4.0

Using the matrix algebra method presented earlier, the determinant of S is found to be:

$aei - bdi$:	$(4.0 * 4.0 * 4.0) - (2.4 * 2.4 * 4.0) =$	40.96
$bfg - ceg$:	$(2.4 * 2.0 * 2.8) - (2.8 * 4.0 * 2.8) =$	−17.92
$cdh - afh$:	$(2.8 * 2.4 * 2.0) - (4.0 * 2.0 * 2.0) =$	−2.56
	Total:	20.48

Note that this is the same value that was calculated for S by SAS PROC IML in Table 3.2. Using the same method, the determinant of Σ is found to be 30.72 (the determinants of S and Σ could also be computed by Venn diagrams, although this would require more time-consuming steps of solving for the unique and overlapping areas in x, y, and z; e.g., regressing each variable onto the remaining two variables). Because Σ has more generalized variance (30.72) than S (20.48), a poor-fitting model may result. The additional nonredundant variance in Σ is due to the fact that the Figure 3.8 model parameters predicted less covariance between x and y (σ_{xy} = 1.2) than was observed in the sample data (σ_{xy} = 2.8).

DIFFERENCE OF NATURAL LOGS (LN|S| − LN|Σ|)

Returning to the equation for F_{ML}, note that the natural logs (ln) of the determinants are used rather than the raw values of the determinants themselves. The reason for this is as much computational as theoretical. Specifically, logs have several properties that make working with and interpreting them easier than other methods. For instance, consider the situation where the probability of a treatment success is 70% (p = .7), and the probability of no treatment success is 30% (q = .3). Converting the probability of success to an odds ratio, OR = $p / (1 - p)$, yields a value of 2.33 (.7/.3); that is, treatment success is 2.3 times more likely than nonsuccess. However, the OR for nonsuccess is .429 (.3/.7). Although it would be helpful to be able to interpret this outcome in a symmetrical fashion (i.e., the odds of success are opposite the odds of nonsuccess), ORs do not have this symmetry (i.e., 2.33, .43). However, taking the natural log of these ORs provides this symmetry, ln(2.33) = .846, ln(.29) = −.846. This is one reason why log odds are commonly used in logistic regression. Logs have other useful properties such as:

if:	$a = bc$	e.g., 20 = 5(4)
then:	$\ln(a) = \ln(b) + \ln(c)$	2.996 = 1.6094 + 1.3863

and thus,

if:	$a/b = c$	e.g., 20/5 = 4
then:	$\ln(c) = \ln(a) - \ln(b)$	1.3863 = 2.996 = 1.6094
		(within rounding error)

The latter equation is reflected in the first part of the formula for F_{ML}. Specifically, taking the difference in $\ln|S|$ and $\ln|\Sigma|$ is equivalent to dividing the determinant of one of these matrices by the determinant of the other; in fact, in a few textbooks, this portion of the equation is written as $\ln(S\Sigma^{-1})$, which produces the same result as $\ln|S| - \ln|\Sigma|$. Thus, the parallel of $\ln|S| - \ln|\Sigma|$ with the F ratio found in analysis of variance (calculated by dividing one variance into another) should be apparent. This is also why the likelihood ratio (LR) is so called, when it fact it is calculated by taking the difference between two log-likelihoods; LR = $2(LL_A - LL_O)$.

Using the results from Figure 3.8, the natural logs of the determinants of S and Σ are as follows:

$\ln(S) = \ln(20.48) = 3.01945$
$\ln(\Sigma) = \ln(30.72) = 3.42491$

The difference between the natural logs of these determinants is .40546, which equals the value of F_{ML} obtained by SAS PROC IML (Table 3.2). The last part of the equation does not alter this value because the trace of $[(S)(\Sigma^{-1})]$ and order of the matrix (p) both equal 3 (actually, the result of $\ln|S| - \ln|\Sigma|$ is a negative, $-.40546$, but the absolute value is the obviously the same). Equivalently, we could solve for F_{ML} by dividing the raw score values of the determinants of S and Σ, and then derive the natural log of the result; for example, 30.72/20.48 = 1.5, $\ln(1.5)$ = .40546; 20.48/30.72 = 0.66667, $\ln(.66667)$ = $-.40546$ (again illustrating the advantage of natural logs shown in the discussion of log odds).

MODEL χ^2

Finally, note from Chapter 3 that the model χ^2 is calculated by $F_{ML}(N - 1)$. This equation parallels a χ^2 formula commonly found in introductory statistics books for testing whether one sample variance (σ_2^2) differs from another variance or a population variance (σ_1^2):

$$\chi^2 = \frac{(N-1)\sigma_2{}^2}{\sigma_1{}^2}$$

This equation can be readily re-expressed as $(\sigma_2{}^2/\sigma_1{}^2)(N-1)$, of which the first part is equivalent to $\ln(\sigma_2{}^2) - \ln(\sigma_1{}^2)$, and which, in its entirety, is equivalent to $\chi^2 = (\ln|S| - \ln|\Sigma|)(N-1)$ or $F_{ML}(N-1)$ (again, this holds in situations where the latter half of the F_{ML} equation equals 0).

4

Specification and Interpretation of CFA Models

This chapter uses a complete, applied example to demonstrate and extend the concepts of CFA presented in the previous chapters. A two-factor measurement model of personality (neuroticism, extraversion) is evaluated using five popular latent variable software programs (LISREL, Mplus, Amos, EQS, CALIS). The example is used to illustrate how the acceptability of the CFA solution should be evaluated (e.g., overall goodness-of-fit, areas of strain in the solution, interpretability/ strength of parameter estimates). Moreover, the results of this analysis are used to demonstrate how to interpret and compute the unstandardized and standardized parameter estimates of the CFA solution. The aims and procedures of CFA models that include single indicators are described. The chapter concludes with a discussion and illustration of the material that should be included when reporting the results of a CFA study.

AN APPLIED EXAMPLE OF A CFA MEASUREMENT MODEL

The concepts introduced in Chapters 2 and 3 are now illustrated and extended in the context of a full example of a CFA measurement model. The hypothesized model and sample data (correlations, *SD*s) are presented in Figure 4.1 (drawn loosely from the vast literature on the five-factor model of personality; e.g., Wiggins, 1996). In this example, a researcher has collected eight measures (subscales of a widely used personality mea-
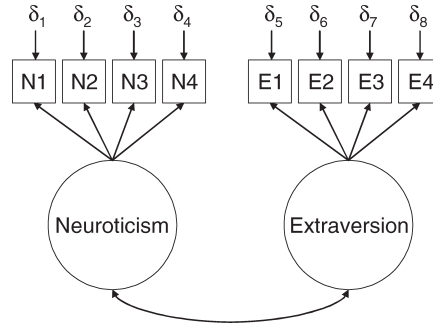
Correlations/Standard Deviations (SD):

```
            N1       N2       N3       N4       E1       E2       E3       E4
N1    1.000
N2    0.767   1.000
N3    0.731   0.709   1.000
N4    0.778   0.738   0.762   1.000
E1   -0.351  -0.302  -0.356  -0.318   1.000
E2   -0.316  -0.280  -0.300  -0.267   0.675   1.000
E3   -0.296  -0.289  -0.297  -0.296   0.634   0.651   1.000
E4   -0.282  -0.254  -0.292  -0.245   0.534   0.593   0.566   1.000

SD:   5.700   5.600   6.400   5.700   6.000   6.200   5.700   5.600
```

FIGURE 4.1. Path diagram and input data for two-factor CFA model of Neuroticism and Extraversion. N1 = anxiety; N2 = hostility; N3 = depression; N4 = self-consciousness; E1 = warmth; E2 = gregariousness; E3 = assertiveness; E4 = positive emotions. All indicators measured on scales ranging from 0 to 32 (higher scores reflect higher levels of the assessed dimension); $N = 250$.

sure) from a sample of 250 psychotherapy outpatients: anxiety (N1), hostility (N2), depression (N3), self-consciousness (N4), warmth (E1), gregariousness (E2), assertiveness (E3), and positive emotions (E4). A two-factor model is posited whereby the observed measures of anxiety, hostility, depression, and self-consciousness are conjectured to load on a latent dimension of Neuroticism, and the observed measures of warmth, gregariousness, assertiveness, and positive emotions are predicted to load onto a distinct latent factor of Extraversion.

Although the Figure 4.1 model is basic, numerous predictions underlie this model specification. For instance, all measurement error is presumed to be unsystematic—that is, there are no correlated measurement

errors for any pairs of indicators. This implies that for indicators loading on the same latent factor, the observed covariance among these measures can be explained entirely by the underlying construct; that is, there is no reason for these observed relationships other than the latent factor. For example, anxiety (N1) and hostility (N2) are correlated (r = .767; see Figure 4.1) because they are manifest symptoms influenced (caused) by the same latent construct (Neuroticism); in other words, if Neuroticism was somehow partialed out, the relationship between anxiety and hostility would be reduced to zero (r = .00). In addition, this measurement model asserts that Neuroticism and Extraversion are correlated (as depicted by the bidirectional, curved arrow in Figure 4.1), although the nature of this relationship is unanalyzed. In CFA, latent factors are almost always permitted to be intercorrelated and there is no claim about the directionality of such relationships (e.g., that Neuroticism has a direct effect on Extraversion or vice versa). Indeed, as noted in Chapter 3, if the researcher specifies directional relationships among latent factors, he or she has moved out of the CFA measurement model framework and into the realm of structural models where any sources of ill fit may stem from both the measurement and structural portions of the solution (see Figure 3.2 in Chapter 3).[1] Allowing Neuroticism and Extraversion to be correlated also implies that there may be some relationship between indicators that load on separate factors. However, because indicators do not load on more than one factor, it is predicted that any such relationships can be accounted for by the correlation between the latent factors. For example, the observed correlation between anxiety (N1) and warmth (E1) was –.351 (see Figure 4.1). As illustrated by a formula introduced in Chapter 3 (Eq. 3.8), this model specification implies that the observed correlation of –.351 should be reproduced in the completely standardized solution by multiplying the factor loadings of N1 and E1 and the correlation between Neuroticism and Extraversion (i.e., $\lambda_{x11}\phi_{21}\lambda_{52}$). In other words, no cross-loadings (e.g., Neuroticism $\rightarrow$ E1) or error covariances are needed to reproduce these relationships; any covariation between N1–N4 and E1–E4 can be explained by overlap in Neuroticism and Extraversion in tandem with the relationships of the indicators with their latent factors.

Using a simple formula provided in Chapter 3 (Eq. 3.14), it can be readily determined that the input matrix (S) contains 36 pieces of information: 8 variances (p) and 28 covariances [$p(p-1)/2$]; that is, 8(9) / 2 = 36. The measurement model presented in Figure 4.1 contains 17 freely estimated parameters: 6 factor loadings (N1 and E1 will serve as marker indi-

cators and thus their factor loadings will be fixed), 8 error variances, 2 factor variances, and 1 factor covariance. The model is overidentified with 19 *df* (*df* = 36 − 17) and goodness-of-fit evaluation will apply.

MODEL SPECIFICATION

Substantive Justification

A few important steps precede the actual CFA analysis. As noted in preceding chapters, CFA requires specification of a measurement model that is well grounded by prior empirical evidence and theory. This is because CFA entails more constraints than other approaches such as EFA (e.g., pre-specification of indicator–factor relationships, fixing cross-loadings and error covariances to zero), thus requiring the researcher to have a firm substantive and empirical basis to guide the model specification. The CFA specification depicted in Figure 4.1 would likely be based on a strong conceptual framework (e.g., the Big Five model of personality) and prior research of a more exploratory nature (e.g., EFAs that indicate the number of factors and pattern of indicator–factor loadings and that may have led to the refinement of an assessment instrument such as removal of poorly behaved indicators).

Defining the Metric of Latent Variables

Model specification also entails defining the metric of the latent factors. As shown in Chapter 3, this can be accomplished by setting one observed measure on each factor as a marker indicator or by fixing the variance of the factors to a specific value (most commonly to 1.0). In applied research, the marker indicator approach is more frequently used. When this method is used, the researcher must decide which observed measures will serve as marker indicators. In practice, marker indicators are often selected with little consideration or are determined by software defaults (e.g., unless the default is overridden by the user, Mplus automatically selects the first indicator to be the reference indicator; see Figure 4.2). Although this selection may be relatively trivial in some instances (e.g., CFA of a questionnaire with tau equivalent items), there are many circumstances in which marker indicators should be chosen carefully. For example, consider the scenario where the researcher wishes to define a latent construct of Depression with the following three indicators: (1) a single 0–8 clinical rating of depression

severity (with little available evidence supporting its interrater reliability), (2) a widely used and psychometrically well-established and normed questionnaire (range of possible scores = 0–64), and (3) a newly developed but psychometrically promising self-report screening measure comprised of dimensional ratings of the nine constituent symptoms of DSM-IV major depression (range of possible scores = 0–36). In this instance, a strong case would be made for using the second measure as the marker indicator because it is the psychometrically strongest measure, it is the most widely known in the applied literature, its metric is most meaningful (e.g., normative data exist). And, it yields the most units of discrimination (e.g., 0–64 vs. 0–36 or 0–8). As a result of choosing the second indicator, more variance is apt to be passed onto the latent factor of Depression, and the unstandardized solution should be more interpretable. The latter aspect would be particularly germane to MIMIC models (Chapter 7) and structural models that might entail regressive effects involving the latent dimension of Depression; for example, a one unit increase in the predictor is associated with an x-unit change in Depression—this unstandardized path x is more easily interpreted if the metric of the marker indicator for Depression is widely known and understood.

DATA SCREENING AND SELECTION OF THE FITTING FUNCTION

The vast majority of CFA and SEM analyses in the applied research literature are conducted using maximum likelihood (ML) estimation. As discussed in Chapter 3, ML rests on several key assumptions, including sufficient sample size and the use of indicators that approximate interval-level scales and multivariate normality. In some instances, ML estimation is able to withstand mild violations of these assumptions (see Chapter 9). Nevertheless, it is important to evaluate the sample size and sample data in regard to their suitability for CFA and ML estimation because marked departures from these assumptions would point to the need for alternative analytic approaches or estimators (e.g., robust ML, weighted least squares). Sample size considerations are fully discussed in Chapter 10. With the exception of Mplus, the major latent variable software programs (e.g., Amos, EQS, LISREL) contain routines for screening the sample data for univariate and multivariate normality (i.e., skewness and kurtosis) and outliers (i.e., aberrant cases). Broader statistical packages such as SAS and SPSS also have these capabilities (cf. Tabachnick & Fidell, 2001). The spe-

cifics of these data screening procedures are not detailed here, given that this topic falls somewhat outside the scope of this book and that thorough presentations of this issue are available in other SEM sources (e.g., Kline, 2005). Data screening is conducted on the raw sample data. If the data are deemed suitable for ML analysis, the researcher has the option of using the raw data, a correlation matrix, or a variance–covariance matrix as input data for the CFA. If a correlation matrix is used, the indicator *SD*s must also be provided in order for the program to convert the matrix into variances and covariances. If a fitting function other than ML is employed, a simple correlation or covariance matrix cannot be used (e.g., WLS requires that either the raw data or correlation and asymptotic covariance matrices be inputted; see Chapter 9). At this stage of the process, the researcher must also decide how missing data will be handled, a topic discussed in Chapter 9.

RUNNING THE CFA ANALYSIS

After model specification and data screening issues have been settled, the CFA model can be fit to the data. Table 4.1 provides syntax programs for the Figure 4.1 model in LISREL, Mplus, EQS, Amos (Basic), and CALIS. The reader is encouraged to focus initially on the longhand LISREL programming, given its direct correspondence to the matrices, symbols, and equations introduced in Chapter 3; for example, lambda-X (LX; Λ_x, λ_x), theta-delta (TD; Θ_δ, δ) and phi (PH; Φ, ϕ). Although the particular aspects of each programming language are not detailed here, a few comparisons are merited. In each analysis, anxiety and warmth were used as marker indicators for Neuroticism and Extraversion, respectively. However, in addition to many other syntax differences, the method of setting the marker indicator varies across software programs: (1) LISREL uses the Value (VA) command to fix the unstandardized loading to 1.0; (2) by default, Mplus selects the first indicator listed after the BY keyword as the marker variable (this default can be overridden by additional programming); (3) in EQS, an asterisk (*) is used to specify that a parameter should be freely estimated; if the asterisk is omitted, the EQS default is to fix the unstandardized parameter to 1.0 (as in Mplus, this and any other system default can be overridden by the user); and (4) in Amos and CALIS, a "1" is placed before the factor name in the equation involving the marker indicator.

TABLE 4.1. Computer Syntax (LISREL, Mplus, EQS, Amos, CALIS) for a Two-Factor CFA Model of Neuroticism and Extraversion

LISREL 8.72

```
TITLE TWO FACTOR MODEL OF NEUROTICISM AND EXTRAVERSION
DA NI=8 NO=250 MA=CM
LA
N1 N2 N3 N4 E1 E2 E3 E4
KM
1.000
 0.767  1.000
 0.731  0.709  1.000
 0.778  0.738  0.762  1.000
-0.351 -0.302 -0.356 -0.318  1.000
-0.316 -0.280 -0.300 -0.267  0.675  1.000
-0.296 -0.289 -0.297 -0.296  0.634  0.651  1.000
-0.282 -0.254 -0.292 -0.245  0.534  0.593  0.566  1.000
SD
5.7 5.6 6.4 5.7 6.0 6.2 5.7 5.6
MO NX=8 NK=2 PH=SY,FR LX=FU,FR TD=SY,FR
LK
NEUROT EXTRAV
PA LX
0 0
1 0
1 0
1 0
0 0
0 1
0 1
0 1
VA 1.0 LX(1,1) LX(5,2)
PA TD
1
0 1
0 0 1
0 0 0 1
0 0 0 0 1
0 0 0 0 0 1
0 0 0 0 0 0 1
0 0 0 0 0 0 0 1
PA PH
1
1 1
OU ME=ML RS MI SC ND=4
```

(cont.)

TABLE 4.1. *(cont.)*

Mplus 3.11

```
TITLE:      TWO FACTOR MODEL OF NEUROTICISM AND EXTRAVERSION
DATA:       FILE IS NEUROT.DAT;
            TYPE IS STDEVIATIONS CORRELATION;
            NOBSERVATIONS ARE 250;
VARIABLE:   NAMES ARE N1-N4 E1-E4;
ANALYSIS:   ESTIMATOR=ML;
MODEL:      NEUROT BY N1-N4;
            EXTRAV BY E1-E4;
OUTPUT:     SAMPSTAT MODINDICES(3.84) STANDARDIZED RESIDUAL;
```

EQS 5.7b

```
/TITLE
 two-factor model of neuroticism and extraversion
/SPECIFICATIONS
 CASES=250; VARIABLES=8; METHODS=ML; MATRIX=COR; ANALYSIS=COV;
/LABELS
 v1=anxiety; v2=hostil; v3=depress; v4=selfcon; v5=warmth; v6=gregar;
 v7=assert; v8=posemot; f1 = neurot; f2 = extrav;
/EQUATIONS
 V1 =    F1+E1;
 V2 =   *F1+E2;
 V3 =   *F1+E3;
 V4 =   *F1+E4;
 V5 =    F2+E5;
 V6 =   *F2+E6;
 V7 =   *F2+E7;
 V8 =   *F2+E8;
/VARIANCES
 F1 TO F2 = *;
 E1 TO E8 = *;
/COVARIANCES
 F1 TO F2 = *;
/MATRIX
1.000
 0.767  1.000
 0.731  0.709  1.000
 0.778  0.738  0.762  1.000
-0.351 -0.302 -0.356 -0.318  1.000
-0.316 -0.280 -0.300 -0.267  0.675  1.000
-0.296 -0.289 -0.297 -0.296  0.634  0.651  1.000
-0.282 -0.254 -0.292 -0.245  0.534  0.593  0.566  1.000
/STANDARD DEVIATIONS
5.7 5.6 6.4 5.7 6.0 6.2 5.7 5.6
/PRINT
 fit=all;
```

TABLE 4.1. *(cont.)*

```
/LMTEST
/WTEST
/END
```

Amos Basic 5.0

```
' Example of CFA in Amos 5.0
' Two-factor model
`
Sub Main ()
Dim sem As New AmosEngine

sem.TextOutput
sem.Standardized
sem.Smc

sem.BeginGroup "Ninput2.txt"

sem.Structure "x1 =   (1) NEUROT + (1) th1"
sem.Structure "x2 =       NEUROT + (1) th2"
sem.Structure "x3 =       NEUROT + (1) th3"
sem.Structure "x4 =       NEUROT + (1) th4"
sem.Structure "x5 =   (1) EXTRAV + (1) th5"
sem.Structure "x6 =       EXTRAV + (1) th6"
sem.Structure "x7 =       EXTRAV + (1) th7"
sem.Structure "x8 =       EXTRAV + (1) th8"
sem.Structure "NEUROT <-> EXTRAV"

End Sub
```

SAS 8.2 PROC CALIS

```
Title "CFA of Two-Factor Model of Neuroticism and Extraversion";
Data NEO (type=CORR);
 input _TYPE_ $ _NAME_ $ V1-V8;
 label V1 = 'anxiety'
       V2 = 'hostil'
       V3 = 'depress'
       V4 = 'selfcons'
       V5 = 'warmth'
       V6 = 'gregar'
       V7 = 'assert'
       V8 = 'posemots';
cards;
```

TABLE 4.1. *(cont.)*

```
mean  .    0       0      0      0      0      0      0      0
 std  .   5.7     5.6    6.4    5.7    6.0    6.2    5.7    5.6
   N  .   250     250    250    250    250    250    250    250
corr V1  1.000    .       .      .      .      .      .      .
corr V2  0.767   1.000    .      .      .      .      .      .
corr V3  0.731   0.709  1.000    .      .      .      .      .
corr V4  0.778   0.738  0.762  1.000    .      .      .      .
corr V5 -0.351  -0.302 -0.356 -0.318  1.000    .      .      .
corr V6 -0.316  -0.280 -0.300 -0.267  0.675  1.000    .      .
corr V7 -0.296  -0.289 -0.297 -0.296  0.634  0.651  1.000    .
corr V8 -0.282  -0.254 -0.292 -0.245  0.534  0.593  0.566  1.000
;
run;
proc calis data=NEO cov method=ml pall pcoves;
var = V1-V8;
lineqs
 V1 = 1.0  f1 + e1,
 V2 = lam2 f1 + e2,
 V3 = lam3 f1 + e3,
 V4 = lam4 f1 + e4,
 V5 = 1.0  f2 + e5,
 V6 = lam6 f2 + e6,
 V7 = lam7 f2 + e7,
 V8 = lam8 f2 + e8;
std
  f1-f2 = ph1-ph2,
  e1-e8 = td1-td8;
cov
  f1-f2 = ph3;
run;
```

Mplus requires the fewest programming lines because the sample data cannot be embedded in the syntax file (same with Amos) and because the program includes a variety of defaults that correspond to frequently employed CFA specifications. For instance, Mplus automatically sets the first observed variable to be the marker indicator and freely estimates the factor loadings for the remaining indicators in the list. By default, all error variances are freely estimated and all error covariances and indicator cross-loadings are fixed to 0; the factor variances and covariances are also freely estimated by default. These and other convenience features in Mplus are no doubt quite appealing to the experienced latent variable researcher. However, novice users are advised to become fully aware of these system defaults to ensure that their models are specified as intended. Note that starting values have not been provided in any of the syntax files. Thus, by

default, each program is instructed to automatically generate initial estimates to begin the iterations to minimize F_{ML}.

MODEL EVALUATION

One of the most important aspects of model evaluation occurs prior to the actual statistical analysis—that is, providing a compelling rationale that the model is meaningful and useful on the basis of prior research evidence and theory. After substantive justification of the model is established, the acceptability of the fitted CFA solution should be evaluated on the basis of three major aspects: (1) overall goodness of fit; (2) the presence or absence of localized areas of strain in the solution (i.e., specific points of ill fit); and (3) the interpretability, size, and statistical significance of the model's parameter estimates. A common error in applied CFA research is to evaluate models exclusively on the basis of overall goodness of fit. However, descriptive fit indices are best viewed as providing information on the extent of a model's *lack of fit*. In other words, although these indices may provide conclusive evidence of a misspecified model, they cannot be used in isolation of other information to support the conclusion of a good-fitting model. Goodness-of-fit indices provide a global descriptive summary of the ability of the model to reproduce the input covariance matrix, but the other two aspects of fit evaluation (localized strain, parameter estimates) provide more specific information about the acceptability and utility of the solution.

Overall Goodness of Fit

Overall goodness of fit in model evaluation is discussed in Chapter 3, where several goodness-of-fit indices are recommended (e.g., SRMR, RMSEA, CFI). As noted in Chapter 3, at least one index from each fit class (absolute, parsimony, comparative) should be considered because each provides different information about the fit of the CFA solution. After ensuring that the model was specified as intended (e.g., verifying model *df* and freely estimated, fixed, and constrained parameters), goodness-of-fit indices are then examined to begin evaluating the acceptability of the model. If these indices are consistent with good model fit, this provides initial (tentative) support for the notion that the model was properly specified. The remaining aspects of model evaluation can be evaluated in this context; for example, is goodness of fit also supported by the lack of local-

ized areas of strain in the solution? However, if indices point to poor fit, subsequent aspects of fit evaluation would be focused on diagnosing the sources of model misspecification (e.g., inspection of modification indices and standardized residuals; see the next section of this chapter). In addition, it would be erroneous to interpret the model's parameter estimates (e.g., size and significance of factor loadings and factor correlations) if the solution fit the data poorly, because misspecified models produce biased parameter estimates. Occasionally, fit indices will provide inconsistent information about the fit of the model. In these instances, greater caution is needed in determining the acceptability of the solution and in detecting the potential sources of misspecification. For example, the SRMR and CFI may suggest that fit is acceptable at an absolute level and in relation to a null solution, but an RMSEA > .08 may indicate a lack of parsimony (i.e., look for the inclusion of freely estimated parameters that are unnecessary).

In the Figure 4.1 example, each of the overall goodness-of-fit indices suggest that the two-factor model does fit these data well: $\chi^2(19) = 13.23$, $p = .83$, SRMR = .019, RMSEA = 0.00 (90% CI = 0.00 − .018), CFit = .99, TLI = 1.007, CFI = 1.00 (Mplus users will obtain a χ^2 value of 13.285; cf. $F_{ML}(N)$, Eq. 3.17, Chapter 3). Although providing initial support for the acceptability of the two-factor model, this tentative judgment must be verified by considering other aspects of the results.

Localized Areas of Strain

A limitation of goodness-of-fit statistics (e.g., SRMR, RMSEA, CFI) is that they provide a *global*, descriptive indication of the ability of the model to reproduce the observed relationships among the indicators in the input matrix. However, in some instances, overall goodness-of-fit indices suggest acceptable fit despite the fact that some relationships among indicators in the sample data have not been reproduced adequately; or alternatively, some model-implied relationships may markedly exceed the associations seen in the data. This outcome is more apt to occur in complex models (e.g., models that entail an input matrix consisting of a large set of indicators) where the sample matrix is reproduced reasonably well on the whole, and the presence of a few poorly reproduced relationships have less impact on the global summary of model fit, as reflected by overall goodness-of-fit statistics. Yet overall goodness-of-fit indices may indicate that a model poorly reproduced the sample matrix. However, these indices do not provide information on the reasons why the model fit the data poorly (various forms of model misspecification are considered in Chapter 5). Two statis-

tics that are frequently used to identify focal areas of misfit in a CFA solution are *residuals* and *modification indices*.

Residuals

As discussed in Chapter 3 (e.g., see Figure 3.8), there are three matrices associated with the typical CFA model: the sample variance–covariance matrix (S), the model-implied (predicted) variance–covariance matrix (Σ), and the residual variance–covariance matrix, which reflects the difference between the sample and model-implied matrices (i.e., residual matrix = $S - \Sigma$). Table 4.2 presents these matrices for the two-factor model depicted in Figure 4.1. Although a correlation matrix and indicator *SD*s were used as input (see Table 4.1), the latent variable software program uses this information to create the sample variance–covariance matrix; for example, the covariance of N1 and N2 = .767(5.7)(5.6) = 24.48. As discussed in Chapter 3, the predicted (model-implied) variance–covariance matrix is generated on the basis of the parameter estimates obtained in the minimization of F_{ML} (or some other fitting function such as weighted least squares). For example, as will be shown later in this chapter, the completely standardized factor loadings of N1 and N2 were estimated to be .8848 and .8485, respectively. The model-implied correlation between these two indicators is .751; that is, $\sigma_{21} = \lambda_{x11}\phi_{11}\lambda_{x21} = .8848(1)(.8485)$. Thus, the predicted covariance of N1 and N2 is 23.97; i.e., .751(5.7)(5.6) (cf. Table 4.2). The fitted residual for N1 and N2 is .518, the difference between their sample and model-implied covariance; using the unrounded values from Table 4.2: 24.4826 − 23.965 = .5176.

Thus, while goodness-of-fit statistics such as the SRMR provide a global summary of the difference between the sample and model-implied matrices (see Chapter 3), the residual matrix provides specific information about how well each variance and covariance was reproduced by the model's parameter estimates. Indeed, there exists one residual for each pair of indicators. However, fitted residuals can be difficult to interpret because they are affected by the raw metric and dispersion of the observed measures. It is particularly difficult to determine whether a fitted (unstandardized) residual is large or small when the units of measurement of the indicators are markedly disparate. This problem is addressed by *standardized residuals*, which are computed by dividing the fitted residuals by their estimated standard errors.[2] Accordingly, standardized residuals are analogous to standard scores in a sampling distribution and can thus be interpreted along the lines of z scores. Stated another way, these values can be concep-

TABLE 4.2. Sample, Model-Implied, Fitted Residual, and Standardized Residual Matrices for the Two-Factor CFA Model of Neuroticism and Extraversion

Sample Variance–Covariance Matrix (S)

	N1	N2	N3	N4	E1	E2	E3	E4
N1	32.4900							
N2	24.4826	31.3600						
N3	26.6669	25.4106	40.9600					
N4	25.2772	23.5570	27.7978	32.4900				
E1	-12.0042	-10.1472	-13.6704	-10.8756	36.0000			
E2	-11.1674	-9.7216	-11.9040	-9.4358	25.1100	38.4400		
E3	-9.6170	-9.2249	-10.8346	-9.6170	21.6828	23.0063	32.4900	
E4	-9.0014	-7.9654	-10.4653	-7.8204	17.9424	20.5890	18.0667	31.3600

Model-Implied Variance–Covariance Matrix (Σ; also referred to as the Fitted or Predicted Matrix)

	N1	N2	N3	N4	E1	E2	E3	E4
N1	32.4900							
N2	23.9650	31.3600						
N3	27.2314	25.6558	40.9600					
N4	25.3548	23.8878	27.1437	32.4900				
E1	-10.5542	-9.9436	-11.2988	-10.5202	36.0000			
E2	-11.3400	-10.6839	-12.1401	-11.3035	24.8700	38.4400		
E3	-9.8715	-9.3003	-10.5680	-9.8397	21.6493	23.2613	32.4900	
E4	-8.5876	-8.0907	-9.1934	-8.5599	18.8335	20.2358	17.6153	31.3600

Fitted Residual Matrix (also referred to as the Unstandardized Residual Matrix)

	N1	N2	N3	N4	E1	E2	E3	E4
N1	0.0000							
N2	0.5177	0.0000						
N3	-0.5645	-0.2452	0.0000					
N4	-0.0776	-0.3309	0.6541	0.0000				
E1	-1.4500	-0.2036	-2.3716	-0.3554	0.0000			
E2	0.1726	0.9623	0.2361	1.8678	0.2400	0.0000		
E3	0.2545	0.0755	-0.2666	0.2227	0.0335	-0.2550	0.0000	
E4	-0.4139	0.1253	-1.2718	0.7395	-0.8911	0.3531	0.4515	0.0000

Standardized Residual Matrix

	N1	N2	N3	N4	E1	E2	E3	E4
N1	- -							
N2	1.7927	- -						
N3	-1.6542	-0.5570	- -					
N4	-0.3476	-1.1170	1.8689	- -				
E1	-1.2191	-0.1624	-1.6409	-0.2970	- -			
E2	0.1515	0.7896	0.1678	1.6279	0.6530	- -		
E3	0.2194	0.0620	-0.1901	0.1909	0.0754	-0.6788	- -	
E4	-0.3126	0.0923	-0.8157	0.5566	-1.4052	0.6405	0.7080	- -

Note. Matrices obtained from LISREL 8.72 printout.

117

tually considered as the number of standard deviations by which the fitted residuals differ from the zero-value residuals that would be associated with a perfectly fitting model.

The standardized residuals of the two-factor CFA model are presented in Table 4.2. As seen in this table, standardized residuals can have either positive or negative values. A *positive* standardized residual suggests that the model's parameters *underestimate* the zero-order relationship between two indicators to some degree. For example, the standardized residual for the N1–N2 relationship was 1.79; the sign of this residual is consistent with the fact that the sample covariance of these indicators (24.48) was larger than the model-implied covariance (23.96). Large, positive standardized residuals may indicate that additional parameters are needed in the model to better account for the covariance between the indicators. Conversely, a *negative* standardized residual suggests that the model's parameters *overestimate* the relationship between two indicators to some extent. As seen in Table 4.2, the standardized residual for the N1–N3 relationship is –1.65 due to the fact that the sample covariance (26.67) was smaller than the covariance predicted by the two-factor model (27.23).

Of course, CFA models specified in applied research will rarely produce standardized residuals that uniformly approximate zero. Thus, the question arises as to how large standardized residuals should be to be considered salient. Because standardized residuals can be roughly interpreted as z scores, the z score values that correspond to conventional statistical significance levels are often employed as practical cutoffs. For instance, researchers may scan for standardized residuals that are equal to or greater than the absolute value of 1.96 because this value corresponds to a statistically significant z score at $p < .05$. In practice, this critical value is often rounded up to 2.00. However, the size of standardized residuals is influenced by sample size. In general, larger Ns are associated with larger standardized residuals because the size of the standard errors of the fitted residuals is often inversely related to sample size. For this reason, some methodologists recommend the use of larger cutoff values (e.g., 2.58, which corresponds to the .01 alpha level; Byrne, 1998). In any case, the researcher should be mindful of the potential impact of sample size when interpreting the salience of the standardized residual. Other information provided by the software output, such as expected parameter change, can often assist in these considerations (see Modification Indices section). While a cutoff of 2.00 or 2.58 provides a general guideline for this aspect of model fit diagnostics, the researcher should especially look for standardized residuals with outlying values. This process is fostered in the LISREL

software by the output of stemleaf and Q plots, which can also be useful for detecting non-normality or nonlinear associations in the indicators, in addition to specification errors.

As seen in Table 4.2, the standardized residuals ranged from –1.65 to 1.87. Using the guidelines just presented, this outcome would be consistent with the conclusion for the absence of localized areas of ill fit in the solution. No relationships among the indicators are substantially under- or overestimated by the model's parameter estimates.

Modification Indices

Another aspect of model evaluation that focuses on specific relationships in the solution is the *modification index* (Sörbom, 1989; referred to as the univariate *Lagrange multiplier* in the EQS program). Modification indices can be computed for each fixed parameter (e.g., parameters that are fixed to zero such as indicator cross-loadings and error covariances) and constrained parameter in the model (e.g., parameters that are constrained to equal magnitudes as in the test of tau equivalence; see Chapter 7). The modification index reflects an *approximation* of how much the overall model χ^2 would decrease if the fixed or constrained parameter was freely estimated. Indeed, if the parameter is freely estimated in a subsequent analysis, the actual decrease in model χ^2 may be somewhat smaller or larger than the value of the modification index. In other words, the modification index is roughly equivalent to the difference in the overall χ^2 between two models, where in one model the parameter is fixed or constrained and in the other model the parameter is freely estimated. Thus, modification indices are analogous to the χ^2 difference (with a single *df*) of *nested models* (nested models were introduced in Chapter 3 and are discussed more fully in subsequent chapters, e.g., Chapter 5 and Chapter 7).

To illustrate this concept, modification indices from the Figure 4.1 solution are presented in Table 4.3. Consider the modification index for the E1 indicator (1.32) listed under the heading "Modification Indices for Lambda-X." This value indicates that if the E1 indicator was freely estimated to cross-load on the Neuroticism latent factor, the overall model χ^2 is estimated to drop by 1.32 units (e.g., from 13.23 to 11.91). If this cross-loading was in fact specified in a subsequent analysis, the model *df* would decrease from 19 to 18 because an additional parameter has been freely estimated (Neuroticism $\rightarrow$ E1). Therefore, this modification index estimates a χ^2 difference (with 1 *df*) of 1.32 if this cross-loading is freely estimated.

TABLE 4.3. Modification Indices and (Completely Standardized) Expected Parameter Change Values for the Two-Factor CFA Model of Neuroticism and Extraversion

Modification Indices for LAMBDA-X

	NEUROT	EXTRAV
N1	– –	0.2723
N2	– –	0.1424
N3	– –	0.9001
N4	– –	1.0623
E1	1.3228	– –
E2	1.1167	– –
E3	0.0143	– –
E4	0.0125	– –

Completely Standardized Expected Change for LAMBDA-X

	NEUROT	EXTRAV
N1	– –	-0.0226
N2	– –	0.0175
N3	– –	-0.0443
N4	– –	0.0449
E1	-0.0614	– –
E2	0.0552	– –
E3	0.0065	– –
E4	-0.0065	– –

No Non-Zero Modification Indices for PHI

Modification Indices for THETA-DELTA

	N1	N2	N3	N4	E1	E2	E3	E4
N1	- -							
N2	3.2137	- -						
N3	2.7363	0.3102	- -					
N4	0.1208	1.2477	3.4929	- -				
E1	0.2541	0.6555	1.0965	0.0000	- -			
E2	0.4976	0.0043	0.1069	1.2708	0.4264	- -		
E3	0.9158	0.2208	0.2526	0.8349	0.0057	0.4608	- -	
E4	0.0162	0.0021	0.5849	0.4783	1.9747	0.4102	0.5013	- -

Completely Standardized Expected Change for THETA-DELTA

	N1	N2	N3	N4	E1	E2	E3	E4
N1	- -							
N2	0.0484	- -						
N3	-0.0444	-0.0147	- -					
N4	-0.0098	-0.0301	0.0501	- -				
E1	-0.0119	0.0206	-0.0269	-0.0001	- -			
E2	-0.0160	-0.0016	0.0081	0.0257	0.0289	- -		
E3	0.0229	-0.0121	0.0131	-0.0220	0.0031	-0.0295	- -	
E4	-0.0034	0.0013	-0.0220	0.0184	-0.0546	0.0252	0.0274	- -

Maximum Modification Index is 3.49 for Element (4, 3) of THETA-DELTA

Note. Matrices obtained from LISREL 8.72 printout. For space reasons, only completely standardized expected parameter change values are presented.

121

In general, a good-fitting model should also produce modification indices that are small in magnitude. Because the modification index can be conceptualized as a χ^2 statistic with 1 *df*, indices of 3.84 or greater (which reflects the critical value of χ^2 at *p* < .05, 1 *df*) suggest that the overall fit of the model could be significantly improved (*p* < .05) if the fixed or constrained parameter was freely estimated (although in practice, this critical value is often rounded up to 4.00; cf. Jaccard & Wan, 1996). Thus, for example, the modification index corresponding to the E1 cross-loading (1.32) would argue against freely estimating this parameter because this respecification would not result in a significant improvement in model fit.

Like overall model χ^2 and standardized residuals, modification indices are sensitive to sample size. For instance, when *N* is very large, a large modification index may suggest the need to add a given parameter despite the fact that the magnitude of the parameter in question, if freely estimated, is rather trivial (e.g., a large modification index associated with a completely standardized cross-loading of .10). To address this problem, latent variable software programs provide expected parameter change (EPC) values for each modification index. The EPC values provided by these programs may be unstandardized, standardized, or completely standardized. Some programs, such as Mplus and LISREL, provide all three forms of EPC values. As the name implies, EPC values provide an estimate of how much the parameter is expected to change in a positive or negative direction if it were freely estimated in a subsequent analysis. Unstandardized EPC values are proportional to the scale of the observed measures in question. Thus, fit diagnostics typically focus on completely standardized EPC values. For example, as shown in Table 4.3, the completely standardized cross-loading of E1 onto Neuroticism is predicted to be –.06, which further supports the contention that warmth (E1) has no meaningful direct relationship with the dimension of Neuroticism. Especially when sample size is large, the size and direction of EPC values should be considered in tandem with modification indices to assist in the determination of whether the respecification is conceptually and statistically viable (cf. Kaplan, 1989, 1990).

Indeed, a key principle of model respecification is that modification indices and standardized residuals should prompt the researcher to relax a fixed or constrained parameter only when there exists a compelling substantive basis for doing so (i.e., freeing the parameter is supported by empirical, conceptual, or practical considerations) and when the expected value (EPC) of the parameter can be clearly interpreted. Especially when

sample size is large, modification indices and standardized residuals will often indicate the presence of parameters that, if freed, would improve the fit of the model. However, such parameters should not be freed with the sole intent of improving model fit. Rather, all model respecifications must be justified on the basis of prior research or theory.

Several studies have highlighted the problems that occur when models are respecified solely on the basis of modification indices or standardized residuals. For example, MacCallum (1986) examined how often and under what conditions *specification searches* (i.e., post hoc revisions to a fitted model as determined by modification indices) are likely to lead to the discovery of the correct population model. This question was examined by generating simulated data for which the "true" model was known, fitting various misspecified models to the data, and then determining whether the specification search (e.g., freeing parameters with the highest modification indices one at a time in sequential order) led to finding the correct model. In many instances, the model revisions implied by the specification searches were incorrect; for example, they led to model specifications that did not correspond closely to the true model. This was particularly apparent when the initial model was markedly misspecified and when sample size was modest (e.g., $N = 100$). In a follow-up to this study, Silvia and MacCallum (1988) found that restricting modifications to those that could be justified on the basis of prior theory greatly improved the success of the specification search.

For these reasons, Jöreskog (1993) has recommended that model modification begin by freely estimating the fixed or constrained parameter with the largest modification index (and EPC) *if this parameter can be interpreted substantively*. If there does not exist a substantive basis for relaxing the parameter with the largest modification index, consider the parameter associated with the second largest modification index, and so on. However, it is important to note that a poorly reproduced relationship may be reflected by high modification indices in more than one matrix (e.g., suggestive of a cross-loading in lambda-X; suggestive of an error covariance in theta-delta). Thus, several high modification indices may be remedied by freeing a single parameter. Because the model-implied covariance matrix (Σ) is dependent on all model parameters, there are countless ways a single model revision can impact Σ. Therefore, revisions of a model should always focus exclusively on parameters justified by prior evidence or theory. This point is further underscored by the fact that some modification indices output by latent variable software programs appear nonsensi-

cal; for instance, a modification index suggesting a direct effect of a disturbance on a latent variable, or a method effect involving the measurement error of a single indicator.

Other potential adverse consequences of atheoretical specification searches are overfitting (i.e., adding unnecessary parameters to the model) and capitalization on chance associations in the sample data (i.e., accounting for weak effects in the data that stem from sampling error and are not apt to be replicated in independent data sets). MacCallum, Roznowski, and Necowitz (1992) have argued against modifying a good-fitting model to achieve even better fit because this practice is likely to be "fitting small idiosyncratic characteristics of the sample" (p. 501). Thus, sampling error, in tandem with the sensitivity of modification indices and standardized residuals to sample size, heightens the risk for model respecifications of a trivial nature. Researchers are perhaps most prone to model overfitting in instances where overall goodness-of-fit indices are on the border of conventional guidelines for acceptable fit. For example, as the result of examining modification indices, the researcher might determine that freeing some correlated errors in a CFA solution would improve a "borderline" value of a given fit index. However, in addition to lacking a substantive basis, this practice may introduce other problems such as biasing other parameters in the model (e.g., the magnitude of the factor loadings) and their standard errors.

Even when revisions to the initial solution can be defended with compelling empirically or theoretically based arguments, it is important to note that by pursuing respecifications of the initial model, one has moved out of a confirmatory analytic framework. Indeed, as the name implies, a "specification search" (MacCallum, 1986) entails an exploratory endeavor to determine the nature of the misfit of the initial, hypothesized model. Accordingly, respecified models should be interpreted with caution. Especially in instances where substantial changes have been made to the initial model, modified solutions should be replicated in independent samples. Or if size of the original sample is sufficiently large, the sample could be split into derivation and holdback samples to cross-validate the revised model.

Unnecessary Parameters

Thus far, the discussion of model modification has focused primarily on the issue of adding new freely estimated parameters to the initial model. Although this is the most common form of model respecification, models

can be revised by eliminating statistically nonsignificant parameters. It was noted earlier that the presence of unnecessary parameters may be reflected by large, negative standardized residuals that indicate that the model is overestimating the observed relationship between a pair of indicators. A similar statistic found in the EQS software program is the *univariate Wald test*. Converse to the modification index (i.e., univariate Lagrange multiplier), the univariate Wald test provides an estimate of how much the overall model χ^2 would increase if a freely estimated parameter were fixed to zero.[3] A nonsignificant Wald test value (e.g., < 3.84 in the case of a single parameter) would indicate that removing the freely estimated parameter (e.g., fixing it to zero) would not result in a significant decrease in model fit. In programs such as LISREL, Mplus, and Amos, the necessity of existing parameters can be evaluated by examining their statistical significance. As discussed in the next section, the statistical significance of a freely estimated parameter is indicated by a test statistic (which can be interpreted as a z statistic) that is calculated by dividing the unstandardized parameter estimate by its standard error. At the .05 alpha level (two-tailed), parameters associated with z values of ±1.96 or greater are statistically significant. Thus, parameters with z values less than 1.96 are statistically nonsignificant and might be considered unnecessary to the solution. The z and χ^2 distributions are closely related—for example, the critical value of χ^2 with 1 *df*, α = .05 is 3.842, which equals the squared critical value (α = .05) of z (i.e., 1.96^2 = 3.842). Accordingly, although the univariate Wald test is not provided by programs other than EQS, it can be calculated from the output of all software packages by squaring the z value associated with the unstandardized parameter estimate; that is, the squared z value provides an estimate of how much model χ^2 would increase if the freed parameter was removed. However, CFA and SEM users should be aware that the results of Wald tests and parameter significance tests are sensitive to how the metric of the latent variable is identified (e.g., selection of marker indicator, fixing variance of the latent variable to 1.0). This issue is discussed in Appendix 4.1.

In structural models, Wald tests suggest that some directional paths among latent variables may not be necessary. In CFA measurement models, nonsignificant parameters or Wald tests may point to a variety of model modifications such as the removal of indicators (e.g., nonsignificant factor loadings) or specific parameters that have been freely estimated in the initial solution (e.g., correlated errors, cross-loadings, factor correlations). The various forms of CFA model misspecification are discussed in Chapter 5.

Interpretability, Size, and Statistical Significance of the Parameter Estimates

Returning to the Figure 4.1 example, recall that each of the overall goodness-of-fit indices suggested that the two-factor model fit the data well. Goodness of model fit was further verified by the absence of large modification indices and standardized residuals (see Tables 4.2 and 4.3), thereby indicating no focal areas of ill fit in the solution. Thus, model evaluation can proceed to inspecting the direction, magnitude, and significance of the parameter estimates—namely, the factor loadings, factor variances and covariance, and indicator errors. These results are presented in Table 4.4.

An initial step in this process is to determine whether the parameter estimates make statistical and substantive sense. From a statistical perspective, the parameter estimates *should not take on out-of-range values* such as completely standardized factor correlations that exceed 1.0, negative factor variances, or negative indicator error variances.[4] In Chapter 3 it was noted that such out-of-range values, which are often referred to as "Heywood cases" (or "offending estimates"), may be indicative of model specification error or problems with the sample or model-implied matrices (e.g., a nonpositive definite matrix, small *N*). Thus, the model and sample data must be viewed with caution to rule out the existence of more serious causes of these outcomes. A classic reference for dealing with improper solutions and nonpositive definite matrices is Wothke (1993). This issue is also discussed at length in Chapter 5. As seen in Table 4.4, all parameter estimates from the Figure 4.1 solution are statistically viable (e.g., no negative variances). Readers will note that the estimate of the factor covariance/correlation is negative (e.g., completely standardized estimate = −.435). However, the sign of this relationship is in accord with theory and the scaling of the latent factors; that is, Neuroticism is inversely related to Extraversion. Moreover, the unstandardized factor loadings of the N1 and E1 are 1.0 because these observed measures were used as marker indicators; their loadings were fixed to 1.0 in order to pass the metric of N1 and E1 onto the latent factors of Neuroticism and Extraversion, respectively. Consequently, the standard errors of these estimates are 0.0, and thus no significance test (i.e., *z* value) is available for the marker indicators (in Table 4.4, note that the S.E. and Est./S.E. estimates for N1 and E1 are 0.0).

From a substantive perspective, the *direction of the parameter estimates* provided in Table 4.4 are in accord with prediction. The N1, N2, N3, and N4 indicators are positively related to the latent construct of Neuroticism,

TABLE 4.4. Parameter Estimates from the Two-Factor CFA Model of Neuroticism and Extraversion

```
MODEL RESULTS
                Estimates   S.E.   Est./S.E.    Std      StdYX
NEUROT BY
  N1              1.0000   0.0000    0.0000    5.0435    0.8848
  N2              0.9421   0.0525   17.9452    4.7517    0.8485
  N3              1.0706   0.0603   17.7754    5.3993    0.8436
  N4              0.9968   0.0517   19.2737    5.0272    0.8820

EXTRAV BY
  E1              1.0000   0.0000    0.0000    4.8111    0.8018
  E2              1.0745   0.0790   13.6089    5.1693    0.8338
  E3              0.9353   0.0725   12.9011    4.4999    0.7895
  E4              0.8137   0.0725   11.2246    3.9146    0.6990

EXTRAV WITH
  NEUROT        -10.5542   1.9313   -5.4647   -0.4350   -0.4350

Variances
  NEUROT         25.4367   2.9175    8.7186    1.0000    1.0000
  EXTRAV         23.1466   3.2068    7.2179    1.0000    1.0000

Residual Variances
  N1              7.0533   0.9123    7.7309    7.0533    0.2171
  N2              8.7816   1.0050    8.7376    8.7816    0.2800
  N3             11.8074   1.3361    8.8374   11.8074    0.2883
  N4              7.2168   0.9217    7.8299    7.2168    0.2221
  E1             12.8534   1.5900    8.0837   12.8534    0.3570
  E2             11.7182   1.6118    7.2702   11.7182    0.3048
  E3             12.2411   1.4670    8.3441   12.2411    0.3768
  E4             16.0359   1.6760    9.5678   16.0359    0.5113

R-SQUARE

  Observed
  Variable    R-Square
  N1            0.7829
  N2            0.7200
  N3            0.7117
  N4            0.7779
  E1            0.6430
  E2            0.6952
  E3            0.6232
  E4            0.4887
```

Note. Estimates, unstandardized parameter estimate; S.E., standard error; Est./S.E., test statistic (z value); Std, standardized parameter estimate; StdYX, completely standardized parameter estimate; Mplus 3.11 output.

and the E1, E2, E3, and E4 indicators are positively related to the latent construct of Extraversion. For example, the unstandardized factor loading for N2 (hostility) is 0.942, which can be interpreted as indicating that a one-unit increase in the latent dimension of Neuroticism is associated with a .942-unit increase in the observed measure of hostility (see Table 4.4, in the "Estimates" column).

The results provided in Table 4.4 also indicate that every freely estimated parameter is *statistically significant*. As noted earlier, statistical significance is determined by dividing the unstandardized parameter estimate by its standard error (see "S.E." column in Table 4.4; for a discussion of how standard errors are calculated in ML estimation, see Kaplan, 2000). Because this ratio can be interpreted as a z score, ± 1.96 would be the critical value at an alpha level of .05 (two-tailed). For example, the z value for the E2 factor loading is 13.6 (i.e., 1.074 / .079 = 13.6; see "Est./S.E." column of Table 4.4), which can thus be interpreted as indicating that the gregariousness indicator loads significantly on the latent factor of Extraversion. In addition to the factor loadings, the factor covariance (Extraversion with Neuroticism), factor variances, and indicator error variances (listed under the "Residual Variances" heading in Table 4.4) all differ significantly from zero (but see Appendix 4.1).[5]

The z tests of the statistical significance of the unstandardized parameter estimates cannot be relied upon to determine the statistical significance of corresponding (completely) standardized estimates (see Bollen, 1989, for discussion). Although not identical, the z test results for standardized and unstandardized parameter estimates are usually close; that is, especially in the case of strong effects, these tests will more times than not yield the same conclusion regarding statistical significance. If significance tests of standardized coefficients are desired, standard errors of these estimates must be estimated (from which z test statistics can be computed). Most of the major latent variable software programs do not readily provide the standard errors of standardized coefficients, although they can be computed using the delta method or by the bootstrapping capabilities of these programs (see Chapter 9; e.g., in version 3.0 and later of Mplus, these standard errors can be computed using the MODEL INDIRECT command).

In addition, it is important to evaluate the *standard errors of the parameter estimates* to determine if their magnitude is appropriate, or problematically too large or too small. Standard errors represent estimates of how much sampling error is operating in the model's parameter estimates (i.e., how closely the model's parameter estimates approximate the true popula-

tion parameters). Stated another way, standard errors provide an estimate of how stable the model's parameter estimates would be if we were able to fit the model repeatedly by taking multiple samples from the population of interest. The 95% confidence interval of an unstandardized parameter estimate can be calculated by adding and subtracting the estimate by the product of 1.96 times the standard error. For example, the 95% confidence interval of the E2 factor loading is .919 to 1.229; that is, 1.074 ±1.96(.079) (see Table 4.4). This is interpreted as indicating we can be 95% certain that the true population value of this parameter is between .919 and 1.229.

Although small standard errors might imply considerable precision in the estimate of the parameter, the significance test of the parameter (i.e., z statistic) cannot be calculated if the standard error approximates zero. Conversely, excessively large standard errors indicate problematically imprecise parameter estimates (i.e., very wide confidence intervals) and hence are associated with low power to detect the parameter as statistically significant from zero. Problematic standard errors may stem from a variety of difficulties, such as a misspecified model, small sample size, or use of non-normal data, an improper estimator, or matrix type, or some combination. Unfortunately, there are no specific guidelines available to assist the researcher in determining if the magnitude of standard errors is problematic in a given data set. This is because the size of standard errors is determined in part by the metric of the indicators and latent variables (cf. Appendix 4.1) and the size of the actual parameter estimate, which vary from data set to data set. However, keeping the metric of the variables in mind, the researcher should be concerned about standard errors that have standout values or approach zero, as well as parameter estimates that appear reasonably large but are not statistically significant.

Assuming the problems related to insufficient sample size and inappropriate standard errors can be ruled out, the researcher must consider parameter estimates in the model that fail to reach statistical significance (further reflected by Wald tests; see the preceding section). For example, a nonsignificant factor loading in a congeneric CFA solution indicates that an observed measure is not related to its purported latent dimension, and would typically suggest that the indicator should be eliminated from the measurement model. In a noncongeneric solution, a nonsignificant crossloading would suggest that this parameter is not important to the model and can be dropped. Likewise, a nonsignificant error covariance suggests that the parameter does not assist in accounting for the relationship between two indicators (beyond the covariance explained by the latent factors). As in model respecifications where parameters are added to the

initial solution, nonsignificant parameters should be deleted from the model one at a time as guided by a conceptual or empirical rationale. A factor variance that does not significantly differ from zero typically signals significant problems in the solution, such as the use of a marker variable that does not have a relationship with the latent factor (another potential consequence of this is empirical underidentification; see Chapter 3), substantial non-normality in the input matrix, or use of a sample that is too small. Whether or not nonsignificant factor covariances (correlations) should be of concern depends on the theoretical context of the CFA solution (e.g., is the lack of overlap among latent factors in accord with prediction?). Error variances are inversely related to the size of their respective factor loadings; that is, the more indicator variance explained by a latent factor, the smaller the error variance will be. Thus, assuming no other problems with the solution (e.g., a leptokurtotic indicator, small N), nonsignificant error variances should not prompt remedial action and in fact may signify that an indicator is very strongly related to its purported latent factor. In applied research, however, indicator error variances almost always differ significantly from zero because an appreciable portion of an indicator's variance is usually not explained by the factor, in tandem with the use of large sample sizes (see below).

The acceptability of parameter estimates should not be determined solely on the basis of their direction and statistical significance. Because CFA is typically conducted in large samples, the analysis is often highly powered to detect rather trivial effects as statistically significant. Thus, it is important not only to demonstrate that the specified model reproduces the relationships in the input data well, but that the resulting parameter estimates are of a *magnitude that is substantively meaningful*. For an illustration of how goodness of model fit may be unrelated to the reasonability of the CFA parameter estimates, see Appendix 4.2. For example, the size of the factor loadings should be considered closely to determine if all indicators can be regarded as reasonable measures of their latent constructs. The issue of what constitutes a sufficiently large parameter estimate varies across empirical contexts. For example, in applied factor analytic research of questionnaires, completely standardized factor loadings of .30 (or .40) and above are commonly used to operationally define a "salient" factor loading or cross-loading. However, such guidelines may be viewed as too liberal in many forms of CFA research, such as construct validation studies where scale composite scores, rather than individual items, are used as indicators.

In CFA models where there are no cross-loading indicators, the completely standardized factor loading can be interpreted as the correlation

between the indicator and the latent factor.[6] Accordingly, squaring the completely standardized factor loading provides the proportion of variance of the indicator that is explained by the latent factor; that is, a *communality* (Eq. 3.5, Chapter 3). For example, the completely standardized loading of N4 (self-consciousness) is estimated to be .882, which can be interpreted as indicating that a standardized score increase in Neuroticism is associated with a .882 standardized score increase in self-consciousness. The squared factor loading is .778, indicating that 77.8% of the variance in the observed measure of self-consciousness is accounted for by the latent factor of Neuroticism. These values are provided in Mplus, as well as in EQS, printouts under the "R-Square" heading (e.g., see Table 4.4); in LISREL and Amos output, these values are provided under the heading "Squared Multiple Correlations." Especially in the context of psychometric research (e.g., evaluating the latent structure of a test instrument), these squared factor loadings can be considered as estimates of the indicator's *reliability*; that is, the proportion of the indicator's variance that is estimated to be "true score" variance (see Chapter 8, for further discussion). Accordingly, the proportion of explained variance in the indicators can be quite useful in formulating conclusions about whether the measures are meaningfully related to their purported latent dimensions.

As noted earlier, the interpretability of the size and statistical significance of factor intercorrelations depends on the specific research context. For instance, in the Figure 4.1 solution, the correlation between the latent factors of Neuroticism and Extraversion was estimated to be –.435. The corresponding factor covariance was statistically significant, $p < .05$, as reflected by an absolute value of z greater than 1.96; that is, $z = –5.46$ (see Table 4.4). The substantive sensibility of this relationship would be evaluated by considering prior evidence and theory that bear on this analysis. In general, the size of the factor correlations in multifactorial CFA solutions should also be interpreted with regard to the *discriminant validity* of the latent constructs. Small, or statistically nonsignificant, factor covariances are usually not considered problematic and are typically retained in the solution (i.e., they provide evidence that the discriminant validity of the factors is good). However, if factor correlations approach 1.0, there is strong evidence to question the notion that the latent factors represent distinct constructs. In applied research, a factor correlation that exceeds .80 or .85 is often used as the criterion to define poor discriminant validity. When two factors are highly overlapping, a common research strategy is to respecify the model by collapsing the dimensions into a single factor and determine whether this modification results in a significant degradation in model fit (see Chapter 5). If the respecified model provides an acceptable

fit to the data, it is usually favored because of its superior parsimony, although substantive considerations must be brought to bear (i.e., consider the implications to the literature that guided the formation of the initial, less parsimonious model).

INTERPRETATION AND CALCULATION OF CFA MODEL PARAMETER ESTIMATES

This section reviews and expands on how the various parameter estimates in the two-factor CFA model of Neuroticism and Extraversion are calculated and interpreted. Figure 4.2 presents this model with the unstandardized and completely standardized parameter estimates. Recall that in an *unstandardized solution*, all parameter estimates (factor loadings, factor variances and covariances, indicator error variances and covariances) are
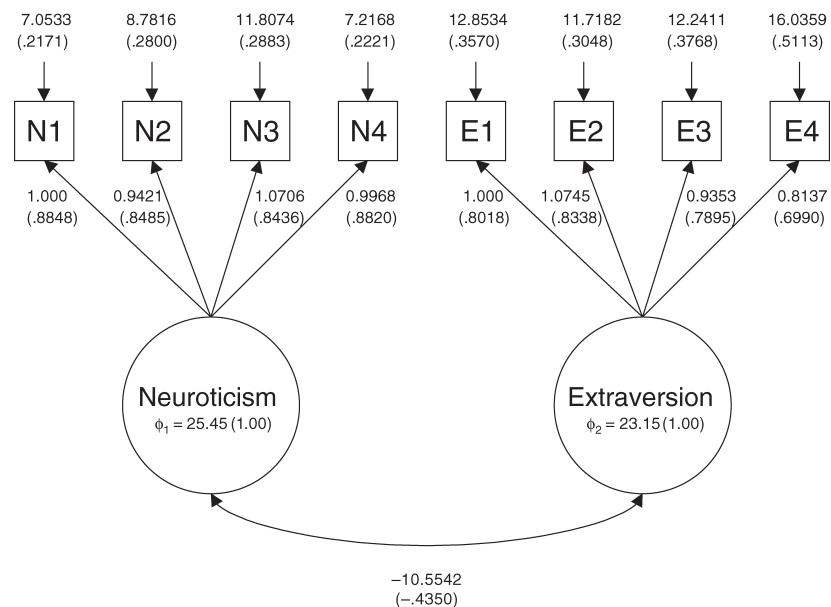
FIGURE 4.2. Unstandardized and completely standardized parameter estimates from the two-factor CFA model of Neuroticism and Extraversion. N1 = anxiety; N2 = hostility; N3 = depression; N4 = self-consciousness; E1 = warmth; E2 = gregariousness; E3 = assertiveness; E4 = positive emotions. Completely standardized parameter estimates are presented in parentheses. All freely estimated unstandardized parameter estimates are statistically significant ($p < .001$).

based on the original metrics of the indicators and latent factors when the marker indicator method is used to define the scales of the latent variables (i.e., the metric of the marker indicator is passed on to the latent factor). For instance, in unstandardized solutions, factor loadings can be interpreted as unstandardized regression coefficients; for example, a one unit increase in Neuroticism is associated with a 1.07 increase in depression (N3; see Figure 4.2).

In *completely standardized solutions*, the metrics of both the indicators and latent factors are standardized (i.e., $M = 0.0$, $SD = 1.0$). Thus, factor loadings in a completely standardized solution can be interpreted as standardized regression coefficients; for example, a one standardized score increase in Neuroticism is associated with a .844 standardized score increase in depression (N3; see Figure 4.2). However, when the measurement model contains no double-loading indicators, a completely standardized factor loading can also be interpreted as the correlation of the indicator with the latent factor because the latent factor is the sole predictor of the indicator. Accordingly, squaring a completely standardized factor loading provides the proportion of variance in the indicator that is explained by the latent factor; for example, Neuroticism accounts for 71.2% of the variance in the indicator (N3) of depression ($.8436^2 = .712$). Thus, it follows that the basic equation of "1.0 minus the squared factor loading" (Eq. 3.6, Chapter 3) provides the proportion of variance in an indicator that is not explained by the latent factor; for example, $1 - .8436^2 = .288$, indicating that 28.8% of the observed variance in the depression indicator (N3) is estimated to be unique or error variance (see Figure 4.2).

By means of simple calculations, it is straightforward to transform a completely standardized CFA solution into an unstandardized solution, and vice versa. The following calculations transform a completely standardized solution into a unstandardized solution. When the marker indicator approach to model identification is employed, the variance of the latent factor is calculated by squaring the completely standardized factor loading of the marker indicator and multiplying this result by the observed variance of the marker indicator. To illustrate, the variance of the latent variable of Neuroticism was estimated to be 25.44 (see Figure 4.2). In Chapter 3 (Eq. 3.10), it was shown that this unstandardized parameter estimate can be calculated by squaring the completely standardized factor loading of N1 (anxiety; $\lambda_{11} = .8848$) and multiplying this result by the observed variance of this indicator ($\sigma_1^2 = 32.49$; see Table 4.2):

$$\phi_{11} = .8848^2(32.49) = 25.44 \qquad (4.1)$$

The factor variance of Extraversion (ϕ_{22} = 23.15; Figure 4.2) can be computed in the same fashion by multiplying the squared completely standardized factor loading of E1 (warmth; λ_{52} = .8018) by the observed variance of E1 (σ^2_5 = 36.00; see Table 4.2):

$$\phi_{22} = .8018^2(36.00) = 23.15 \qquad (4.2)$$

The standard deviations (*SDs*) of the latent factors are calculated by simply taking the square root of the factor variances; that is, Neuroticism *SD* = SQRT(25.44) = 5.04, Extraversion *SD* = SQRT(23.15) = 4.81 (some readers may notice that these values equal the *standardized* factor loadings for the marker indicators N1 and E1; see "Std" column in Table 4.4).

The unstandardized error variances of the indicators can be calculated by multiplying the completely standardized residuals (δ) by the observed variance of the indicators. For example, the error variance of N2 (hostility) was estimated by the two-factor model to be 8.78 (see Figure 4.2). This estimate can be reproduced by multiplying N2's completely standardized residual (δ_2 = .28; Figure 4.2) by its observed variance (σ_2^2 = 31.36; Table 4.2):

$$\delta_2 = .28(31.36) = 8.78 \qquad (4.3)$$

Alternatively, error variances can be computed using the squared completely standardized factor loadings and the observed variances of indicators:

$$\begin{aligned}
\delta_2 &= \sigma_2^2 - \sigma_2^2(\lambda_{21}^2) \qquad (4.4)\\
&= 31.36 - (.8485^2)(31.36)\\
&= 8.78
\end{aligned}$$

Factor covariances can be hand calculated using the *SDs* and correlations of the factors, using the same equation used for calculating covariances among the input indicators:

$$\phi_{21} = r_{21}(SD_1)(SD_2) \qquad (4.5)$$

In the two-factor model, the correlation between Neuroticism and Extraversion was estimated to be –.435 (see Figure 4.2) and the *SDs* of these factors were estimated to be 5.04 and 4.81, respectively. Thus, the factor

covariance of −10.55 (see Figure 4.2) can be calculated by multiplying these three estimates:

$$\phi_{21} = -.435(5.04)(4.81) = -10.55 \tag{4.6}$$

Although not relevant to the current two-factor solution, indicator error covariances can be calculated in the same fashion (i.e., multiply the error correlation by the indicator error *SD*s).

　　As noted earlier, factor loadings are regression coefficients. Therefore, the equations used for calculating unstandardized and standardized regression coefficients can be applied to the transformation of a completely standardized factor loading into an unstandardized loading, and vice versa. In a regression equation involving a single predictor (*x*) and criterion variable (*y*), the unstandardized regression coefficient (*b*) can be computed by the following formula:

$$b = (r_{yx}SD_y) / (SD_x) \tag{4.7}$$

This equation readily generalizes to the CFA analysis of congeneric indicator sets whereby b = unstandardized factor loading, r_{yx} = completely standardized factor loading, SD_y = SD of the indicator, and SD_x = SD of the latent factor. For example, the completely standardized factor loading of the E2 indicator (gregariousness; λ_{62} = .8338) can be converted into an unstandardized loading (λ_{62} = 1.075; see Figure 4.2) as follows:

$$\lambda_{62} = .8338(6.2) / 4.81 = 1.075 \tag{4.8}$$

(The observed *SD* of E2 is 6.2, and the *SD* of the latent factor of Extraversion is 4.81.)

　　It is also straightforward to transform an unstandardized solution into a completely standardized solution. As a function of standardization, the factor variance in a completely standardized solution must be 1.0 (see Table 4.4). A completely standardized indicator error can be readily calculated by dividing the model-estimated error variance (δ) by the observed variance (σ^2) of the indicator. For example, the completely standardized error estimate of .28 for N2 (hostility) is reproduced by dividing its unstandardized error variance (δ_2 = 8.78; Figure 4.2) by its observed variance (σ^2_2 = 31.36; Table 4.2):

$$\delta_2 = 8.78 / 31.36 = .280 \tag{4.9}$$

Again, this value reflects the proportion of the variance in N2 that is not explained by the latent factor of Neuroticism.

A factor intercorrelation can be calculated by dividing a factor covariance by the product of the *SD*s of the factors. Indeed, the factor correlation between Neuroticism and Extraversion was estimated to be −.435. This coefficient can be computed by dividing the factor covariance (ϕ_{21} = −10.55; Figure 4.2) by the product of the *SD*s of Neuroticism and Extraversion (5.04 and 4.81, respectively):

$$\phi_{21} = -10.55 \, / \, [(5.04)(4.81)] = -.435 \qquad (4.10)$$

(This formula generalizes to the calculation of the correlations between indicator errors.) Squaring a factor correlation provides the proportion of overlapping variance between two factors; for example, Neuroticism and Extraversion share 18.9% of their variance ($-.435^2$ = .189).

As was the case for unstandardized factor loadings, an equation from multiple regression can be readily applied to the computation of completely standardized loadings from unstandardized estimates. In multiple regression, an standardized regression coefficient (b^*) can be computed by the following formula:

$$b^* = (bSD_x) \, / \, (SD_y) \qquad (4.11)$$

where *b* is the unstandardized regression coefficient, SD_x is the *SD* of the predictor, and *SDy* is the *SD* of the criterion variable. The parameters of a CFA solution can be substituted in this equation: b^* = completely standardized factor loading, *b* = unstandardized factor loading, SD_x = the *SD* of the latent factor, and SD_y = the *SD* of the indicator. For example, the unstandardized factor loading of the E2 indicator (gregariousness; λ_{62} = 1.075) can be converted into a completely standardized factor loading (λ_{62} = .8338; see Figure 4.2) as follows:

$$\lambda_{62} = 1.075 \, (4.81) \, / \, 6.2 = .8338 \qquad (4.12)$$

(The *SD* of the latent factor of Extraversion is 4.81, the observed *SD* of E2 is 6.2.)

In addition to unstandardized and completely standardized solutions, some popular latent variable software packages (e.g., LISREL, Mplus) provide *standardized solutions*. Table 4.4 provides the standardized parameter estimates under the column heading "Std" (in LISREL, these estimates are

provided under the heading "Standardized Solution"). However, the term "standardized solution" is frequently used in applied and even basic quantitative research in reference to a completely standardized solution. Although standardized solutions are rarely reported in applied SEM and CFA research, the reader must take care to determine whether a completely standardized solution or a standardized solution is being presented by the study's authors as the "standardized solution." Strictly speaking, in a standardized solution, only the latent variables are standardized (i.e., the indicator is expressed by its original metric). For example, the standardized factor loading of N2 (hostility) is estimated to be 4.75. This estimate is interpreted as indicating that for every one standardized score unit increase in Neuroticism, N2 is predicted to increase by 4.75 unstandardized units. This parameter is calculated by multiplying the unstandardized factor loading of N2 (λ_{21} = 0.9421) by the *SD* of the Neuroticism latent factor (5.04):

$$\lambda_{21} = .9421(5.04) = 4.75 \qquad (4.13)$$

Although standardized estimates often do not provide useful additional information for common forms of CFA analyses, they are helpful to the interpretation of models where latent factors are regressed onto categorical background variables (e.g., a dummy code representing Gender: 0 = female, 1 = male), as is done in MIMIC models (see Chapter 7). In these instances, it does not make interpretative sense to use completely standardized estimates of the relationships between the latent factors and the observed score covariates. Consider a scenario where the latent factor of Neuroticism is regressed onto a dummy code for gender (Gender → Neuroticism). Because gender is a binary observed variable (0/1), it makes little substantive sense to consider the completely standardized regressive path that reflects the predicted standardized score change in Neuroticism for every *standardized score increase in gender*. Instead, it is better to interpret the path using estimates that involve the original metric of the dummy code for gender. This standardized parameter estimate would indicate how many standardized scores Neuroticism is predicted to change as a function of an *unstandardized unit increase in gender*; in other words, it reflects the standardized score difference between males and females on the latent dimension of Neuroticism (cf. Cohen's *d*; Cohen, 1988).

Thus far, parameter estimates have been considered under the assumption that marker indicators are used to define the metric of the latent factors. However, as discussed in Chapter 3, the metric of latent variables

can also be defined by fixing the factor variance to a specific value. When this method is used, factor variances are usually fixed to 1.0, consistent with traditional forms of standardization (i.e., $\sigma^2 = SD = 1.0$). When factor variances are fixed to 1.0, many aspects of the CFA solution do not change (relative to the marker indicator approach), including overall model fit indices, fit diagnostic information (e.g., modification indices, standardized residuals), standardized and completely standardized parameter estimates, and the unstandardized estimates of indicator error variances and their corresponding standard errors and significance tests (z statistics). The primary impact that this approach will have is on the unstandardized estimates of the factor loadings, factor variances, and factor covariances. Factor variances will equal 1.0 as the result of the method used to define the latent variable metric, and there will be no corresponding standard errors and z statistics for these fixed parameters (as occurred in the marker indicator approach for indicators N1 and E1; see Table 4.4). Because all factor variances are standardized, the factor covariances will reflect completely standardized estimates (e.g., the factor covariance would equal –.435 in the Table 4.4 solution). In addition, the "unstandardized" factor loadings will take on the same values of the standardized estimates that were produced by a solution using marker indicators; that is, the "unstandardized" and standardized parameter estimates of the CFA model will be identical. Accordingly, the standard errors and z statistics associated with the factor loadings (and factor covariances) will also change.

CFA MODELS WITH SINGLE INDICATORS

Another advantage of CFA over EFA is the ability to include single indicator variables in the analysis. It may be necessary to use single indicators in a CFA or SEM analysis, given the nature of the variable (e.g., sex, age) or the unavailability of multiple measures of a construct. Variables assessed by a single measure should not be interpreted as factors (i.e., a factor accounts for shared variance of multiple observed measures). However, the inclusion of single indicators in CFA is very useful in many situations. For instance, single indicators are used in MIMIC models where latent variables and the indicators of the factors are regressed onto covariates (see Chapter 7). Moreover, CFA is routinely used as a precursor to SEM because the measurement model must be worked out before evaluating a more parsimonious structural solution (cf. Figure 3.2, Chapter 3). If the structural model will contain a mix of single indicators and latent vari-

ables, it is important to include the single indicators in the measurement model. If not, specification error may occur when the single indicators are added to the SEM model; for example, the measure might have poor discriminant validity with a latent variable or might better operate as an indicator of a latent variable (see Chapter 5 for a detailed discussion of specification errors). Including single indicators in the CFA allows the researcher to examine the correlations among the latent variables and single indicators before a structural model is imposed on these relationships. These estimates will provide important information about the viability of the structural model; for example, if a single indicator is weakly correlated with a latent variable, this bodes poorly for a proposed structural model in which the single indicator is specified to have a significant indirect effect on the latent variable.

When observed measures are used as indicators of latent variables, the CFA provides the estimate of their measurement error; that is, amount of variance in the indicator that is not explained by the factor. Obviously, this cannot be done for variables assessed by single indicators. However, an error theory can be invoked for these variables by fixing the unstandardized error of the indicator to some predetermined value. If the error variance is fixed to zero, then it is assumed that the indicator is perfectly reliable (i.e., entirely free of measurement error). This assumption may be reasonable for some variables such as age, height, or weight. If the single indicator is a measure of a construct (e.g., a self-esteem questionnaire), the assumption of perfect reliability is less tenable. Fortunately, measurement error can be readily incorporated into a dimensional indicator by fixing its unstandardized error to some non-zero value, calculated on the basis of the measure's sample variance estimate and known psychometric information (e.g., internal consistency estimate):

$$\delta_x = \mathrm{VAR}(X)(1 - \rho) \qquad (4.14)$$

where $\mathrm{VAR}(X)$ is the sample variance of the single indicator and ρ is the reliability estimate of the indicator. Ideally, ρ would be derived from prior psychometric research that generalizes well to the sample for which the current analysis is being undertaken.[7]

To illustrate, if the sample variance of an indicator is 42.00 and the best estimate of its scale reliability is .85, its error variance would be calculated as

$$\delta_x = (42)(1 - .85) = 6.3 \qquad (4.15)$$

The value of 6.3 could then be used to fix the error variance of this indicator in the CFA model.

A CFA model entailing a combination of latent variables and single indicators is presented in Figure 4.3. In this example, an epidemiological researcher is interested in examining whether dimensions of health status are predictive of various public health outcomes (e.g., quality of life, use of medical services); the example is loosely based on the SF-36 Health Survey (e.g., Ware & Sherbourne, 1992). For the sake of illustration, only a portion of the variable set is included in the example, but in practice, all predictor and outcome variables to be used in the SEM analysis should be
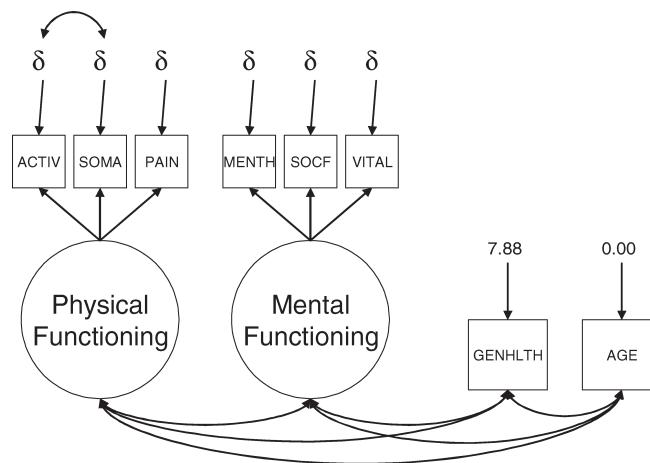
Correlations/Standard Deviations (SD):

	ACTIV	SOMA	PAIN	MENTH	SOCF	VITAL	GENHLTH	AGE
ACTIV	1.000							
SOMAT	0.779	1.000						
PAIN	0.412	0.463	1.000					
MENTH	0.366	0.311	0.285	1.000				
SOCF	0.280	0.273	0.269	0.784	1.000			
VITAL	0.227	0.193	0.218	0.696	0.748	1.000		
GENHLTH	0.417	0.452	0.365	0.539	0.605	0.456	1.000	
AGE	-0.118	-0.122	-0.146	-0.082	-0.073	-0.119	-0.096	1.000
SD:	19.961	15.851	3.270	5.545	9.366	13.838	8.462	11.865

FIGURE 4.3. CFA model of health status containing latent variables and single indicators. ACTIV, physical activity; SOMA, somatic complaints; PAIN, bodily pain; MENTH, mental health; SOCF, social functioning; VITAL, vitality; GENHLTH, general health; N = 500.

included in the measurement model. There are two latent variables, Physical Functioning and Mental Functioning, defined by three observed measures each. A third health status variable, General Well-Being, is assessed by one self-report measure of general health. Prior population-based studies indicate that the best reliability estimate (ρ) of this measure is .89; thus, 11% of the total variance of this measure is estimated to be error (i.e., $1 - .89$). As shown in Figure 4.3, the *SD* of the general health measure is 8.462 and thus its sample variance is 71.61 (i.e., 8.462^2). Because participant age will be an important covariate in the SEM analysis, it is included in the measurement model.

In Table 4.5, Mplus syntax and selected output are presented. The correlation matrix and indicator *SD*s are presented in Figure 4.3, but a raw data file is used as input in this syntax example ($N = 500$ adults). The syntax programming for Physical Functioning and Mental Functioning is conducted in the usual manner. Because two indicators of Physical Functioning are subscales from the same measure, a correlated error is specified; that is, ACTIV WITH SOMA. Although it might be substantively plausible to expect that a method effect exists for the indicator of general health (e.g., it is also a subscale from the measure used to obtain indicators of physical functioning), correlated errors cannot be specified for single indicators because their error variances are fixed (see Chapter 5). Although General Well-Being and Age are not latent variables, they are programmed as such; for example, GWB BY GENHLTH (Table 4.5). The unstandardized factor loading of the single indicator on its "pseudofactor" is fixed to 1.0 (in Mplus, this is a default). Next, the appropriate error variance constraints are imposed. It is assumed that age was measured without measurement error and thus the error variance of this indicator is fixed to zero; that is, AGE@0. Using Equation 4.14, the error variance of the general health indicator (GENHLTH) is calculated to be $(71.61)(1 - .11) = 7.88$; this value is added to the model; that is, GENHLTH@7.88. Programming would follow similar lines in other software packages; for example, LISREL: VA 7.88 TD(7,7); *EQS*: /VARIANCES E7 = 7.88; Amos: sem.Structure "E7 (7.88)."

The Figure 4.3 model fits the data reasonably well, $\chi^2(15) = 45.00$, $p < .001$, SRMR = .028, RMSEA = 0.063, CFit = .137, TLI = 0.971, CFI =.984. To ascertain that the error variance constraints were modeled as intended, one should first inspect the appropriate section of the output; in Mplus, this section is under the heading "Residual Variances." As can be seen in Table 4.5, the unstandardized and completely standardized error variances of GENHLTH are 7.88 and .11, respectively, in accord with the

TABLE 4.5. Mplus Syntax and Selected Results for a Measurement Model of Health Status Involving Latent Variables and Single Indicators

Mplus Syntax

```
TITLE: MPLUS 3.11 CFA MODEL WITH TWO FACTORS, TWO SINGLE INDICATORS
DATA:
      FILE IS CH4SI.DAT;
      FORMAT IS F3,8F3;
VARIABLE:
      NAMES ARE SUBJID ACTIV SOMA PAIN MENTH SOCF VITAL GENHLTH AGE;
      USEV = ACTIV SOMA PAIN MENTH SOCF VITAL GENHLTH AGE;
ANALYSIS: ESTIMATOR=ML;
MODEL:
      PHYSF BY ACTIV SOMA PAIN;
      MENTF BY MENTH SOCF VITAL;
      GWB BY GENHLTH;
      AGEF BY AGE;
      GENHLTH@7.88; AGE@0;
      ACTIV WITH SOMA;
OUTPUT:  SAMPSTAT STANDARDIZED MODINDICES(5);
```

Selected Results

MODEL RESULTS

	Estimates	S.E.	Est./S.E.	Std	StdYX
PHYSF BY					
ACTIV	1.000	0.000	0.000	13.496	0.677
SOMA	0.868	0.049	17.631	11.714	0.740
PAIN	0.150	0.018	8.566	2.027	0.620
MENTF BY					
MENTH	1.000	0.000	0.000	4.705	0.849
SOCF	1.850	0.072	25.787	8.705	0.930
VITAL	2.357	0.109	21.599	11.089	0.802
GWB BY					
GENHLTH	1.000	0.000	0.000	7.974	0.943
AGEF BY					
AGE	1.000	0.000	0.000	11.853	1.000
MENTF WITH					
PHYSF	28.376	4.215	6.732	0.447	0.447
GWB WITH					
PHYSF	69.058	7.881	8.763	0.642	0.642
MENTF	25.179	2.283	11.027	0.671	0.671

(cont.)

TABLE 4.5. *(cont.)*

AGEF WITH					
PHYSF	-30.309	9.033	-3.355	-0.189	-0.189
MENTF	-5.291	2.622	-2.018	-0.095	-0.095
GWB	-9.664	4.502	-2.147	-0.102	-0.102
ACTIV WITH					
SOMA	87.988	17.716	4.967	87.988	0.279
Variances					
PHYSF	182.153	29.368	6.203	1.000	1.000
MENTF	22.141	1.936	11.439	1.000	1.000
GWB	63.582	4.520	14.068	1.000	1.000
AGEF	140.490	8.885	15.811	1.000	1.000
Residual Variances					
ACTIV	215.495	24.525	8.787	215.495	0.542
SOMA	113.527	15.653	7.253	113.527	0.453
PAIN	6.567	0.581	11.311	6.567	0.615
MENTH	8.547	0.751	11.383	8.547	0.279
SOCF	11.781	1.874	6.287	11.781	0.135
VITAL	68.146	5.263	12.948	68.146	0.357
GENHLTH	7.880	0.000	0.000	7.880	0.110
AGE	0.000	0.000	0.000	0.000	0.000

R-SQUARE

Observed Variable	R-Square
ACTIV	0.458
SOMA	0.547
PAIN	0.385
MENTH	0.721
SOCF	0.865
VITAL	0.643
GENHLTH	0.890
AGE	1.000

reliability information used to specify the model (e.g., proportion of error variance = 1 – .89). Similarly, the error variance of age is 0.00 as intended. Observe that factor loadings are estimated by Mplus for the single indicator and their "pseudofactor." These loadings are a byproduct of the error variance constraints. For instance, the completely standardized loading of GENHLTH on the general well-being factor (GWB) is .943. Squaring this loading produces the reliability coefficient used to calculate the error variance of GENHLTH; that is, $.943^2 = .89$ (cf. Eq. 3.5, Chapter 3).

Correlations involving the general well-being factor (GWB) reflect the relationship of the GENHLTH indicator with other variables in the model, adjusting for measurement error; for example, the correlation of GENHLTH with Physical Functioning is .642. The factor loading of age on the Age "factor" is 1.00, reflecting a perfect relationship between the observed measure and its underlying "true" score.

REPORTING A CFA STUDY

Good-fitting CFA solutions are presented in this chapter to illustrate the specification and interpretation of latent variable measurement models. In addition to understanding how to conduct and interpret these analyses properly, applied researchers must be aware of what information should be presented when reporting the results of a CFA study. Indeed, although many excellent guides have been published for reporting SEM and CFA results (e.g., Hoyle & Panter, 1995; McDonald & Ho, 2002; Raykov, Tomer, & Nesselroade, 1991), recent surveys have indicated that applied research articles continue to often omit key aspects of the analyses such as the type of input matrix used, the identifiability and exact specification of the model, and the resulting parameter estimates (MacCallum & Austin, 2000; McDonald & Ho, 2002). Recommended information to include in a CFA research report is listed in Table 4.6. Appendix 4.3 provides a sample write-up of some of this suggested information using the example of the two-factor CFA model of Neuroticism and Extraversion.

Although Table 4.6 and Appendix 4.3 are fairly self-explanatory, a few elaborations are warranted. The parameter estimates from the two-factor model used in this example lent themselves to presentation in a path diagram (Figure 4.2), but many applied CFA models are too complex to be presented in this fashion (e.g., models often contain a large number of indicators and latent factors). Thus, applied CFA findings are frequently presented in tabular formats; specifically, a p (indicator) by m (factor) matrix of factor loadings. While a tabular approach may be preferred for large CFA models (cf. McDonald & Ho, 2002), the researcher must be sure to provide the remaining parameter estimates (e.g., factor and error correlations) in the text, a table note, or a companion figure. Although methodologists have underscored the importance of providing standard errors (and/or confidence intervals) of parameter estimates, this information is rarely reported in CFA research (MacCallum & Austin, 2000;

TABLE 4.6. Information to Report in a CFA Study

Model Specification

- Conceptual/empirical justification for the hypothesized model
- Complete description of the parameter specification of the model
  —List the indicators for each factor
  —Indicate how the metric of the factors was defined (e.g, specify which observed variables were used as marker indicators)
  —Describe all freely estimated, fixed, and constrained parameters (e.g., factor loadings and cross-loadings, random and correlated indicator errors, factor correlations, intercepts and factor means[a])
- Demonstrate that the model is identified (e.g., positive model *df*, scaling of latent variables, absence of empirical underidentification)

Input Data

- Description of sample characteristics, sample size, and sampling method
- Description of the type of data used (e.g., nominal, interval; scale range of indicators)
- Tests of estimator assumptions (e.g., multivariate normality of input indicators)
- Extent and nature of missing data, and the method of missing data management (e.g., direct ML, multiple imputation[b])
- Provide sample correlation matrix and indicator *SD*s (and means, if applicable[a]), or make such data available on request

Model Estimation

- Indicate the software and version used (e.g., LISREL 8.72)
- Indicate the type of data/matrices analyzed (e.g., variance–covariance, tetrachoric correlations/asymptotic covariances[b])
- Indicate the estimator used (e.g., ML, weighted least squares[b]; as justified by properties of the input data)

Model Evaluation

- Overall goodness-of-fit
  —Report model χ^2 along with its *df* and *p* value
  —Report multiple fit indices (e.g., SRMR, RMSEA, CFI) and indicate cutoffs used (e.g., RMSEA ≤ .06); provide confidence intervals, if applicable (e.g., RMSEA)
- Localized areas of ill fit
  —Report strategies used to assess for focal strains in the solution (e.g., modification indices/Lagrange multipliers, standardized residuals, Wald tests, EPC values)
  —Report absence of areas of ill fit (e.g., largest modification index) or indicate the areas of strain in the model (e.g., modification index, EPC value)
- If model is respecified, provide a compelling substantive rationale for the added or removed parameters and clearly document (improvement in) fit of the modified models

(cont.)

TABLE 4.6. *(cont.)*

- Parameter estimates
  —Provide all parameter estimates (e.g., factor loadings, error variances, factor variances), including any nonsignificant estimates
  —Consider the clinical as well as the statistical significance of the parameter estimates (e.g., are all indicators meaningfully related to the factors?)
  —Ideally, include the standard errors or confidence intervals of the parameter estimates
- If necessary (e.g., suitability of N could be questioned), report steps taken to verify the power and precision of the model estimates (e.g., Monte Carlo evaluation using the model estimates as population values[c])

Substantive Conclusions

- Discuss CFA results in regard to their substantive implications, directions for future research, and so on.
- Interpret the findings in context of study limitations (e.g., range and properties of the indicators and sample) and other important considerations (e.g., equivalent CFA models[d])

[a]see Chapter 7; [b]see Chapter 9; [c]see Chapter 10; [d]see Chapter 5.

McDonald & Ho, 2002). These data could also be readily presented in a table or figure note, or in the body of the table or figure itself when only unstandardized estimates are provided.

It should be noted that there is no "gold standard" for how a path diagram should be prepared. Although some constants exist (e.g., representing observed and latent variables by rectangles and circles, respectively; depicting a direct effect by a unidirectional arrow), the reader will encounter many variations in the applied and quantitative literature; for instance, an indicator error variance may be represented by "e," "θ," "δ," "ε," a circle, a double-headed curved arrow, or some combination. No approach is considered to be more correct than another. One particular method is used throughout this book, but the reader is encouraged to peruse other sources to decide which approach best suits his or her own tastes and purposes.

Another consideration is whether to present an unstandardized solution, completely standardized solution, or both (also, as noted in the previous section, standardized estimates may be relevant for some parameters such as in MIMIC models). In applied CFA research, the convention has been to report completely standardized parameters. This may be due in part to the fact that CFA is frequently employed after EFA in the psycho-

metric analysis of test instruments (e.g., multiple-item questionnaires). As noted in Chapter 2, the tradition of EFA is to standardize both the observed and latent variables. Often, neither the observed nor latent variables structure is standardized in the CFA analysis (e.g., a variance–covariance matrix is inputted, a marker indicator is specified). Although the completely standardized CFA solution can be informative (e.g., a completely standardized error indicates the proportion of variance in the indicator that is not accounted for by the latent factor), the unstandardized solution (or both) may be preferred in some instances such as in measurement invariance evaluations where constraints are placed on the unstandardized parameters (Chapter 7), in construct validity studies where the indicators are composite measures with readily interpretable metrics, and in analyses using item parcels (Chapter 9) where information regarding the relative magnitudes of the relationships to the factors has little substantive importance or does not convey information about the original items. As discussed in Chapter 3, there are some CFA and SEM scenarios where exclusive reliance on completely standardized solutions can result in misleading or erroneous conclusions (Bollen, 1989; Willett et al., 1998).

If possible, the sample input data used in the CFA should be published in the research report. In instances where this is not feasible (e.g., a large set of indicators are used), it is helpful for authors to make these data available upon request (or post the data on a website). Inclusion of these data provides a wealth of information (e.g., magnitudes and patterns of relationships among variables) and allows the reader to replicate the study's models and to explore possible conceptually viable alternative models (equivalent or better fitting; see Chapter 5). In general, the sample correlation matrix (accompanied with *SD*s and *M*s, if applicable) should be provided, rather than a variance–covariance matrix. This is because the reader will be able to analyze a variance–covariance matrix that contains less rounding error than the typical variance–covariance matrix published in research reports by creating it directly from the sample correlations and *SD*s. Of course, there are some situations where the CFA analysis cannot be reproduced from a published table of data. Whereas tabled data work fine for ML analysis of data sets that contain no missing data (or where pairwise or listwise deletion was used to manage missing data), this is not the case for analyses that are conducted on raw data (e.g., as in analyses that use direct ML for missing data; see Chapter 9) or that require companion matrices (e.g., asymptotic covariance as well as tetrachoric correlation matrices in WLS-estimated models; see Chapter 9). In these cases, it may

be preferable to make the data available by request or possibly by downloading from the author's or journal's website.

Finally, it should be emphasized that the suggestions provided in Table 4.6 are most germane to a measurement model conducted in a single group and thus must be adapted on the basis of the nature of the particular CFA study. For example, many CFA investigations entail the comparison of a target model to a substantively viable competing solution (see Chapter 5). In such studies, the recommendations listed in Table 4.6 should be extended to the alternative models as well (e.g., also provide conceptual/empirical justification for the competing models). In multiple-group CFA studies (see Chapter 7), it is important to evaluate the CFA solutions separately in each group before conducting the simultaneous analysis. In addition, the multiple-groups analysis typically entails invariance evaluation of the CFA parameters (e.g., are the factor loadings invariant, consistent with the notion that the indicators measure the latent construct in comparable ways in all groups?). Thus, although all of the steps listed in Table 4.6 are relevant, the report of a multiple-groups CFA study is considerably more extensive than the sample write-up provided in Appendix 4.3 (e.g., data normality screening and model fit estimation/evaluation in each group, nested invariance evaluation; see Chapter 7).

SUMMARY

The fundamental concepts and procedures of CFA were illustrated using a full example. As can be seen in this chapter, the proper conduct of CFA requires a series of steps and decisions including the specification of the measurement model (based on prior evidence and theory), selection of a statistical estimator appropriate for the type and distributional properties of the data (e.g., ML), choice of a latent variable software program (e.g., EQS, LISREL), evaluation of the acceptability of the model (e.g., overall goodness of fit, focal areas of strain in the solution, interpretability/strength of parameter estimates), and the interpretation and presentation of results. Although this material was presented in the context of a well-specified measurement model, some of the complications and issues often encountered in applied CFA research were introduced (e.g., potential sources of ill fit, Heywood cases, significance testing of parameter estimates). These issues are considered at much greater length in the next chapter, which focuses on the respecification and comparison of CFA models.

NOTES

1. Although in the case of the Figure 4.1 model, specification of a direct effect between Neuroticism and Extraversion would produce the same fit as a solution that simply allows these factors to be intercorrelated. This is because the potential structural component of the model (e.g., Neuroticism $\rightarrow$ Extraversion, or Extraversion $\rightarrow$ Neuroticism) is just-identified.

2. However, in EQS, the term *standardized residual* is used differently. In EQS, a standardized residual reflects the difference between the observed correlation and the model-implied correlation; for example, for the N1 and N2 relationship in the Figure 4.1 model, the EQS standardized residual is .016 (i.e., .767 − .751).

3. Like Lagrange multipliers, Wald tests can be used as multivariate statistics that estimate the change in model χ^2 if *sets* of freed parameters are fixed.

4. Although completely standardized loadings > 1.0 are generally considered to be Heywood cases, Jöreskog (1999) has demonstrated instances where such estimates are valid (i.e., models that contain double-loading indicators).

5. One-tailed (directional) tests are appropriate for parameter estimates involving variances (i.e., indicator error variances, factor variances) because these parameters cannot have values below zero (z_{crit} = 1.645, α = .05, one-tailed).

6. For indicators that load on more than one factor, factor loadings should be interpreted as partial regression coefficients; for example, a given factor loading for a double-loading indicator would be interpreted as how much the indicator is predicted to change given a unit increase in one factor, holding the other factor constant.

7. The question often arises as to which reliability estimate should be selected for the error variance constraint in situations where the measure in question has an extensive psychometric history or the quality of the extant psychometric evidence is poor. Although qualitative considerations are important (e.g., selection guided by the quality and generalizability of psychometric studies), it is often useful to conduct a sensitivity analysis in which the stability of the results is examined using a range of viable reliability estimates.

Appendix 4.1

Model Identification Affects the Standard Errors of the Parameter Estimates

Chapter 9 discusses the fact that the standard error of the parameter estimates of the CFA solution should not be trusted when the ML estimator is used when the data are non-normal (i.e., an estimator robust to non-normality should be employed). If the standard errors are incorrect, significance testing and confidence interval estimation of the parameter estimates are under-mined. However, even when the assumptions of the fitting function hold, usu-ally a problem still exists in the estimation of standard errors and their corre-sponding z tests, confidence intervals, and Wald tests. Specifically, *standard errors are not invariant to the method used to define the scale of the latent variable*. In other words, the magnitude of the standard error and corresponding z test of a parameter estimate tend to vary based on the selection of the marker indi-cator or when the scale of the latent variable is defined by fixing its variance to 1.0.

To illustrate, the factor covariance from the two-factor model of Neuroti-cism and Extraversion (Figure 4.1) will be used as the example (although the illustration would apply to any free parameter in the solution). The model is estimated three times, using differing approaches to scale the latent factors: (1) using N1 and E1 as the marker indicators (as in Figure 4.2); (2) using N4 and E4 as the marker indicators; and (3) fixing the variances of Neuroticism and Extraversion to 1.0.

It is important to note that the manner in which the scale of the latent variables is identified has no impact on overall goodness of fit. Indeed, the three models produce identical goodness-of-fit indices, for example, $\chi^2(19) = 13.23$, $p = .83$. However, each model produces different standard errors (*SE*) and test statistics (z):

	Factor Covariance Estimate	*SE*	*z*
Model 1 (N1, E1 as markers)	−0.431 (−0.453)	0.079	5.465
Model 2 (N4, E4 as markers)	−0.430 (−0.453)	0.081	5.288
Model 3 ($\phi_{11} = \phi_{22} = 1.0$)	−0.453 (−0.453)	0.059	7.395

(Completely standardized estimates are in parentheses; that is, these estimates are not affected by the scaling approach.)

In this example, the results are consistent in their indication that the null hypothesis is false (i.e., the factor covariance significantly differs from zero). However, this will not always be the case in applied data sets—that is, *different approaches to scaling the latent variable may yield different conclusions regarding the statistical significance of freely estimated parameters*. Because standard errors are used in the computation of confidence intervals and Wald tests, the same problem applies to these statistics as well.

The explanation for why standard errors vary across identification methods is very technical (e.g., *SE*s are determined on the basis of the curvature of fit function when the maximum height is reached; curvature at the maximum differs depending on the identification method). Interested readers are referred to Bollen (1989) and Gonzalez and Griffin (2001) to learn more about this issue. However, a more general question is: *How can one test the statistical significance of CFA parameters if standard errors are not invariant to the latent variable scaling method*? Fortunately, the standard χ^2 difference procedure can be implemented for this purpose.

As noted in Chapter 3, a *nested model* has a subset of the free parameters of another model (often referred to as the *parent model*). Gonzalez and Griffin (2001) note that instead of approximating the value of the fitting function by a measure of curvature that is sensitive to the scaling of the latent variable (i.e., as in typical *SE*s and the corresponding z tests), it is better to evaluate statistical significance by comparing model fit when the parameter of interest is freely estimated (i.e., maximum height) to when the parameter is constrained to the value of the null hypothesis. In the current example, this entails comparing the fit of the model where the factor covariance is freely estimated (the parent model) to the model that constrains the factor variance to zero (the nested model). The resulting χ^2 difference test (also referred to as the *likelihood ratio test*) is not affected by the manner in which the metric of the latent variable is set (cf. Figure 3 and associated text in Gonzalez & Griffin, 2001). Using Model 1 as the parent model (Model P), the χ^2 difference test would be as follows:

	χ^2	df
Model N (ϕ_{21} fixed to 0)	54.198	20
Model P (ϕ_{21} free)	13.232	19
χ^2 difference	40.966	1

As the critical value of χ^2 with $df = 1$ is 3.84 ($\alpha = .05$), it can be concluded that the parent model provides a significantly better fit to the data than the nested model. This is because the difference in model χ^2 follows a χ^2 distribution with one degree of freedom; $\chi^2_{diff}(1) = 40.966$, $p < .001$. More germane to the present discussion is the relationship between the z and χ^2 distributions

that was noted in Chapters 3 and 4: that is, at $df = 1$, the square root of the χ^2 is normally distributed. Thus, in the current example, the z test of the null hypothesis (that the factor covariance of Neuroticism and Extraversion does not differ from zero) can be obtained by taking the square root of the χ^2 difference value:

$$Z = SQRT(40.966) = 6.40 \qquad (p < .001)$$

Observe that the z value of 6.40 does not correspond to any of the z values derived from the three models that were estimated above. Although this example used Model 1 as the parent model, the same result would be obtained using either Model 2 or Model 3 instead.

Although this approach is recommended over the typical z tests provided by software program outputs for determining the statistical significance of parameter estimates, a few limitations should be noted. First, the method is not feasible in situations where the null (nested) model would be under-identified. For instance, we could not employ this method to test the significance of the factor covariance if Neuroticism and Extraversion were measured by two indicators each (cf. Model D, Figure 3.7 in Chapter 3). Second, this procedure is not appropriate when the null hypothesis entails fixing a parameter to a value that is on the border of inadmissibility (e.g., setting a factor variance, indicator error variance, or disturbance to a value of 0). As discussed in later chapters (e.g., Chapter 5), models that contain borderline values do not yield proper χ^2 distributions (cf. Self & Liang, 1987). Third, this approach addresses only significance testing and does not address calculation of confidence intervals (but see Neale & Miller, 1997). There are alternative ways of obtaining standard errors in each of the above three scenarios, but these methods are somewhat complex and cumbersome. Finally, applied users may balk at this approach because performing nested χ^2 tests for each parameter is time-consuming. Gonzalez and Griffin (2001) recommend that, at the very least, researchers perform these tests on the parameters of the model that are most important to the research question (e.g., structural relationships among latent variables).

Goodness of Model Fit Does Not Ensure Meaningful Parameter Estimates

Novice CFA researchers occasionally confuse the notions of goodness of model fit and the meaningfulness of the model's parameter estimates. Even if a model is very successful at reproducing the observed relationships in the input matrix (S), this does not ensure that the latent variables are substantively interrelated or account for meaningful variance in the indicators. Thus, it is just as important to consider the size of the model's parameter estimates as it is to consider the model's goodness of fit when determining the acceptability of the solution. This point is illustrated below using a somewhat exaggerated example.

Keeping with the example presented in Figure 4.1, a researcher has developed a brief questionnaire designed to measure the constructs of neuroticism and extraversion (three items each). Thus, a two-factor model is anticipated in which the first three items (X1–X3) load onto a latent factor of Neuroticism, and the remaining three items (X4–X6) load onto a distinct factor of Extraversion. The questionnaire is administered to 300 participants and the two-factor model is fit to the following sample data (for ease of interpretation, a completely standardized matrix is presented):

```
        X1       X2        X3       X4      X5     X6
X1   1.0000
X2   0.0500   1.0000
X3   0.0600   0.0750    1.0000
X4  -0.0100  -0.0125   -0.0150   1.0000
X5  -0.0120  -0.0150   -0.0180   0.0750 1.0000
X6  -0.0100  -0.0125   -0.0150   0.0875 0.0750 1.0000
```

The following completely standardized solution results from this sample matrix and two-factor model specification:
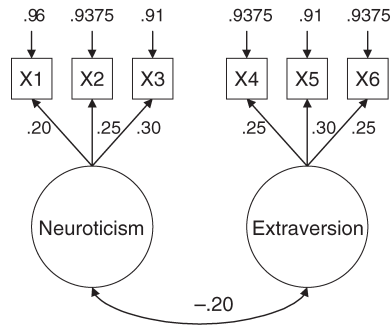
In fact, this solution provides a perfect fit to the data, $\chi^2(8) = 0.00$; for example, multiplying the factor loadings of X1 and X2 perfectly reproduces their observed relationship (.20 × .25 = .05), multiplying the factor loading of X1 and X5 with the factor correlation perfectly reproduces the observed relationship between X1 and X5 (.20 × .30 × −.20 = −.012). However, goodness of fit of the two-factor solution is not determined by the *absolute magnitude* of the sample correlations, but by whether the *differential magnitude* of the correlations can be reproduced by the specified model and its resulting parameter estimates. Accordingly, because the X1–X3 indicators are more strongly intercorrelated with each other than they are with the X4–X6 indicators (and vice versa), the two-factor model specification fits the data well. Nevertheless, the size of the resulting parameter estimates would lead the sensible researcher to reject the two-factor model and the questionnaire on which it is based. For example, the latent factor of Neuroticism accounts for only 4% of the variance in the X1 indicator ($.20^2 = .04$); that is, the greatest part of its variance (96%) is not explained by the latent variable. Indeed, the largest percentage of variance explained in the items is only 9% (X3 and X5). Thus, the questionnaire items are very weakly related to the latent factors and should not be considered to be reasonable indicators of their purported constructs.

Appendix 4.3

Example Report of the Two-Factor CFA Model of Neuroticism and Extraversion

Based on prior evidence and theory bearing on the Big Five model of personality, a two-factor model was specified in which anxiety (N1), hostility (N2), depression (N3), and self-consciousness (N4) loaded onto the latent variable of Neuroticism, and in which warmth (E1), gregariousness (E2), assertiveness (E3), and positive emotions (E4) loaded onto the latent variable of Extraversion. The indicators were subscales of the NEO Personality Inventory and had a range of scores from 0 to 32, with higher scores reflecting higher levels of the personality dimension. Figure 4.1 depicts the complete specification of the two-factor model. Anxiety (N1) and warmth (E1) were used as marker indicators for Neuroticism and Extraversion, respectively. The measurement model contained no double-loading indicators and all measurement error was presumed to be uncorrelated. The latent factors of Neuroticism and Extraversion were permitted to be correlated based on prior evidence of a moderate inverse relationship between these dimensions. Accordingly, the model was overidentified with 19 *df*.

As noted in the Method section, the NEO was administered to 250 college undergraduates who participated in the study for course credit (see Method section for a description of sample demographics). All 250 cases had complete NEO data. Prior to the CFA analysis, the data were evaluated for univariate and multivariate outliers by examining leverage indices for each participant. An outlier was defined as a leverage score that was five times greater than the sample average leverage value. No univariate or multivariate outliers were detected. Normality of the indicators was examined using PRELIS 2.30 (Jöreskog & Sörbom, 1996b). The standardized skewness score was 1.17 and the standardized kurtosis score was 1.32 (*p*s > .05). The joint test of non-normality in terms of skewness and kurtosis was not significant, $\chi^2 = 8.35$, $p = .38$.

Thus, the sample variance–covariance matrix was analyzed using LISREL 8.72, and a maximum likelihood minimization function (sample correlations and *SD*s are provided in Figure 4.1). Goodness of fit was evaluated using the standardized root mean square residual (SRMR), root mean square error of approximation (RMSEA) and its 90% confidence interval (90% CI) and test of close fit (CFit), comparative fit index (CFI), and the Tucker–Lewis index (TLI). Guided by suggestions provided in Hu and Bentler (1999), acceptable

model fit was defined by the following criteria: RMSEA ($\leq$ .06, 90% CI $\leq$ .06, CFit *ns*), SRMR ($\leq$ .08), CFI ($\geq$ .95), and TLI ($\geq$ .95). Multiple indices were used because they provide different information about model fit (i.e., absolute fit, fit adjusting for model parsimony, fit relative to a null model); used together, these indices provide a more conservative and reliable evaluation of the solution.

Each of the overall goodness-of-fit indices suggested that the two-factor model fit the data well, $\chi^2(19) = 13.23$, $p = .83$, SRMR = .019, RMSEA = 0.00 (90% CI = 0.00 – .018; CFit = .99), TLI = 1.007, CFI = 1.00. Inspection of standardized residuals and modification indices indicated no localized points of ill fit in the solution (e.g., largest modification index = 3.49, largest standardized residual = 1.87). Unstandardized and completely standardized parameter estimates from this solution are presented in Figure 4.2 (standard errors of the estimates are provided in Table 4.4). All freely estimated unstandardized parameters were statistically significant (*p*s < .001). Factor loading estimates revealed that the indicators were strongly related to their purported latent factors (range of R^2s = .49–.78), consistent with the position that the NEO scales are reliable indicators of the constructs of neuroticism and extraversion. Moreover, estimates from the two-factor solution indicate a moderate relationship between the dimensions of Neuroticism and Extra-aversion (–.435), in accord with previous evidence and theory.

<Discussion of other implications, limitations, and future directions would follow>

5

CFA Model Revision and Comparison

In Chapter 4, many of the procedures and issues associated with model specification were introduced in the context of a good-fitting, two-factor CFA solution of Neuroticism and Extraversion. In this chapter, these concepts are further illustrated and extended, using initially poor-fitting solutions. Although poor-fitting models are frequently encountered in applied research, SEM sourcebooks rarely deal with this topic in detail because of the numerous potential sources of ill-fit, and the fact that proper model specification and respecification hinges directly on the substantive context of the analysis. Thus, although some examples are provided, readers must adapt these general guidelines and principles to the specific aspects of their own data sets and models. This chapter discusses the fact that model respecification can also be carried out to improve the parsimony and interpretability of a CFA solution. The sources of improper solutions, and ways to deal with them, are also discussed. In addition, the technique of EFA within the CFA framework (E/CFA) is presented as an intermediate step between EFA and CFA. This methodology allows the researcher to explore measurement structures more fully to develop more realistic (i.e., better-fitting) CFA solutions. The chapter concludes with a discussion of equivalent CFA models, an issue that is germane to virtually all measurement models but is largely unrecognized in applied research.

GOALS OF MODEL RESPECIFICATION

Often a CFA model will need to be revised. The most common reason for respecification is to improve the fit of the model. In this case, the results of an initial CFA indicate that one or more of the three major criteria used to

evaluate the acceptability of the model are not satisfied; that is, the model does not fit well on the whole, does not reproduce some indicator relationships well, or does not produce uniformly interpretable parameter estimates (see Chapter 4). Based on fit diagnostic information (e.g., modification indices) and substantive justification, the model is revised and fit to the data again in the hope of improving its goodness of fit. The sources of CFA model misspecification, and the methods of detecting and rectifying those sources, are discussed in the next section of this chapter.

In addition, respecification is often conducted to improve the parsimony and interpretability of the CFA model. Rarely do these forms of respecification improve the fit of the solution; in fact, they may worsen overall fit to some degree. For example, the results of an initial CFA may indicate that some factors have poor discriminant validity—that is, two factors are so highly correlated that the notion that they represent distinct constructs is untenable. Based on this outcome, the model may be respecified by collapsing the highly overlapping factors; that is, the indicators that loaded on separate, overlapping factors are respecified to load on a single factor. Although this respecification may foster the parsimony and interpretability of the measurement model, it will lead to some decrease in goodness of fit relative to the more complex initial solution.

Two other types of model respecification frequently used to improve parsimony are multiple-groups solutions and higher-order factor models. These respecifications are conducted after an initial CFA model has been found to fit the data well. For instance, equality constraints are placed on parameters in multiple-groups CFA solutions (e.g., factor loadings) to determine the equivalence of a measurement model across groups (e.g., do test items show the same relationships to the underlying construct of cognitive ability in men and women?). With the possible exception of parsimony fit indices (e.g., RMSEA, TLI), these constraints will not improve the fit of the model as compared with a baseline solution where the parameters are freely estimated in all groups. In higher-order CFA models, the goal is to reproduce the correlations among the factors of an initial CFA solution with a more parsimonious higher-order factor structure; for example, can the six correlations among the four factors of an initial CFA model be reproduced by a single higher-order factor? As with the previous examples, this respecification cannot improve model fit because the number of parameters in the higher-order factor structure is less than the number of freely estimated factor correlations in the initial CFA model. Multiple-groups solutions and higher-order factor models are discussed in Chapter 7 and Chapter 8, respectively.

SOURCES OF POOR-FITTING CFA SOLUTIONS

In a CFA model, the main potential sources of misspecification are the number of factors (too few or too many), the indicators (e.g., selection of indicators, patterning of indicator–factor loadings), and the error theory (e.g., uncorrelated vs. correlated measurement errors). As discussed in Chapter 4, a misspecified CFA solution may show itself in several aspects of the results: (1) overall goodness-of-fit indices that fall below accepted thresholds (e.g., CFI, TLI < .95); (2) large standardized residuals or modification indices; and (3) unexpectedly large or small parameter estimates or "Heywood cases," which are estimates with out-of-range values. Standardized residuals and modification indices are often useful for determining the particular sources of strain in the solution. However, these statistics are most apt to be helpful when the solution contains minor misspecifications. When the initial model is grossly misspecified, specification searches are not nearly as likely to be successful (MacCallum, 1986).

Number of Factors

In practice, misspecifications resulting from an improper number of factors should occur rarely. When this occurs, it is likely that the researcher has moved into the CFA framework prematurely. CFA hinges on a strong conceptual and empirical basis. Thus, in addition to a compelling substantive justification, CFA model specification is usually supported by prior (but less restrictive) exploratory analyses (i.e., EFA) that have established the appropriate number of factors, and pattern of indicator–factor relationships. Accordingly, gross misspecifications (e.g., specifying too many or too few factors) should be unlikely when the proper groundwork for CFA has been conducted.

However, there are some instances where EFA has the potential to provide misleading information regarding the appropriate number of factors in CFA. This is particularly evident in scenarios where the relationships among indicators are better accounted for by correlated errors than separate factors. A limitation of EFA is that its identification restrictions preclude the specification of correlated indicator errors (see Chapters 2 and 3). Thus, the results of EFA may suggest additional factors when in fact the relationships among some indicators are better explained by correlated errors from *method effects*. A method effect exists when some of the differential covariance among items is due to the measurement approach rather than the substantive latent factors. Method effects may stem from

the general assessment modality (e.g., questionnaires, behavioral observation ratings, clinical interview ratings). Or more specifically, these effects may be due to similarly or reverse-worded assessment items, or other sources such as items with differential proneness to response set, acquiescence, or social desirability.[1] A comprehensive review of the potential sources of method effects is provided in Podsakoff, MacKenzie, Lee, and Podsakoff (2003).

The influence of method effects has been illustrated in factor analyses of the Self-Esteem Scale (SES; Rosenberg, 1965) (e.g., Marsh, 1996; Tomás & Oliver, 1999; Wang, Siegal, Falck, & Carlson, 2001) and the Penn State Worry Questionnaire (PSWQ; Meyer, Miller, Metzger, & Borkovec, 1990) (e.g., Brown, 2003; Fresco, Heimberg, Mennin, & Turk, 2002). Specifically, this research has shown the impact of method effects in questionnaires comprised of some mixture of positively and negatively worded items (the SES contains 4 positively worded items, e.g., "I feel good about myself," and 3 negatively worded items, e.g., "At times I think I am no good at all"; the PSWQ contains 11 items worded in the symptomatic direction, e.g., "I worry all the time," and 5 items worded in the nonsymptomatic direction, e.g., "I never worry about anything"). In other words, the differential covariance among these items is not based on the influence of distinct, substantively important latent dimensions. Rather, this covariation reflects an artifact of response styles associated with the wording of the items (cf. Marsh, 1996).

Nevertheless, studies that conducted EFAs with these measures routinely reported two-factor solutions, with one factor comprised of positively worded items (SES: "Positive Self-Evaluation;" PSWQ: "Absence of Worry") and the second factor comprised of negatively worded items (SES: "Negative Self-Evaluation"; PSWQ: "Worry Engagement"). However, subsequent CFA studies challenged the conceptual utility of these two-factor models (e.g., Brown, 2003; Hazlett-Stevens, Ullman, & Craske, 2004; Marsh, 1996). For example, because the PSWQ was designed to measure the trait of pathological worry, what is the practical and conceptual importance of a dimension of "absence of worry" (Brown, 2003)? In both lines of research, the CFA studies (e.g., Brown, 2003; Marsh, 1996) demonstrated the substantive (i.e., interpretability) and empirical (i.e., goodness of fit) superiority of single-factor solutions where the additional covariance stemming from the directionality of item wording was accounted for by correlated measurement errors (for a detailed discussion of this approach, see Chapter 6). This also highlights the importance of keeping substantive issues firmly in mind when formulating and interpreting EFA

and CFA solutions. In the above examples, two-factor EFA solutions for SES and PSWQ items are apt to provide a better fit (in terms of χ^2, RMSEA, etc.) than a one-factor model. Although the viability of a one-factor model with correlated errors could not be explored within the EFA framework, the acceptability of these two-factor solutions could be challenged on substantive grounds, despite their superior fit.

In the case of a typical measurement model of congeneric indicator sets (a model in which there are no double-loading items and no correlated errors), a CFA solution with too few factors will fail to adequately reproduce the observed relationships among several indicators. For instance, consider the scenario where the two-factor model of Neuroticism and Extraversion in Chapter 4 is specified as a one-factor solution with no correlated indicator errors (ML estimation). The overall fit of this solution is poor, $\chi^2(20) = 373.83$, $p < .001$, SRMR = .187, RMSEA = .306 (90% CI = .283 to .330; CFit $p < .001$), TLI = .71, CFI = .79. As seen in Table 5.1, both the standardized residuals and modification indices indicate that, as a consequence of forcing the four Extraversion indicators (E1–E4) to load on the same latent variable as the four indicators of Neuroticism (N1–N4), the parameter estimates of the solution markedly underestimate the observed relationships among the E1–E4 measures (see Chapter 4 for guidelines on interpreting modification indices and standardized residuals). Specifically, the range of standardized residuals for the Extraversion indicators range from 7.17 to 9.75, and modification indices range from 51.43 to 91.09. It is noteworthy that in the case of a one-factor solution, modification indices can appear only in sections of the results that pertain to indicator measurement errors (e.g., the Theta-Delta matrix in LISREL; see Table 5.1). Although the "true" model in this instance is a two-factor solution (see Figure 4.2, Chapter 4), the fact that fit diagnostics appear only in this fashion might lead novice CFA researchers to conclude that correlated measurement errors are required. Modification indices can point to problems with the model that are not the real source of mis-fit. Again, this underscores the importance of an explicit substantive basis (both conceptual and empirical) for model (re)specification; for example, specifying a model with correlated errors among the Extraversion indicators is not well founded in relation to a model entailing two distinct factors. In this example, the pattern of relationships in the input matrix (see Figure 4.1, Chapter 4), in addition to the aggregation of standardized residuals and modification indices associated with a set of indicators (E1–E4; see Table 5.1), would provide clear empirical evidence against a simple one-factor solution. However, the determination of whether such

TABLE 5.1. Standardized Residuals and Modification for a One-Factor CFA Model of Indicators of Neuroticism and Extraversion

Standardized Residuals

	N1	N2	N3	N4	E1	E2	E3	E4
N1	– –							
N2	2.5394	– –						
N3	-1.1329	-0.0994	– –					
N4	1.2744	0.3203	2.5275	– –				
E1	1.5849	2.5164	0.5238	2.7600	– –			
E2	1.5119	2.0128	1.3017	3.3384	9.7512	– –		
E3	2.4183	1.7140	1.4378	2.1656	8.9485	9.3798	– –	
E4	1.6615	1.8858	0.5442	2.9684	7.1715	8.3930	7.8846	– –

Modification Indices for THETA-DELTA

	N1	N2	N3	N4	E1	E2	E3	E4
N1	– –							
N2	6.4488	– –						
N3	1.2834	0.0099	– –					
N4	1.6243	0.1026	6.3884	– –				
E1	2.5118	6.3322	0.2744	7.6176	– –			
E2	2.2858	4.0513	1.6945	11.1448	95.0860	– –		
E3	5.8479	2.9376	2.0673	4.6898	80.0762	87.9802	– –	
E4	2.7604	3.5562	0.2962	8.8113	51.4310	70.4417	62.1671	– –

Note. Results obtained from LISREL 8.72.

outcomes signify distinct factors versus method effects (or *minor factors*) is not always clear.

The nested model comparison methodology is often used in the applied literature to statistically compare the fit of CFA models that differ in terms of the number of latent factors (e.g., does a two-factor model of Neuroticism and Extraversion provide a better fit to the data than a one-factor model?). Recall that a *nested model* contains a subset of the freed parameters of another solution (Chapters 3 and 4). Consider the following one-factor models involving the same set of 5 indicators (number of elements of the input variance–covariance matrix = 5(6)/2 = 15 = 5 variances, 10 covariances). *Model N* (the nested model) is a simple one-factor solution with no correlated measurement errors; thus, it consists of 10 freed parameters (5 factor loadings, 5 indicator errors, factor variance is fixed to 1.0 to define latent variable metric), and the model has 5 *df* (15 − 10). *Model P* (the parent model) is identical to Model N, with the exception that a correlated error is specified for the fourth and fifth indicators; thus, it consists of 11 freely estimated parameters (5 factor loadings, 5 indicator errors, 1 correlated error), and this model's *df* = 4 (15 − 11). In this scenario, Model N is nested under Model P; if a path is dropped from Model P—the correlated residual for indicators 4 and 5—Model N is formed. In other words, Model N contains a subset of the freed parameters of the parent model, Model P; a nested model will possess a larger number of *df*s than the parent model, in this case the *df* difference is 1, that is, 5 − 4. In Model N, the correlations among all indicator errors are fixed to zero. In Model P, this is not the case because the correlation between the residuals of two indicators is freely estimated.

Models that differ in the number of latent factors are considered nested models. For example, consider the one- and two-factor solutions for the 8 indicators of Neuroticism and Extraversion (*b* = 36 = 8 variances and 28 covariances). The two-factor solution discussed in Chapter 4 contains 17 freely estimated parameters: 6 loadings, 8 indicator errors, 2 factor variances, 1 factor covariance (the loadings of N1 and E1 were fixed as marker indicators). Thus, this two-factor model has *df* = 19 (36 − 17). In contrast, the one-factor solution presented above contains 16 freed parameters: 7 loadings, 8 errors, 1 factor variance (the loading of N1 was fixed as the marker indicator). The one-factor solution has *df* = 20 (36 − 16). Although structurally more discrepant than the Model N versus Model P example described previously, the one-factor model could be construed as nested under the two-factor solution, again with a difference of *df* = 1 (i.e., 20 − 19). This single *df* difference relates to factor correlation in the two-

factor solution. Specifically, the two-factor solution would provide a model fit identical to the one-factor model if the correlation between Neuroticism and Extraversion was fixed to 1.0. Under this specification, the factors would be identical (i.e., 100% shared variance) and thus the indicators would relate to the factors in the same fashion as in the one-factor solution. Accordingly, the one-factor solution can be viewed as a more constrained version of the two-factor model; that is, in the two-factor model the factor correlation is freely estimated, and thus has one less *df*. This principle generalizes to solutions with larger numbers of factors (e.g., a four-factor solution versus a three-factor solution in which two of the factors from the former model are collapsed).

When models are nested, the χ^2 statistic can be used to statistically compare the fit of the solutions. Used in this fashion, χ^2 is often referred to as the χ^2 difference test (χ^2_{diff}) or the nested χ^2 test. If a model is nested under a parent model, the simple difference in the model χ^2s is also distributed as χ^2 in many circumstances (for exceptions to this guideline, see Chapter 9). For example, in the case of the one- versus two-factor model of Neuroticism and Extraversion, the χ^2 difference test would be calculated as follows:

	df	χ^2
One-factor model	20	373.83
Two-factor model	19	13.23 (see Chapter 4)
χ^2 difference (χ^2_{diff})	1	360.60

Thus, χ^2_{diff} (1) = 360.60. In this example, the χ^2 difference test has 1 *df*, which reflects the difference in model *df*s for the one- and two-factor solutions (20 − 19). Therefore, the critical value for χ^2_{diff} in this example is 3.84 (α = .05, *df* = 1). Because the χ^2_{diff} test value exceeds 3.84 (360.60), it would be concluded that the two-factor model provides a significantly better fit to the data than the one-factor model. It is also important that the two-factor model fit the data well. Use of the χ^2 difference test to compare models is not justified when neither solution provides an acceptable fit to the data.

However, some methodologists would argue that models that differ in regard to the number of latent factors are not nested. This is because the restriction required to make the two-factor model equivalent to the one-factor model (or a three-factor model equivalent to a two-factor model, etc.) entails a fixed parameter that is on the border of admissible parameter space: a factor correlation of 1.0 (i.e., a factor correlation > 1.0 constitutes an out-of-range parameter). In other words, nested models that contain

such borderline values (e.g., unit factor correlations, indicator error variances fixed to zero) may not yield proper χ^2 distributions, and thus the χ^2 difference test would also be compromised. Thus, the information criterion indices (e.g., AIC, ECVI; see Eqs. 5.2 and 5.4) could also be considered and reported in instances where the researcher wishes to compare CFA models that differ in the number of factors. These indices are discussed later in this chapter.

Based on the above illustration, the relationship between χ^2_{diff} and modification indices should be more apparent. As discussed in Chapter 4, modification indices represent the predicted decrease in model χ^2 if a fixed or constrained parameter was freely estimated. Accordingly, modification indices (and univariate Lagrange multipliers, cf. EQS) reflect expected χ^2 changes associated with a single *df*. However, nested models often involve solutions that differ by more than a single *df*. For example, a two- and three-factor measurement model of the same data set would differ by two degrees of freedom; that is, the two-factor model contains one factor covariance, the three-factor model contains three factor covariances. Measurement invariance evaluation (Chapter 7) typically entails simultaneously placing constraints on multiple parameters (e.g., constraining factor loadings to equality in two or more groups). Such models are also nested, but differ by more than a single *df*. Moreover, modification indices differ from χ^2_{diff} test values in that they represent an *estimate* of how much model χ^2 will decrease after freeing a fixed or constrained parameter. Quite often, the *actual* difference in model χ^2 (reflected by χ^2_{diff}) produced by a single *df* model modification differs from the estimate of χ^2 change provided by the modification index (this is also true for expected parameter change values; see Chapter 4).

Although applied researchers often compare CFA measurement models that differ in number of latent factors, it is important that a strong conceptual rationale exist for doing so. Occasionally, such analyses are "straw man" comparisons, where the models specified as competing solutions to the hypothesized model have dubious substantive bases and little likelihood of providing an equivalent or superior fit to the data; for example, in the psychometric evaluation of a questionnaire, comparing a three-factor model to a one-factor model, when in fact the measure was designed to be multifactorial and prior EFA research has supported this structure. Whereas the one- versus two-factor models of Neuroticism and Extraversion were presented in this chapter to discuss some of the concepts and issues of model comparison, this criticism would apply; that is, there was no basis for the one-factor model, in view of compelling theory and scien-

tific evidence for the distinctiveness of Neuroticism and Extraversion. Thus, the researcher should provide justification for both the hypothesized and alternative CFA models.

If too many factors have been specified in the CFA model, this is likely to be detected by correlations between factors that approximate ±1.0, and so the latent dimensions have poor *discriminant validity*. In applied research, a factor correlation that equals or exceeds .85 is often used as the cutoff criterion for problematic discriminant validity (cf. guidelines for multicolinearity in regression; Cohen et al., 2003; Tabachnick & Fidell, 2001). When factors overlap to this degree, it may be possible to combine factors to acquire a more parsimonious solution. The goal of such respecification is not to improve the overall fit of the model (e.g., a model with fewer factors entails a smaller number of freely esti- mated parameters), but ideally the fit of the more parsimonious solution will be similar to the initial model, assuming that overall fit of the initial model was satisfactory, except for the excessive correlation between two factors. Again, revisions of this nature require a clear rationale.

Earlier in this chapter, studies on method effects arising from reverse- worded items were briefly discussed. A paper from this literature provides an applied example of excessive overlap between latent factors. Recall that in the Brown (2003) study, a one-factor model of the PSWQ was hypothe- sized in which all 16 items were specified to load on a single dimension of pathological worry, with correlated errors to account for method effects from reversed items. Because a two-factor solution ("Worry Engagement," "Absence of Worry"; cf. Fresco et al., 2002) had prevailed in prior EFA and CFA studies of the PSWQ, it was also fit to the data in Brown (2003) to serve as a competing solution to the hypothesized one-factor model. Results of this CFA indicated poor discriminant validity of the Worry Engagement and Absence of Worry latent dimensions; that is, these factors were highly correlated (e.g., r = .87; Brown, 2003). This finding, along with other evidence and considerations (e.g., superior fit of the hypothe- sized one-factor model, no substantive basis for an "absence of worry" dimension), strongly challenged the acceptability of a two-factor solution.

An alternative to combining two highly correlated factors is to drop one of the factors and its constituent indicators. Although the best reme- dial approach depends on the specific scientific context, dropping a factor might be favored if one of the factors is defined by only a few indicators or has limited variance, or if substantive and practical considerations support this strategy (e.g., there are clear advantages to developing a briefer ques- tionnaire).

Indicators and Factor Loadings

Another potential source of CFA model misspecification is an incorrect designation of the relationships between indicators and the latent factors. This can occur in the following manners (assuming the correct number of factors was specified): (1) the indicator was specified to load on one factor, but actually has salient loadings on two or more factors; (2) the indicator was specified to load on the wrong factor; and (3) the indicator was specified to load on a factor, but actually has no salient relationship to any factor. Depending on the problem, the remedy will be either to respecify the pattern of relationships between the indicator and the factors or to eliminate the indicator from the model altogether.

Fit diagnostics for these forms of misspecifications are presented in context of the CFA measurement model presented in Figure 5.1. The path diagram in this figure represents the latent structure of the population measurement model; a sample variance–covariance matrix for this model (N = 500) was generated using the Monte Carlo utility in Mplus (see Chapter 10). In this example, a researcher has developed a 12-item questionnaire (items are rated on 0–8 scales) designed to assess young adults' motives to consume alcoholic beverages (cf. Cooper, 1994). The measure was intended to assess three facets of this construct (4 items each): (1) coping motives (items 1–4), (2) social motives (items 5–8), and (3) enhancement motives (items 9–12). All of the items were phrased in the positive direction (e.g., item 5: "Because you feel more self-confident about yourself"), with the exception of items 11 and 12, which are reverse-worded and scored. Although the three-factor model was intended to possess congeneric item sets, the "true" model contains one double-loading item (item 4) and one correlated error resulting from the reverse wording (cf. Marsh, 1996). Figure 5.1 presents the completely standardized parameter estimates and overall fit of this model; LISREL syntax and the input matrix are provided in Table 5.2. As seen in this figure, fit indices were consistent with good model fit; this outcome was supported by standardized residuals below 2.00 and modification indices below 4.00. All unstandardized parameter estimates were statistically significant ($ps < .001$) and of a magnitude in accord with expectation.

Now consider the various forms of indicator–factor misspecifications previously mentioned. For instance, a poor-fitting model may result from specification of congeneric indicator sets, when in fact some indicators load on more than one factor. In the current example, the researcher is predicting such a model in which four items each load on the three latent
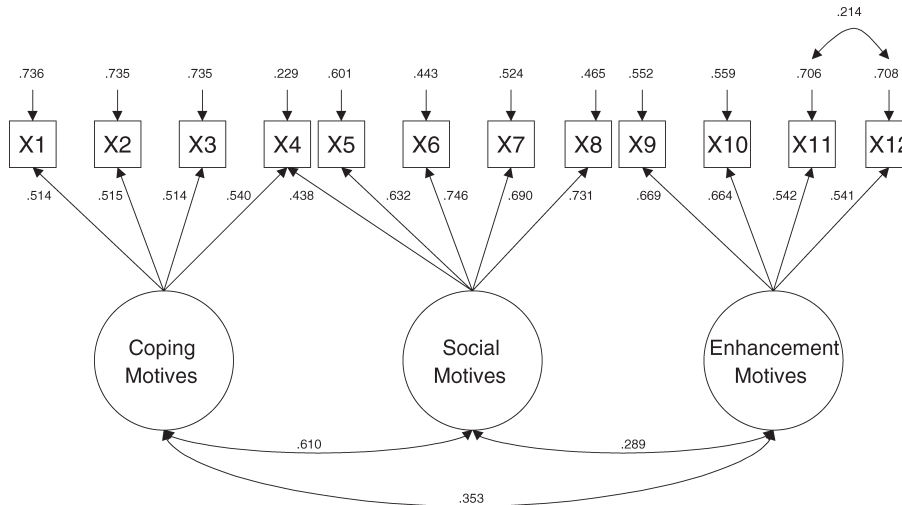
FIGURE 5.1. Completely standardized parameter estimates from the three-factor CFA model of a 12-item Alcohol Drinking Motives Questionnaire (N = 500). Overall fit of the model: $\chi^2(49)$ = 44.865, p = .641, SRMR = .025, RMSEA = 0.00 (90% CI = 0.00 to 0.025; CFit p = 1.00) TLI = 1.002, CFI = 1.00. All freely estimated unstandardized parameter estimates are statistically significant (p < .001).

dimensions of drinking motives (including item 4, which is expected to load only on the Social Motives factor). When this model is fit to the data (the correlated error of items 11 and 12 is included), the following fit indices result: $\chi^2(50)$ = 61.535, p = .127, SRMR = .032, RMSEA = .025 (90% CI = 0.00 to .040; CFit p = .998), TLI = .994, CFI = .996. Note that this solution has one more *df* than the "true" model because the relationship between item 4 and Social Motives has been fixed to zero; hence, this solution is nested under the "true" model.[2]

Selected results of this solution are presented in Table 5.3: standardized residuals, modification indices, completely standardized expected parameter change values (EPCs), and completely standardized estimates of factor loadings and factor correlations. These data exemplify several points made earlier in this chapter and in Chapter 4. For instance, researchers (e.g., MacCallum, 1986) have found that specification searches based on modification indices are more likely to be successful when the model contains only minor misspecifications. Thus, following the recommendation (cf. Jöreskog, 1993) of freeing a fixed or constrained parameter with the largest modification index (provided that this parameter can be interpreted

TABLE 5.2. LISREL Syntax for a Three-Factor Model of a Alcohol Drinking
Motives Questionnaire

```
TITLE THREE FACTOR MODEL FOR DRINKING MOTIVES
DA NI=12 NO=500 MA=CM
LA
X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12
KM
1.000
0.300 1.000
0.229 0.261 1.000
0.411 0.406 0.429 1.000
0.172 0.252 0.218 0.481 1.000
0.214 0.268 0.267 0.579 0.484 1.000
0.200 0.214 0.241 0.543 0.426 0.492 1.000
0.185 0.230 0.185 0.545 0.463 0.548 0.522 1.000
0.134 0.146 0.108 0.186 0.122 0.131 0.108 0.151 1.000
0.134 0.099 0.061 0.223 0.133 0.188 0.105 0.170 0.448 1.000
0.160 0.131 0.158 0.161 0.044 0.124 0.066 0.061 0.370 0.350 1.000
0.087 0.088 0.101 0.198 0.077 0.177 0.128 0.112 0.356 0.359 0.507 1.000
SD
2.06 1.52 1.92 1.41 1.73 1.77 2.49 2.27 2.68 1.75 2.57 2.66
MO NX=12 NK=3 PH=SY,FR LX=FU,FR TD=SY,FR
LK
Coping Social Enhance
PA LX
0 0 0
1 0 0
1 0 0
1 1 0    !double loading item
0 0 0
0 1 0
0 1 0
0 1 0
0 0 0
0 0 1
0 0 1
0 0 1
VA 1.0 LX(1,1) LX(5,2) LX(9,3)
PA TD
1
0 1
0 0 1
0 0 0 1
0 0 0 0 1
0 0 0 0 0 1
0 0 0 0 0 0 1
0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 0 1 1   !correlated residual
PA PH
1
1 1
1 1 1
OU ME=ML RS MI SC AD=OFF IT=100 ND=4
```

TABLE 5.3. Selected Results for the Drinking Motives Three-Factor Model When a Double-Loading Item Has Not Been Specified

Standardized Residuals

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
X1	--											
X2	3.22	--										
X3	1.01	1.89	--									
X4	-0.17	-2.65	-0.26	--								
X5	-1.36	0.96	-0.29	-0.08	--							
X6	-1.44	0.29	-0.05	0.80	0.60	--						
X7	-1.16	-0.79	-0.21	1.24	-0.53	-1.57	--					
X8	-2.15	-0.75	-2.56	-0.82	0.07	0.21	1.20	--				
X9	0.94	1.22	0.19	-1.14	-0.03	-0.45	-0.77	0.29	--			
X10	0.94	0.04	-1.00	0.39	0.27	1.34	-0.88	0.86	0.36	--		
X11	1.99	1.27	1.87	-0.35	-1.43	0.19	-1.12	-1.46	0.86	-0.64	--	
X12	0.21	0.22	0.47	0.81	-0.60	1.64	0.51	-0.09	-0.34	-0.22	--	--

Modification Indices for LAMBDA-X

	Coping	Social	Enhance
X1	--	6.63	1.89
X2	--	0.03	0.89
X3	--	1.92	0.02
X4	--	**18.04**	2.16
X5	0.03	--	0.19
X6	0.44	--	1.37
X7	0.81	--	1.06
X8	1.94	--	0.01
X9	0.76	0.35	--
X10	0.18	0.53	--
X11	0.16	1.53	--
X12	0.77	1.16	--

Completely Standardized Expected Change for LAMBDA-X

	Coping	Social	Enhance
X1	- -	-0.27	0.07
X2	- -	-0.02	0.05
X3	- -	-0.15	0.01
X4	- -	1.00	-0.10
X5	-0.01	- -	-0.02
X6	0.06	- -	0.05
X7	0.08	- -	-0.05
X8	-0.13	- -	0.01
X9	-0.04	-0.03	- -
X10	0.02	0.04	- -
X11	-0.02	-0.06	- -
X12	0.04	0.05	- -

Modification Indices for THETA-DELTA

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
X1	- -											
X2	10.37	- -										
X3	1.02	3.58	- -									
X4	0.03	7.04	0.07	- -								
X5	0.45	1.45	0.03	0.10	- -							
X6	0.47	0.11	0.23	0.08	0.37	- -						
X7	0.28	1.12	0.01	2.22	0.28	2.45	- -					
X8	1.53	0.36	5.21	0.14	0.00	0.04	1.43	- -				
X9	0.50	2.18	0.27	2.59	0.45	1.28	0.01	0.93	- -			
X10	0.04	0.70	3.39	0.49	0.10	0.83	1.71	0.68	0.13	- -		
X11	4.98	2.06	5.92	1.37	0.72	0.12	0.41	1.58	0.67	0.24	- -	
X12	1.93	0.87	0.32	1.14	0.74	1.53	0.68	0.23	0.26	0.00	- -	- -

(cont.)

171

TABLE 5.3. *(cont.)*

Completely Standardized Solution

LAMBDA-X

	Coping	Social	Enhance
X1	0.4311	--	--
X2	0.4349	--	--
X3	0.4502	--	--
X4	0.9548	--	--
X5	--	0.6333	--
X6	--	0.7478	--
X7	--	0.6895	--
X8	--	0.7291	--
X9	--	--	0.6663
X10	--	--	0.6688
X11	--	--	0.5370
X12	--	--	0.5414

PHI

	Coping	Social	Enhance
Coping	1.0000		
Social	0.7978	1.0000	
Enhance	0.3348	0.2917	1.0000

substantively), the correct model can be readily obtained. As seen in Table 5.3, the fixed parameter with by far the highest modification index is λ_{42} (18.04), corresponding to a double-loading of item 4 with the Social Motives factor (cf. Figure 5.1).

Second, these results demonstrate that the acceptability of the model should not be based solely on indices of overall model fit, although this practice is somewhat common in the literature. Note that all descriptive fit indices were consistent with good model fit. However, these global indices masked the fact that at least one relationship was not well represented by the model (i.e., the double-loading of item 4).

Third, the modification index (18.04) and completely standardized EPC values (1.00) do not correspond exactly to the actual change in model χ^2 and parameter estimate, respectively, when the relationship between item 4 and the Social Motives factor is freely estimated. As seen in Figure 5.1, the completely standardized parameter estimate of this double-loading is .438. This revision produces a significant improvement in model fit, $\chi^2_{\text{diff}}(1) = 16.67$, $p < .01$ (i.e., $61.535 - 44.865 = 16.67$), but the χ^2 difference is slightly smaller than the modification index associated with this parameter (18.04). Again, this underscores the fact that modification indices and EPC values are approximations of model change if a fixed or constrained parameter is freed.

Fourth, aspects of these results illustrate that parameter estimates should not be interpreted when the model is poor fitting.[3] For instance, it can be seen in Table 5.3 that some of the consequences of failing to specify the item 4 cross-loading include an inflated estimate of the factor correlation between Coping Motives and Social Motives (.798), an inflated estimate of the loading of item 4 on Coping Motives (.955), and underestimates of the loadings of items 1–3 on Coping Motives. Because item 4 was not specified to load on Social Motives, its moderate relationships with the Social Motives indicators (items 5–8, rs = .48 to .58; see input correlation matrix in Table 5.2) had to be reproduced primarily through its factor loading with Coping Motives (λ_{41} = .955) and the factor correlation of Coping and Social Motives (ϕ_{21} = .798); e.g., the model-predicted relationship of X4 and X5 (completely standardized) = $\lambda_{41}\phi_{21}\lambda_{52}$ = .955(.798)(.633) = .482 (observed r of X4 and X5 was .481; see Table 5.2). Thus, as the result of the iterative process of minimizing F_{ML}, these parameters (λ_{41}, ϕ_{21}) were inflated (and λ_{11}, λ_{21}, and λ_{31} were underestimated) in order to acquire a model that best reproduced the observed relationships ($S - \Sigma$). This also exemplifies why Heywood cases (offending estimates) may arise from a misspecified solution—through the iterative process, the parame-

ters may take on out-of-range values to minimize F_{ML}. In this example, a solution with Heywood cases can be produced by fitting the Figure 5.1 model, by specifying X7 to load on Coping Motives instead of Social Motives. Also see in Table 5.3 that the standardized residuals are not clearly diagnostic of the source of misspecification in this instance; for instance, these values suggest that the relationships between item 4 and the Social Motives indicators have been reproduced adequately. This is because the ML iterations produced parameter estimates that were able to reproduce most relationships reasonably well. Focal areas of strains are nonetheless evident; e.g., some relationships are overestimated (e.g., standardized residual for X3, X8 = –2.56), and one relationship is underestimated (standardized residual for X1, X2 = 3.22), each relating to strains in the model for reproducing the relationship of the Coping Motives and Social Motives indicators. In addition to emphasizing the need to examine multiple aspects of fit (i.e., overall fit, standardized residuals, modification indices, parameter estimate values), this illustrates that respecifying one parameter may successfully eliminate what seem to be multiple strains, because all relationships in the model are interdependent.

Moreover, an indicator–factor relationship may be misspecified when an indicator loads on the wrong factor. Table 5.4 presents selected results of an analysis of the drinking motives model in which item 12 was specified to load on the Social Motives factor (its relationship to Enhancement Motives was fixed to 0, but its correlated error with item 11 was freely estimated). As in the prior example, overall fit statistics suggest acceptable model fit, $\chi^2(49) = 126.337$, $p < .001$, SRMR = .063, RMSEA = .054 (90% CI = 0.042 to 0.066; CFit $p = .302$), TLI = .954, CFI = .971, based on the guidelines recommended by Hu and Bentler (1999). However, as shown in Table 5.4, modification indices and standardized residuals indicate localized points of strain in the solution. As before, simply freeing the fixed parameter ($\lambda_{12,3}$) with the largest modification index (77.01) would result in the proper measurement model. This misspecification is evident in other ways, such as (1) large standardized residuals indicating that the observed relationships among the Enhancement Motives indicators (e.g., item 12 with items 9–11) are being underestimated by the model parameter estimates; and (2) a factor loading of item 12 with Social Motives ($\lambda_{12,2}$ = .20) that, while statistically significant ($z = 4.46$, not shown in Table 5.4), is well below conventional guidelines for a "salient" indicator–factor relationship. Note that another impact of this misspecification is an elevated model estimate of the correlated error of items 11 and 12 ($\delta_{12,11}$ = .40 vs. $\delta_{12,11}$ = .21 in the properly specified solution). This inflated estimate

reflects the attempts of the iterations to reproduce the observed correlation between items 11 and 12 (r = .507), although the model still underestimates this relationship (standardized residual = 7.37). In the correct solution, the sample correlation is reproduced by the sum of the product of the item 11 and item 12 factor loadings [$\lambda_{11,3}\lambda_{12,3}$ = .542 (.541) = .293] and the correlated error of these indicators; that is, .293 + .214 = .507. Because the misspecified model cannot use the product of the item 11 and item 12 factor loadings (because item 12 loads on a different factor than item 11), the solution must rely more on the correlated error ($\delta_{12,11}$) to reproduce this observed relationship. A very small portion of this relationship is also estimated in the solution by $\lambda_{12,2}\phi_{32}\lambda_{11,3}$.

Unlike some previous examples, the Figure 5.1 model and the current model are not nested. Both models entail 29 freely estimated parameters (df = 49), and thus one does not contain a subset of the freed parameters of the other. Therefore, the χ^2_{diff} test cannot be used to statistically compare these two solutions. In this scenario, a strategy that can be used to compare solutions is to qualitatively evaluate each with regard to the three major aspects of model acceptability: overall goodness of fit, focal areas of ill fit, and interpretability/strength of parameter estimates. For instance, if one model satisfies each of these criteria and the other does not, the former model would be favored.

In addition, methodologists have developed procedures for using χ^2 in the comparison of non-nested models. Two popular methods are the Akaike Information Criterion (AIC; Akaike, 1987) and the Expected Cross-Validation Index (ECVI; Browne & Cudeck, 1989). These indices are closely related in that they both take into account model fit (as reflected by χ^2) and model complexity/parsimony (as reflected by model df or the number of freely estimated parameters; cf. the RMSEA). The ECVI also incorporates sample size—specifically, a greater penalty function for fitting a nonparsimonious model in a smaller sample.

Latent variable software programs differ in their computation of the AIC. For instance, in EQS, the AIC is computed as follows:

$$AIC = \chi^2 - 2df \qquad (5.1)$$

Thus, for the Figure 5.1 model, AIC would be –53.13 (i.e., 44.8654 – 98). In LISREL and Amos, AIC is computed as

$$AIC = \chi^2 + 2a \qquad (5.2)$$

TABLE 5.4. Selected Results for the Drinking Motives Three-Factor Model When Item 12 Has Been Specified to Load on the Wrong Factor

Standardized Residuals

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
X1	—											
X2	1.32	—										
X3	-1.49	-0.24	—									
X4	-0.36	-1.13	1.89	—								
X5	-0.76	1.57	0.60	-0.21	—							
X6	-0.68	1.08	1.07	0.37	0.77	—						
X7	-0.50	-0.09	0.77	0.94	-0.37	-1.47	—					
X8	-1.42	0.01	-1.42	-1.30	0.20	0.23	1.22	—				
X9	0.35	0.68	-0.38	-1.61	-0.40	-1.01	-1.26	-0.20	—			
X10	0.39	-0.60	-1.64	0.21	-0.03	0.97	-1.28	0.49	1.54	—		
X11	2.35	1.62	2.30	1.15	-0.77	0.96	-0.40	-0.66	6.07	5.18	8.78	
X12	0.55	0.57	0.88	1.48	-1.54	0.85	-0.41	-1.26	7.32	7.40	7.37	0.41

Modification Indices for LAMBDA-X

	Coping	Social	Enhance
X1	—	1.98	1.03
X2	—	0.67	0.29
X3	—	0.34	0.28
X4	—	—	0.64
X5	0.13	—	0.09
X6	0.78	—	0.09
X7	0.18	—	2.60
X8	3.61	—	0.00
X9	0.98	1.21	—
X10	0.34	0.67	—
X11	4.59	0.17	—
X12	1.11	—	**77.01**

Completely Standardized Expected Change for LAMBDA-X

	Coping	Social	Enhance
X1	--	-0.12	0.06
X2	--	0.07	0.03
X3	--	0.05	-0.03
X4	--	--	-0.04
X5	0.03	--	-0.02
X6	0.06	--	0.01
X7	0.03	--	-0.08
X8	-0.14	--	0.00
X9	-0.07	-0.07	--
X10	-0.04	0.05	--
X11	0.12	0.02	--
X12	0.08	--	**0.53**

Modification Indices for THETA-DELTA

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
X1	--											
X2	1.73	--										
X3	2.23	0.06	--									
X4	0.81	3.16	1.56	--								
X5	0.26	2.44	0.17	0.39	--							
X6	0.15	0.60	0.83	0.03	0.60	--						
X7	0.01	0.43	0.37	1.15	0.14	2.15	--					
X8	0.37	0.02	2.95	0.12	0.04	0.05	1.48	--				
X9	0.03	1.05	0.02	1.78	0.03	1.31	0.05	0.26	--			
X10	0.01	1.19	3.84	0.69	0.07	0.72	2.30	0.07	2.37	--		
X11	2.77	0.76	3.97	0.20	0.30	0.66	0.04	0.68	0.00	1.34	--	
X12	0.59	0.38	0.53	0.36	1.85	0.12	0.20	1.15	20.96	22.72	--	--

(*cont.*)

177

TABLE 5.4. *(cont.)*

Completely Standardized Solution

LAMBDA-X

	Coping	Social	Enhance
X1	0.5163	– –	– –
X2	0.5178	– –	– –
X3	0.5158	– –	– –
X4	0.5301	0.4468	– –
X5	– –	0.6297	– –
X6	– –	0.7471	– –
X7	– –	0.6886	– –
X8	– –	0.7293	– –
X9	– –	– –	0.6683
X10	– –	– –	0.6599
X11	– –	– –	0.3816
X12	– –	0.2048	– –

```
                    PHI

             Coping      Social     Enhance
             --------    --------   --------
Coping       1.0000
Social       0.6084      1.0000
Enhance      0.3524      0.3222     1.0000


THETA-DELTA

          X1       X2       X3       X4       X5       X6       X7       X8       X9      X10      X11      X12
          --       --       --       --       --       --       --       --       --      ---      ---      ---
X1       0.73
X2        --      0.73
X3        --       --      0.73
X4        --       --       --      0.23
X5        --       --       --       --      0.60
X6        --       --       --       --       --      0.44
X7        --       --       --       --       --       --      0.53
X8        --       --       --       --       --       --       --      0.47
X9        --       --       --       --       --       --       --       --      0.55
X10       --       --       --       --       --       --       --       --       --      0.56
X11       --       --       --       --       --       --       --       --       --       --      0.85
X12       --       --       --       --       --       --       --       --       --       --      0.40     0.96
```

179

where a is the number of freely estimated parameters in the model (in Figure 5.1, $a = 29$). Using this equation, AIC would equal 102.87 (44.87 + 58). Mplus output presents the AIC as

$$AIC = -2(\text{loglikelihood}) + 2a \qquad (5.3)$$

For a demonstration of the equivalence of these three equations, see Kaplan (2000).

Although Mplus and EQS do not provide the ECVI (the ECVI is calculated in LISREL and Amos), its approximation is straightforward:

$$ECVI = (\chi^2/n) + 2(a/n) \qquad (5.4)$$

where $n = N - 1$ and a is the number of freely estimated parameters in the model. Thus, the ECVI for the Figure 5.1 solution is .21; that is, (44.8654/499) + 2(29/499) = .09 + .12 = .21. For the current misspecified model involving the item 12 loading, AIC and ECVI are 184.34 and .37, respectively (AIC calculated using Eq. 5.2).

Whereas the AIC and ECVI can be used in tandem with χ^2_{diff} for the comparison of nested solutions, these indices are more often considered in the evaluation of competing, non-nested models. Generally, models with the lowest AIC and ECVI values are judged to fit the data better in relation to alternative solutions (regardless of the method by which AIC is calculated). From this standpoint, the Figure 5.1 solution would be favored over the current model because it is associated with lower AIC and ECVI values (although in the present case, this could be determined simply by comparing model χ^2s because the two solutions do not differ in df). It must be emphasized that, unlike χ^2_{diff}, the AIC and ECVI do not provide a *statistical* comparison of competing models. Rather, these indices foster the comparison of the overall fit of models, adjusting for the complexity of each.[4]

Another possible problematic outcome of a CFA solution is that an indicator does not load on any latent factor. This problem is readily diagnosed by results showing that the indicator has a nonsignificant or nonsalient loading on the conjectured factor, as well as modification indices (and EPC values) suggesting that the fit of the model could not be improved by allowing the indicator to load on a different factor. This conclusion should be further supported by inspection of standardized residuals (2.00) and sample correlations that point to the fact that the indicator is weakly related to other indicators in the model. Unlike the prior exam-

ples, this scenario does not substantially degrade the fit of the model (assuming that the model is well specified otherwise). Thus, although the proper remedial action is to eliminate the problematic indicator, the overall fit of the model will usually not improve. In addition, the revised model is not nested with the initial solution because the input variance–covariance matrix has changed.

Although most of this section has described various ways a model might be respecified retaining the original set of indicators, it is often the case that a much better fitting solution can be obtained by simply dropping bad indicators from the model. For example, an indicator may be associated with several large modification indices and standardized residuals, reflecting that the indicator is rather nonspecific in that it evidences similar relationships to all latent variables in the solution. Simply dropping this indicator from the model will eliminate multiple strains in the solution.

Correlated Errors

A CFA solution can also be misspecified with respect to the relationships among the indicator error variances. When no correlated errors (error covariances) are specified, the researcher is asserting that all of the covariation among indicators that load on a given factor is due to that latent dimension and that all measurement error is random. Correlated errors between indicators are specified on the basis of the notion that some of the covariance in the indicators not explained by the latent variable is due to another exogenous common cause; that is, some of the shared variance is due to the latent factor, some of the shared variance is due to an outside cause. It is also possible that correlated errors exist for indicators that load on separate latent factors. In this case, most of the shared variance may be due to an outside cause (some of the observed covariation may also be reproduced by the product of the indicators' factor loadings and the factor correlation). Occasionally, variables measured by single indicators are included in a CFA model (see Chapters 4 and 7). Because such models would be underidentified, it is not possible to correlate the measurement errors of single indicators with the errors of other indicators in the solution.

As discussed earlier in this chapter, in the case of the analysis of multiple questionnaire items, correlated errors may arise from items that are very similarly worded, reverse-worded, or differentially prone to social desirability, and so forth. In CFA construct validation studies, correlated

errors may be needed to account for method covariance, such as in the analysis of indicators collected from different assessment modalities (e.g., self-report, behavioral observation, interview rating; cf. the multitrait–multimethod approach, Chapter 6).

Unnecessary correlated errors can be readily detected by results indicating their statistical or clinical nonsignificance (e.g., z values below 1.96, or very small parameter estimates that reflect trivial shared variance of the errors). The next step is simply to refit the model by fixing these covariances to zero and verify that this respecification does not result in a significant decrease in model fit. The χ^2_{diff} test can be used in this situation (but see below). The more common difficulty is the failure to include salient correlated errors in the solution. As in the prior examples, the omission of these parameters is typically manifested by large standardized residuals, modification indices, and EPC values.

Table 5.5 presents selected results of the drinking motives CFA solution where the correlated error between items 11 and 12 has not been specified (all other aspects of the solution were properly specified). Although the overall fit of the model is good, for example, $\chi^2(50) = 69.30$, $p = .037$, CFI = .99, RMSEA = .029, SRMR = .031, standardized residuals (5.10) and modification indices ($\delta_{12,11} = 26.02$) indicate that the relationship between these items has not been adequately reproduced by the model's parameter estimates. The need for a correlated error for these items is further evidenced by a rather large completely standardized expected change value ($\delta_{12,11} = .25$). Because this correlated error can be defended substantively (i.e., item 11 and item 12 are the only reverse-worded items in this questionnaire), this parameter is freed in the re-specified solution. This modification significantly improves model fit, $\chi^2_{\text{diff}}(1) = 24.43$, $p < .001$, and the completely standardized estimate of this correlated error is .214 (see Figure 5.1).

As can be seen in Table 5.5, consequences of this misspecification include higher factor loadings for items 11 and 12, and lower loadings for items 9 and 10, on the Enhancement Motives factor. The factor loadings of items 11 and 12 were inflated in the iterative process because a considerable portion of the observed correlation between these indicators could not be reproduced by the correlated error; for example, $\lambda_{x11,3} = .665$ versus .542 in the Table 5.5 and Figure 5.1 solutions, respectively. However, in attempt to avoid marked overestimation of the relationships between items 9 and 10 with items 11 and 12, the magnitude of the item 9 and item 10 factor loadings is attenuated; for example, $\lambda_{x10,3} = .595$ versus .664 in the Table 5.5 and Figure 5.1 solutions, respectively. In other words, because

TABLE 5.5. Selected Results for the Drinking Motives Three-Factor Model without the Correlated Error of Item 11 and Item 12

Standardized Residuals

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
X1	--											
X2	1.40	--										
X3	-1.43	-0.19	--									
X4	-0.34	-1.07	1.61	--								
X5	-0.76	1.58	0.56	-0.22	--							
X6	-0.65	1.11	1.06	0.56	0.71	--						
X7	-0.50	-0.09	0.74	0.95	-0.51	-1.55	--					
X8	-1.42	0.02	-1.45	-1.27	0.07	0.18	1.08	--				
X9	0.67	0.97	-0.02	0.08	0.56	0.34	-0.06	0.97	--			
X10	0.70	-0.21	-1.20	1.36	0.88	2.00	-0.10	1.54	4.48	--		
X11	1.09	0.30	1.02	-1.51	-1.84	-0.22	-1.59	-2.01	-1.91	-2.76	--	
X12	-0.86	-0.85	-0.50	-0.14	-0.93	1.39	0.20	-0.48	-2.69	-2.12	5.10	--

Modification Indices for LAMBDA-X

	Coping	Social	Enhance
X1	--	1.94	0.31
X2	--	0.72	0.00
X3	--	0.30	0.04
X4	--	--	0.11
X5	0.08	--	0.46
X6	1.41	--	2.70
X7	0.24	--	0.47
X8	3.66	--	0.09
X9	0.24	0.31	--
X10	0.56	3.16	--
X11	0.12	4.92	--
X12	0.60	0.02	--

Completely Standardized Expected Change for LAMBDA-X

	Coping	Social	Enhance
X1	--	-0.12	0.03
X2	--	0.07	0.00
X3	--	0.05	-0.01
X4	--	--	-0.01
X5	0.02	--	-0.03
X6	0.09	--	0.07
X7	0.04	--	-0.03
X8	-0.14	--	-0.01
X9	0.03	0.03	--
X10	0.04	0.08	--
X11	-0.02	-0.10	--
X12	-0.04	0.01	--

(cont.)

TABLE 5.5. *(cont.)*

Modification Indices for THETA-DELTA

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
X1	- -											
X2	1.96	- -										
X3	2.04	0.04	- -									
X4	0.77	3.02	1.09	- -								
X5	0.24	2.50	0.17	0.34	- -							
X6	0.19	0.51	0.68	0.03	0.50	- -						
X7	0.02	0.47	0.31	1.05	0.26	2.40	- -					
X8	0.36	0.01	3.07	0.07	0.00	0.03	1.17	- -				
X9	0.15	1.29	0.02	0.54	0.85	0.98	0.02	1.80	- -			
X10	0.10	0.47	3.82	1.13	0.29	0.52	1.59	1.26	20.06	- -		
X11	2.02	0.34	3.49	1.31	1.02	0.34	0.18	1.79	3.65	7.60	- -	
X12	1.91	0.97	0.08	0.32	1.22	1.52	0.58	0.59	7.21	4.50	**26.02**	- -

Completely Standardized Expected Change for THETA-DELTA

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
X1	- -											
X2	0.06	- -										
X3	-0.06	-0.01	- -									
X4	0.03	-0.07	0.04	- -								
X5	-0.02	0.05	0.01	-0.02	- -							
X6	-0.01	0.02	0.03	0.00	0.02	- -						
X7	0.00	-0.02	0.02	0.03	-0.02	-0.05	- -					
X8	-0.02	0.00	-0.05	-0.01	0.00	0.01	0.03	- -				
X9	0.01	0.04	-0.01	-0.02	0.03	-0.03	0.00	0.04	- -			
X10	0.01	-0.02	-0.07	0.03	0.02	0.02	-0.04	0.03	0.18	- -		
X11	0.05	0.02	0.06	-0.03	-0.03	0.02	-0.01	-0.04	-0.08	-0.12	- -	
X12	-0.05	-0.03	-0.01	0.01	-0.03	0.03	0.02	-0.02	-0.12	-0.09	**0.25**	- -

Completely Standardized Solution

        LAMBDA-X

                 Coping   Social   Enhance
                 ------   ------   -------
X1   0.5137   --   --
X2   0.5149   --   --
X3   0.5160   --   --
X4   0.5380   0.4390   --
X5   --   0.6319   --
X6   --   0.7465   --
X7   --   0.6906   --
X8   --   0.7309   --
X9   --   --   0.6026
X10   --   --   0.5947
X11   --   --   0.6649
X12   --   --   0.6629

        PHI

                 Coping   Social   Enhance
                 ------   ------   -------
Coping   1.0000
Social   0.6096   1.0000
Enhance   0.3500   0.2645   1.0000

        THETA-DELTA

          X1   X2   X3   X4   X5   X6
        ------   ------   ------   ------   ------   ------
          0.74   0.73   0.73   0.23   0.60   0.44

          X7   X8   X9   X10   X11   X12
        ------   ------   ------   ------   ------   ------
          0.52   0.47   0.64   0.65   0.56   0.56

185

the factor loadings for items 11 and 12 had to be increased to better approximate the observed relationship of these indicators, the factor loadings of items 9 and 10 were lowered in the iterations to better approximate the relationships of item 9 with items 11 and 12, and item 10 with items 11 and 12. However, these relationships were still overestimated by the parameter estimates (e.g., standardized residual for items 10 and 11 = –2.76, Table 5.5), and the reduction in the magnitude of the item 9 and item 10 factor loadings resulted in a model underestimate of the observed relationship of these indicators (i.e., standardized residual for items 9 and 10 = 4.48, Table 5.5). This again illustrates how correcting one misspecified parameter ($\delta_{12,11}$) may resolve several strains in the solution.

Because of the large sample sizes typically involved in CFA, the researcher will often encounter "borderline" modification indices (e.g., larger than 3.84, but not of particularly strong magnitude) that suggest that the fit of the model could be improved if correlated errors were added to the model. As with any type of parameter specification in CFA, correlated errors must be supported by a substantive rationale and should not be freely estimated simply to improve model fit. In addition, the magnitude of EPC values should also contribute to the decision about whether to free these parameters. As discussed in Chapter 4, the researcher should resist any temptation to use borderline modification indices to overfit the model. These trivial additional estimates usually have minimal impact on the key parameters of the CFA solution (e.g., factor loadings) and are apt to be highly unstable (i.e., reflect sampling error rather than an important relationship; cf. MacCallum, 1986).

It is also important to be consistent in the decision rules used to specify correlated errors; that is, if there is a plausible reason for correlating the errors of two indicators, then all pairs of indicators for which this reasoning applies should also be specified with correlated errors. For instance, if it is believed that method effects exist for questionnaire items that are reverse worded (e.g., Marsh, 1996), correlated errors should be freely estimated for all such indicators, not just a subset of them. Similarly, if the errors of indicators X1 and X2 and indicators X2 and X3 are correlated for the same reason, then the errors of X1 and X3 should also be estimated. Earlier in this chapter where the one- versus two-factor solution of Neuroticism and Extraversion was considered, it was shown that the patterning of standardized residuals and δ modification indices may suggest the existence of a distinct factor (see Table 5.1). However, in some situations, such as in the analysis of questionnaires with reverse-worded items (e.g., Brown, 2003; Marsh, 1996), this patterning may not reflect important

latent dimensions but rather the impact of substantively irrelevant method effects. Theoretical considerations must strongly be brought to bear in this determination, as this decision may have far-reaching implications for the future measurement and conceptualization of the construct.

Improper Solutions and Nonpositive Definite Matrices

A measurement model should not be deemed acceptable if the solution contains one or more parameter estimates that have out-of-range values. As noted earlier, such estimates are usually referred to as *Heywood cases* or *offending estimates*. The most common type of Heywood case in a CFA model is a negative error variance. Moreover, a completely standardized factor loading with a value greater than 1.0 is problematic if the CFA consists of congeneric indicator sets (i.e., each indicator loads on one factor only). In CFA models with indicators that load on more than one factor (and the factors are specified to be intercorrelated), the factor loadings of such indicators are regression coefficients, *not* correlations between the indicators and the factors. A completely standardized factor loading above 1.0 may be admissible in such models, although this result might be indicative of multicollinearity in the sample data (see below). This section discusses the various sources of improper solutions and their remedies.

A necessary condition for obtaining a proper CFA solution is that both the input variance–covariance matrix and the model-implied variance–covariance matrix are *positive definite*. Appendix 3.3 (Chapter 3) introduced the concept of positive definiteness in context of the calculation of determinants and F_{ML}. As noted in Appendix 3.3, a determinant is a single number (*scalar*) that conveys the amount of nonredundant variance in a matrix (i.e., the extent to which variables in the matrix are free to vary). When a determinant equals 0, the matrix is said to be *singular*, meaning that one or more rows or columns in the matrix are linearly dependent on other rows and columns. An example of a singular matrix is one comprised of three test scores: Subscale A, Subscale B, and Total Score (sum of Subscale A and Subscale B). The resulting matrix is singular because the third variable (Total Score) is redundant; that is, it is linearly dependent on the other two variables and thus has no freedom to vary. A singular matrix is problematic because it has no inverse. Consequently, multivariate statistics that require the inverse of a matrix cannot be computed (e.g., F_{ML}; cf. Appendix 3.3). Singularity is one reason why a matrix will not be positive definite. For the condition of positive definiteness to hold, the input and model-implied matrices and every principal submatrix

of these matrices must have determinants greater than zero ("principal submatrices" are all possible subsets of the original matrix created by removing variables from the original matrix).[5]

The condition of positive definiteness can be evaluated by submitting the variance–covariance matrix in question to principal components analysis (PCA; see Chapter 2). PCA will produce as many eigenvalues as the number of variables (p) in the input matrix. If all eigenvalues are greater than zero, the matrix is positive definite. If one or more eigenvalues are less than zero, the matrix is *indefinite*. The term *semidefinite* is used in reference to matrices that produce at least one eigenvalue that equals zero, but no negative eigenvalues (Wothke, 1993).

There are several potential causes of a nonpositive definite input variance–covariance matrix. As noted above, this problem may stem from high multicollinearities or linear dependencies in the sample data. This problem is usually addressed by eliminating collinear variables from the input matrix or combining them (cf. parceling, Chapter 9). Often, a nonpositive definite input matrix is due to a minor data entry problem, such as typographical errors in preparing the input matrix (e.g., a negative or zero sample variance, a correlation > 1.0) or errors in reading the data into the analysis (e.g., formatting errors in the syntax file). Large amounts of missing data, in tandem with use of an inferior approach to missing data management, can create a nonpositive definite input matrix. In Chapter 9, it is demonstrated that the range of possible values that a correlation (or covariance) may possess is dependent on all other relationships in the input matrix. For example, if $r_{x,z}$ = .80 and $r_{y,z}$ = .80, then $r_{x,y}$ must not be below .28 (see Eq. 9.1, Chapter 9) or this submatrix would not be positive definite. Input matrices created from complete data of a large sample are usually positive definite. However, pairwise deletion of missing data can cause definiteness problems (out-of-bound correlations) because the input matrix is computed on different subsets of the sample. Listwise deletion can produce a nonpositive definite matrix by decreasing sample size (see below) or other problems (e.g., creating constants in the sample data that have zero variance). Here, the best solution would be to use a state-of-the-art approach to missing data (i.e., direct ML, multiple imputation; see Chapter 9).

There are several other problems that may lead to improper solutions (i.e., nonpositive definite model matrices, Heywood cases). Perhaps the most common cause is a misspecified model: improper solutions frequently occur when the specified model is very different from models that the data would support. Structurally and empirically underidentified mod-

els will also lead to nonconverging or improper solutions (e.g., see Figure 3.7, Chapter 3; "Model Identification Revisited" section of this chapter). In these situations, it is often possible to revise the model using the fit diagnostic procedures described in this chapter. If the model is grossly misspecified, the researcher may need to move back into a purely exploratory analytic framework (i.e., EFA) to revamp the measurement model.[6] Bad starting values can be the root of improper solutions (see Chapter 3). However, with the possible exception of very complex models, starting values are rarely the cause of improper solutions in today's latent variable software programs, which have become very sophisticated in the automatic generation of such numbers.

Problems often arise with the use of small samples. For instance, an input matrix may not be positive definite owing to sampling error. Small samples often work in concert with other problems to create model estimation difficulties. For instance, small samples are more prone to the influence of outliers (i.e., cases with aberrant values on one or more variables). Outliers can cause collinearities and non-normality in the sample data and can lead to Heywood cases such as negative variance estimates (e.g., an indicator error less than zero). Moreover, some non-normal theory estimators (e.g., weighted least squares) perform very poorly with small or moderate-size samples because their associated weight matrices cannot be inverted (see Chapter 9). Anderson and Gerbing (1984) reported how sample size may interact with other aspects of a model to affect the risk for improper solutions. Specifically, these authors found that the risk of negative variance estimates is highest in small samples when there are only two or three indicators per latent variable and when the communalities of the indicators are low (see also Chen, Bollen, Paxton, Curran, & Kirby, 2001). As discussed in Chapter 10, each of these aspects also contributes to statistical power. In addition to reducing the risk of empirical underidentification, having multiple indicators per factor also decreases the likelihood of improper solutions. If an improper solution is caused by these issues, additional cases or data must be obtained; for example, collect a larger sample, or obtain a larger set of indicators with stronger relationships to the latent variables.

An estimator appropriate to the sample data must be used. For example, if the indicators are binary (e.g., Yes/No items), it would be inappropriate to conduct the CFA on a matrix of Pearson correlations or covariances using the ML estimator. This analysis would provide incorrect results. The proper procedures for conducting CFA with non-normal and categorical data are presented in Chapter 9.

In addition, the risk of nonconvergence and improper solutions is positively related to model complexity (i.e., models that have a large number of freely estimated parameters). A good example of an overparameterized model is the correlated methods CFA of multitrait–multimethod data (see Figure 6.1, Chapter 6). As discussed in Chapter 6, these models produce improper solutions most of the time. In other types of models, it may be possible to rectify the inadmissibility problem by removing some freely estimated parameters (e.g., if substantively justified, drop nonsignificant parameters or place other restrictions on the model, such as equality constraints; see Chapter 7). However, the estimation problems associated with model complexity are magnified by small sample size.

Some examples of nonpositive definite matrices and improper solutions are presented in Figure 5.2. In Model A, the input matrix is semidefinite because of a linear dependency in the sample data (i.e., Y1 is the sum of the other indicators). The fact that this matrix is not positive definite is upheld by the results of a PCA, which shows that one eigenvalue equals zero. Model B analyzes an indefinite input matrix; that is, one of its associated eigenvalues is negative. In this example, the correlation of Y1–Y2 cannot be –.20, given the relationships of Y1 and Y2 with other indicators in the sample data. This problem may have been produced by a data entry error or the use of pairwise deletion as the missing data strategy. Depending on the latent variable software program, Models A and B either will not be executed (i.e., the program will issue the warning that the matrix is not positive definite and will stop) or will produce a factor loading above 1.0 and a negative indicator error variance.

In Model C, the input matrix is positive definite: PCA indicates that all eigenvalues are positive. However, the two-factor measurement model is misspecified. The proposed model will not support the pattern of relationships in the sample data; for instance, Y1 and Y2 are specified as indicators of η_1, yet Y1 is much more strongly related to Y3 than Y2. In an effort to reproduce the sample relationships, given the model that is specified, the ML estimation process is "forced to push" one or more parameter estimates out of the range of admissibility. In this case, ML estimation produces an indefinite factor correlation matrix; that is, the correlation of η_1 and η_2 is estimated to be 1.368. When the model is properly specified (Model D), the results indicate a good-fitting model, for example, χ^2 (1, N = 500) = 2.78, p = .096, and reasonable parameter estimates; for example, factor correlation = .610, range of factor loadings = 0.778–0.867.

In practice, the problem of improper solutions is often circumvented by a "quick fix" method. For example, a negative indicator error variance

Model A: Semidefinite Input Matrix
        (Linear dependency: Y1 is the sum of Y2, Y3, and Y4)

Input Matrix   Eigenvalues

```
          Y1        Y2        Y3        Y4
Y1      1.0000                                           2.877
Y2      0.8035    1.0000                                 0.689
Y3      0.6303    0.4379    1.0000                       0.435
Y4      0.8993    0.5572    0.3395    1.0000             0.000
```

Consequence: factor loading > 1.0, negative error variance
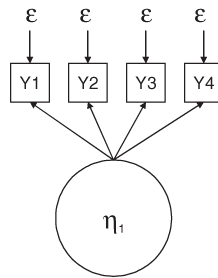
Model B: Indefinite Input Matrix
        (Correlation of Y1–Y2 is out of possible range given other relationships)

Input Matrix   Eigenvalues

```
          Y1        Y2        Y3        Y4
Y1       1.0000                                          2.432
Y2      -0.2000    1.0000                                1.209
Y3       0.7000    0.7000    1.0000                      0.513
Y4       0.4500    0.6000    0.5000    1.0000           -0.155
```

Consequence: factor loading > 1.0, negative error variance
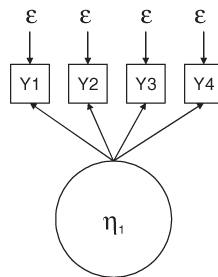
(cont.)

FIGURE 5.2. Examples of nonpositive definite matrices and improper solutions.

Model C: Positive Definite Input Matrix, Indefinite Model-Implied Matrix
        (Specification problem: incorrect specification of indicator–factor relationships)

Input Matrix   Eigenvalues

	Y1	Y2	Y3	Y4	
Y1	1.0000				2.501
Y2	0.4000	1.0000			0.851
Y3	0.7000	0.4000	1.0000		0.360
Y4	0.4500	0.6500	0.4000	1.0000	0.289

Consequence: Factor correlation > 1.0
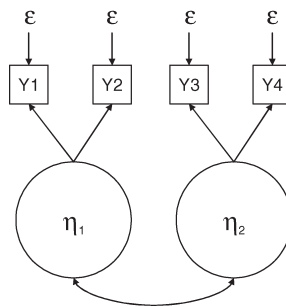
Model D: Positive Definite Input and Model-Implied Matrices
        (Correctly specified model)

Input Matrix   Eigenvalues

	Y1	Y2	Y3	Y4	
Y1	1.0000				2.501
Y2	0.4000	1.0000			0.851
Y3	0.7000	0.4000	1.0000		0.360
Y4	0.4500	0.6500	0.4000	1.0000	0.289

Consequence: Good-fitting model, χ^2 (1, N = 500) = 2.78, reasonable
parameter estimates
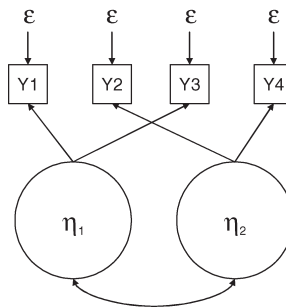
FIGURE 5.2. (*cont.*)

might be addressed by respecifying the model with additional constraints (e.g., fixing the error variance to zero or a small positive value). Although overall fit will worsen somewhat, a proper solution may be obtained.[7] In fact, a default in the EQS program prevents negative variances by setting the lower bound of these estimates to zero. When the LISREL program encounters an indefinite or semidefinite matrix, it invokes a *ridge option*— a smoothing function to eliminate negative or zero eigenvalues—to make the input data suitable for the analysis. These remedial strategies are not recommended. As noted in this section, a negative indicator error variance may be due to a variety of problems, such as sampling fluctuation (small sample size), non-normality (e.g., outliers), multicollinearity, and model misspecification. In addition to causing the Heywood case, these issues signal other problems with the analysis and the sample data (e.g., low statistical power, poorly screened data, redundant indicators). Thus, it is better to diagnose and correct the true source of the problem rather than sidestep it with one of these quick fix remedies.

EFA IN THE CFA FRAMEWORK

A common sequence in scale development and construct validation is to conduct CFA as the next step after latent structure has been explored using EFA. However, the researcher frequently encounters a poor-fitting CFA solution because of the potential sources of misfit that are not present in EFA. For example, unlike the situation in EFA, indicator cross-loadings and residual covariances are usually fixed to zero in initial CFA models. The researcher is then faced with potentially extensive post hoc model testing subject to the criticisms of specification searches in a single data set (MacCallum, 1986).

Although underutilized in the applied literature, the procedure of "exploratory factor analysis within the CFA framework" (E/CFA; Jöreskog, 1969; Jöreskog & Sörbom, 1979; Muthén & Muthén, 1998) can be a useful precursor to CFA that allows the researcher to explore measurement structures more fully before moving into a confirmatory framework. The E/CFA approach represents an intermediate step between EFA and CFA that provides substantial information important in the development of realistic confirmatory solutions. In this strategy, the CFA applies the same number of identifying restrictions used in EFA (m^2) by fixing factor variances to unity, freely estimating the factor covariances, and by selecting an anchor item for each factor whose cross-loadings are fixed to zero (the

loadings of nonanchor items are freely estimated on each factor). Whereas this specification produces the same model fit as maximum likelihood EFA, the CFA estimation provides considerably more information, including the statistical significance of cross-loadings and the potential presence of salient error covariances. Thus, the researcher can develop a realistic measurement structure prior to moving into the more restrictive CFA framework (for applied examples of this approach, see Brown, White, & Barlow, 2005; Brown, White, Forsyth, & Barlow, 2004; Campbell-Sills, Liverant, & Brown, 2004). In addition, E/CFA can be used to bring other variables (i.e., predictors or distal outcomes of the factors) into an EFA-type solution, eliminating the need for factor scores (see Chapters 2 and 3 for an overview of the limitations of factor scores).

To illustrate this strategy, the data from the drinking motives questionnaire are again used. However, in this example, suppose that the psychometric development of this measure is in the early stages and the researcher has only a sense of the correct number of common factors and the hypothesized pattern of item–factor relationships based on theory-driven item development and preliminary exploratory research (which may have led to the elimination of some poorly behaved items). As the next exploratory step, the researcher conducts an EFA in a larger sample (i.e., the current sample with an $N = 500$) with the intent of verifying that a three-factor solution provides acceptable fit, and that the primary loadings of the items are generally in accord with prediction (e.g., items 1 through 4 have their highest loadings on the latent dimension of Coping Motives). At this stage of psychometric evaluation, use of CFA is premature. Although the researcher has a firm conceptual sense of this measure (i.e., number of factors, conjectured pattern of item–factor relationships, as supported by preliminary research), the initial EFA findings are limited in their ability to fully guide the CFA specification (e.g., reasonability of fixing all cross-loadings and error covariances to zero).

Table 5.6 presents Mplus syntax and selected results of an EFA using the $N = 500$ sample. As can be seen in this table, the three-factor EFA solution provided a good fit to the data, $\chi^2(33) = 55.546$, $p = .008$, RMSEA = .037 (90% CI = .019 to .053, CFit = .898). Although these results support the viability of a three-factor model, the researcher wishes to further explore the latent structure of this measure before specifying a CFA solution in an independent sample.

Table 5.7 provides the syntax from several programs (LISREL, Mplus, EQS, CALIS, Amos) for specifying an E/CFA in the $N = 500$ data set. The programs appear very similar to the CFAs presented in earlier examples

TABLE 5.6. Mplus Syntax and Selected Results for an Exploratory Factor Analysis of a Drinking Motives Questionnaire

```
TITLE:  DRINKING MOTIVES EFA
DATA:
  FILE IS "C:\efa.sav"; ! the correlation matrix in Table 5.2 could be
VARIABLE:              ! used as input here
  NAMES ARE X1-X12 GP;
  USEV ARE X1-X12;
ANALYSIS:
  TYPE IS EFA 3 3;
  ESTIMATOR IS ML;
  ITERATIONS = 1000;
  CONVERGENCE = 0.00005;
OUTPUT: SAMPSTAT;
```

EIGENVALUES FOR SAMPLE CORRELATION MATRIX

	1	2	3	4	5
1	3.876	1.906	1.150	0.837	0.722

	6	7	8	9	10
1	0.669	0.576	0.557	0.487	0.471

	11	12
1	0.426	0.323

EXPLORATORY ANALYSIS WITH 3 FACTOR(S) :

```
        CHI-SQUARE VALUE              55.546
        DEGREES OF FREEDOM               33
        PROBABILITY VALUE            0.0083
        RMSEA (ROOT MEAN SQUARE ERROR OF APPROXIMATION) :
        ESTIMATE (90 PERCENT C.I.) IS   0.037 ( 0.019   0.053)
        PROBABILITY RMSEA LE 0.05 IS    0.898
```

PROMAX ROTATED LOADINGS

	1	2	3	
X1	0.573	-0.037	0.031	*to be used as anchor item
X2	0.502	0.065	0.000	
X3	0.475	0.071	0.009	
X4	0.514	0.494	0.006	
X5	0.095	0.590	-0.039	
X6	0.107	0.671	0.041	
X7	0.103	0.641	-0.034	
X8	-0.011	0.759	-0.001	*to be used as anchor item
X9	-0.003	0.043	0.575	
X10	-0.044	0.107	0.567	
X11	0.108	-0.134	0.689	
X12	-0.041	0.031	0.679	*to be used as anchor item

TABLE 5.7. Syntax for Conducting an EFA within the CFA Framework
for the Drinking Motives Questionnaire

LISREL 8.72

```
TITLE E/CFA OF DRINKING MOTIVES ITEMS
DA NI=13 NO=500 MA=CM
LA
X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 GP
RA FI = C:\EFA.SAV
SE
X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 /
MO NX=12 NK=3 PH=SY,FR LX=FU,FR TD=DI
LK
COPING SOCIAL ENHANCE
PA LX
1 0 0                          ! ITEM 1 IS AN ANCHOR ITEM FOR COPING
1 1 1
1 1 1
1 1 1
1 1 1
1 1 1
1 1 1
0 1 0                          ! ITEM 8 IS AN ANCHOR ITEM FOR SOCIAL
1 1 1
1 1 1
1 1 1
0 0 1                          ! ITEM 12 IS AN ANCHOR ITEM FOR ENHANCE
PA PH
0
1 0
1 1 0                                ! FACTOR COVARIANCES FREELY ESTIMATED
VA 1.0 PH(1,1) PH(2,2) PH(3,3)   ! FACTOR VARIANCES FIXED TO 1.0
OU ME=ML RS MI SC ND=4
```

Mplus 3.11

```
TITLE:  DRINKING MOTIVES ECFA
DATA:
  FILE IS "C:\efa.sav";
VARIABLE:
  NAMES ARE X1-X12 GP;
  USEV ARE X1-X12;
ANALYSIS: ESTIMATOR IS ML;
          ITERATIONS=1000;
MODEL:
  COPING BY  X1-X12*.5 X8@0 X12@0;  !X1 is ANCHOR ITEM
  SOCIAL BY  X1-X12*.5 X1@0 X12@0;  !X8 is ANCHOR ITEM
  ENHANCE BY X1-X12*.5 X1@0 X8@0;   !X12 is ANCHOR ITEM
  COPING-ENHANCE@1;                 !FACTOR VARIANCES FIXED TO 1.0
OUTPUT:  STANDARDIZED MODINDICES(10.00) SAMPSTAT;
```

TABLE 5.7. *(cont.)*

EQS 5.7b

```
/TITLE
 E/CFA OF DRINKING MOTIVES ITEMS
/SPECIFICATIONS
 CASES=500; VAR=13; ME=ML; MA=RAW;
 DA = 'C:\EFA.SAV';
/LABELS
 v1=item1; v2= item2; v3= item3; v4= item4; v5= item5; v6= item6;
 v7= item7; v8= item8; v9= item9; v10=item10; v11=item11; v12=item12;
 v13=gp; f1 = coping; f2 = social; f3 = enhance;
/EQUATIONS
 V1  =  *F1+E1;            ! ITEM 1 IS ANCHOR ITEM FOR COPING
 V2  =  *F1+*F2+*F3+E2;
 V3  =  *F1+*F2+*F3+E3;
 V4  =  *F1+*F2+*F3+E4;
 V5  =  *F1+*F2+*F3+E5;
 V6  =  *F1+*F2+*F3+E6;
 V7  =  *F1+*F2+*F3+E7;
 V8  =  *F2+E8;            ! ITEM 8 IS ANCHOR ITEM FOR SOCIAL
 V9  =  *F1+*F2+*F3+E9;
 V10 = *F1+*F2+*F3+E10;
 V11 = *F1+*F2+*F3+E11;
 V12 = *F3+E12;            ! ITEM 12 IS ANCHOR ITEM FOR ENHANCE
/VARIANCES
 F1 TO F3 = 1.0;          ! FACTOR VARIANCES FIXED TO 1.0
 E1 TO E12 = *;
/COVARIANCES
 F1 TO F3 = *;            ! FACTOR COVARIANCES FREELY ESTIMATED
/PRINT
 fit=all;
/LMTEST
/END
```

SAS PROC CALIS 8.2

```
proc calis data=EFADATA cov method=ml pall pcoves;
var = X1-X12;
lineqs
 X1  =  lm11 f1 + e1,
 X2  =  lm21 f1 + lm22 f2 + lm23 f3 + e2,
 X3  =  lm31 f1 + lm32 f2 + lm33 f3 + e3,
 X4  =  lm41 f1 + lm42 f2 + lm43 f3 + e4,
 X5  =  lm51 f1 + lm52 f2 + lm53 f3 + e5,
 X6  =  lm61 f1 + lm62 f2 + lm63 f3 + e6,
 X7  =  lm71 f1 + lm72 f2 + lm73 f3 + e7,
 X8  =  lm82 f2 + e8,
 X9  =  lm91 f1 + lm92 f2 + lm93 f3 + e9,
 X10 = lm101 f1 + lm102 f2 + lm103 f3 + e10,
 X11 = lm111 f1 + lm112 f2 + lm113 f3 + e11,
 X12 = lm123 f3 + e12;
```

(cont.)

TABLE 5.7. *(cont.)*

```
std
  f1-f3 = 1.0,
  e1-e12 = td1-td12;
cov
  f1-f3 = ph21 ph31 ph32;
run;
```

Amos Basic 5

```
' Example of E/CFA in Amos 5.0
Sub Main ()
Dim sem As New AmosEngine
sem.TextOutput
sem.Standardized
sem.Smc

sem.BeginGroup "ecfa.txt"

sem.Structure "x1 =   COPING + (1) E1"               ' X1 IS ANCHOR
sem.Structure "x2 =   COPING + SOCIAL + ENHANCE + (1) E2"
sem.Structure "x3 =   COPING + SOCIAL + ENHANCE + (1) E3"
sem.Structure "x4 =   COPING + SOCIAL + ENHANCE + (1) E4"
sem.Structure "x5 =   COPING + SOCIAL + ENHANCE + (1) E5"
sem.Structure "x6 =   COPING + SOCIAL + ENHANCE + (1) E6"
sem.Structure "x7 =   COPING + SOCIAL + ENHANCE + (1) E7"
sem.Structure "x8 =   SOCIAL + (1) E8"               ' X8 IS ANCHOR
sem.Structure "x9 =   COPING + SOCIAL + ENHANCE + (1) E9"
sem.Structure "x10 =  COPING + SOCIAL + ENHANCE + (1) E10"
sem.Structure "x11 =  COPING + SOCIAL + ENHANCE + (1) E11"
sem.Structure "x12 =  ENHANCE + (1) E12"             ' X12 IS ANCHOR
sem.Structure "COPING (1)"
sem.Structure "SOCIAL (1)"
sem.Structure "ENHANCE (1)"
sem.Structure "COPING <–> SOCIAL"
sem.Structure "COPING <–> ENHANCE"
sem.Structure "SOCIAL <–> ENHANCE"

End Sub
```

Note. N = 500.

(e.g., see Table 4.1, Chapter 4), except for two major differences: (1) all factor loadings and cross-loadings are freely estimated, with the exception of the cross-loadings of the items selected as anchor indicators (X1, X6, X8); and (2) as in EFA, the metric of the latent factors is specified by fixing the factor variances to 1.0 (but the three-factor correlations are freely estimated). Note that items 1 (X1), 6 (X6), and 8 (X8) were selected as anchor indicators for Coping Motives, Social Motives, and Enhancement Motives, respectively. Anchor items are selected on the basis of EFA results (Table 5.6) and entail one item from each factor that has a high (or the highest) primary loading on the factor and low (or the lowest) cross-loadings on the remaining factors. For example, item 8 (X8) was chosen as the anchor indicator for Social Motives because it had the highest loading on this factor (.759) and the lowest cross-loadings on Coping Motives and Enhancement Motives (.011 and .001, respectively; see Table 5.6).

Selected Mplus output of the E/CFA is provided in Table 5.8. First, note that although this analysis was conducted in the CFA framework, the degrees of freedom and overall fit are the same as in the EFA solution presented in Table 5.6; $\chi^2(33) = 55.546$, $p = .008$, RMSEA = .037 (90% CI = .019 to .053, CFit = .898). Nonetheless, this analysis provides considerably more information than the EFA. The E/CFA provides z tests (labeled under the "Est./S.E." column in Mplus) to determine the statistical significance of primary and secondary loadings (except for the cross-loadings of the three anchor items).[8] For example, items 5 through 8 (X5–X8) have statistically significant ($ps < .001$) loadings on Social Motives (range of zs = 10.44 to 17.22), but none of these items have statistically significant cross-loadings on Coping Motives or Enhancement Motives (range of zs = 0.67 to 1.69). Conversely, whereas items 1 through 4 have statistically significant ($ps < .001$) loadings on Coping Motives, the E/CFA results indicated that item 4 also has a large (λ_{42} = .546) and statistically significant ($z = 8.21$, $p < .001$) loading on Social Motives. This suggests that a CFA model of congeneric indicator sets may not be viable; that is, item 4 should be specified to load on both Coping Motives and Social Motives, although its loading on Enhancement Motives could be fixed to zero (i.e., $z = 0.35$, Table 5.8). Other than item 4, the E/CFA results are generally supportive of fixing all other cross-loadings at zero in a subsequent CFA model. In addition, unlike the EFA, the E/CFA provides significance tests of the factor covariances; as seen in Table 5.8, the three drinking motives factors are significantly interrelated (range of zs = 2.99 to 3.98).

TABLE 5.8. Selected Mplus Output of the E/CFA of the Drinking Motives
Questionnaire Items

TESTS OF MODEL FIT

Chi-Square Test of Model Fit

Value	55.546
Degrees of Freedom	33
P-Value	0.0083

CFI/TLI

CFI	0.986
TLI	0.972

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.037	
90 Percent C.I.	0.019	0.053
Probability RMSEA <=.05	0.898	

SRMR (Standardized Root Mean Square Residual)

Value	0.023

MODEL RESULTS

	Estimates	S.E.	Est./S.E.	Std	StdYX
COPING BY					
X1	0.613	0.060	10.274	0.613	0.567
X2	0.514	0.075	6.888	0.514	0.497
X3	0.487	0.074	6.598	0.487	0.471
X4	0.536	0.068	7.873	0.536	0.515
X5	0.097	0.068	1.427	0.097	0.100
X6	0.114	0.067	1.691	0.114	0.117
X7	0.110	0.071	1.556	0.110	0.109
X8	0.000	0.000	0.000	0.000	0.000
X9	0.032	0.075	0.431	0.032	0.031
X10	-0.008	0.076	-0.112	-0.008	-0.008
X11	0.150	0.076	1.957	0.150	0.146
X12	0.000	0.000	0.000	0.000	0.000
SOCIAL BY					
X1	0.000	0.000	0.000	0.000	0.000
X2	0.101	0.076	1.338	0.101	0.098
X3	0.105	0.074	1.413	0.105	0.102
X4	0.546	0.067	8.209	0.546	0.525
X5	0.574	0.055	10.441	0.574	0.594
X6	0.651	0.053	12.252	0.651	0.672
X7	0.649	0.056	11.492	0.649	0.645
X8	0.791	0.046	17.225	0.791	0.753
X9	0.020	0.065	0.310	0.020	0.020

(cont.)

TABLE 5.8. *(cont.)*

X10	0.083	0.065	1.282	0.083	0.081
X11	-0.158	0.069	-2.305	-0.158	-0.154
X12	0.000	0.000	0.000	0.000	0.000
ENHANCE BY					
X1	0.000	0.000	0.000	0.000	0.000
X2	-0.028	0.067	-0.410	-0.028	-0.027
X3	-0.018	0.066	-0.263	-0.018	-0.017
X4	-0.022	0.063	-0.353	-0.022	-0.021
X5	-0.042	0.053	-0.789	-0.042	-0.043
X6	0.034	0.051	0.668	0.034	0.035
X7	-0.039	0.054	-0.723	-0.039	-0.039
X8	0.000	0.000	0.000	0.000	0.000
X9	0.582	0.056	10.447	0.582	0.570
X10	0.580	0.056	10.315	0.580	0.564
X11	0.693	0.057	12.106	0.693	0.676
X12	0.671	0.048	13.929	0.671	0.675
SOCIAL WITH					
COPING	0.400	0.100	3.983	0.400	0.400
ENHANCE WITH					
COPING	0.313	0.105	2.992	0.313	0.313
SOCIAL	0.255	0.081	3.162	0.255	0.255

MODEL MODIFICATION INDICES
Minimum M.I. value for printing the modification index   10.000

		M.I.	E.P.C.	Std E.P.C.	StdYX E.P.C.
WITH Statements					
X10	WITH X9	24.856	0.213	0.213	0.203
X12	WITH X11	23.393	0.283	0.283	0.277

Although EFA may also furnish evidence of the presence of double-loading items (see promax rotated loadings of item 4 in Table 5.6), EFA does not provide any direct indications of the potential existence of salient correlated errors (as noted in Chapters 2 and 3, EFA identification restrictions prevent the specification of correlated indicator errors). This is another area where the results of E/CFA can be quite valuable. Because the analysis is conducted within the CFA framework, modification indices and other fit diagnostic information (e.g., standardized residuals) are available to examine whether the observed correlations among indicators can be adequately reproduced by the latent factors alone. Modification indices are provided only for the measurement error portion of the model (e.g., the theta-delta matrix in LISREL) because all other portions of the solution are saturated (e.g., factor loadings and cross-loadings).

As shown in Table 5.8, large modification indices were obtained in regard to the correlated errors of items 11 and 12, and items 9 and 10 (completely standardized EPCs = .28 and .20, respectively). The correlated error between items 11 and 12 may be grounded substantively by the fact that these are the only reverse-worded items in the questionnaire. It was also described earlier in this chapter why the failure to specify this correlated error may negatively impact the ability of the solution to reproduce the observed relationship between items 9 and 10. Accordingly, the researcher would likely take these results as evidence for the need to freely estimate an error covariance between items 11 and 12 in a subsequent CFA solution. The absence of other large modification indices suggests that other measurement errors can be presumed to be random. Collectively, these findings may foster the refinement of the solution initially suggested by EFA—namely, that the CFA model should not be specified as congeneric (i.e., item 4 should load on two factors) and that all error covariances may be fixed to zero with the exception of items 11 and 12.

MODEL IDENTIFICATION REVISITED

General rules and guidelines for CFA model identification were discussed in Chapter 3. That discussion included basic principles such as the need to define a metric for the latent variables, the need for model *df* to be positive (i.e., the number of elements of the input matrix should equal or exceed the number of freely estimated parameters), and the issue of empirical underidentification (e.g., the marker indicator must be significantly related to its latent factor). However, in the current chapter, more complex measurement models have been considered that entail double-loading indicators and correlated indicator errors. Although these general rules still apply, researchers are more likely to encounter identification problems with more complex solutions. Thus, identification issues with such models are now discussed briefly.

Because latent variable software programs are capable of evaluating whether a given model is identified, it is often most practical to simply try to estimate the solution and let the computer determine the model's identification status. Nevertheless, it is helpful for the researcher to be aware of general model identification guidelines to avoid pursuing structurally or empirically underidentified solutions (such as proposing them in research grant applications). In addition to the guidelines presented in Chapter 3, the researcher should be mindful of the fact that specification of a large number of correlated errors may produce an underidentified model (even

when model *df* is positive). As a general rule, for every indicator there should be at least one other indicator in the solution (which may or may not load on the same latent factor) with which it does not share an error covariance (although in some cases, models that satisfy this rule may still be underidentified; cf. Kenny, Kashy, & Bolger, 1998). In addition, models specified with double-loading indicators may be more prone to under-identification. As discussed in Chapter 6, empirical underidentification is a serious problem in *correlated methods* approaches to CFA multitrait–multimethod analyses where each indicator loads on both a trait factor and a methods factor. The risk for underidentification is increased in models that contain some mixture of double-loading indicators *and* correlated errors. For instance, a solution would be underidentified if an indicator, X1, was specified to load on Factors A and B, but was also specified to have correlated errors with each of the Factor B indicators.

EQUIVALENT CFA SOLUTIONS

Another important consideration in model specification and evaluation is the issue of *equivalent solutions*. Equivalent solutions exist when different model specifications produce identical goodness of fit (with the same number of *df*) and predicted covariance matrices (Σ) in any given data set. Consider the two mediation models depicted below. Both models are overidentified with one *df* corresponding to the nontautological relation-ship between A and C. Although these models differ greatly in terms of their substantive meaning (Model 1: A is a direct cause of B and an indirect cause of C; Model 2: C is a direct cause of B and an indirect cause of A), they generate the same predicted covariance matrix; that is, in both solu-tions, the model-implied relationship of A and C is .30.[9]
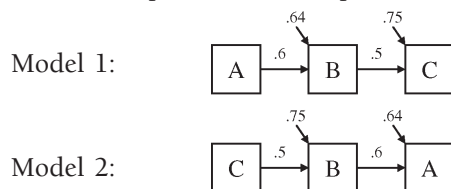
Interested readers are referred to Stelzl (1986), Lee and Hershberger (1990), and Hershberger (1994), who have developed rules for explicating equivalent solutions for various types of SEM models. A noteworthy aspect of this work is that the number of equivalent solutions is related to the sat-uration of the model. Stated another way, models with fewer *df*s have a

greater number of alternative, equivalent solutions than more parsimonious models.

Figure 5.3 presents four equivalent CFA solutions. In this example, the researcher is interested in examining the latent dimensionality of situational social anxiety (i.e., anxiety in social situations due to concerns about being negatively evaluated by others, embarrassment, etc.). Of the six indicators, four could be construed as measuring anxiety in one-on-one social interactions (S1–S4), and the remaining two pertain to anxiety in public speaking situations (S5, S6). Each of the four CFA models in Figure 5.3 is overidentified, with df = 13, and would fit the data equally well.

Models A and B exemplify a quandary often faced by applied CFA researchers—in this case, should the additional covariance existing between the two public speaking indicators be accounted for as a distinct latent factor (Model A), or within a unidimensional solution containing an error covariance between S5 and S6 (Model B)? These alternative specifications may have substantial conceptual implications in the researcher's applied area of study. For instance, the Model A specification would forward the conceptual notion that the construct of Social Anxiety is multidimensional. Model B would assert that Social Anxiety is a broader unitary construct and that the differential covariance of indicators S5 and S6 represents a correlated error, perhaps due to a substantively trivial method effect (unlike S1–S4, these two indicators are more specifically worded to assess anxiety in large groups). Because Models A and B provide the same fit to the data, the procedures of CFA cannot be employed to resolve the question of which model is more acceptable. This is particularly problematic in instances where two or more equivalent models are substantively plausible.

This example also illustrates how model fit and equivalent solutions can be impacted by the composition of the indicator set. The latent dimensionality of a collection of indicators may be strongly impacted by potentially artifactual issues, such as the inclusion of individual items with very similar or reverse wordings or the use of highly overlapping (multicollinear) measures. In the present case, it might be decided prior to specifying a unidimensional CFA model of social anxiety (along the lines of Model B) that the S5 and S6 indicators are over-redundant (i.e., both assess anxiety in speaking in front of large groups). If one of these indicators was omitted from the analyses (or if the S5 and S6 were combined; cf. parceling in Chapter 9), the researcher would avert the problems of poor model fit (if a correlated error had not been specified between S5 and S6 in an initial one-factor model) and equivalence with a two-factor specification (Model A).

Model A
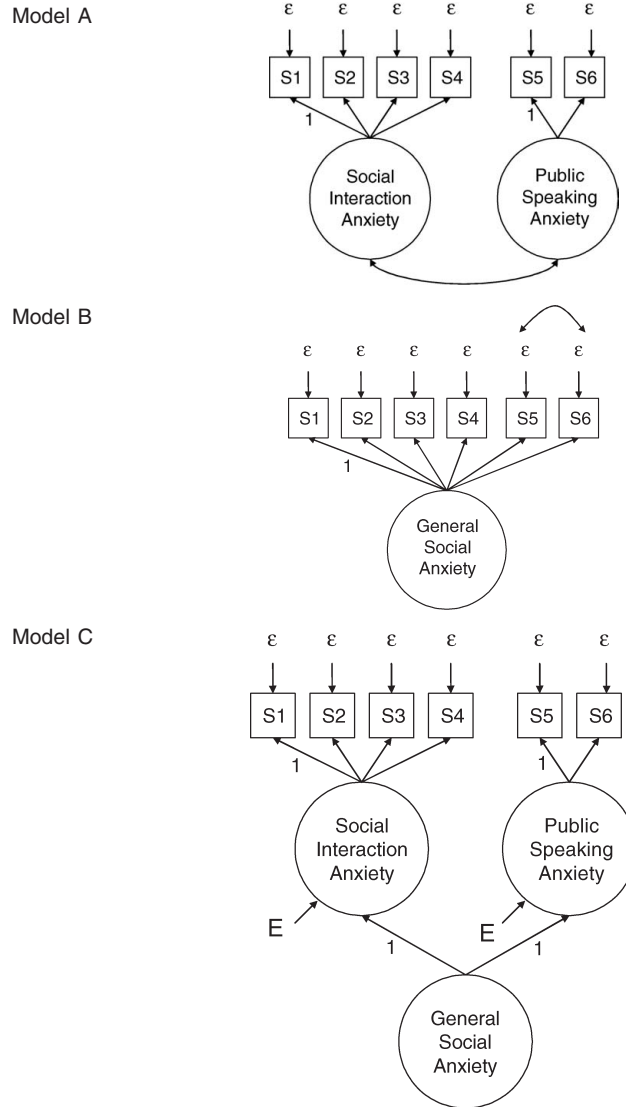
Model B

Model C

FIGURE 5.3. Examples of equivalent CFA models of social anxiety. S1, making eye contact; S2, maintaining a conversation; S3, meeting strangers; S4, speaking on the telephone; S5, giving a speech to a large group; S6, introducing yourself to large groups. All path diagrams use latent-Y notation. In order for Model C to be identified, the two higher-order factor loadings (i.e., General Social Anxiety → Social Interaction Anxiety, General Social Anxiety → Public Speaking Anxiety) must be constrained to equality. To identify Model D, the factor loadings of S5 and S6 on the latent factor of Public Speaking Anxiety must be constrained to equality.
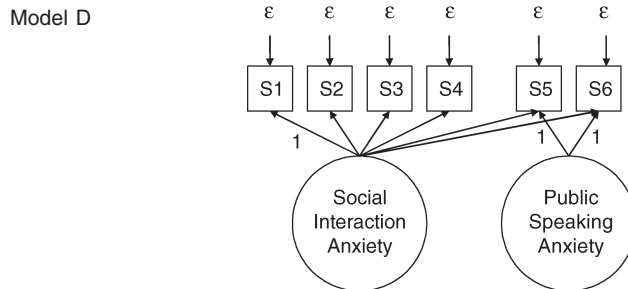
Model D

FIGURE 5.3. *(cont.)*

Figure 5.3 presents two additional equivalent CFA models of social anxiety (assuming that a two-factor solution is conceptually viable). Model C depicts a higher-order factor model (see Chapter 8) in which the correlation between the lower-order factors of Social Interaction Anxiety and Public Speaking Anxiety (Model A) is accounted for by a second-order factor of General Social Anxiety. This specification would imply that social interaction anxiety and public speaking anxiety represent distinct sub-dimensions influenced by the broader construct of general social anxiety. It should be noted that because there are only two lower-order factors, the higher-order portion of Model C would be underidentified if the loadings of General Anxiety → Social Interaction Anxiety and General Social Anxiety → Public Speaking Anxiety were freely estimated (i.e., it is analogous to the underidentified model depicted by Model A in Figure 3.6, Chapter 3). Thus, to identify Model C, the two higher-order factor loadings must be constrained to be equal (equality constraints are discussed in detail in Chapter 7).[10]

Methodologists have noted that some equivalent solutions can be readily dismissed based on logic or theory (e.g., MacCallum, Wegener, Uchino, & Fabrigar, 1993). Model D could be regarded as one such example. In this model, S5 and S6 are purported to be indicators of both Social Interaction Anxiety and Public Speaking Anxiety (this specification requires an equality constraint for the S5 and S6 loadings on Public Speaking Anxiety). However, Social Interaction Anxiety and Public Speaking Anxiety are assumed to be unrelated (i.e., the factor covariance is fixed to zero). It is likely that both of these assertions (S5 and S6 as indicators of anxiety in one-on-one social interactions; orthogonal nature of social interaction and public speaking anxiety) can be quickly rejected on conceptual grounds. Other examples of nonsensical equivalent solutions include lon-

gitudinal models with directional relationships among temporal variables (e.g., regressing a Time 1 variable onto a Time 2 variable; i.e., Time 1 → Time 2 vs. Time 2 → Time 1) or models with direct effects of predetermined exogenous indicators such as gender or age (e.g., Gender → Job Satisfaction vs. Job Satisfaction → Gender).

Figure 5.4 presents three equivalent solutions that entail a single latent factor. In this example, the four indicators are observed measures related to the construct of depression: hopelessness (D1), depressed mood (D2), feelings of guilt (D3), and loss of interest in usual activities (D4). In Model A, a simple one-factor measurement model is specified in which the four indicators are presumed to be interrelated because they are each influenced (caused) by the underlying dimension of Depression. In Model B, the latent factor of Depression (now defined by indicators D2–D4) is regressed onto the single indicator of hopelessness (D1). Although providing the same fit to the data, Models A and B are profoundly different in terms of their conceptual implications. Unlike Model A, which presumes that hopelessness is just another manifest symptom of depression, Model B is in accord with a conceptualization that a sense of hopelessness is a cause of depression. Note that Model A could be respecified such that any of the four indicators are designated as having direct effects on the Depression latent variable (e.g., Model C is similar to Model B, except that guilt, D3, is specified as the cause of Depression). Models B and C can be regarded as examples of CFA with covariates (covariates are sometimes referred to as background variables; cf. Chapter 7). In such models, the latent factors are endogenous (i.e., Latent Y variables) because the solution attempts to explain variance in them with an exogenous indicator (e.g., D1 in Model B). Accordingly, in Models B and C, the residual variance in Depression (often referred to as a *disturbance*; see Chapters 3 and 7) is freely estimated in the psi matrix (cf. Model A, where the factor variance is estimated in the phi matrix). The residual variance in Depression is depicted as "E" in the path diagrams of Models B and C (i.e., "E" = φ_1). As will be seen in Chapter 9, the issue of equivalent models is also germane to solutions that contain formative constructs (i.e., latent variables "caused" by a composite set of indicators), because such models can often be alternatively parameterized as MIMIC models (Chapter 7).

Perhaps the most widely cited article on the issue of equivalent SEM models was published by MacCallum and colleagues in 1993. These authors found that although equivalent models were quite common in applied SEM research, this issue was virtually ignored by investigators. For example, of the 20 SEM articles published in the *Journal of Personality and*
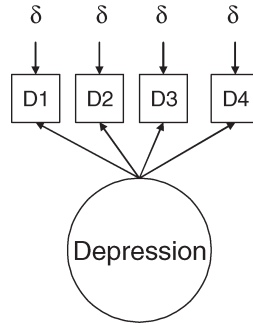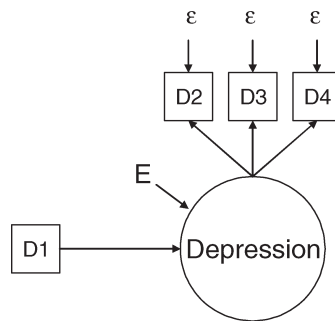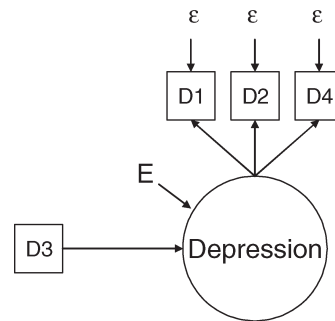
FIGURE 5.4. Examples of equivalent CFA models of depression. D1, hopelessness; D2, depressed mood; D3, guilt; D4, loss of interest in activities.

Social Psychology between 1988 and 1991, 100% reported a model where there existed three or more equivalent models. The median number of equivalent models in these studies was 12 (range of equivalent models = 3 to 33,925!). However, none of these studies contained an acknowledgment of the existence of equivalent solutions. In the 10+ years since the publication of MacCallum et al. (1993), little has changed in regard to applied researchers' explicit recognition of equivalent models. This may be due in part to the unavailability of utilities in latent variable software packages to generate the equivalent models associated with the models specified by the researcher. If such utilities were available, the equivalent models revealed by the software could be evaluated by the researcher in terms of their substantive plausibility. In some cases, this process may lend further support for the hypothesized model (i.e., in situations where there exist no zero equivalent models, or where each of the equivalent models is conceptually not viable). In other instances, the process may have considerable heuristic value (e.g., reveal theoretically plausible alternative models that were not recognized by the researcher).

SUMMARY

Poor-fitting models are common in applied data sets. It is important to have a sound knowledge of the sample data and model before proceeding with a CFA. Data screening procedures (e.g., normality, outliers), in tandem with a proper missing data strategy (see Chapter 9), should be conducted to help ensure that the sample data are appropriate for CFA and the statistical estimator (e.g., ML). Principal components analysis is another helpful procedure in this phase to verify that the input matrix is positive definite (i.e., all resulting eigenvalues are positive). EFA and E/CFA are important precursors to CFA for developing and refining a measurement model. Although use of these methods will foster the chances of success in the restrictive CFA framework (e.g., most or all cross-loadings and error covariances are fixed to zero), the solution still may not satisfy one or more of the three major criteria for model acceptability (i.e., overall goodness of fit, areas of localized strain in the solution, interpretability/strength or parameter estimates). Thus, the researcher must be adept at diagnosing and rectifying the sources of ill fit (interpretation of modification indices, standardized residuals, and reasonability of parameter estimates), keeping in mind both substantive issues (i.e., conceptual justification for the

respecified model) and methodological considerations pertaining to the principles of model identification, equivalent models, model comparison, and improper solutions (e.g., knowing when additional parameters will lead to underidentification).

The first five chapters of this book have covered what could be considered the "fundamentals" of CFA; that is, assuming no complications such as non-normality or missing data, what every researcher "needs to know" to conduct CFA properly with applied data sets. Beginning with Chapter 6, which focuses on the topic of CFA with multitrait–multimethod matrices, the remaining chapters of this book address more specialized applications of CFA (e.g., higher-order factor models, multiple-groups solutions, MIMIC models, scale reliability evaluation) and other issues that frequently arise in the analysis of real data (e.g., missing data, non-normality, categorical indicators, statistical power analysis).

NOTES

1. Correlated errors are sometimes referred to as *minor factors*, especially in instances where three or more indicators loading a broader latent factor have intercorrelated measurement errors. This terminology is most appropriate when the error covariances are presumed to reflect a substantively salient dimension (rather than a byproduct of the measurement approach such as reverse-worded items) that is subsumed by a broader latent construct.

2. Unlike factor correlations = 1.0 or error variances = 0.0 (see discussion of comparing one- and two-factor models), factor loadings, factor covariances, and indicator error covariances that are fixed to 0.0 are not restrictions on the edge of inadmissible parameter spaces (i.e., theoretically, these unstandardized parameters can have values of $\pm \infty$).

3. Interested readers are referred to Yuan, Marshall, and Bentler (2003) for a detailed demonstration of how model misspecification affects parameter biases.

4. An alternative, yet less frequently used, approach to comparing non-nested models involves tests based on "nested tetrads" (Bollen & Ting, 1993, 2000). Tetrads are differences in the products of pairs of covariances (e.g., $\tau_{1234} = \sigma_{12}\sigma_{34} - \sigma_{13}\sigma_{24}$). Depending on the model specification, some tetrads will equal zero (referred to as "vanishing tetrads"). Bollen (1990) derived a χ^2 statistic to perform a simultaneous test of vanishing tetrads of a model to assess its fit to data (*df* = the number of vanishing tetrads). Bollen and Ting have demonstrated scenarios where two models are not structurally nested (i.e., one model does not contain a subset of the freed parameters of the other) but are nested in their vanishing tetrads (i.e., the vanishing tetrads of one model are a subset of the other's). In these instances, two models can be compared by taking the difference in their χ^2s.

5. This rule also applies to weight matrices associated with non-normal theory estimators (e.g., weighted least squares, Chapter 9) and variance–covariance matrices within the model (i.e., Θ_δ, Θ_ε, Ψ, and Φ).

6. On a related note, Heywood cases may occur in EFA when too many factors have been extracted from the data.

7. However, constraining a variance to zero places the estimate on the border of inadmissibility (see the discussion in this chapter on the use of χ^2 difference testing to compare CFA models that differ in number of factors).

8. Although the most widely used software programs for EFA (e.g., SPSS) do not provide significance testing of factor loadings, this feature is available in the Comprehensive Exploratory Factor Analysis freeware program developed by Michael Browne (this can be downloaded at quantrm2.psy.ohio-state.edu/browne/software.htm), and in LISREL 8.72.

9. On rare occasions, equal goodness of fit (in terms of χ^2, CFI, etc.) may occur by chance for two models that do not produce the sample model-implied covariance matrix. This is not an instance of equivalent models.

10. As shown in Chapter 8, a hierarchical model involving a single higher-order factor and three lower-order factors produces the same goodness of fit as a three-factor model where the factors are freely intercorrelated. This is because the higher-order portion of the hierarchical model is just-identified (i.e., analogous to Model B in Figure 3.6, Chapter 3).

CFA of Multitrait–Multimethod Matrices

As seen in prior chapters, CFA provides an elegant analytic framework for evaluating the validity of constructs, and for examining the interrelations of constructs adjusting for measurement error and an error theory. The application of CFA to multitrait–multimethod (MTMM) matrices offers an even more sophisticated methodology for estimating convergent validity, discriminant validity, and method effects in the evaluation of the construct validity of constructs in the social and behavioral sciences. This chapter discusses the various ways an MTMM CFA model can be parameterized and extended, and the strengths and drawbacks of each approach. In addition to bolstering the concepts of convergent and discriminant validity in the context of CFA, the chapter illustrates the deleterious consequences that may result from failing to account for measurement error and method effects.

CORRELATED VERSUS RANDOM MEASUREMENT ERROR REVISITED

In Chapter 5, the need for specifying correlated errors was discussed in instances where some of the observed covariance of two or more indicators is believed to be the effect of the measurement approach (i.e., method covariance), over and above covariance explained by the substantive latent factors. Thus, the error theory of a CFA measurement model can entail some combination of *random measurement error* (i.e., the unexplained vari-

ance of one indicator does not covary with the unexplained variance of another) and *correlated measurement error* or uniqueness (i.e., the unexplained variance of one indicator covaries with the unexplained variance of another), as guided by the conceptual or empirical basis of the model specification. As discussed in Chapter 5, correlated errors should not be specified solely for the purpose of improving model fit. In the majority of models involving latent variables defined by multiple indicators, measurement errors are freely estimated (and correlated errors, if applicable). However, in some instances it is justified to impose constraints on these parameters (e.g., constraining error variances to equality, as in the evaluation of parallel tests) or to fix these estimates to predetermined values (e.g., prespecifying the amount of measurement error in a variable measured by a single indicator; Chapter 4). These alternative specifications for measurement errors are discussed in other chapters (Chapter 4 and Chapter 7).

THE MULTITRAIT–MULTIMETHOD MATRIX

A common limitation of applied research is that the dimensionality and validity of constructs are evaluated in a cross-sectional fashion using a single measurement scale. For example, a researcher may hypothesize that the negative symptoms of schizophrenia are comprised of three components: flat affect, alogia (poverty of speech), social amotivation. To examine this notion, he or she develops a multi-item clinical observation rating system that is subsequently used to rate the behavior of patients admitted to a state hospital. After a sufficient sample has been collected, the ratings are submitted to factor analysis. The results indicate a three-factor solution, which the researcher interprets as supporting the conjectured tripartite model of negative symptoms. He or she concludes that the findings attest to favorable convergent and discriminant validity in that features of flat affect, alogia, and social amotivation are differentially intercorrelated to such an extent that they load on distinct latent factors. Further support for validity is obtained by results showing that the three latent factors are more strongly related to measures of schizophrenia severity than indicators of other disorders (e.g., bipolar disorder).

The aforementioned scenario is a common empirical sequence in scale development and construct validation. Although a useful part of such endeavors, this sequence provides an incomplete evaluation of construct validity. It is not clear as to what extent the multidimensionality of the

negative symptom rating scale could be attributed to artifacts of the indicator set. For example, the chances of obtaining a distinct latent factor of Flat Affect may be fostered by creation of a rating scale containing several similarly worded items that assess this feature. This issue is particularly salient in instances where factor analysis uncovers more latent dimensions than initially predicted; that is, are these additional dimensions conceptually and practically useful, or do they stem from artifacts of scale development (cf. Models A and B in Figure 5.3 in Chapter 5)? In addition, method effects may obscure the discriminant validity of the constructs. That is, when each construct is assessed by the same measurement approach (e.g., observer rating), it cannot be determined how much of the observed overlap (i.e., factor correlations) is due to method effects as opposed to "true" covariance of the traits. In sum, construct validation is limited in instances where a single assessment method is employed.

Campbell and Fiske (1959) developed the multitrait–multimethod (MTMM) matrix as a method for establishing the construct validity of psychological measures. This methodology entails a matrix of correlations arranged in a manner that fosters the evaluation of construct validity. *Construct validity* is the overarching principle of validity, referring to the extent to which a psychological measure in fact measures the concept it purports to measure. This approach requires that several *traits* (*T*; e.g., attitudes, personality characteristics, behaviors) are each assessed by several *methods* (*M*; e.g., alternative test forms, alternative assessment modalities such as questionnaires and observer ratings, or separate testing occasions). The result is a $T \times M$ correlation matrix that is interpreted with respect to convergent validity, discriminant validity, and method effects.

An example of the MTMM matrix is presented in Table 6.1. In this illustration, the researcher wishes to examine the construct validity of the DSM-IV Cluster A personality disorders, which are enduring patterns of symptoms characterized by odd or eccentric behaviors (American Psychiatric Association, 1994). Cluster A is comprised of three personality disorder constructs: (1) paranoid (an enduring pattern of distrust and suspicion such that others' motives are interpreted as malevolent); (2) schizoid (an enduring pattern of detachment from social relationships and restricted range of emotional expression); and (3) schizotypal (an enduring pattern of acute discomfort in social relationships, cognitive and perceptual distortions, and behavioral eccentricities). In a sample of 500 patients, each of these three traits is measured by three assessment methods: (1) a self-report inventory of personality disorders; (2) dimensional ratings from a structured clinical interview of personality disorders; and (3) observa-

TABLE 6.1. Multitrait–Multimethod Matrix of Cluster A Personality Disorder Constructs

	Inventory			Clinical Interview			Observer Ratings		
	PAR	SZT	SZD	PAR	SZT	SZD	PAR	SZT	SZD
Inventory									
Paranoid (PAR)	(.930)								
Schizotypal (SZT)	.290	(.900)							
Schizoid (SZD)	.372	.478	(.910)						
Clinical Interview									
Paranoid (PAR)	**.587**	.238	.209	(.800)					
Schizotypal (SZT)	.201	**.586**	.126	.213	(.870)				
Schizoid (SZD)	.218	.281	**.681**	.195	.096	(.830)			
Observer Ratings									
Paranoid (PAR)	**.557**	.228	.195	**.664**	.242	.232	(.820)		
Schizotypal (SZT)	.196	**.644**	.146	.261	**.641**	.248	.383	(.790)	
Schizoid (SZD)	.219	.241	**.676**	.290	.168	**.749**	.361	.342	(.800)
Standard Deviations:	3.610	3.660	3.590	2.940	3.030	2.850	2.220	2.420	2.040

Note. Values on the diagonal in parentheses are indicator reliabilities; correlations in boldface type represent convergent validities; single-line boxes are heteromethod blocks; double-line boxes are monomethod blocks; correlations in monomethod blocks represent heterotrait–monomethod coefficients; nonbolded correlations in heteromethod blocks represent heterotrait–heteromethod coefficients.

215

tional ratings made by paraprofessional staff. Thus, Table 6.1 is a 3 (T) × 3 (M) matrix, arranged such that the correlations among the different traits (personality disorders: paranoid, schizotypal, schizoid) are nested within each method (assessment type: inventory, clinical interview, observer ratings). The MTMM is a symmetric correlation matrix with one exception: reliability estimates (e.g., Cronbach's alphas) of the measures are inserted in the diagonal in place of ones (e.g., in Table 6.1, the internal consistency estimate of the inventory measure of paranoid personality is .93). As Campbell and Fiske (1959) note, ideally the reliability diagonal should contain the largest coefficients in the matrix; that is, the measure should be more strongly correlated with itself than with any other indicator in the MTMM matrix.

The MTMM matrix consists of two general types of blocks of coefficients: (1) *monomethod blocks*, which contain correlations among indicators derived from the same assessment method; and (2) *heteromethod blocks*, which contain correlations among indicators assessed by different methods (see Table 6.1). Of central interest is the *validity diagonal*, which corresponds to the diagonal within each heteromethod block—the bolded correlations in Table 6.1. Correlations on the validity diagonal represent estimates of *convergent validity*: different measures of theoretically similar or overlapping constructs should be strongly interrelated. In the MTMM matrix, convergent validity is evidenced by strong correlations among methods measuring the same trait (i.e., monotrait–heteromethod coefficients). For example, the findings in Table 6.1 indicate that the three different measures of schizotypal personality are strongly interrelated (range of rs = .676 to .749). The off-diagonal elements of the heteromethod blocks reveal *discriminant validity*: measures of theoretically distinct constructs should not be highly intercorrelated. Discriminant validity in the MTMM matrix is evidenced by weaker correlations between different traits measured by different methods (i.e., heterotrait–heteromethod coefficients) in relation to correlations on the validity diagonal (monotrait–heteromethod coefficients). In Table 6.1, support for discriminant validity is obtained by the finding that correlations in the off-diagonal elements of the heteromethod blocks are uniformly lower (range of rs = .126 to .290) than the validity coefficients (range of rs = .557 to .749).

Finally, evidence of *method effects* is obtained by an examination of the off-diagonal elements of the monomethod blocks. The extent of method effects is reflected by the differential magnitude of correlations between

different traits measured by the same method (heterotrait–monomethod coefficients) relative to the correlations between the same two traits measured by different methods. As shown in Table 6.1, although not extreme, some method variance is evident, especially for the inventory and observer rating measures. For example, the observer ratings of the traits of paranoid and schizotypal personality are more highly correlated (r = .383) than heteromethod measures of these traits (e.g., the correlation between paranoid and schizotypal personality traits measured by inventory and observer rating, respectively, is .196; see Table 6.1). As in the present example, when the collective results indicate that convergent and discriminant validity are high and method effects are negligible, construct validity is supported.

CFA APPROACHES TO ANALYZING THE MTMM MATRIX

Despite the fact that the Campbell and Fiske (1959) methodology represented a significant advance in the conceptualization and evaluation of construct validity, the MTMM approach was not widely used over the years immediately following its inception. In addition, several limitations of the MTMM methodology were noted, including the subjective nature of its interpretation (e.g., ambiguity in terms of what patterns of correlations reflect satisfactory convergent and discriminant validity), its reliance on correlations among fallible observed measures to draw inferences about trait and methods factors (cf. Schmitt & Stults, 1986), and the failure of EFA to obtain meaningful solutions of MTMM data (e.g., identification restrictions of EFA prevent specification of correlated errors; see Chapter 2).

Interest in the MTMM matrix increased with the realization that the procedures of CFA could be readily applied to its analysis (cf. Cole, 1987; Flamer, 1983, Marsh & Hocevar, 1983; Widaman, 1985). In other words, MTMM matrices, like any other form of correlation or covariance matrix, can be analyzed by CFA to make inferences about potential underlying dimensions such as trait and methods factors. Although several different types of CFA models can be applied to MTMM data (cf. Marsh & Grayson, 1995; Widaman, 1985), two forms of CFA specification have been predominant. Using more contemporary terminology, these two types of solutions have been referred to as *correlated methods* and *correlated uniqueness* models (Marsh & Grayson, 1995).

Correlated Methods Models

The correlated methods parameterization reflects the traditional CFA approach to analyzing MTMM matrices. There are five major aspects of correlated methods model specifications (Widaman, 1985): (1) to be identified, there must be at least three traits (T) and three methods (M); (2) $T \times M$ indicators are used to define $T + M$ latent factors (i.e., the number of trait factors = T; the number of method factors = M); (3) each indicator is specified to load on two factors—its trait factor and its method factor (all other cross-loadings are fixed to zero); (4) correlations among trait factors and among method factors are freely estimated, but the correlations between trait and method factors are usually fixed to zero; and (5) indicator uniquenesses (i.e., variance in the indicators not explained by the trait and method factors) are freely estimated but cannot be correlated with the uniquenesses of other indicators. Accordingly, in this specification, each indicator is considered to be a function of trait, method, and unique factors.

Figure 6.1 depicts the path diagram of the correlated methods CFA specification for the MTMM matrix of Cluster A personality disorders (Table 6.1). Table 6.2 provides the LISREL syntax for this model. For reasons discussed later in this chapter (e.g., propensity for improper solu-
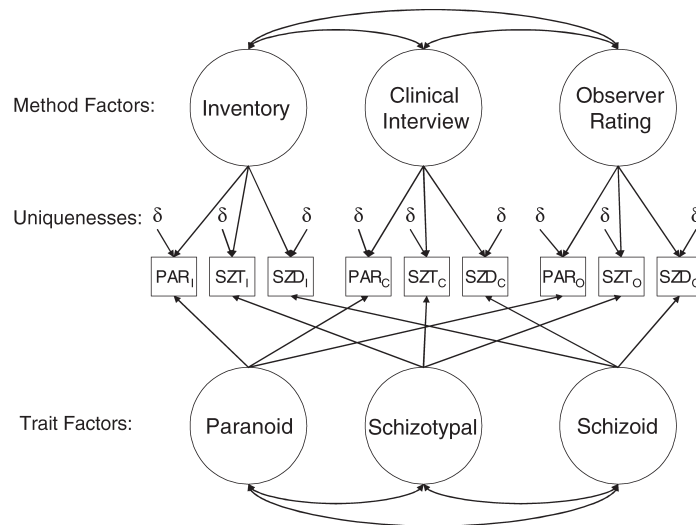
FIGURE 6.1. Correlated methods CFA specification of the MTMM matrix of Cluster A personality disorders.

TABLE 6.2. LISREL Syntax for Specification of a Correlated Methods CFA
of the MTMM Matrix of Cluster A Personality Disorders

```
TITLE LISREL SYNTAX FOR CORRELATED METHODS MTMM SPECIFICATION
DA NI=9 NO=500 MA=CM
LA
PARI SZTI SZDI PARC SZTC SZDC PARO SZTO SZDO
KM
1.000
0.290  1.000
0.372  0.478  1.000
0.587  0.238  0.209  1.000
0.201  0.586  0.126  0.213  1.000
0.218  0.281  0.681  0.195  0.096  1.000
0.557  0.228  0.195  0.664  0.242  0.232  1.000
0.196  0.644  0.146  0.261  0.641  0.248  0.383  1.000
0.219  0.241  0.676  0.290  0.168  0.749  0.361  0.342  1.000
SD
3.61 3.66 3.59 2.94 3.03 2.85 2.22 2.42 2.04
MO NX=9 NK=6 PH=SY,FR LX=FU,FR TD=SY,FR
LK
PARANOID SCHIZOTYP SCHIZOID INVENTRY INTERVW OBSERVE
PA LX
1 0 0 1 0 0
0 1 0 1 0 0
0 0 1 1 0 0
1 0 0 0 1 0
0 1 0 0 1 0
0 0 1 0 1 0
1 0 0 0 0 1
0 1 0 0 0 1
0 0 1 0 0 1
PA TD
1
0 1
0 0 1
0 0 0 1
0 0 0 0 1
0 0 0 0 0 1
0 0 0 0 0 0 1
0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 1
PA PH
0
1 0
1 1 0
0 0 0 0
0 0 0 1 0
0 0 0 1 1 0
VA 1.0 PH(1,1) PH(2,2) PH(3,3) PH(4,4) PH(5,5) PH(6,6)
OU ME=ML RS MI SC AD=OFF IT=500 ND=4
```

tions), results of the correlated methods CFA specification for the Cluster A personality disorders example are not provided.

An important special case of correlated methods solutions is the *uncorrelated methods* model. Its specification is identical to that of correlated methods, except that the covariances of the methods factors are fixed to zero (Widaman, 1985); that is, in the LISREL syntax provided in Table 6.2, this would simply entail fixing the PH(5,4), PH(6,4), and PH(6,5) elements to zero instead of freely estimating these parameters. Because these two models are nested, this comparison provides a statistical evaluation of whether the effects associated with the different assessment methods are correlated; for example, a lack of correlated method effects would be indicated by a nonsignificant χ^2 difference test.

Correlated Uniqueness Models

The correlated uniqueness CFA model (Kenny, 1979; Marsh, 1989) was introduced as an alternative approach to analyzing MTMM data. Figure 6.2 depicts the path diagram of the correlated uniqueness CFA specification for the MTMM matrix of Cluster A personality disorders (Table 6.1). In order for the correlated uniqueness model to be identified, there must be at least two traits (*T*) and three methods (*M*) (although a $2T \times 2M$ model can be fit to the data if the factor loadings of indicators loading on the same trait factor are constrained to equality). Specification of the trait por-
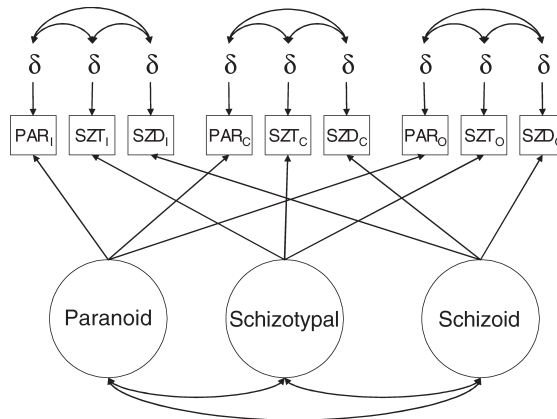
FIGURE 6.2. Correlated uniqueness CFA specification of the MTMM matrix of Cluster A personality disorders.

tion of the correlated uniqueness model is the same as that of the correlated methods approach: (1) each indicator is specified to load on one trait factor (all other cross-loadings are fixed to zero); and (2) correlations among trait factors are freely estimated. Thus, the key difference between these parameterizations is the manner in which method effects are estimated. In the correlated uniqueness model, method effects are estimated by specifying correlated uniquenesses (errors) among indicators based on the same assessment method rather than by method factors. Table 6.3 provides the programming syntax for this model in the LISREL, Mplus, EQS, CALIS, and Amos languages.

A strong practical advantage of the correlated uniqueness approach is that, unlike correlated methods models, this parameterization rarely results in improper solutions (Kenny & Kashy, 1992; Marsh & Bailey, 1991; Tomás, Hontangas, & Oliver, 2000). Indeed, the correlated uniqueness MTMM model of Cluster A personality disorders converged and produced a proper solution that provided an acceptable fit to the data, χ^2 (15) = 14.34, p = .50, SRMR = .025, RMSEA = 0.00 (90% CI = 0.00 to 0.04, CFit = .99), TLI = 1.00, CFI = 1.00. Inspection of standardized residuals and modification indices indicated a reasonable solution. The completely standardized parameter estimates of this solution are presented in Table 6.4. With the exception of one correlated uniqueness (δ_{54}), all unstandardized parameter estimates are significantly different from zero (p < .05).

In regard to the parameter estimates produced by a correlated uniqueness CFA model specification, trait factor loadings that are large and statistically significant would be viewed in support of convergent validity. However, large correlations among the trait factors would be indicative of poor discriminant validity. The presence of appreciable method effects is reflected by correlated uniquenesses among indicators assessed by the same method that are moderate or greater in magnitude. The results in Table 6.4 support the construct validity of the Cluster A personality disorders. The trait factor loadings are consistently large (range = .712–.872), providing evidence that the indicators are strongly related to their purported latent constructs (convergent validity), adjusting for the effects of assessment method. Adequate discriminant validity is evidenced by the modest correlations among trait factors (range = .310–.381). Although significant method effects were obtained in all but one instance (SZT$_C$ with PAR$_C$), the size of these effects is modest (standardized values range = –.037–.293). As in the MTMM matrix presented in Table 6.1, method effects estimated by the CFA model are smallest for the clinical ratings indicators.

TABLE 6.3. Computer Syntax (LISREL, Mplus, EQS, CALIS, Amos) for Specification of a Correlated Uniquenesses CFA of the MTMM Matrix of Cluster A Personality Disorders

LISREL 8.72

```
TITLE LISREL SYNTAX FOR CORRELATED UNIQUENESS MTMM SPECIFICATION
DA NI=9 NO=500 MA=CM
LA
PARI SZTI SZDI PARC SZTC SZDC PARO SZTO SZDO
KM
1.000
0.290  1.000
0.372  0.478  1.000
0.587  0.238  0.209  1.000
0.201  0.586  0.126  0.213  1.000
0.218  0.281  0.681  0.195  0.096  1.000
0.557  0.228  0.195  0.664  0.242  0.232  1.000
0.196  0.644  0.146  0.261  0.641  0.248  0.383  1.000
0.219  0.241  0.676  0.290  0.168  0.749  0.361  0.342  1.000
SD
3.61 3.66 3.59 2.94 3.03 2.85 2.22 2.42 2.04
MO NX=9 NK=3 PH=SY,FR LX=FU,FR TD=SY,FR
LK
PARANOID SCHIZOTYP SCHIZOID
PA LX
1 0 0
0 1 0
0 0 1
1 0 0
0 1 0
0 0 1
1 0 0
0 1 0
0 0 1
PA TD
1
1 1
1 1 1
0 0 0 1
0 0 0 1 1
0 0 0 1 1 1
0 0 0 0 0 0 1
0 0 0 0 0 0 1 1
0 0 0 0 0 0 1 1 1
PA PH
0
1 0
1 1 0
VA 1.0 PH(1,1) PH(2,2) PH(3,3)
OU ME=ML RS MI SC AD=OFF IT=500 ND=4
```

TABLE 6.3. *(cont.)*

```
TITLE:      MPLUS PROGRAM FOR CORRELATED UNIQUENESS MTMM MODEL
DATA:       FILE IS "C:\INPUT5.DAT";
            TYPE IS STD CORR;
            NOBS ARE 500;
VARIABLE:   NAMES ARE PARI SZTI SZDI PARC SZTC SZDC PARO SZTO SZDO;
ANALYSIS:   ESTIMATOR=ML;
MODEL:      PARANOID BY PARI* PARC PARO;
            SCHZOTYP BY SZTI* SZTC SZTO;
            SCHIZOID BY SZDI* SZDC SZDO;
            PARANOID@1.0; SCHZOTYP@1.0; SCHIZOID@1.0;
            PARI SZTI WITH SZDI; PARI WITH SZTI; PARC SZTC WITH SZDC;
            PARC WITH SZTC; PARO SZTO WITH SZDO; PARO WITH SZTO;
            PARANOID SCHZOTYP WITH SCHIZOID; PARANOID WITH SCHZOTYP;
OUTPUT:     SAMPSTAT MODINDICES(3.84) STAND RESIDUAL;
```

EQS 5.7b

```
/TITLE
 EQS SYNTAX FOR CORRELATED UNIQUENESS MTMM SPECIFICATION
/SPECIFICATIONS
 CASES=500; VARIABLES=9; METHODS=ML; MATRIX=COR; ANALYSIS=COV;
/LABELS
 v1=PARI; v2=SZTI; v3=SZDI; v4=PARC;
 v5=SZTC; v6=SZDC; v7=PARO; v8=SZTO; v9=SZDO;
 f1 = paranoid; f2 = schizotyp; f3 = schizoid;
/EQUATIONS
 V1 =   *F1+E1;
 V2 =   *F2+E2;
 V3 =   *F3+E3;
 V4 =   *F1+E4;
 V5 =   *F2+E5;
 V6 =   *F3+E6;
 V7 =   *F1+E7;
 V8 =   *F2+E8;
 V9 =   *F3+E9;
/VARIANCES
 F1 TO F3 = 1.0;
 E1 TO E9= *;
/COVARIANCES
 F1 TO F3 = *;
 E1 TO E3= *; E4 TO E6= *; E7 TO E9= *;
/MATRIX
1.000
0.290  1.000
0.372  0.478  1.000
0.587  0.238  0.209  1.000
0.201  0.586  0.126  0.213  1.000
0.218  0.281  0.681  0.195  0.096  1.000
0.557  0.228  0.195  0.664  0.242  0.232  1.000
0.196  0.644  0.146  0.261  0.641  0.248  0.383  1.000
0.219  0.241  0.676  0.290  0.168  0.749  0.361  0.342  1.000
```

(cont.)

223

TABLE 6.3. *(cont.)*

```
/STANDARD DEVIATIONS
3.61 3.66 3.59 2.94 3.03 2.85 2.22 2.42 2.04
/PRINT
 fit=all;
/LMTEST
/WTEST
/END
```

SAS 8.2 PROC CALIS

```
Title "CALIS SYNTAX FOR CORRELATED UNIQUENESS MTMM SPECIFICATION";
Data CLUSTA (type=CORR);
 input _TYPE_ $ _NAME_ $ V1-V9;
 label V1 = 'pari'
       V2 = 'szti'
       V3 = 'szdi'
       V4 = 'parc'
       V5 = 'sztc'
       V6 = 'szdc'
       V7 = 'paro'
       V8 = 'szto'
       V9 = 'szdo';
cards;
mean .      0      0      0      0      0      0      0      0      0
 std .    3.61   3.66   3.59   2.94   3.03   2.85   2.22   2.42   2.04
   N .     500    500    500    500    500    500    500    500    500
corr V1  1.000    .      .      .      .      .      .      .      .
corr V2  0.290  1.000    .      .      .      .      .      .      .
corr V3  0.372  0.478  1.000    .      .      .      .      .      .
corr V4  0.587  0.238  0.209  1.000    .      .      .      .      .
corr V5  0.201  0.586  0.126  0.213  1.000    .      .      .      .
corr V6  0.218  0.281  0.681  0.195  0.096  1.000    .      .      .
corr V7  0.557  0.228  0.195  0.664  0.242  0.232  1.000    .      .
corr V8  0.196  0.644  0.146  0.261  0.641  0.248  0.383  1.000    .
corr V9  0.219  0.241  0.676  0.290  0.168  0.749  0.361  0.342  1.000
;
run;

proc calis data= CLUSTA cov method=ml pall pcoves;
var = V1-V9;
lineqs
 V1 = lam1 f1 + e1,
 V2 = lam2 f2 + e2,
 V3 = lam3 f3 + e3,
 V4 = lam4 f1 + e4,
 V5 = lam5 f2 + e5,
 V6 = lam6 f3 + e6,
 V7 = lam7 f1 + e7,
 V8 = lam8 f2 + e8,
 V9 = lam9 f3 + e9;
```

(cont.)

TABLE 6.3. *(cont.)*

```
std
  f1-f3 = 1.0,
  e1-e9 = td1-td9;
cov
  f1-f3 = ph1-ph3,
  e1-e3 = td10-td12,
  e4-e6 = td13-td15,
  e7-e9 = td16-td18;
run;
```

Amos Basic 5.0

```
' Example of Correlated Uniqueness MTMM in Amos 5.0

Sub Main ()
Dim sem As New AmosEngine

sem.TextOutput
sem.Standardized
sem.Mods 4                 ' same as Amos default, MIs  4.0 are printed
sem.Smc

sem.BeginGroup "mtmm.txt"

sem.Structure "x1 =   PARANOID + (1) E1"
sem.Structure "x2 =   SCHIZOTP + (1) E2"
sem.Structure "x3 =   SCHIZOID + (1) E3"
sem.Structure "x4 =   PARANOID + (1) E4"
sem.Structure "x5 =   SCHIZOTP + (1) E5"
sem.Structure "x6 =   SCHIZOID + (1) E6"
sem.Structure "x7 =   PARANOID + (1) E7"
sem.Structure "x8 =   SCHIZOTP + (1) E8"
sem.Structure "x9 =   SCHIZOID + (1) E9"

sem.Structure "PARANOID (1)"
sem.Structure "SCHIZOTP (1)"
sem.Structure "SCHIZOID (1)"

sem.Structure "PARANOID  SCHIZOTP"
sem.Structure "PARANOID  SCHIZOID"
sem.Structure "SCHIZOTP  SCHIZOID"

sem.Structure "E1  E2"
sem.Structure "E1 <-> E3"
sem.Structure "E2 <-> E3"
sem.Structure "E4 <-> E5"
sem.Structure "E4 <-> E6"
sem.Structure "E5 <-> E6"
sem.Structure "E7 <-> E8"
sem.Structure "E7 <-> E9"
sem.Structure "E8 <-> E9"

End Sub
```

TABLE 6.4. Completely Standardized Estimates for the Correlated Uniqueness Model of the MTMM Matrix of Cluster A Personality Disorders

	Trait Factor Loadings			SMC	Uniqueness	Correlated Uniquenesses		
	Paranoid	Schizotypal	Schizoid					
PAR_I	.712			.507	.493	1.000		
SZT_I		.788		.621	.379	.094	1.000	
SZD_I			.769	.592	.408	.198	.293	1.000
PAR_C	.841			.707	.293	1.000		
SZT_C		.768		.589	.411	-.037	1.000	
SZD_C			.860	.740	.260	-.073	-.108	1.000
PAR_O	.788			.620	.380	1.000		
SZT_O		.843		.711	.289	.137	1.000	
SZD_O			.872	.760	.240	.111	.126	1.000

Trait Factor Correlations

	Paranoid	Schizotypal	Schizoid
Paranoid	1.000		
Schizotypal	.381	1.000	
Schizoid	.359	.310	1.000

Note. SMC, squared multiple correlation (i.e., λ^2). All unstandardized parameters are significantly different ($p < .05$) from zero except for the correlation between the unique variances of SZT_C and PAR_C.

ADVANTAGES AND DISADVANTAGES OF CORRELATED METHODS AND CORRELATED UNIQUENESS MODELS

As Kenny and Kashy (1992) note, an appealing feature of the correlated methods model is that it corresponds directly to Campbell and Fiske's (1959) original conceptualization of the MTMM matrix. Under this specification, each indicator is considered to be a function of trait, method, and unique variance. For instance, using the completely standardized solution, the squared trait factor loading (squared multiple correlations or communalities), squared method factor loading, and uniqueness for any given indicator sum to 1.0. Thus, these estimates can be interpreted as the proportions of trait, method, and unique variance of each indicator. Moreover, the parameter estimates produced by correlated methods solutions provide seemingly straightforward interpretations with regard to construct validity; for example, large trait factor loadings suggest favorable convergent validity, small or nonsignificant method factor loadings imply an absence of method effects, and modest trait factor intercorrelations suggest favorable discriminant validity. The specification of method factors fosters the substantive interpretation of method effects. Because the covariance associated with a given method is (it is hoped) accounted for by a single latent factor (e.g., a method factor for Observer Ratings; see Figure 6.1), method effects are assumed to be unidimensional (although, as discussed later, this assumption does not always hold). Moreover, unlike the correlated uniqueness model, the correlated methods approach allows for the evaluation of the extent to which method factors are intercorrelated.

However, an overriding drawback of the correlated methods model is that it is usually empirically underidentified. Consequently, a correlated methods solution will typically fail to converge. If it does converge, the solution will usually be associated with Heywood cases (Chapter 5) and large standard errors. For instance, using more than 400 MTMM matrices derived from real and simulated data, Marsh and Bailey (1991) found that the correlated methods model resulted in improper solutions 77% of the time. These authors noted that improper solutions were most probable when the MTMM design was small (e.g., $3T \times 3M$), when sample size was small, and when the assumption of unidimensional method effects was untenable. Moreover, Kenny and Kashy (1992) demonstrated that the correlated methods model was empirically underidentified in two special cases: (1) when the loadings on a trait or methods factor are equal; and (2) when there is no discriminant validity between two or more factors. Although these conditions are never perfectly realized, Kenny and Kashy

(1992) have shown that research data frequently approximate these cases (e.g., factor loading estimates that are roughly equal in magnitude). When factor loadings are roughly equal or if discriminant validity is poor, this empirical underidentification results in severe estimation difficulties. Because of these problems, Kenny and Kashy (1992) and other methodologists (e.g., Marsh & Grayson, 1995) have recommended the use of the correlated uniqueness model over the correlated methods approach for the analysis of MTMM data. However, other researchers (Lance, Noble, & Scullen, 2002) have concluded that, given the substantive strengths of the correlated methods model, the correlated uniqueness model should be used only if the correlated methods model fails. These authors underscore the various design features that may foster the chances of obtaining an admissible correlated methods solution (e.g., increased sample size, larger MTMM designs).[1]

Although their propensity for estimation problems is the primary disadvantage of correlated methods models, researchers have noted another limitation with this specification approach. In particular, these models do not allow for multidimensional method effects. A method effect would be multidimensional if there exist two or more systematic sources of variability (aside from the underlying trait) that affect some or all of the indicators in the model (e.g., response set, reverse-worded questionnaire items, indicators differentially affected by social desirability or the tendency for over- or underreporting). As noted earlier, correlated methods models attempt to explain all the covariance associated with a given assessment method by a single method factor; hence, the methods effects are assumed to be unidimensional. Although these models do not permit correlations between trait and method factors, researchers have demonstrated instances where this assumption is unrealistic (e.g., Kumar & Dillon, 1992). Moreover, methodologists have illustrated scenarios in which the partitioning of observed measure variance into trait and method components does not produce trait-free and method-free interpretations (e.g., Bagozzi, 1993).

Unlike the correlated methods model, correlated uniqueness models rarely pose estimation problems. For instance, in the Marsh and Bailey (1991) study, correlated uniqueness model specifications resulted in proper solutions 98% of the time. In addition, these models can accommodate both unidimensional and multidimensional method effects because method covariance is reproduced by freely estimated correlations (which may differ greatly in magnitude) among indicators based on the same assessment approach. However, the interpretation of correlated uniquenesses as method effects is not always straightforward (e.g., Bagozzi,

1993). Although this parameterization allows for multidimensional method effects, the resulting solution does not provide information on the interpretative nature of these effects (e.g., does not foster a substantive explanation for why the magnitude of the correlated errors may vary widely), and so, for instance, it can be difficult to determine which method has the most method variance.

Another potential drawback of the correlated uniqueness model is its assumption that the correlations among methods, and the correlations between traits and methods, are zero. In fact, when the number of traits (T) is three, the correlated uniqueness model and the uncorrelated methods model virtually always produce equivalent solutions (e.g., model *df* and goodness of fit are identical; the parameter estimates from one solution can be transformed into the other). Methodologists (e.g., Byrne & Goffin, 1993; Kenny & Kashy, 1992) have shown how the parameter estimates of a correlated uniqueness solution may be biased when the assumption of zero correlations among methods and between methods and traits does not hold. If these zero correlation constraints are not tenable, the amount of trait variance and covariance between trait factors will be overestimated, resulting in inflated estimates of convergent validity and lower estimates of discriminant validity, respectively (Kenny & Kashy, 1992; Marsh & Bailey, 1991). Kenny and Kashy (1992) have illustrated this biasing effect in situations where the covariances among methods are mistakenly assumed to be zero and the true factor loadings are all equal:

> The average method–method covariance is added to each element of the trait–trait covariance matrix. So, if the methods are similar to one another, resulting in positive method–method covariances, the amount of trait variance will be overestimated as will be the amount of trait–trait covariance. (pp. 169–170)

Although the size of these biases is usually trivial (Marsh & Bailey, 1991), Kenny and Kashy (1992) nonetheless recommend that researchers using correlated uniqueness models to analyze MTMM data should try to employ assessment methods that are as independent as possible.

OTHER CFA PARAMETERIZATIONS OF MTMM DATA

A number of other CFA-based strategies for analyzing MTMM data have been developed. One of the more prominent alternative approaches is the *direct product model* (Browne, 1984a; Cudeck, 1988; Wothke & Browne,

1990). Unlike the preceding CFA approaches, the direct product model addresses the possibility that method factors interact with trait factors in a multiplicative manner rather than additively (cf. Campbell & O'Connell, 1967). In other words, method effects may augment the correlations of strongly related traits more than they augment the correlations of weakly related constructs; for example, the higher the correlation between traits, the greater the effects of methods. When the data conform to the direct product model, this parameterization provides an elegant test of the criteria outlined by Campbell and Fiske (1959) for evaluating convergent and discriminant validity (for applied examples, see Bagozzi & Yi, 1990; Coovert, Craiger, & Teachout, 1997; and Lievens & Conway, 2001). Another advantage of this method is that it estimates a correlation matrix for the methods. This matrix can be inspected to evaluate the similarity of methods; for example, high correlations between purportedly distinct methods would challenge their discriminant validity (i.e., the methods are in effect the same). However, the direct product model is relatively difficult to program and interpret (the reader is referred to Bagozzi & Yi, 1990, and Wothke, 1996, for fully worked through LISREL parameterizations of direct product solutions). The direct product model often produces improper solutions, although much less so than correlated methods approaches (e.g., Lievens & Conway, 2001). A number of other disadvantages have been noted. For instance, some researchers (e.g., Podsakoff et al., 2003) have concluded that the extant evidence indicates that trait × method interactions are not very common or potent and thus a simpler CFA parameterization will usually be just as good. Nevertheless, more extensive evaluation of the direct product model is needed.

A newer alternative is the *correlated trait–correlated method minus one model*, CT–C(M–1) (Eid, 2000; Eid, Lischetzke, Nussbeck, & Trierweiler, 2003). The CT-C(M-1) model is very similar to the correlated methods model (Figure 6.1), except that it contains one method factor less than the methods included (M–1); for example, the path diagram in Figure 6.1 could be converted into a CT–C(M–1) model by eliminating the Inventory method factor.[2] This parameterization resolves some identification problems of the correlated methods model. The omitted method factor becomes the "comparison standard." The principal notion behind the CT–C(M–1) model is that, for each trait, the true-score variables of indicators of the comparison standard are regressors in a latent regression analysis in which the dependent measures are the true-score variables of the "nonstandard" methods ("nonstandard" indicates methods other than the comparison standard). Thus, a method factor is a residual factor common

to all variables measured by the same method. In other words, it represents the portion of a trait measured by a nonstandard method that cannot be predicted by the true-score variable of the indicators measured by the comparison standard method. A step-by-step illustration of the CT–C(M–1) model is presented in Eid et al. (2003); for an applied example, see Lischetzke and Eid (2003). The CT–C(M–1) has many of the strengths of the correlated methods model (e.g., straightforward decomposition of trait, method, and error effects) but avoids the serious problems of underidentification and improper solutions. The CT–C(M–1) model has several other advantages. When multiple indicators are used (see note 2), the CT–C(M–1) model can test for trait-specific method effects (i.e., there exists a source of variance specific to each trait–method combination), allowing the researcher to examine the generalizability of method effects across traits. Moreover, the model also provides correlations among methods factors, and to some extent, information on the relationships between method factors and trait factors. One limitation is the asymmetry of the CT–C(M–1) model because a comparison standard method must be selected. Choice of the comparison standard may be clear-cut in some situations (e.g., when a "gold standard" exists), but not others (e.g., when all methods are somewhat similar or randomly chosen). Like the direct product model, the performance of the CT–C(M–1) model in applied and simulated data sets has yet to receive extensive study.

CONSEQUENCES OF NOT MODELING METHOD VARIANCE AND MEASUREMENT ERROR

It was noted in earlier chapters (e.g., Chapters 3 and 5) that the ability to model method effects is an important advantage of CFA over EFA. For instance, EFAs of questionnaires that were designed to measure unidimensional constructs (e.g., self-esteem) tend to produce two-factor solutions (Positive Self-Esteem, Negative Self-Esteem). These multidimensional latent structures stem from method effects introduced by including reverse-worded items. A conceptually more viable unidimensional structure can be upheld by CFA through specification of a single trait factor (e.g., Self-Esteem) and method effects (e.g., correlated errors) among the reverse-worded items (e.g., Marsh, 1996). In CFA, specification of a one-factor model without method effects would result in a poor-fitting solution. The correlated errors are needed to account for the additional covariance among items that is due to nonrandom measurement error (e.g.,

acquiescent response style). In addition, this specification yields better estimates of the relationships between the indicators and the latent construct (i.e., reduces bias in the factor loadings).

These points also apply to the CFA of MTMM data. For instance, when the data in Table 6.1 are reanalyzed without modeling method variance (i.e., three trait factors are specified, all measurement error is presumed to be random), a poorly fitting solution results, χ^2 (24) = 454.41, p < .001, SRMR = .063, RMSEA = 0.183 (90% CI = 0.168 to 0.199, CFit < .001), TLI = .765, CFI = .843. Inspection of standardized residuals and modification indices reveal that relationships among indicators obtained from the same assessment method are not well explained by the model; for example, standardized residuals (LISREL) for observer ratings:

	PARO	SZTO	SZDO
PARO	– –		
SZTO	4.3866	– –	
SZDO	3.9154	3.6211	– –

The failure to account for method variance may lead to the false conclusion that the constructs under study have poor discriminant validity. This is because the correlations among the factors (traits) are apt to be inflated by the CFA estimation process. Because method effects have not been taken into account in the model specification, the estimation process attempts to reproduce the additional covariance of indicators sharing the same assessment method by increasing the magnitude of the factor correlations (cf. Eq. 3.8, Chapter 3). This is exemplified in the current data set, where each of the correlations among the Cluster A personality traits are inflated when method effects are not taken into account; for example, correlation of Schizotypal and Schizoid = .31 and .36 in CFA solutions with and without method effects, respectively. This is less problematic in the Table 6.1 data since the parameter estimates would not be interpreted in the context of a poorly fitting model (e.g., RMSEA = 0.183), and the factor correlations in both solutions are not large (e.g., .31 vs. .36). However, this will not always be the case in applied data sets (e.g., inflation of factor correlations can be more extreme).

Another good example in the applied literature on the effects of (not) modeling measurement error is in the area of emotion. Investigators have long debated whether positive and negative mood (e.g., happiness and sadness) are largely independent (i.e., separate constructs) or bipolar (i.e., represent opposite ends of a single dimension). Although original theories

of emotion assumed bipolarity, much of the early research did not support this position. For instance, the initial evidence showed that single indicators of happy and sad moods were not strongly correlated. Consequently, EFAs typically revealed two-factor solutions, along the lines of EFAs of questionnaires containing reversed items (described above). However, researchers later came to learn that the low correlations between indicators of positive and negative moods may be due to the failure to consider and model the effects of random and nonrandom response error (e.g., Green, Goldman, & Salovey, 1993; Russell, 1979). A thorough description of the sources of systematic error in the measurement of affect can be found in Russell (1979) and Green et al. (1993).[3] But when random and systematic error is adjusted for, the correlation between positive and negative emotions increases to an extent that severely challenges the differentiation (discriminant validity) of these dimensions.

To illustrate, a sample correlation matrix from Green et al. (1993) is presented in Figure 6.3 (N = 304 undergraduates). The emotions of happiness and sadness were assessed by four different types of self-report formats: adjective checklist, item response options ranging from strong agreement to strong disagreement, self-description response options ranging from very well to not at all, and a semantic differential Likert scale. Some of the correlations seen in Figure 6.3 might be mistakenly interpreted in support of the notion that happiness and sadness are distinct dimensions; for example, the correlation between adjective checklist indicators of happiness and sadness is –.10. When a CFA model with the appropriate error theory is fit to these data (see path diagram in Figure 6.3), a different picture emerges. The two-factor measurement model, which takes random and nonrandom error into account, provides a reasonable fit to the data, $\chi^2(15)$ = 26.13, p = .036, SRMR = .026, RMSEA = 0.049 (90% CI = 0.01 to 0.08, CFit = .487), TLI = .993, CFI = .996. However, the factor correlation of Happiness and Sadness is found to be –.87, a result that seriously challenges the discriminant validity of these constructs and upholds the contention of the bipolarity of emotion. These findings strongly underscore the importance of multimethod research designs and analytic approaches (i.e., CFA) that model measurement error in construct validation.

SUMMARY

CFA approaches to MTMM matrices (and the MTMM approach in general) continue to be somewhat underutilized in many areas of the applied litera-
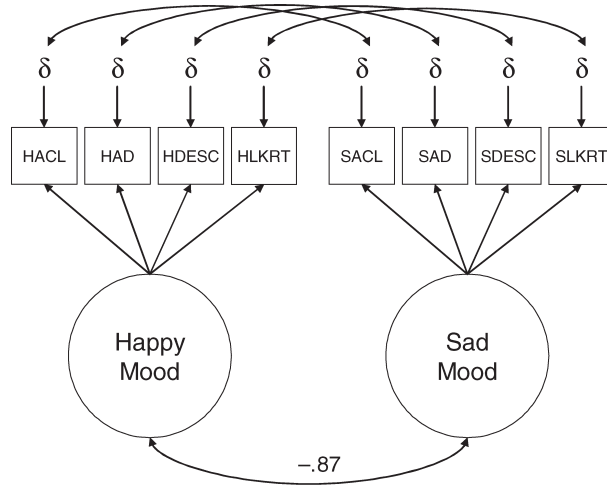
Correlations (*N* = 304):

	HACL	HAD	HDESC	HLKRT	SACL	SAD	SDESC	SLKRT
HACL	1.00							
HAD	.60	1.00						
HDESC	.59	.69	1.00					
HLKRT	.59	.67	.74	1.00				
SACL	-.10	-.46	-.53	-.50	1.00			
SAD	-.44	-.53	-.63	-.64	.68	1.00		
SDESC	-.44	-.48	-.59	-.59	.52	.60	1.00	
SLKRT	-.47	-.59	-.65	-.66	.60	.69	.61	1.00

FIGURE 6.3. Path diagram and input data for CFA model of positive and negative mood. Correlation matrix is taken from Green et al. (1993, Table 7). The first letter of each indicator refers to the construct it purportedly measures: H, happy; S, sad. The remaining letters of each indicator pertain to the self-report assessment format: ACL, adjective checklist; AD, agree–disagree format; DESC, self-descriptive format; LKRT, semantic differential Likert scale.

ture. This is presumably due in part to the fact that, in many cases, multiple measures are not available for a given construct (e.g., the correlated uniqueness approach ordinarily requires at least three assessment methods). Although the examples used in this chapter involved relatively disparate assessment methods, these approaches can be employed for constructs assessed by a single assessment modality; for example, Cluster A personality dimensions assessed by three different self-report inventories (also see Green et al., 1993).

MTMM models can be extended in useful ways. First, the MTMM model could be embedded in a larger structural equation model, such as

one that relates the trait factors to background variables or distal outcomes that represent external validators; for example, are the Cluster A personality dimensions differentially related to salient clinical variables such as comorbidity, overall functioning, or long-term adjustment? Such analyses could strongly bolster the importance of the substantive constructs of the MTMM model (e.g., establish their predictive validity). A second possible extension is analyzing the MTMM model within a multiple-groups solution to determine whether salient parameters (e.g., trait factor loadings) are invariant across salient subgroups (e.g., is each observed measure of Cluster A personality related to the substantive trait in an equivalent manner for men and women?). Approaches to evaluating invariance of CFA models are presented in Chapter 7, along with the analysis of mean structures (i.e., indicator intercepts, latent factor means).

NOTES

1. For instance, Tomás et al. (2000) found that, under some conditions, the correlated methods model works reasonably well (and perhaps better than the correlated uniqueness approach) when more than two indicators per trait–method combination are available. However, because the typical MTMM study is a $3T \times 3M$ design with a single indicator per each trait–method combination (Marsh & Grayson, 1995), evaluation of the correlated methods model under the conditions studied by Tomás et al. (2000) may not be practical in many applied research scenarios.

2. The CT–C(M–1) model can be identified with as few as two traits and two methods. However, to separate trait, method, and error effects in the model, at least two indicators per each trait–method combination are required (e.g., for each trait in Figure 6.1, at least two inventory indicators, two clinical interview indicators, and two observer rating indicators are needed; $p = 18$).

3. In addition, a comprehensive review of the most common types of method effects found in social and behavioral sciences research is provided by Podsakoff et al. (2003).

CFA with Equality Constraints, Multiple Groups, and Mean Structures

The previous CFA examples presented in this book have been estimated within a single group, used a variance–covariance matrix as input, and entailed model parameters that were either freely estimated or fixed. In this chapter, these analyses are extended in several ways. For instance, some CFA specifications will place equality constraints on selected parameters of the measurement model. Such constraints may be applicable in CFA analyses involving a single group (e.g., do the items of a questionnaire assess the same latent construct in equivalent units of measurement?) or two or more groups (e.g., do males and females respond to items of a measuring instrument in a similar manner?). In addition, two different approaches to CFA with multiple groups are presented (multiple-groups CFA, MIMIC models) in context of the analysis of measurement invariance and population heterogeneity. Finally, CFA is extended to the analysis of mean structures involving the estimation of indicator intercepts and latent factor means and the evaluation of their equivalence in multiple groups. Consequently, the input matrix must include the sample means of the indicators, in addition to their variances and covariances. The substantive applications of each approach are discussed.

OVERVIEW OF EQUALITY CONSTRAINTS

As first noted in Chapter 3, parameters in a CFA solution can be freely estimated, fixed, or constrained. A free parameter is unknown, and the researcher allows the analysis to find its optimal value that, in tandem with other model estimates, minimizes the differences between the observed and predicted variance–covariance matrices (e.g., in a one-factor CFA model, obtain the set of factor loadings that best reproduce the observed correlations among four input indicators). A fixed parameter is prespecified by the researcher to be a specific value, most commonly either 1.0 (e.g., in the case of marker indicators or factor variances to define the metric of a latent variable) or 0 (e.g., the absence of cross-loadings or error covariances). As with a free parameter, a constrained parameter is also unknown. However, the parameter is not free to be any value, but rather the specification places restrictions on the values it may assume. The most common form of constrained parameters are *equality constraints*, in which unstandardized parameters are restricted to be equal in value (i.e., see the section in Chapter 8 on scale reliability evaluation for a different type of constrained parameter). Consider a CFA model in which four indicators are specified to load on a single factor. If the CFA specification entails an equality constraint on the four factor loadings, the specific value of these loadings is unknown a priori (as in models with freely estimated factor loadings), but the analysis must find a single estimate (applied to all four loadings) that best reproduces the observed relationships among the four indicators. This is unlike the models with freely estimated parameters where the factor loadings are free to take on any set of values that maximize the fit of the solution. Two important principles apply to the various examples of equality constraints discussed in this chapter. First, as with fixed parameters, these constraints are placed on the unstandardized solution. Accordingly, the indicators whose parameters are to be held equal should have the same metric; for example, the unstandardized loading of an indicator defined by a 0–100 scale will inherently differ from the unstandardized loading of an indicator measured on a 1–8 scale. Second, because a CFA model with equality constraints is nested under the measurement model without these constraints (i.e., it entails a subset of the parent model's freely estimated parameters), χ^2 difference testing can be employed as a statistical comparative evaluation of the constrained solution. For example, if the CFA model where the four factor loadings are held to equality does not produce a significant reduction in fit relative to

the corresponding unconstrained solution, it can be concluded that the four indicators have equivalent relationships to the latent factor.

EQUALITY CONSTRAINTS WITHIN A SINGLE GROUP

Congeneric, Tau-Equivalent, and Parallel Indicators

In Chapter 3, it was noted that most CFA specifications in applied research entail congeneric indicator sets. Congeneric indicators are presumed to measure the same construct, and the size of their factor loadings and measurement errors are free to vary; in addition, the assumption of independent measurement errors must hold. Figure 7.1 provides an example of a
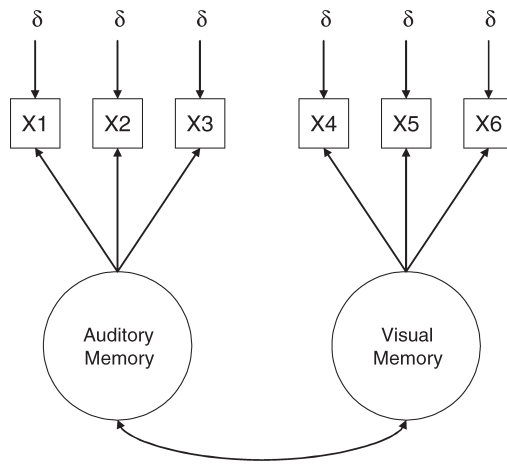
Sample Correlations and Standard Deviations (*SD*s); *N* = 200

	X1	X2	X3	X4	X5	X6
X1	1.000					
X2	0.661	1.000				
X3	0.630	0.643	1.000			
X4	0.270	0.300	0.268	1.000		
X5	0.297	0.265	0.225	0.805	1.000	
X6	0.290	0.287	0.248	0.796	0.779	1.000
SD:	2.610	2.660	2.590	1.940	2.030	2.050

FIGURE 7.1. Two-factor model of memory. X1, Logical Memory; X2, Verbal Paired Association; X3, Word List; X4, Faces; X5, Family Pictures; X6, Visual Reproduction.

congeneric model in which the first three observed measures (X1, X2, X3) are indicators of one latent construct (Auditory Memory), and the second set of measures (X4, X5, X6) are indicators of another latent construct (Visual Memory). In addition to congeneric models, the psychometric literature distinguishes more restrictive solutions that test for the conditions of *tau-equivalent* and *parallel* indicators. A tau-equivalent model entails a congeneric solution in which the indicators of a given factor have equal loadings but differing error variances. As noted previously, when the condition of equal factor loadings holds, it can be asserted that the indicators have equivalent relationships with the underlying construct they measure. Stated another way, a one-unit change in the latent variable is associated with the same amount of change in each indicator that loads on that factor. The most restrictive solution treats indicators as parallel, in which the observed measures are posited to have equal factor loadings and equal error variances. Thus, in addition to the assumption that indicators measure the latent construct in the same units of measurement (tau equivalence), parallel indicators are assumed to measure the latent construct with the same level of precision (i.e., reflected by equivalent error variances).[1] These distinctions have psychometric implications. For instance, Raykov (2001a) has shown that Cronbach's coefficient alpha is a misestimator of the scale reliability of a multicomponent measuring instrument (e.g., multiple-item questionnaire) when the assumption of tau equivalence does not hold (in Chapter 8, a CFA-based approach to estimating scale reliability is presented that does not rest on this assumption). If the conditions of parallel indicators are met, this lends support for the notion that the measures in question are psychometrically interchangeable, a finding that is germane to the endeavor of establishing parallel test forms or justifying the practice of operationalizing a latent construct by summation of its indicators' observed scores (McDonald, 1999).

The test for tau-equivalent and parallel indicators begins with the evaluation of the congeneric measurement model in which the factor loadings and residual variances are free to vary. If the conditions of a congeneric solution are not met (e.g., an indicator loads on more than one factor), the analysis would not proceed to the evaluation of tau equivalence unless, perhaps, substantive considerations allow for the measurement model to be revised to conform to a congeneric model (e.g., removal of a cross-loading indicator). In the context of a congeneric measurement model, the test for tau equivalence is conducted by placing the appropriate restrictions on the solution (i.e., equality constraints on indicators that load on the same factor) and evaluating the resulting change in model χ^2

using the less restricted congeneric model as the baseline solution. If the results are in accord with tau equivalence, that is, constraining the factor loadings to equality does not result in a significant increase in χ^2, the analysis can proceed to the evaluation of parallel indicators.

To illustrate this procedure, a two-factor model of memory is used in which three observed measures are posited to be indicators of the construct of Auditory Memory (X1 = logical memory, X2 = verbal paired association, X3 = word list) and three additional measures are posited to be indicators of the construct of Visual Memory (X4 = faces, X5 = family pictures, X6 = visual reproduction) (the example is loosely based on the structure of the Wechsler Memory Scale, Wechsler, 1997; for applied CFA examples in this research domain, see Price, Tulsky, Millis, & Weiss, 2002, and Tulsky & Price, 2003). The measurement model is presented in Figure 7.1 along with the sample (N = 200) correlations and standard deviations of the six indicators.

The congeneric, two-factor solution provides a good fit to the data, $\chi^2(8) = 4.88$, $p = .77$, SRMR = .012, RMSEA = 0.00 (90% CI = 0.00 to 0.06, CFit = .93), TLI = 1.01, CFI = 1.00. The inspection of modification indices and standardized residuals reveals no areas of strain in the solution. The model parameter estimates are presented in Table 7.1. All six factor loadings are statistically significant (ps < .001) and sufficiently large (range of completely standardized estimates = .78 to .91) (in this example, the metric of the latent variables was defined by fixing factor variances to 1.0). As expected, the constructs of Auditory Memory and Visual Memory are significantly correlated ($p < .001$), $\phi_{21} = .38$.

First, the tau equivalence of the indicators that load on Auditory Memory are examined. This is performed by placing an equality constraint on the factor loadings of the X1, X2, and X3 indicators. Table 7.2 provides the programming syntax for this model in the LISREL, Mplus, EQS, Amos, and CALIS languages. As shown in this table, the software programs differ in their syntax for imposing equality constraints on the model parameters. In LISREL, this restriction is accomplished by adding the line

    EQ LX(1,1) LX(2,1) LX(3,1)

where EQ is the keyword for holding the parameters that follow to equality. The syntax also illustrates a programming shortcut in LISREL, TD = DI, where the theta-delta matrix is specified to be diagonal (DI), meaning that the indicator error variances are freely estimated and the error covariances are fixed to 0 (i.e., the only freed parameters within TD are on the diagonal). In Mplus, the equality constraints are reflected in the line

TABLE 7.1. Mplus Results of the Two-Factor Model of Memory

```
TESTS OF MODEL FIT

Chi-Square Test of Model Fit
        Value                          4.877
        Degrees of Freedom                 8
        P-Value                        0.7706

CFI/TLI
        CFI                            1.000
        TLI                            1.008

RMSEA (Root Mean Square Error Of Approximation)
        Estimate                       0.000
        90 Percent C.I.                0.000   0.057
        Probability RMSEA <= .05       0.929

SRMR (Standardized Root Mean Square Residual)
        Value                          0.012


MODEL RESULTS
            Estimates      S.E.    Est./S.E.       Std      StdYX
AUDITORY BY
 X1           2.101       0.166     12.663        2.101      0.807
 X2           2.182       0.168     12.976        2.182      0.823
 X3           2.013       0.166     12.124        2.013      0.779

VISUAL BY
 X4           1.756       0.108     16.183        1.756      0.907
 X5           1.795       0.115     15.608        1.795      0.887
 X6           1.796       0.117     15.378        1.796      0.878

VISUAL WITH
 AUDITORY     0.382       0.070      5.464        0.382      0.382

Variances
 AUDITORY     1.000       0.000      0.000        1.000      1.000
 VISUAL       1.000       0.000      0.000        1.000      1.000

Residual Variances
 X1           2.366       0.372      6.365        2.366      0.349
 X2           2.277       0.383      5.940        2.277      0.323
 X3           2.620       0.373      7.027        2.620      0.393
 X4           0.662       0.117      5.668        0.662      0.177
 X5           0.877       0.134      6.554        0.877      0.214
 X6           0.956       0.139      6.866        0.956      0.229


R-SQUARE
Observed
Variable   R-Square
X1           0.651
X2           0.677
X3           0.607
X4           0.823
X5           0.786
X6           0.771
```

LISREL 8.72

```
TITLE LISREL PROGRAM FOR TAU EQUIVALENT AUDITORY INDICATORS
DA NI=6 NO=200 MA=CM
LA
X1 X2 X3 X4 X5 X6
KM
1.000
0.661 1.000
0.630 0.643 1.000
0.270 0.300 0.268 1.000
0.297 0.265 0.225 0.805 1.000
0.290 0.287 0.248 0.796 0.779 1.000
SD
2.61 2.66 2.59 1.94 2.03 2.05
MO NX=6 NK=2 PH=SY,FR LX=FU,FR TD=DI
LK
AUDITORY VISUAL
PA LX
1 0
1 0
1 0
0 1
0 1
0 1
PA PH
0
1 0
VA 1.0 PH(1,1) PH(2,2)
EQ LX(1,1) LX(2,1) LX(3,1)              ! EQUALITY CONSTRAINT
OU ME=ML RS MI SC AD=OFF IT=100 ND=4
```

Mplus 3.11

```
TITLE:    MPLUS PROGRAM FOR TAU EQUIVALENT AUDITORY INDICATORS
DATA:     FILE IS "C:\input6.dat";
          TYPE IS STD CORR;
          NOBS ARE 200;
VARIABLE: NAMES ARE X1-X6;
ANALYSIS: ESTIMATOR=ML;
MODEL:    AUDITORY BY X1* X2 X3 (1);   ! EQUALITY CONSTRAINT
          VISUAL BY X4* X5 X6;
          AUDITORY@1.0; VISUAL@1.0;
OUTPUT:   SAMPSTAT MODINDICES(4.00) STAND RESIDUAL;
```

EQS 5.7b

```
/TITLE
 EQS SYNTAX FOR TAU EQUIVALENT AUDITORY INDICATORS
/SPECIFICATIONS
 CASES=200; VARIABLES=6; METHODS=ML; MATRIX=COR; ANALYSIS=COV;
```

TABLE 7.2. *(cont.)*

```
/LABELS
 v1=logical; v2=verbal; v3=word; v4=faces; v5=family; v6=visrep;
 f1 = auditory; f2 = visual;
/EQUATIONS
 V1 = *F1+E1;
 V2 = *F1+E2;
 V3 = *F1+E3;
 V4 = *F2+E4;
 V5 = *F2+E5;
 V6 = *F2+E6;
/VARIANCES
 F1 TO F2 = 1.0;
 E1 TO E6= *;
/COVARIANCES
 F1 TO F2 = *;
/CONSTRAINT
 (V1,F1) = (V2,F1) = (V3,F1);
/MATRIX
1.000
0.661 1.000
0.630 0.643 1.000
0.270 0.300 0.268 1.000
0.297 0.265 0.225 0.805 1.000
0.290 0.287 0.248 0.796 0.779 1.000
/STANDARD DEVIATIONS
2.61 2.66 2.59 1.94 2.03 2.05
/PRINT
 fit=all;
/LMTEST
/END
```

SAS 8.2 PROC CALIS

```
Title "CALIS SYNTAX FOR TAU EQUIVALENT AUDITORY INDICATORS";
Data WMS (type=CORR);
 input _TYPE_ $ _NAME_ $ V1-V6;
 label V1 = 'logical'
 V2 = 'verbal'
 V3 = 'word'
 V4 = 'faces'
 V5 = 'family'
 V6 = 'visrep';
cards;
mean  .     0      0      0      0      0      0
 std  .   2.61   2.66   2.59   1.94   2.03   2.05
   N  .    200    200    200    200    200    200
corr V1  1.000    .      .      .      .      .
corr V2  0.661  1.000    .      .      .      .
corr V3  0.630  0.643  1.000    .      .      .
corr V4  0.270  0.300  0.268  1.000    .      .
corr V5  0.297  0.265  0.225  0.805  1.000    .
corr V6  0.290  0.287  0.248  0.796  0.779  1.000
```

TABLE 7.2. *(cont.)*

```
;
run;

proc calis data=WMS cov method=ml pall pcoves;
var = V1-V6;
lineqs
 V1 = lam1 f1 + e1,
 V2 = lam1 f1 + e2,
 V3 = lam1 f1 + e3,
 V4 = lam4 f2 + e4,
 V5 = lam5 f2 + e5,
 V6 = lam6 f2 + e6;
std
 f1-f2 = 1.0,
 e1-e6 = td1-td6;
cov
 f1-f2 = ph3;
run;
```

Amos Basic 5.0

```
' Example of Tau Equivalence in Amos 5.0

Sub Main ()
Dim sem As New AmosEngine

sem.TextOutput
sem.Standardized
sem.Smc

sem.BeginGroup "memory.txt"

sem.Structure "x1 <- AUDITORY (lam1)"
sem.Structure "x2 <- AUDITORY (lam1)"
sem.Structure "x3 <- AUDITORY (lam1)"
sem.Structure "x4 <- VISUAL"
sem.Structure "x5 <- VISUAL"
sem.Structure "x6 <- VISUAL"

sem.Structure "x1 <- E1 (1)"
sem.Structure "x2 <- E2 (1)"
sem.Structure "x3 <- E3 (1)"
sem.Structure "x4 <- E4 (1)"
sem.Structure "x5 <- E5 (1)"
sem.Structure "x6 <- E6 (1)"

sem.Structure "AUDITORY (1)"
sem.Structure "VISUAL (1)"
sem.Structure "AUDITORY <-> VISUAL"

End Sub
```

244

AUDITORY BY X1* X2 X3 (1);

where parameters are constrained to be equal by placing the same number in parentheses following the parameters that are to be held equal. An "*" is placed directly after the X1 indicator to override the Mplus default of setting the first indicator in the list as the marker indicator. This was done because the metric of Auditory Memory was defined by fixing its variance to 1.0. In EQS, the same constraints are imposed by adding a /CONSTRAINT section to the command file as follows:

/CONSTRAINT
(V1, F1) = (V2,F1) = (V3,F1);

In CALIS, these constraints are made in the following lines:

V1 = lam1 f1 + e1,
V2 = lam1 f1 + e2,
V3 = lam1 f1 + e3,

in which the factor loadings for the first factor (f1; Auditory Memory) are each given the same parameter name (lam1; the user may provide any name for these or other model parameters). The same strategy is used in Amos Basic (parameters held to equality given same label, lam1).

As shown in Table 7.3, this restriction results in an increase in χ^2 to 5.66 with df = 10 (p = .84); for a more in-depth discussion of model estimation under this equality constraint, see Appendix 7.1. This solution has 2 more dfs than the congeneric model (df = 8) because of the equality constraint on the factors loadings of Auditory Memory; that is, the congeneric model contains three separate estimates of the X1, X2, and X3 factor loadings, and the tau-equivalent model contains one factor loading estimate applied to each of these three indicators. The difference in χ^2 of the current solution and the congeneric solution is 0.78, with df = 2. Because this χ^2 difference is below the critical value of the χ^2 distribution at df = 2 (i.e., χ^2_{crit} = 5.99, at α = .05), it can be concluded that the three indicators of Auditory Memory are tau equivalent; that is, a unit increase in the latent construct of Auditory Memory is associated with the same amount of change in each of the X1, X2, and X3 indicators. The selected Mplus output provided below shows that these equality constraints apply to both the unstandardized factor loadings and their standard errors (and hence the z tests of significance as well).

TABLE 7.3. Model Fit of Congeneric, Tau-Equivalent, and Parallel Solutions of a Two-Factor Model of Memory ($N = 200$)

	χ^2	df	χ^2_{diff}	Δdf	RMSEA (90% CI)	CFit	SRMR	CFI	TLI
Congeneric solution	4.88	8			.000 (.000–.057)	.93	.012	1.00	1.01
Tau equivalent: Auditory Memory	5.66	10	0.78	2	.000 (.000–.044)	.96	.021	1.00	1.01
Tau equivalent: Auditory & Visual Memory	5.88	12	0.22	2	.000 (.000–.025)	.99	.022	1.00	1.01
Parallel: Auditory Memory	5.97	14	0.09	2	.000 (.000–.000)	1.00	.021	1.00	1.01
Parallel: Auditory & Visual Memory	9.28	16	3.31	2	.000 (.000–.028)	.99	.027	1.00	1.01

Note. χ^2_{diff}, nested χ^2 difference; RMSEA, root mean square error of approximation; 90% CI, 90% confidence interval for RMSEA; CFit, test of close fit (probability RMSEA ≤ .05); SRMR, standardized root mean square residual; CFI, comparative fit index; TLI, Tucker–Lewis Index. All χ^2 and χ^2_{diff} values nonsignificant ($ps > .05$).

```
MODEL  RESULTS
           Estimates     S.E.   Est./S.E.       Std     StdYX
AUDITORY  BY
  X1           2.100    0.125      16.800      2.100     0.807
  X2           2.100    0.125      16.800      2.100     0.804
  X3           2.100    0.125      16.800      2.100     0.798

VISUAL  BY
  X4           1.756    0.108      16.183      1.756     0.907
  X5           1.795    0.115      15.608      1.795     0.887
  X6           1.796    0.117      15.378      1.796     0.878
```

Because this restriction of factor loading equality holds, it will be retained in subsequent tests of the two-factor measurement model of memory.

Next, the model is respecified with the added constraint of holding the factor loadings of Visual Memory to equality, to evaluate the tau equivalence of these indicators. As seen in Table 7.3, this restriction also does not significantly degrade the fit of the solution, $\chi^2_{\text{diff}}(2) = 0.22$, *ns*; therefore, the indicators of Visual Memory can also be considered to be tau equivalent. Because tau equivalence has been established, the analysis can proceed to evaluating the condition of parallel indicators.

The current example defined the metric of Auditory Memory and Visual Memory by fixing their variances to 1.0. It is noteworthy that the same results would be obtained if the unstandardized factor loadings of all indicators were set to 1.0 (with the factor variances freely estimated). This is because when an indicator is selected to be a marker, its unstandardized loading is fixed to 1.0. Thus, if all factor loadings are to be tested for equality (tau equivalence), they must also equal the value of the marker (1.0).

In addition to tau equivalence, the condition of parallelism requires that the error variances of indicators loading on the same factor are the same. This added restriction is tested for Auditory Memory by placing equality constraints on the measurement errors of X1, X2, and X3. The findings presented in Table 7.3 show that this restriction does not result in a significant increase in χ^2, $\chi^2_{\text{diff}}(2) = 0.09$, *ns*. Finally, the results in this table also indicate that the indicators of Visual Memory can be regarded as parallel, $\chi^2_{\text{diff}}(2) = 3.31$, *ns*. The collective findings can be interpreted as being consistent with the notion that X1–X3 and X4–X6 are interchangeable indicators of the latent constructs of Auditory Memory and Visual Memory, respectively.

For the reader's information, the model estimates for the final solution are presented in Table 7.4. Table 7.5 provides the programming syntax for the final model in LISREL, Mplus, EQS, Amos, and CALIS. Before the dis-

TABLE 7.4. Selected Mplus Results of the Final Two-Factor Measurement Model of Memory

```
TESTS OF MODEL FIT

Chi-Square Test of Model Fit
        Value                                9.277
        Degrees of Freedom                      16
        P-Value                             0.9016

CFI/TLI
        CFI                                  1.000
        TLI                                  1.009

RMSEA (Root Mean Square Error Of Approximation)
        Estimate                             0.000
        90 Percent C.I.                      0.000    0.028
        Probability RMSEA <= .05             0.989

SRMR (Standardized Root Mean Square Residual)
        Value                                0.027

MODEL RESULTS
             Estimates      S.E.    Est./S.E.      Std      StdYX
AUDITORY BY
X1              2.099       0.125      16.795      2.099     0.803
X2              2.099       0.125      16.795      2.099     0.803
X3              2.099       0.125      16.795      2.099     0.803

VISUAL BY
X4              1.782       0.097      18.364      1.782     0.890
X5              1.782       0.097      18.364      1.782     0.890
X6              1.782       0.097      18.364      1.782     0.890

VISUAL WITH
AUDITORY        0.381       0.070       5.431      0.381     0.381

Variances
AUDITORY        1.000       0.000       0.000      1.000     1.000
VISUAL          1.000       0.000       0.000      1.000     1.000

Residual Variances
X1              2.427       0.172      14.142      2.427     0.355
X2              2.427       0.172      14.142      2.427     0.355
X3              2.427       0.172      14.142      2.427     0.355
X4              0.832       0.059      14.142      0.832     0.208
X5              0.832       0.059      14.142      0.832     0.208
X6              0.832       0.059      14.142      0.832     0.208

R-SQUARE
Observed
Variable   R-Square
X1           0.645
X2           0.645
X3           0.645
X4           0.792
X5           0.792
X6           0.792
```

TABLE 7.5. Computer Syntax (LISREL, Mplus, EQS, CALIS, Amos) for Specification of Parallel Indicators of Auditory Memory and Visual Memory within a Two-Factor Measurement Model of Memory

---

LISREL 8.72

```
TITLE LISREL PROGRAM FOR PARALLEL INDICATORS
DA NI=6 NO=200 MA=CM
LA
X1 X2 X3 X4 X5 X6
KM
<insert correlation matrix from Figure 7.1 here>
SD
2.61 2.66 2.59 1.94 2.03 2.05
MO NX=6 NK=2 PH=SY,FR LX=FU,FR TD=DI
LK
AUDITORY VISUAL
PA LX
1 0
1 0
1 0
0 1
0 1
0 1
PA PH
0
1 0
VA 1.0 PH(1,1) PH(2,2)
EQ LX(1,1) LX(2,1) LX(3,1)
EQ LX(4,2) LX(5,2) LX(6,2)
EQ TD(1,1) TD(2,2) TD(3,3)
EQ TD(4,4) TD(5,5) TD(6,6)
OU ME=ML RS MI SC AD=OFF IT=100 ND=4
```

Mplus 3.11

```
TITLE:      MPLUS PROGRAM FOR PARALLEL INDICATORS
DATA:       FILE IS "C:\input6.dat";
            TYPE IS STD CORR;
            NOBS ARE 200;
VARIABLE:   NAMES ARE X1-X6;
ANALYSIS:   ESTIMATOR=ML;
MODEL:      AUDITORY BY X1* X2 X3 (1);
            VISUAL BY X4* X5 X6 (2);
            AUDITORY@1.0; VISUAL@1.0;
            X1 X2 X3 (3);
            X4 X5 X6 (4);
OUTPUT:     SAMPSTAT MODINDICES(4.00) STAND RESIDUAL;
```

EQS 5.7b

```
/TITLE
 EQS SYNTAX FOR PARALLEL INDICATORS
/SPECIFICATIONS
 CASES=200; VARIABLES=6; METHODS=ML; MATRIX=COR; ANALYSIS=COV;
```

TABLE 7.5. *(cont.)*

```
/LABELS
 v1=logical; v2=verbal; v3=word; v4=faces; v5=family; v6=visrep;
 f1 = auditory; f2 = visual;
/EQUATIONS
 V1 = *F1+E1;
 V2 = *F1+E2;
 V3 = *F1+E3;
 V4 = *F2+E4;
 V5 = *F2+E5;
 V6 = *F2+E6;
/VARIANCES
 F1 TO F2 = 1.0;
 E1 TO E6= *;
/COVARIANCES
 F1 TO F2 = *;
/CONSTRAINT
 (V1,F1) = (V2,F1) = (V3,F1);
 (V4,F2) = (V5,F2) = (V6,F2);
 (E1, E1) = (E2, E2) = (E3, E3);
 (E4, E4) = (E5, E5) = (E6, E6);
/MATRIX
<insert correlation matrix from Figure 7.1 here>
/STANDARD DEVIATIONS
2.61 2.66 2.59 1.94 2.03 2.05
/PRINT
 fit=all;
/LMTEST
/END
```

SAS 8.2 PROC CALIS

```
Title "CALIS SYNTAX FOR PARALLEL INDICATORS";
Data WMS (type=CORR);
 input _TYPE_ $ _NAME_ $ V1-V6;
 label V1 = 'logical'
 V2 = 'verbal'
 V3 = 'word'
 V4 = 'faces'
 V5 = 'family'
 V6 = 'visrep';
cards;
 mean .   0      0      0      0      0      0
  std .  2.61   2.66   2.59   1.94   2.03   2.05
   N  .  200    200    200    200    200    200
corr V1 1.000   .      .      .      .      .
corr V2 0.661  1.000   .      .      .      .
corr V3 0.630  0.643  1.000   .      .      .
corr V4 0.270  0.300  0.268  1.000   .      .
corr V5 0.297  0.265  0.225  0.805  1.000   .
corr V6 0.290  0.287  0.248  0.796  0.779  1.000
;
run;
```

TABLE 7.5. *(cont.)*

```
proc calis data=WMS cov method=ml pall pcoves;
var = V1-V6;
lineqs
 V1 = lam1 f1 + e1,
 V2 = lam1 f1 + e2,
 V3 = lam1 f1 + e3,
 V4 = lam2 f2 + e4,
 V5 = lam2 f2 + e5,
 V6 = lam2 f2 + e6;
std
 f1-f2 = 1.0,
 e1-e3 = td1,
 e4-e6 = td2;
cov
 f1-f2 = ph3;
run;
```

Amos Basic 5.0

```
' Example of Parallel Indicators in Amos 5.0

Sub Main ()
Dim sem As New AmosEngine

sem.TextOutput
sem.Standardized
sem.Smc

sem.BeginGroup "memory.txt"
sem.Structure "x1 <- AUDITORY (lam1)"
sem.Structure "x2 <- AUDITORY (lam1)"
sem.Structure "x3 <- AUDITORY (lam1)"
sem.Structure "x4 <- VISUAL (lam2)"
sem.Structure "x5 <- VISUAL (lam2)"
sem.Structure "x6 <- VISUAL (lam2)"
sem.Structure "x1 <- E1 (1)"
sem.Structure "x2 <- E2 (1)"
sem.Structure "x3 <- E3 (1)"
sem.Structure "x4 <- E4 (1)"
sem.Structure "x5 <- E5 (1)"
sem.Structure "x6 <- E6 (1)"
sem.Structure "E1 (err1)"
sem.Structure "E2 (err1)"
sem.Structure "E3 (err1)"
sem.Structure "E4 (err2)"
sem.Structure "E5 (err2)"
sem.Structure "E6 (err2)"
sem.Structure "AUDITORY (1)"
sem.Structure "VISUAL (1)"
sem.Structure "AUDITORY <-> VISUAL"

End Sub
```

cussion continues with the next section, a couple of additional points are made about the previous analyses. First, the nested χ^2 procedure was employed by comparing the model in question to the previous, slightly less restricted solution; for example, a model in which the indicators of both Auditory Memory and Visual Memory are held to be parallel versus the model in which only the indicators of Auditory Memory are held parallel. It should be noted that any two models presented in Table 7.3 are nested and thus any less restricted solution could serve as the comparison model. For instance, the final solution (in which both the indicators of Auditory Memory and Visual Memory are parallel) could be tested against the congeneric solution. In this case, the χ^2_{diff} value would be 4.40 with df = 8; that is, χ^2 = 9.28 – 4.88 = 4.40, df = 16 – 8 = 8. This comparison would produce the equivalent result that the fully parallel model did not significantly degrade the fit of the solution because the χ^2 difference of 4.40 is less than the critical value of the χ^2 distribution at df = 8 (i.e., χ^2_{crit} = 15.51, at α = .05). It might appear that it would be more efficient to move straight from the congeneric model to the most restrictive solution. However, the conditions of tau-equivalent and parallel indicators often do not hold in applied data sets (in particular, the condition of parallel indicators is quite restrictive). Thus, it is usually better to employ an incremental strategy that will allow one to more readily detect the sources on noninvariance if significant degradations in model fit are encountered, because the restrictions are placed on a single set of parameters at a time rather than all at once.

Second, in the case of a multifactorial, congeneric solution, the evaluation of tau-equivalent and parallel indicators of one factor does not rely on the respective findings for indicators loading on different factors. For instance, if the condition of tau equivalence for the indicators of Auditory Memory was not met (i.e., these equality constraints led to a significant increase in model χ^2), the researcher could still proceed to evaluating whether the indicators of Visual Memory were tau equivalent and parallel.

Longitudinal Measurement Invariance

Another type of invariance evaluation that can be conducted on CFA models within a single group concerns the equality of construct measurement over time. Although rarely addressed in the applied literature, longitudinal measurement invariance is a fundamental aspect of evaluating temporal change in a construct. In the absence of such evaluation, it cannot be determined whether temporal change observed in a construct is due to true change or to changes in the structure or measurement of the construct

over time. Drawing on the work of Golembiewski, Billingsley, and Yeager (1976), Chan (1998) outlined three types of change that may be encountered in repeated measurements: *alpha*, *beta*, and *gamma* change (these terms do not correspond to parameters of structural equation models). Alpha change refers to true score change in a construct given a constant conceptual domain and constant measurement. Alpha change (true score change) can only be said to occur in the context of longitudinal measurement invariance (i.e., evidence that the measurement of the construct does not change over time). Chan (1998) notes that longitudinal measurement invariance can be construed as an absence of beta and gamma change. Beta change occurs in instances where the construct of interest remains constant, but the measurement properties of the indicators of the construct are temporally inconsistent (e.g., numerical values across assessment points are not on the same measurement scale). Gamma change occurs when the meaning of the construct changes over time (e.g., the number of factors that represent the construct vary across assessment waves). In applied longitudinal research, measurement invariance is often simply (implicitly) assumed and not examined. However, when measurement is not invariant over time, it is misleading to analyze and interpret the temporal change in observed measures or latent constructs; that is, change may be misinterpreted as alpha change when in fact the precision of measurement of the construct, or the construct itself, varies across time. Thus, the examination of measurement invariance should precede applications of SEM procedures with longitudinal data (e.g., latent growth curve models, autoregressive/cross-lagged panel models; Bollen & Curran, 2004; Curran & Hussong, 2003; Duncan, Duncan, Strycker, Li, & Alpert, 1999).

These procedures are illustrated using the longitudinal measurement model presented in Figure 7.2. In this example, the researcher wishes to evaluate whether an intervention was successful in improving employees' job satisfaction. Employees ($N = 250$) of a large company were administered four measures of job satisfaction (Measures A–D, which varied in assessment modality, e.g., questionnaires, supervisor ratings) immediately before and after the intervention (the pre- and posttest interval was 4 weeks). For each measure, higher scores reflect higher job satisfaction. Prior to examining whether the intervention resulted in an overall increase in job satisfaction, the researcher wishes to verify that the construct of job satisfaction, and its measurement, remain stable over time.

Figure 7.2 presents the hypothesized path model in which the construct of job satisfaction is posited to be structurally the same (i.e., unidimensional) at both assessment points (cf. gamma change). If the factor structure is temporally equivalent, additional tests can be performed to
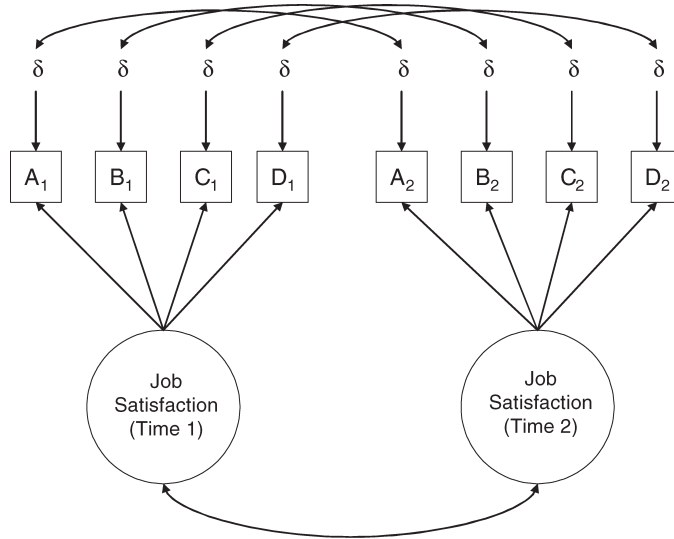
Sample Correlations, Means (*M*), and Standard Deviations (*SD*); *N* = 250

	A1	B1	C1	D1	A2	B2	C2	D2
A1	1.000							
B1	0.736	1.000						
C1	0.731	0.648	1.000					
D1	0.771	0.694	0.700	1.000				
A2	0.685	0.512	0.496	0.508	1.000			
B2	0.481	0.638	0.431	0.449	0.726	1.000		
C2	0.485	0.442	0.635	0.456	0.743	0.672	1.000	
D2	0.508	0.469	0.453	0.627	0.759	0.689	0.695	1.000
M:	1.500	1.320	1.450	1.410	6.600	6.420	6.560	6.310
SD:	1.940	2.030	2.050	1.990	2.610	2.660	2.590	2.550

FIGURE 7.2. Longitudinal measurement model of job satisfaction. Indicators A through D are various measures (e.g., questionnaires, supervisor ratings) of job satisfaction. Rectangles denote within-time portions of the input correlation matrix.

examine the other aspects of measurement invariance (e.g., equality of factor loadings) in a manner similar to the procedures used for evaluating tau-equivalent and parallel indicators. In addition, note that correlated errors have been specified for each pair of repeated measures (e.g., the residual of indicator A1 is allowed to freely covary with the residual of indicator A2). Recall from earlier chapters (e.g., Chapter 2) that the uniqueness of an indicator is comprised of some combination of random

measurement error and indicator-specific variance. Specification of correlated errors is based on the premise that these indicator-specific variances are temporally stable. For instance, such parameters might be posited to reflect method effects associated with repeated administrations of the same measure (e.g., in addition to the underlying dimension of job satisfaction, some of the variance in the A indicator is due to the influence of social desirability, present at each testing occasion). Depending on the research scenario, it may or may not be necessary to specify autocorrelations among indicators' error variances; for example, these effects may be less likely evident in designs employing wide assessment intervals or monomethod indicators.

It should be noted that longitudinal measurement invariance can also be evaluated using the multiple-group approach discussed later in this chapter; that is, each "group" is represented by a different wave of assessment (e.g., Group 1 = Time 1, Group 2 = Time 2). In reviewing the measurement invariance literature, Vandenberg and Lance (2000) discussed the advantages and disadvantages of assessing longitudinal measurement invariance using a single sample (all assessment waves combined in a single input matrix) versus a multiple-group approach (assessment waves represented by separate input matrices for each "group"). The one-sample approach takes into account the complete data structure; that is, the lagged relationships among indicators in addition to the within-time covariances. The entire matrix presented in Figure 7.2 is used in the analysis including the between-time correlations, such as A1 with A2, B2, C2, and D2. In the multiple-group approach, only the within-time portions of the matrix (denoted by boxes in Figure 7.2) are input to the analysis (each box represents a different "group"). Accordingly, a primary advantage of employing a one-sample approach is that correlated errors of the repeated measurements can be estimated and controlled for in the estimation of other model parameters. The main disadvantage of the single-sample method pertains to the use of an input matrix larger than that used in the multiple-group approach (cf. Figure 7.2). Because a larger matrix is used as input, the single-sample approach may be more prone to poor model fit or possibly improper solutions resulting from the greater complexity of the model (cf. Chapter 5).

The longitudinal measurement example is also used to introduce the reader to the analysis of mean structures. All examples been presented in this book thus far have entailed the analysis of covariance structures. That is, only variances and covariances have been included in the input matrix, and thus the indicators and resulting model parameters were deviations

from their means (i.e., their means are presumed to be zero). For example, as with coefficients in multiple regression, a factor loading can be interpreted as the amount of predicted change in an indicator (the Y variable) given a unit change in the latent factor (the X variable). These coefficients do not reflect the exact predicted score of the Y variable, but estimate how much this variable is predicted to change, given a unit change in X. However, a typical application of multiple regression involves the prediction of specific scores of the Y variable, as shown in the simple regression equation below:

$$\hat{Y} = a + bX \tag{7.1}$$

where $\hat{Y}$ is the predicted Y score; a is the intercept; b is the unstandardized regression coefficient; and X is a given score of the predictor variable. The actual (observed value) of Y can be reproduced by adding the residual (e) to the sum of $a + bX$:

$$Y = a + bX + e \tag{7.2}$$

A basic equation in regression solves for the intercept (a) on the basis of knowledge of the sample means of X and Y (M_y, M_x) and the regression coefficient (b):

$$a = M_y - bM_x \tag{7.3}$$

Through a simple algebraic manipulation, this formula can be reexpressed as

$$M_y = a + bM_x \tag{7.4}$$

A fundamental difference between these equations and the equations presented in the context of CFA thus far is its inclusion of the intercept parameter. In the multiple regression framework, the intercept represents the position where the least squares regression line crosses the Y axis; that is, the predicted value of Y when all X variables are zero. In Chapter 3 (Eq. 3.4), it was shown that the observed variance of an indicator could be calculated from the model estimates of a covariance structure CFA using the formula (Latent X notation)

$$VAR(X) = \lambda_x^2 \phi + \delta \tag{7.5}$$

where λ_x is the unstandardized factor loading; ϕ is the factor variance; and δ is the indicator's error variance. With the exception of the intercept parameter, this equation is very similar to the previous regression equation (e.g., $\lambda_x = b$; $\delta = e$).

The equations of CFA can be expanded to include intercept (and latent mean) parameters; for example,

$$X = \tau_x + \Lambda_x\xi + \theta_\delta \qquad (7.6)$$

where τ_x is the indicator intercept (τ_x = tau X; cf. Figure 3.3 in Chapter 3). Similarly, the CFA equivalent to the equation, $M_y = a + bM_x$, is

$$M_x = \tau_x + \lambda_x\kappa \qquad (7.7)$$

where κ is the mean of the latent exogenous factor (κ = kappa; cf. Figure 3.3 in Chapter 3). Thus, the mean of a given indicator (M_x) can be reproduced by the CFA model's parameter estimates of the indicator's intercept (τ_x), factor loading (λ_x), and latent factor mean (κ).

Later in this chapter it will be shown that, like marker indicators and latent factor variances, indicator intercepts (means) and latent factor means are closely related. In addition, the analysis of mean structure poses new identification issues that can be addressed in different ways in the single- and multiple-group approaches.

Returning to the example of longitudinal measurement invariance, this analysis can be conducted with or without the inclusion of mean structures. For instance, if the ultimate goal is to examine models based on a covariance structure (e.g., as in autoregressive/cross-lagged panel modeling), the analysis of mean structures is less relevant. However, if the goal is to examine the trajectory of change in the level of a given construct (e.g., as in latent growth curve modeling), the measurement invariance evaluation should include the analysis of indicator means; that is, the comparison of means is meaningful only if the factor loadings and measurement intercepts are found to be invariant.

The first step of the longitudinal analysis of job satisfaction is to establish that the same factor structure is present at both testing occasions (equal form). As shown in Figure 7.2, it is predicted that a unidimensional measurement model of job satisfaction is viable at both assessment points. Correlated errors were specified in anticipation that additional covariance would exist between repeated measures owing to temporally stable indicator-specific variance (method effects). The metric of the latent fac-

tors was defined by using the same observed measure at both testing occasions (A1 and A2) as the marker indicator.

The estimation of indicator intercepts and factor means requires that the observed means of the indicators be included as input data. Thus, the number of elements of the input matrix increases by the number of indicator means that are included in the analysis; that is,

$$b = [p(p + 1) / 2] + p \qquad (7.8)$$

where b is the number of elements of the input matrix and p is the number of indicators. Thus, in the current example ($p = 8$ indicators), there are 36 variances and covariances [$p(p + 1) / 2$] and 8 means (p), totaling to 44 elements of the input matrix; Eq. 7.8 can be alternatively written as $b = [p(p + 3)] / 2$. Although this exceeds the number of freely estimated parameters of the equal form solution, the mean structure portion of the solution is underidentified in the absence of other model restrictions. In this example, there are 8 observed means (knowns) but 10 unknown parameters in the CFA mean structure (8 indicator intercepts, 2 latent factor means). Hence, this aspect of the model is underidentified because the number of unknowns exceeds the number of knowns (cf. Chapter 3). Moreover, just as latent variables must be provided a metric (e.g., by fixing the loading of an indicator to 1.0), they must also be assigned an origin (i.e., a mean). In covariance structure analysis, the latent variable means are assumed to be zero. In mean structure analysis, where the latent variables may take on mean values other than zero, origins must be assigned. In a single-sample analysis, the mean structure aspect of the measurement model may be identified in one of two ways: (1) fixing the latent mean to zero; or (2) assigning the factor mean to take on the mean as one of its indicators (by fixing the intercept of one indicator to zero). In the first method, all the indicator intercepts are freely estimated but will equal the observed means of the indicators. In the second strategy, all but one intercept is freely estimated and these freed intercepts will take on values different from their sample means; the factor mean is freely estimated but will equal the observed mean of the indicator whose intercept was fixed to zero. The restrictions associated with both approaches will result in just-identification of the mean structure portion of the solution. In either approach, additional restrictions can be placed on the model to over-identify its mean structure component (e.g., equality constraints on indicator intercepts). Later in this chapter, a slightly different method is used to identify the mean structure of a multiple-groups CFA solution.

In the current example, Indicator A's intercept was fixed to zero at both testing occasions. Consequently, in the initial, less constrained solutions (e.g., the test of equal factor structure), the mean of the latent factor of Job Satisfaction was equal to the observed mean of Indicator A. In addition, these restrictions led to just-identification of the mean structure aspect of the solution; that is, 8 observed means; 8 freely estimated mean parameters = 6 intercepts + 2 latent factor means. Thus, although the input matrix is expanded to include indicator means, the overall model df is the same as in covariance structure analysis ($df = 15$) because the mean structure component does not provide additional degrees of freedom to the model ($df = 8 - 8 = 0$).

Selected results of the equal form (factor structure) solution are presented in Table 7.6. The overall fit of this solution is presented in Table 7.7. All results are in accord with the conclusion that a unidimensional measurement model of Job Satisfaction is viable at both testing occasions. Each of the overall fit statistics is consistent with good model fit, for example, $\chi^2(15) = 2.09$, $p = 1.0$, and fit diagnostics indicate the absence of significant areas of strain in the solution (e.g., all modification indices < 4.0). At both assessments, the indicators are found to be significantly (ps $< .001$) and strongly related to the latent construct of Job Satisfaction (range of completely standardized factor loadings $= .81$ to $.90$). In addition, the four error covariances are statistically significant (ps $< .001$); completely standardized values range in magnitude from $.14$ to $.195$. The test–retest covariance of the latent construct of Job Satisfaction is statistically significant ($\phi_{21} = 2.68$, $p < .001$). As noted above, the intercept of Indicator A was fixed to zero at both assessments in order to identify the mean structure component of the solution. This is reflected by the values of 0.00 for all estimates of A1 and A2 in the Intercepts portion of the model results. As the result of these restrictions, the latent means of Job Satisfaction equal the observed means of the A1 and A2 indicators (1.5 and 6.6 for SATIS1 and SATIS2, respectively; cf. Table 7.6 and Figure 7.2). To illustrate an equation presented earlier in this chapter, the sample mean of the indicators can be reproduced by inserting the appropriate model estimates into Equation 7.7. For instance, the sample mean of B1 = 1.32 and can be reproduced using the model estimates of its intercept ($\tau_1 = -0.107$), unstandardized factor loading ($\lambda_{21} = 0.951$), and the mean of its latent factor ($\kappa = 1.5$):

$$-0.107 + 0.951(1.5) = 1.32 \qquad (7.9)$$

Given evidence of equal form (cf. gamma change), additional tests of longitudinal measurement invariance may proceed. The next analysis tests

TABLE 7.6. Mplus Results of the Equal Form Longitudinal Model
of Job Satisfaction

```
MODEL  RESULTS

            Estimates    S.E.   Est./S.E.      Std     StdYX
SATIS1 BY
   A1          1.000    0.000      0.000     1.728     0.899
   B1          0.951    0.050     19.212     1.643     0.811
   C1          0.970    0.049     19.616     1.676     0.814
   D1          0.990    0.046     21.394     1.710     0.859

SATIS2 BY
   A2          1.000    0.000      0.000     2.327     0.896
   B2          0.916    0.048     18.973     2.131     0.808
   C2          0.922    0.046     20.085     2.144     0.828
   D2          0.934    0.045     20.770     2.173     0.849

SATIS2 WITH
   SATIS1      2.680    0.353      7.586     0.667     0.667

A1 WITH
   A2          0.701    0.119      5.915     0.701     0.141

B1 WITH
   B2          1.043    0.163      6.407     1.043     0.195

C1 WITH
   C2          1.042    0.158      6.576     1.042     0.195
D1 WITH
   D2          0.770    0.134      5.735     0.770     0.151

Means
   SATIS1      1.500    0.122     12.343     0.868     0.868
   SATIS2      6.600    0.164     40.183     2.836     2.836

Intercepts
   A1          0.000    0.000      0.000     0.000     0.000
   B1         -0.107    0.117     -0.910    -0.107    -0.053
   C1         -0.005    0.118     -0.042    -0.005    -0.002
   D1         -0.075    0.108     -0.692    -0.075    -0.038
   A2          0.000    0.000      0.000     0.000     0.000
   B2          0.375    0.340      1.102     0.375     0.142
   C2          0.477    0.324      1.475     0.477     0.184
   D2          0.146    0.316      0.462     0.146     0.057
```

(cont.)

TABLE 7.6. (cont.)

Variances					
SATIS1	2.985	0.321	9.297	1.000	1.000
SATIS2	5.414	0.586	9.244	1.000	1.000
Residual Variances					
A1	0.707	0.105	6.722	0.707	0.192
B1	1.409	0.153	9.178	1.409	0.343
C1	1.434	0.157	9.137	1.434	0.338
D1	1.038	0.127	8.165	1.038	0.262
A2	1.330	0.195	6.815	1.330	0.197
B2	2.418	0.263	9.183	2.418	0.347
C2	2.113	0.239	8.856	2.113	0.315
D2	1.836	0.219	8.390	1.836	0.280

for the equality of factor loadings over the two assessment points. Unlike the evaluation of tau equivalence, this analysis does not constrain the indicators that load on the same factor to equality (e.g., $\lambda_{21} = \lambda_{31} = \lambda_{41}$). Rather, the equality constraint pertains to the factor loadings of indicators administered repeatedly across testing occasions (e.g., $\lambda_{21} = \lambda_{62}$). As in the analysis of tau equivalence, the nested χ^2 test can be employed to determine whether these constraints significantly degrade model fit. As shown in Table 7.7, the model χ^2 of the equal loadings solution is 3.88 ($df = 18$, $p = 1.0$), resulting in a nonsignificant χ^2 difference test, χ^2_{diff} (3) = 1.79, ns; [critical value of $\chi^2(3) = 7.81$, $\alpha = .05$]. The df of this nested model comparison is equal to 3 (and the equal factor loadings model $df = 18$) because only three pairs of factor loadings (B1 = B2; C1 = C2; D1 = D2) are involved in this equality constraint; that is, does not apply to the loadings of A1 and A2 because these parameters were previously fixed to 1.0 to set the scale of the latent factors.

On the basis of the results of the equal factor loading analysis, it can be concluded that the indicators evidence equivalent relationships to the latent construct of Job Satisfaction over time. Keeping the equality constraints of the factor loadings in place, the next model placed additional equality constraints on the indicators' intercepts, except for indicators A1 and A2, whose intercepts were previously fixed to zero for the purposes of model identification. These restrictions also did not lead to a significant reduction in model fit, χ^2_{diff} (3) = 3.37, ns, suggesting that the indicator's

TABLE 7.7. Longitudinal Invariance of a Measurement Model of Job Satisfaction (N = 250)

	χ^2	df	χ^2_{diff}	Δdf	RMSEA (90% CI)	CFit	SRMR	CFI	TLI
Equal form	2.09	15			.000 (.000–.000)	1.00	.010	1.00	1.01
Equal factor loadings	3.88	18	1.79	3	.000 (.000–.000)	1.00	.014	1.00	1.01
Equal indicator intercepts	7.25	21	3.37	3	.000 (.000–.000)	1.00	.026	1.00	1.01
Equal indicator error variances	90.73***	25	83.48***	4	.103 (.080–.126)	0.00	.037	.96	0.96

Note. χ^2_{diff}, nested χ^2 difference; RMSEA, root mean square error of approximation; 90% CI, 90% confidence interval for RMSEA; CFit, test of close fit (probability RMSEA ≤ .05); SRMR, standardized root mean square residual; CFI, comparative fit index; TLI, Tucker–Lewis Index.
***p < .001.

intercepts were invariant between the two testing occasions. Recall that in multiple regression, an intercept can be regarded as the predicted value of Y when X is at zero. Measurement intercepts in CFA can be interpreted in an analogous manner; that is, τ_x = the predicted value of indicator X when κ = 0. Thus, if the intercept of an indicator is found to be temporally *noninvariant*, the predicted score of the indicator will vary across time at a constant level of the latent construct (i.e., $\kappa_1 = \kappa_2$). Stated another way, even when the "true score" (latent factor) remains unchanged, the observed scores of the indicators will vary over time. Thus, it is erroneous to interpret changes in observed scores as true change (alpha change), because the observed change is due in some part to temporal variation in the measurement properties of the indicator. For example, although factor loading equivalence would suggest the indicator possesses a temporally stable relationship to the underlying construct (i.e., a unit increase in the latent construct is associated with comparable changes in the indicator at all assessment points), noninvariant indicator intercepts would suggest inequality of the indicator's location parameters over time (a spurious shift from using one portion of the indicator's response scale at Time 1 to another portion of the response scale at Time 2, as might occur in various forms of rater drift such as leniency bias; cf. Vandenberg & Lance, 2000). However, in the present illustration both the factor loadings and indicator intercepts were found to be invariant, suggesting the analysis of mean change over time can be attributed to true change in the construct (cf. Eq. 7.7).

The final analysis tests for the equality of the indicator's error variances. Table 7.8 provides Mplus and LISREL syntax for this analysis. This restriction results in a significant decrease in model fit, χ^2_{diff} (4) = 83.48, $p < .001$ (critical χ^2 value at $df = 4$, $\alpha = .05$, is 9.49). Fit diagnostics suggest that each indicator's error variance is temporally noninvariant:

```
MODEL MODIFICATION INDICES
Variances/Residual Variances
                M.I.   E.P.C.   Std E.P.C. StdYX E.P.C.
A1            35.185  -0.421        -0.421       -0.108
B1            24.357  -0.579        -0.579       -0.125
C1            15.381  -0.421        -0.421       -0.093
D1            23.308  -0.459        -0.459       -0.112
A2            35.187   0.474         0.474        0.076
B2            24.358   0.602         0.602        0.087
C2            15.381   0.442         0.442        0.065
D2            23.307   0.483         0.483        0.078
```

TABLE 7.8. Computer Syntax (LISREL, Mplus) for Testing a Fully Invariant Longitudinal Measurement Model of Job Satisfaction (Equal Form, Equal Factor Loadings, Equal Intercepts, Equal Residual Variances)

Mplus 3.11

```
TITLE:     MPLUS PROGRAM FOR TIME1-TIME2 MSMT MODEL OF JOB SATISFACTION
DATA:      FILE IS "C:\INPUT.dat";
           TYPE IS MEANS STD CORR; ! INDICATOR MEANS ALSO INPUTTED
           NOBS ARE 250;
VARIABLE:  NAMES ARE A1 B1 C1 D1 A2 B2 C2 D2;
ANALYSIS:  ESTIMATOR=ML;
           TYPE=MEANSTRUCTURE; ! ANALYSIS OF MEAN STRUCTURE
MODEL:     SATIS1 BY A1 B1 (1)
           C1 (2)
           D1 (3);
           SATIS2 BY A2 B2 (1)
           C2 (2)
           D2 (3);
           A1 WITH A2; B1 WITH B2; C1 WITH C2; D1 WITH D2;
           [A1@0]; [A2@0]; ! FIXES THE A INDICATOR INTERCEPTS TO ZERO
           [SATIS1*]; [SATIS2*]; ! FREELY ESTIMATES FACTOR MEANS
           [B1 B2] (4); [C1 C2] (5); [D1 D2] (6);
           A1 A2 (7); B1 B2 (8); C1 C2 (9); D1 D2 (10);
OUTPUT:    SAMPSTAT MODINDICES(4.00) STAND RESIDUAL;
```

LISREL 8.72

```
TITLE LISREL PROGRAM FOR TIME1-TIME2 MSMT MODEL OF JOB SATISFACTION
DA NI=8 NO=250 MA=CM
LA
A1 B1 C1 D1 A2 B2 C2 D2
KM
1.000
0.736 1.000
0.731 0.648 1.000
0.771 0.694 0.700 1.000
0.685 0.512 0.496 0.508 1.000
0.481 0.638 0.431 0.449 0.726 1.000
0.485 0.442 0.635 0.456 0.743 0.672 1.000
0.508 0.469 0.453 0.627 0.759 0.689 0.695 1.000
ME                                   ! MEANS INCLUDED AS INPUT
1.50 1.32 1.45 1.41 6.60 6.42 6.56 6.31
SD
1.94 2.03 2.05 1.99 2.61 2.66 2.59 2.55
MO NX=8 NK=2 PH=SY,FR LX=FU,FR TD=SY,FR TX=FR KA=FR   ! TAU-X AND KAPPA
LK
SATIS1 SATIS2
PA LX
0 0
1 0
```

(cont.)

TABLE 7.8. *(cont.)*

```
1 0
1 0
0 0
0 1
0 1
0 1
VA 1.0 LX(1,1) LX(5,2)             ! SET THE METRIC OF THE LATENT VARIABLES
PA TD
1
0 1
0 0 1
0 0 0 1
1 0 0 0 1                          ! OFF-DIAGONAL 1s ARE CORRELATED ERRORS
0 1 0 0 0 1
0 0 1 0 0 0 1
0 0 0 1 0 0 0 1
PA PH
1
1 1
FI TX(1) TX(5)
VA 0.0 TX(1) TX(5)                 ! FIX INDICATOR A INTERCEPTS TO ZERO
EQ LX(2,1) LX(6,2)                 ! FACTOR LOADING EQUALITY CONSTRAINT
EQ LX(3,1) LX(7,2)                 ! FACTOR LOADING EQUALITY CONSTRAINT
EQ LX(4,1) LX(8,2)                 ! FACTOR LOADING EQUALITY CONSTRAINT
EQ TX(2) TX(6)                     ! INTERCEPT EQUALITY CONSTRAINT
EQ TX(3) TX(7)                     ! INTERCEPT EQUALITY CONSTRAINT
EQ TX(4) TX(8)                     ! INTERCEPT EQUALITY CONSTRAINT
EQ TD(1,1) TD(5,5)                 ! ERROR VARIANCE EQUALITY CONSTRAINT
EQ TD(2,2) TD(6,6)                 ! ERROR VARIANCE EQUALITY CONSTRAINT
EQ TD(3,3) TD(7,7)                 ! ERROR VARIANCE EQUALITY CONSTRAINT
EQ TD(4,4) TD(8,8)                 ! ERROR VARIANCE EQUALITY CONSTRAINT
OU ME=ML RS MI SC AD=OFF IT=100 ND=4
```

Note. In Mplus, there can be only one number in parentheses on each line. Recall that parameters that are followed by the same number in parentheses are constrained to be equal. Thus, the syntax (1), (2), and (3) hold the factor loadings of B1 and B2, C1 and C2, and D1 and D3 to equality (although A1 and A2 are on the same line as B1 and B2, these parameters are not constrained to equality because they have been fixed to 1.0 by Mplus default). The Mplus language uses brackets, [ ], to represent indicator intercepts and factor means. Thus, the command [B1 B2] (4) holds the intercepts of indicators B1 and B2 to equality. The last line on the MODEL: section of the syntax, for example, A1 A2 (7), instructs the analysis to constrain the indicator error variances to equality.

Heterogeneity of variance is a common outcome in repeated measures designs, such as the current example. Thus, the test of equal residual variances usually fails in actual data sets because of the temporal fanspread of indicator variances. In the present context, this could be reflective of individual differences in response to the intervention to improve job satisfaction. That is, at Time 1 the variances were more homogeneous because individuals were more similar with regard to their level of job satisfaction. By Time 2, individual differences were more pronounced because some participants responded favorably to the intervention whereas others did not. This is reflected in the input matrix, where it can be seen that the *SD*s increase in magnitude from Time 1 to Time 2 (in addition to the *M*s, which reflect overall improvement in satisfaction). Accordingly, methodologists concede that the test of equal indicator residual variances is highly stringent and will rarely hold in realistic data sets (e.g., Chan, 1998). Fortunately, this condition is not as important to the evaluation of measurement invariance as the prior tests (equal form, factor loadings, and intercepts).

CFA IN MULTIPLE GROUPS

Overview of Multiple-Groups Solutions

The themes introduced in this chapter are now extended to the simultaneous analysis of more than one group. As noted in previous chapters, one of the major advantages of CFA over EFA is its capability to examine the equivalence of all measurement and structural parameters of the factor model across multiple groups. The measurement model pertains to the measurement characteristics of the indicators (observed measures) and thus consists of the factor loadings (lambda), intercepts (tau), and residual variances (theta). Hence, the evaluation of across-group equivalence of these parameters reflects tests of *measurement invariance*. The structural parameters of the CFA model involve evaluation of the latent variables themselves, and thus consist of the factor variances (phi), covariances (phi), and latent means (kappa). If latent-Y terminology is used, the corresponding matrices would be psi and alpha (cf. Figure 3.4, Chapter 3). These parameters describe characteristics of the population from which the sample was drawn. Thus, the examination of the group concordance of structural parameters can be considered tests of *population heterogeneity*; that is, do the dispersion, interrelationships, and levels of the latent factors vary across groups?

CFA with multiple groups has many potential practical applications. For instance, the issues addressed in measurement invariance evaluation are key to the psychometric development of psychological tests; for example, do the items of a questionnaire measure the same constructs (same factor structure) and evidence equivalent relationships to these constructs (equal factor loadings) in all subgroups of the population for whom the measure will be used? Or are there sex, ethnic/racial, age, or other subgroup differences that preclude responding to the questionnaire in comparable ways? Does the questionnaire contain items that are biased against a particular subgroup; that is, yield substantially higher or lower observed scores in a group at equivalent levels of the latent or "true" score? The evaluation of measurement invariance is also important to determining the generalizability of psychological constructs across groups; for example, does the construct underlying the formal definition of a given psychiatric diagnosis operate equivalently across cultures, sexes, age groups, and so forth? Tests of structural parameters reveal potential group differences adjusting for measurement error and an error theory. For example, tests of equality of factor covariances can be construed as the CFA counterpart to inferential evaluation of the differential magnitude of independent correlations; that is, are two constructs more strongly correlated in one group than another? Tests of the equality of latent means are analogous to the comparison of observed group means via *t*-test or ANOVA. However, the major strength of the CFA-based approach is that such comparisons are made in the context of a latent variable measurement model, which hence adjusts for measurement errors, correlated residuals, and so forth.

Two methods can be used to evaluate CFA solutions in multiple groups: (1) *multiple-groups CFA*; and (2) *MIMIC modeling*. Multiple-groups CFA entails the simultaneous analysis of CFA in more than one group. For instance, if the analysis involves two groups (e.g., males and females), two *separate* input matrices are analyzed and constraints can be placed on like parameters (e.g., factor loadings) in both groups to examine the equivalence of the measurement (measurement invariance) and structural solution (population heterogeneity). Although somewhat underutilized in applied research (cf. Vandenberg & Lance, 2000), multiple-groups CFA can entail the analysis of mean structures to evaluate the equality of indicator intercepts (measurement invariance) and latent means (population heterogeneity). A key advantage of multiple-groups CFA is that all aspects of measurement invariance and population heterogeneity can be examined (i.e., factor loadings, intercepts, residual variances, factor variances, factor covariances, latent means).

Conversely, MIMIC models entail the analysis of a single covariance matrix that, in addition to the indicators, includes the dummy code(s) that convey group membership. MIMIC, an acronym for "multiple indicators, multiple causes," has also been referred to as *CFA with covariates*. In this approach, both the latent factors and indicators are regressed onto dummy code(s) that denote group membership. A significant direct effect of the dummy code (covariate) on the latent factor indicates population heterogeneity (group differences on latent means) and a significant direct effect of the dummy code on an indicator is evidence of measurement non-invariance (group differences on the indicator's intercept, i.e., *differential item functioning*). Because a single input matrix is used, the advantages of MIMIC models over multiple-groups CFA include their greater parsimony (MIMIC entails fewer freely estimated parameters), their relatively greater ease of implementation when several groups are involved (i.e., depending on the complexity of the measurement model, multiple-groups CFA may be cumbersome when the number of groups exceeds two), and their less restrictive sample size requirements (i.e., multiple-groups CFA requires a sufficiently large sample size for each group). However, the key limitation of MIMIC models relative to multiple-groups CFA is their ability to examine just two potential sources of invariance (indicator intercepts, factor means).

Multiple-Groups CFA

Before considering how multiple-groups CFA is implemented, it is important to be aware of the variability in terminologies and procedures within the methodology literature. As noted above, an advantage of multiple-groups CFA is that all potential aspects of invariance across groups can be examined. Different terminologies exist in the literature for these various tests of invariance (cf. Horn & McArdle, 1992; Meredith, 1993). For example, the test of equal factor structures ("equal form," meaning that the number of factors and pattern of indicator–factor loadings is identical across groups) has been referred to as "configural invariance." Equality of factor loadings has been referred to as "metric invariance" or "weak factorial invariance." The equality of indicator intercepts has been alternatively termed "scalar invariance" or "strong factorial invariance." Finally, evaluation of the equality of indicator residuals has also been referred to as a test of "strict factorial invariance" (Meredith, 1993). Some confusion surrounds the use of these varying terminologies (cf. Vandenberg & Lance, 2000). For this reason, a more descriptive and pedagogically useful termi-

nology is encouraged and used in this book (e.g., equal form, equal factor loadings, equal intercepts).

Moreover, there are some discrepancies in the order in which the model restrictions within multiple-groups CFA are evaluated. Most commonly, the stepwise procedures previously illustrated in the tests of longitudinal invariance are employed, whereby the analysis begins with the least restricted solution (equal form) and subsequent models are evaluated (using nested χ^2 methods) that entail increasingly restrictive constraints; that is, equal factor loadings $\rightarrow$ equal intercepts $\rightarrow$ equal residual variances, and so on. However, some methodologists (e.g., Horn & McArdle, 1992) have proffered a "step-down" strategy whereby the starting model contains all the pertinent invariance restrictions, and subsequent models are then evaluated that sequentially relax these constraints. The former approach is recommended for several reasons. Especially in the context of a complex CFA solution (i.e., multiple factors and indicators; > 2 groups), it may be difficult to determine the (multiple) sources of noninvariance when a model held to full invariance is poor fitting. Any aspects of ill fit that are encountered are more easily identified and adjusted for in a model-building approach, where new restrictions are placed on the solution at each step. In addition, tests of some aspects of invariance rest on the assumption that other aspects of invariance hold. Group comparisons of latent means are meaningful only if the factor loadings and indicator intercepts have been found to be invariant (see the "Longitudinal Measurement Invariance" section). Group comparisons of factor variances and covariances are meaningfully only when the factor loadings are invariant. The viability of the fully constrained model rests on the results of the less restricted solutions. Thus, it is more prudent for model evaluation to work upward from the least restricted solution (equal form) to determine if further tests of measurement invariance and population heterogeneity are warranted. Of relevance here is the issue of *partial invariance*. As discussed later in this chapter, it may be possible for the evaluation of group equivalence to proceed even in instances where some noninvariant parameters have been encountered (cf. Byrne et al., 1989). For instance, it may be possible to compare groups on latent means if some (but not all) of the factor loadings and intercepts are invariant. Again, a "step-up" approach to invariance evaluation would foster this endeavor.

For the aforementioned reasons, the recommended sequence of multiple-groups CFA invariance evaluation is as follows: (1) Test the CFA model separately in each group; (2) conduct the simultaneous test of equal form (identical factor structure); (3) test the equality of factor loadings; (4) test

the equality of indicator intercepts; (5) test the equality of indicator residual variances (optional); and, if substantively meaningful; (6) test the equality of factor variances; (7) test the equality of factor covariances (if applicable, i.e., > 1 latent factor); and (8) test the equality of latent means. Steps 1–5 are tests of measurement invariance; Steps 6–8 are tests of population heterogeneity.

Finally, some debate continues as to whether multiple-groups CFA should be prefaced by an overall test of covariance matrices across groups. Introduced by Jöreskog (1971b), the rationale of this procedure is that if group differences exist in the parameters of the CFA model, then some values within the covariance matrix should also differ across groups. If the overall test of the equality of covariances fails to reject the null hypothesis, no further analyses are conducted and it is concluded that the groups are invariant. Rejection of the null hypothesis (e.g., $\Sigma_1 \neq \Sigma_2$) is interpreted as justification for conducting multiple-groups CFA to identify the source(s) of noninvariance (for an illustration of how to employ this test, see Vandenberg & Lance, 2000). Although some researchers have supported its continued use (e.g., Vandenberg & Lance, 2000), many methodologists have questioned the rationale and utility of the omnibus test of equal covariance matrices (e.g., Byrne, 1998; Byrne et al., 1989; Jaccard & Wan, 1996). For instance, Byrne (1998) notes that this test often produces contradictory findings with respect to equivalencies across groups; that is, occasions where the omnibus test indicates $\Sigma_1 = \Sigma_2$ but subsequent hypothesis tests of the invariance of specific CFA measurement or structural parameters must be rejected, and vice versa. Jaccard and Wan (1996) add that if the researcher has a specific hypothesis regarding group differences on selected parameters, it is better to proceed directly to the multiple-groups CFA framework because this more focused test will have greater statistical power than the omnibus comparison of covariance matrices. Accordingly, these methodologists have concluded that the omnibus test of equal covariance matrices provides little guidance for testing the equivalence of CFA parameters and thus should not be regarded as a prerequisite to multiple-groups CFA.

The multiple-groups CFA methodology is now illustrated using an actual data set of 750 adult outpatients (375 men, 375 women) with current mood disorders. In this example, the researcher is interested in examining the generalizability of the DSM-IV (American Psychiatric Association, 1994) criteria for the diagnostic category of major depressive disorder (MDD) between sexes. The analysis was motivated by questions that had arisen in the field regarding the possibility of salient sex differ-

ences in the expression of mood disorders (e.g., somatic symptoms such as appetite/weight change may be more strongly related to depression in women). Patients were rated by experienced clinicians on the severity of the nine symptoms that comprise the diagnostic criteria of MDD on 0–8 scales (0 = none, 8 = very severely disturbing/disabling; see Figure 7.3 for description of the 9 symptoms). Prior to the CFAs, the data were screened to ensure their suitability for the ML estimator (i.e., normality, absence of multivariate outliers). In accord with its DSM-IV conceptualization, a unidimensional model of MDD was posited (see Figure 7.3). For substantive reasons (cf. DSM-IV), a correlated residual was specified between the first two diagnostic criteria (i.e., depressed mood, loss of interest in usual activities). The first criterion (M1, depressed mood) was used as a marker indicator to define the metric of the latent variable. Accordingly, the covariance structure aspect of the model was overidentified in both groups with *df* = 26 (45 variances/covariances, 19 freely estimated parameters).

Prior to conducting the multiple-groups CFAs, it is important to ensure that the posited one-factor model is acceptable in both groups. If markedly disparate measurement models are obtained between groups, this outcome would contraindicate further invariance evaluation. As shown in Table 7.9, in both men and women, overall fit statistics for the one-factor solution were consistent with good model fit. In both groups, all freely estimated factor loadings were statistically significant (all *p*s < .001) and salient (completely standardized factor loadings ranged from .34 to .73). No remarkable points of strain were noted in either solution, as reflected by small modification indices, expected parameter change values, and standardized residuals.

Next, the simultaneous analysis of equal form was conducted. Table 7.10 provides the LISREL, Mplus, Amos, and EQS syntax for the equal form analysis. In LISREL, Amos, and EQS, the programming essentially entails "stacking" the CFA analysis of one group on top of the other. In LISREL and EQS, the programs are alerted to the multiple groups analysis by an additional command on the second line of syntax (NG = 2 on the DATA line of LISREL; GROUPS = 2 on the /SPECIFICATIONS line of EQS). What follows in the EQS programming is stacking the two CFA analyses for men and women; that is, the programming is identical to the syntax that would be written for conducting the CFA in a single group. The LISREL programming also follows this logic, but some programming shortcuts are implemented. Note that on the MODEL (MO) line for men, the keyword value "PS" is used in the specification of the lambda-X (LX), phi (PH), and theta-delta (TD) matrices. In LISREL, "PS" is used to signify
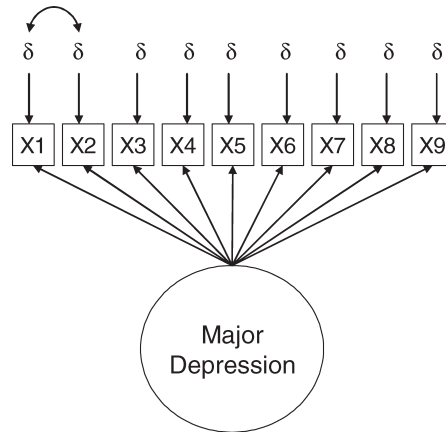
Females: Sample Correlations, Means (*M*), and Standard Deviations (*SD*); *N* = 375

	MDD1	MDD2	MDD3	MDD4	MDD5	MDD6	MDD7	MDD8	MDD9
MDD1	1.000								
MDD2	0.616	1.000							
MDD3	0.315	0.313	1.000						
MDD4	0.349	0.332	0.261	1.000					
MDD5	0.314	0.250	0.270	0.327	1.000				
MDD6	0.418	0.416	0.298	0.328	0.317	1.000			
MDD7	0.322	0.313	0.096	0.117	0.130	0.140	1.000		
MDD8	0.409	0.415	0.189	0.314	0.303	0.281	0.233	1.000	
MDD9	0.318	0.222	0.051	0.115	0.140	0.150	0.217	0.222	1.000
M:	4.184	3.725	1.952	3.589	2.256	3.955	3.869	3.595	1.205
SD:	1.717	2.015	2.096	2.212	2.132	2.005	2.062	2.156	1.791

Males: Sample Correlations, Means (*M*), and Standard Deviations (*SD*); *N* = 375

	MDD1	MDD2	MDD3	MDD4	MDD5	MDD6	MDD7	MDD8	MDD9
MDD1	1.000								
MDD2	0.689	1.000							
MDD3	0.204	0.218	1.000						
MDD4	0.335	0.284	0.315	1.000					
MDD5	0.274	0.320	0.153	0.265	1.000				
MDD6	0.333	0.333	0.221	0.364	0.268	1.000			
MDD7	0.258	0.211	0.114	0.139	0.185	0.132	1.000		
MDD8	0.319	0.346	0.176	0.207	0.231	0.279	0.146	1.000	
MDD9	0.316	0.269	0.111	0.140	0.117	0.131	0.263	0.163	1.000
M:	4.171	3.685	1.739	3.357	2.235	3.661	3.421	3.517	1.259
SD:	1.598	2.018	2.094	2.232	2.108	2.113	2.286	2.174	1.788

FIGURE 7.3. Measurement model of DSM-IV Major Depression. MDD1, depressed mood; MDD2, loss of interest in usual activities; MDD3, weight/appetite change; MDD4, sleep disturbance; MDD5, psychomotor agitation/retardation; MDD6, fatigue/loss of energy; MDD7, feelings of worthlessness/guilt; MDD8, concentration difficulties; MDD9, thoughts of death/suicidality.

TABLE 7.9. Tests of Measurement Invariance and Population Heterogeneity of DSM-IV Major Depressive Disorder in Men and Women

	χ^2	df	χ^2_{diff}	Δdf	RMSEA (90% CI)	CFit[a]	SRMR	CFI	TLI
Single Group Solutions									
Men ($n = 375$)	45.96**	26			.045 (.022–.066)	.62	.041	.97	0.96
Women ($n = 375$)	52.95**	26			.053 (.032–.073)	.39	.044	.96	0.95
Measurement Invariance									
Equal form	98.91***	52			.049 (.034–.064)		.043	.96	0.95
Equal factor loadings	102.84***	60	3.93	8	.043 (.029–.056)		.045	.96	0.96
Equal indicator intercepts	115.31***	68	12.47	8	.043 (.029–.056)		.045	.96	0.96
Equal indicator error variances	125.02***	77	9.71	9	.041 (.027–.053)		.048	.96	0.97
Population Heterogeneity									
Equal factor variance	125.81***	78	0.79	1	.040 (.027–.053)		.050	.96	0.97
Equal latent mean	127.73***	79	1.92	1	.041 (.027–.053)		.052	.96	0.97

Note. N = 750. χ^2_{diff}, nested χ^2 difference; RMSEA, root mean square error of approximation; 90% CI, 90% confidence interval for RMSEA; CFit, test of close fit (probability RMSEA ≤ .05); SRMR, standardized root mean square residual; CFI, comparative fit index; TLI, Tucker–Lewis Index.
[a]Mplus 3.11 does not provide the test of close fit in multiple-groups solutions.
** p < .01, *** p < .001.

273

TABLE 7.10. Computer Syntax (LISREL, Mplus, EQS, Amos) for Equal Form Multiple-Groups Model of Major Depression

LISREL 8.72

```
TITLE LISREL PROGRAM FOR EQUAL FORM OF MAJOR DEPRESSION (FEMALES)
DA NG=2 NI=9 NO=375 MA=CM                                    ! NOTE: NG = 2
LA
M1 M2 M3 M4 M5 M6 M7 M8 M9
KM
1.000
0.616  1.000
0.315  0.313  1.000
0.349  0.332  0.261  1.000
0.314  0.250  0.270  0.327  1.000
0.418  0.416  0.298  0.328  0.317  1.000
0.322  0.313  0.096  0.117  0.130  0.140  1.000
0.409  0.415  0.189  0.314  0.303  0.281  0.233  1.000
0.318  0.222  0.051  0.115  0.140  0.150  0.217  0.222  1.000
SD
1.717 2.015 2.096 2.212 2.132 2.005 2.062 2.156 1.791
MO NX=9 NK=1 PH=SY,FR LX=FU,FR TD=SY,FR
LK
DEPRESS
PA LX
0
1
1
1
1
1
1
1
1
VA 1.0 LX(1,1)                      ! SET THE METRIC OF THE LATENT VARIABLE
PA TD
1
1 1
0 0 1
0 0 0 1
0 0 0 0 1
0 0 0 0 0 1
0 0 0 0 0 0 1
0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 1
OU ME=ML RS MI SC AD=OFF IT=100 ND=4
EQUAL FORM OF MAJOR DEPRESSION (MALES)       ! SEPARATE TITLE FOR MALES
DA NI=9 NO=375 MA=CM
LA
M1 M2 M3 M4 M5 M6 M7 M8 M9
KM
1.000
0.689  1.000
```

(cont.)

274

TABLE 7.10. *(cont.)*

```
0.204  0.218  1.000
0.335  0.284  0.315  1.000
0.274  0.320  0.153  0.265  1.000
0.333  0.333  0.221  0.364  0.268  1.000
0.258  0.211  0.114  0.139  0.185  0.132  1.000
0.319  0.346  0.176  0.207  0.231  0.279  0.146  1.000
0.316  0.269  0.111  0.140  0.117  0.131  0.263  0.163  1.000
SD
1.598  2.018  2.094  2.232  2.108  2.113  2.286  2.174  1.788
MO NX=9 NK=1 PH=PS LX=PS TD=PS                    ! PROGRAMMING SHORTCUTS
OU ME=ML RS MI SC AD=OFF IT=100 ND=4
```

Mplus 3.11

```
TITLE:       MPLUS PROGRAM FOR EQUAL FORM OF MAJOR DEPRESSION
DATA:        FILE IS "C:\MDDALL.DAT";      ! DATA READ FROM RAW DATA FILE
VARIABLE:    NAMES ARE SUBJ SEX M1-M9;
             USEVAR ARE M1-M9;
             GROUPING IS SEX (0=FEMALE 1=MALE); ! SPECIFY GROUPING FACTOR
ANALYSIS:    ESTIMATOR=ML;
MODEL:       DEPRESS BY M1-M9;
             M1 WITH M2;
MODEL MALE:  DEPRESS BY M2-M9;                ! FREELY ESTIMATE PARAMETERS
             M1 WITH M2;                      ! IN MEN AS WELL AS IN WOMEN
OUTPUT:      SAMPSTAT MODINDICES(10.00) STAND RESIDUAL;
```

EQS 5.7b

```
/TITLE
 EQS SYNTAX FOR EQUAL FORM OF MAJOR DEPRESSION (FEMALES)
/SPECIFICATIONS
 CASES=375; VAR=9; ME=ML; MA=COR; ANALYSIS=COV; GROUPS=2;
                                                     ! NOTE: GROUPS=2
/MATRIX
1.000
0.616  1.000
0.315  0.313  1.000
0.349  0.332  0.261  1.000
0.314  0.250  0.270  0.327  1.000
0.418  0.416  0.298  0.328  0.317  1.000
0.322  0.313  0.096  0.117  0.130  0.140  1.000
0.409  0.415  0.189  0.314  0.303  0.281  0.233  1.000
0.318  0.222  0.051  0.115  0.140  0.150  0.217  0.222  1.000
/STANDARD DEVIATIONS
1.717 2.015 2.096 2.212 2.132 2.005 2.062 2.156 1.791
/LABELS
 V1=depmood; V2=anhedon; V3=weight; V4=sleep; V5=motor; V6=fatigue;
 V7=guilt; V8=concent; V9=suicide;
 F1 = DEPRESS;
/EQUATIONS
 V1 = F1+E1;
 V2 = *F1+E2;
```

TABLE 7.10. *(cont.)*

```
 V3 = *F1+E3;
 V4 = *F1+E4;
 V5 = *F1+E5;
 V6 = *F1+E6;
 V7 = *F1+E7;
 V8 = *F1+E8;
 V9 = *F1+E9;
/VARIANCES
 E1 TO E9= *;
 F1 = *;
/COVARIANCES
 E1,E2 = *;
/END
/TITLE
 EQUAL FORM OF MAJOR DEPRESSION (MALES)
/SPECIFICATIONS
 CASES=375; VAR=9; ME=ML; MA=COR; ANALYSIS=COV;
/MATRIX
1.000
0.689  1.000
0.204  0.218  1.000
0.335  0.284  0.315  1.000
0.274  0.320  0.153  0.265  1.000
0.333  0.333  0.221  0.364  0.268  1.000
0.258  0.211  0.114  0.139  0.185  0.132  1.000
0.319  0.346  0.176  0.207  0.231  0.279  0.146  1.000
0.316  0.269  0.111  0.140  0.117  0.131  0.263  0.163  1.000
/STANDARD DEVIATIONS
1.598  2.018  2.094  2.232  2.108  2.113  2.286  2.174  1.788
/LABELS
 V1=depmood; V2=anhedon; V3=weight; V4=sleep; V5=motor; V6=fatigue;
 V7=guilt; V8=concent; V9=suicide;
 F1 = DEPRESS;
/EQUATIONS
 V1 = F1+E1;
 V2 = *F1+E2;
 V3 = *F1+E3;
 V4 = *F1+E4;
 V5 = *F1+E5;
 V6 = *F1+E6;
 V7 = *F1+E7;
 V8 = *F1+E8;
 V9 = *F1+E9;
/VARIANCES
 E1 TO E9= *;
 F1 = *;
/COVARIANCES
 E1,E2 = *;
/PRINT
 fit=all;
/LMTEST
/END
```

(cont.)

TABLE 7.10. *(cont.)*

Amos Basic 5.0

```
' Example of Multi-Grp Equal Form in Amos 5.0

Sub Main ()
Dim sem As New AmosEngine

sem.TextOutput
sem.Standardized
sem.Smc

sem.BeginGroup "DepFEM.txt"
sem.Structure "m1 <- DEPRESS (1)"
sem.Structure "m2 <- DEPRESS"
sem.Structure "m3 <- DEPRESS"
sem.Structure "m4 <- DEPRESS"
sem.Structure "m5 <- DEPRESS"
sem.Structure "m6 <- DEPRESS"
sem.Structure "m7 <- DEPRESS"
sem.Structure "m8 <- DEPRESS"
sem.Structure "m9 <- DEPRESS"
sem.Structure "m1 <- E1 (1)"
sem.Structure "m2 <- E2 (1)"
sem.Structure "m3 <- E3 (1)"
sem.Structure "m4 <- E4 (1)"
sem.Structure "m5 <- E5 (1)"
sem.Structure "m6 <- E6 (1)"
sem.Structure "m7 <- E7 (1)"
sem.Structure "m8 <- E8 (1)"
sem.Structure "m9 <- E9 (1)"
sem.Structure "E1 <-> E2"

sem.BeginGroup "DepMALE.txt"
sem.Structure "m1 <- DEPRESS (1)"
sem.Structure "m2 <- DEPRESS"
sem.Structure "m3 <- DEPRESS"
sem.Structure "m4 <- DEPRESS"
sem.Structure "m5 <- DEPRESS"
sem.Structure "m6 <- DEPRESS"
sem.Structure "m7 <- DEPRESS"
sem.Structure "m8 <- DEPRESS"
sem.Structure "m9 <- DEPRESS"
sem.Structure "m1 <- E1 (1)"
sem.Structure "m2 <- E2 (1)"
sem.Structure "m3 <- E3 (1)"
sem.Structure "m4 <- E4 (1)"
sem.Structure "m5 <- E5 (1)"
sem.Structure "m6 <- E6 (1)"
sem.Structure "m7 <- E7 (1)"
sem.Structure "m8 <- E8 (1)"
sem.Structure "m9 <- E9 (1)"
sem.Structure "E1 <-> E2"

End Sub
```

the same *pattern* and *starting* values of fixed and freed parameters of the corresponding matrix of the prior group; this is not to be confused with *keyword* "PS," which would appear to the left of the equals sign in the model specification of latent-Y variable variances and covariances. Thus, because no invariance constraints are placed in the equal form solution, the PS keyword value allows the user to inform LISREL that that the parameter specification for the measurement model for men and women are identical, without having to repeat the syntax previously written for women (i.e., pattern matrices, marker indicators).

The Mplus example in Table 7.10 runs the analysis on a raw data file instead of variance–covariance matrices (although multiple-groups CFA can also be conducted in Mplus by reading separate input matrices). In this case, the data file contains the subject identification number (SUBJ) and dummy code for sex (SEX: 0 = female, 1 = male), in addition to the nine clinical ratings of MDD (M1–M9). The USEVARIABLE command (USEVAR) selects out the nine indicators that will be used in the CFA. The GROUPING command identifies the variable in the data set that denotes the levels of group that will be used in the multiple-groups CFA. The first MODEL command is the same as would be used to specify the Figure 7.3 measurement model in a single group; in this instance, the first level of GROUP (females) identified by the GROUPING command. Recall that an Mplus default is to fix the first indicator listed to load on a latent factor as the marker indicator; thus, the M1 indicator has been automatically set to be the marker indicator. In multiple-groups analysis, Mplus holds some measurement parameters to equality across groups by default; specifically, the factor loadings as well as intercepts if indicator means are included in the model. Because of the overly stringent nature of this restriction, Mplus does not hold indicator residual variances and covariances to equality by default. Moreover, all structural parameters (factor variances, covariances, latent means) are freely estimated in all groups by default. Thus, because the current model is testing for equal form, the Mplus default for holding factor loadings to equality must be overridden. This is accomplished by the MODEL MALE: command, in which all parameters listed after this keyword are freely estimated in men's solution (however, note that the indicator list omits M1 because it was previously fixed to 1.0 to serve as the marker indicator). Note that while additional programming is necessary to freely estimate that factor loadings in men's solution (i.e., DEPRESS BY M2–M9), the line for correlated residuals between M1 and M2 (i.e., M1 with M2) is included only for clarity (it is redundant with the Mplus default of freely estimating residual variances and covariances in all groups).

As shown in Table 7.9, this solution provides an acceptable fit to the data. This solution will serve as the baseline model for subsequent tests of measurement invariance and population heterogeneity. The parameter estimates for each group are presented in Table 7.11. Inspection of Table 7.9 shows that the *df* and model χ^2 of the equal form solution equal the sum of the *df*s and model χ^2s of the CFAs run separately for men and women; for example, $\chi^2 = 98.91 = 45.96 + 53.95$. Although multiple-groups solutions can be evaluated when the size of the groups vary, if the group sizes differ markedly, interpretation of the analysis may be more complex. This is because many aspects of the CFA are influenced by (sensitive to) sample size. For instance, recall that model χ^2 is calculated as $F_{ML}(N - 1)$. Consider the scenario where the fit function value is the same in two groups (i.e., $F_{ML1} = F_{ML2}$), but the size of groups differ considerably (e.g., $n_1 = 1,000$, $n_2 = 500$). Thus, although the discrepancies between the observed and predicted covariance matrices are the same in both groups, the model χ^2s of the groups will differ greatly, and Group 1 will contribute considerably more to the equal form χ^2 than Group 2. Specifically, in this contrived example, Group 1 will contribute two times as much to the overall χ^2 than Group 2. All other aspects of the CFA model that are based on χ^2 (e.g., overall fit statistics such as the CFI; modification indices) or are influenced by sample size (e.g., standard errors, power to detect parameter estimates as significantly different from zero, standardized residuals) will also be differentially impacted by the unbalanced group sizes. Thus, although it is permissible to conduct multiple-groups CFA with unequal sample sizes, it is preferable for the sizes of the groups to be as balanced as possible. In instances where the group *n*s differ considerably, the researcher must be mindful of this issue when interpreting the results.

The next analysis evaluated whether the factor loadings (unstandardized) of the MDD indicators were equivalent in men and women. The test of equal factor loadings is a critical test in multiple-groups CFA. In tandem with other aspects of measurement invariance evaluation (e.g., equal form), this test determines whether the measures have the same meaning and structure for different groups of respondents. Moreover, this test establishes the suitability of other group comparisons that may be of substantive interest (e.g., group equality of factor variances, factor means, or regressive paths among latent variables). In the current data set, the equal factor loadings model had an overall good fit to the data and did not significantly degrade fit relative to the equal form solution, $\chi^2_{diff}(8) = 3.93$, *ns* (critical value of $\chi^2 = 15.51$, *df* = 8, $\alpha = .05$). The difference in degrees of freedom (*df* = 8) corresponds to the eight factor loadings (M2–M9) that were freely estimated in both groups in the previous analysis. Because M1

TABLE 7.11. Parameter Estimates (Mplus) from the Equal Form Measurement
Model of Major Depression in Men and Women

```
MODEL RESULTS
```

	Estimates	S.E.	Est./S.E.	Std	StdYX
Group FEMALE					
DEPRESS BY					
M1	1.000	0.000	0.000	1.250	0.729
M2	1.107	0.086	12.907	1.384	0.688
M3	0.729	0.101	7.221	0.911	0.435
M4	0.912	0.108	8.406	1.140	0.516
M5	0.812	0.104	7.845	1.016	0.477
M6	0.924	0.100	9.240	1.155	0.577
M7	0.611	0.098	6.220	0.764	0.371
M8	0.979	0.107	9.131	1.224	0.569
M9	0.484	0.085	5.707	0.606	0.339
M1 WITH					
M2	0.394	0.147	2.688	0.394	0.114
Variances					
DEPRESS	1.563	0.224	6.991	1.000	1.000
Residual Variances					
M1	1.376	0.155	8.856	1.376	0.468
M2	2.133	0.223	9.579	2.133	0.527
M3	3.551	0.277	12.837	3.551	0.810
M4	3.583	0.290	12.351	3.583	0.734
M5	3.501	0.278	12.609	3.501	0.772
M6	2.676	0.226	11.822	2.676	0.667
M7	3.658	0.279	13.113	3.658	0.862
M8	3.137	0.264	11.904	3.137	0.677
M9	2.831	0.214	13.223	2.831	0.885
Group MALE					
DEPRESS BY					
M1	1.000	0.000	0.000	1.024	0.642
M2	1.236	0.098	12.580	1.266	0.628
M3	0.786	0.133	5.911	0.805	0.385
M4	1.166	0.152	7.656	1.193	0.535
M5	0.959	0.139	6.915	0.982	0.466
M6	1.132	0.145	7.790	1.159	0.549
M7	0.766	0.143	5.361	0.784	0.344
M8	1.019	0.144	7.075	1.043	0.480
M9	0.632	0.113	5.617	0.647	0.362
M1 WITH					
M2	0.920	0.160	5.743	0.920	0.286

(cont.)

TABLE 7.11. *(cont.)*

Variances					
DEPRESS	1.048	0.183	5.719	1.000	1.000
Residual Variances					
M1	1.499	0.152	9.888	1.499	0.588
M2	2.459	0.244	10.084	2.459	0.605
M3	3.727	0.290	12.830	3.727	0.852
M4	3.547	0.304	11.671	3.547	0.713
M5	3.467	0.282	12.304	3.467	0.783
M6	3.111	0.270	11.516	3.111	0.699
M7	4.599	0.353	13.030	4.599	0.882
M8	3.626	0.297	12.192	3.626	0.769
M9	2.770	0.214	12.943	2.770	0.869

R-SQUARE

Group FEMALE

Observed
Variable   R-Square

Variable	R-Square
M1	0.532
M2	0.473
M3	0.190
M4	0.266
M5	0.228
M6	0.333
M7	0.138
M8	0.323
M9	0.115

Group MALE

Observed
Variable   R-Square

Variable	R-Square
M1	0.412
M2	0.395
M3	0.148
M4	0.287
M5	0.217
M6	0.301
M7	0.118
M8	0.231
M9	0.131

Note. N = 750.

was fixed to 1.0 in both groups to serve as the marker indicator, this measure is not involved in the equality constraints. Incidentally, note in Table 7.9 that the equal factor loading solution produced a slight improvement in the parsimony goodness-of-fit indices as compared with the equal form solution; for example, RMSEA = .043 versus .049 in the equal factor loading and equal form solution, respectively. This is due to the gain in degrees of freedom (60 vs. 52), coupled with the trivial change in model χ^2 (102.84 vs. 98.91) associated with reproducing the observed covariance matrices with fewer freely estimated parameters (i.e., increased model parsimony via constraint of previously free parameters to equality).

Because the constraint of equal factor loadings did not significantly degrade the fit of the solution, it can be concluded that the indicators evidence comparable relationships to the latent construct of Major Depression in men and women. Figure 7.4 graphically illustrates various forms of measurement (non)invariance with respect to factor loadings and indicator intercepts. Although the equality of intercepts has yet to be evaluated, the result of invariant factor loadings is depicted by Figures 7.4A and 7.4B, which show parallel regression slopes for Groups 1 and 2; in other words, a unit change in the underlying dimension (ξ, or Major Depression) is associated with statistically equivalent change in the observed measure (X2, or the indicator of loss of interest in usual activities) in both groups (men and women). However, because the intercepts have not been evaluated, it cannot be concluded that men and women would evidence equivalent observed *scores* on an indicator at a given level of the latent factor (as is shown in Figure 7.4A where both the loading and intercept of an indicator, X2, is equivalent between groups).

The previous analyses were based on covariance structures. To examine the between-group equality of indicator intercepts, the means of the indicators must be input to the analysis along with the indicators' variances and covariances. As in the invariance evaluation of longitudinal measures, the analysis of mean structures poses additional identification issues. In the case of this two-group analysis, there are 18 indicator means (9 for men, 9 for women) but potentially 20 freed parameters of the mean structure solution (18 intercepts, 2 latent means). Moreover, latent variables must be assigned an origin in addition to a metric. Thus, as with the longitudinal invariance example, the mean structure component of the multiple-groups solution is underidentified in the absence of additional restrictions. In addition to holding the indicator intercepts to equality across groups in the measurement invariance solution, identification can be accomplished by fixing the origin (mean) of the latent variable(s) in
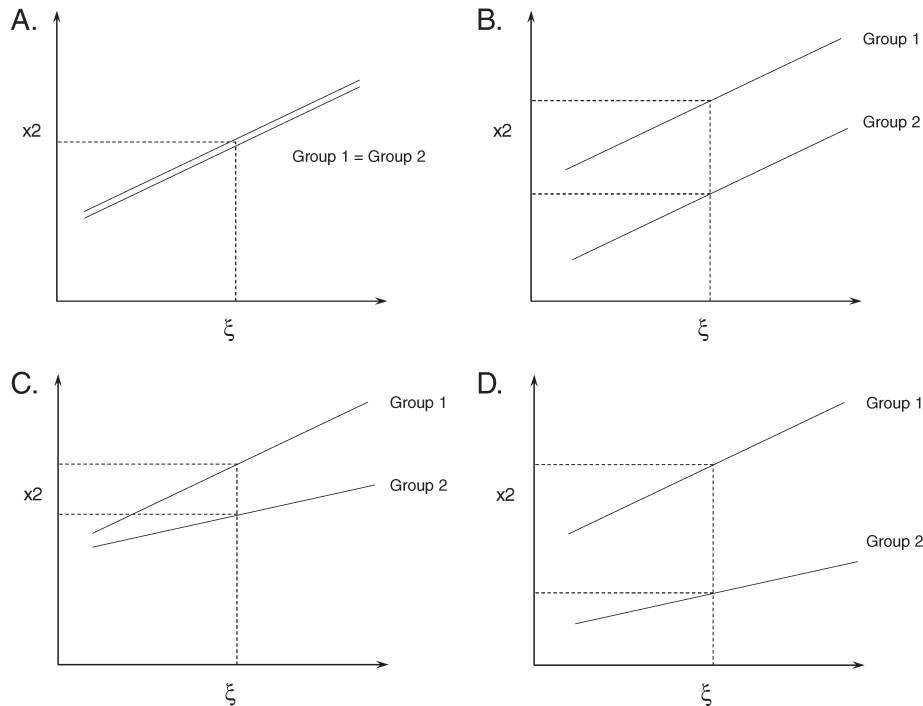
FIGURE 7.4. Graphical depictions of various forms of measurement invariance and noninvariance. (A) Equal loadings and intercepts; (B) equal loadings, unequal intercepts; (C) unequal loadings, equal intercepts; (D) unequal loadings and intercepts.

one group to zero. The group whose latent mean(s) have been fixed to zero becomes the reference group. From a statistical standpoint, selection of the reference group is arbitrary, although this choice might be guided by substantive/interpretative considerations (e.g., reference group = participants who did not receive an intervention/experimental manipulation). The latent means in the remaining groups are freely estimated, but these parameter estimates represent deviations from the reference group's latent mean. For example, if Group 2's latent $M = 1.4$, this is indicates that, on average, this group scores 1.4 units higher than the reference group on the latent dimension, based on the metric of the marker indicator. The rationale of this approach to model specification is that because the measurement intercepts are constrained to equality across groups (to test for intercept invariance), the latent factors have an arbitrary origin (mean). Thus,

latent means are not estimated in the absolute sense but instead reflect average differences in the level of the construct across groups.

Table 7.12 presents the LISREL, Mplus, Amos, and EQS syntax for the invariant intercepts analysis (also holding the factor loadings to equality). In each program, women are used as the reference group by fixing their latent factor mean to zero and the latent mean for men is freely estimated. In the LISREL program, note that the indicator means have been added to the input matrix (under the ME heading), and tau-X (TX) and kappa (KA) matrices have been included on the Model line (MO) for the estimation of indicator intercepts and the latent factor mean, respectively. For females, the LISREL programming looks the same as a typical CFA analysis, except for the TX and KA commands. On the Model line, the command TX=FR informs LISREL that the vector of intercepts are to be freely estimated. The KA=FI command fixes the latent mean of Depression to zero in this group. The remaining salient aspects of the LISREL programming are found on the Model line for males. The factor loadings and indicator intercepts are constrained to equality between groups by the commands LX=IN and TX=IN, respectively (IN = constrain parameters to be invariant with those of the preceding group). Because indicator error variances and factor variances are not held to equality in this analysis, the TD=PS and PH=PS commands are used to inform LISREL that these parameters may be freely estimated using the same pattern matrix and starting values used in the prior group. Finally, the KA=FR command allows the latent mean of Depression to be freely estimated in males. As noted earlier, this estimate will reflect a deviation (difference) of males' latent mean relative to the latent mean of females.

Most of the aforementioned model specifications are handled by default in Mplus (Table 7.12). The syntax now includes the subcommand TYPE=MEANSTRUCTURE to inform Mplus that the analysis will entail both covariance and mean structures. Accordingly, the indicator means must be included in the input matrix, or they can be calculated in Mplus by inputting a raw data file as in the Table 7.12 example. The only syntax provided for men is "M1 with M2" to freely estimate the correlated error of these indicators (which in fact is redundant with another Mplus default of freely estimating residual variances and covariances in all groups). By default, the factor loadings and indicator intercepts are held to equality in Mplus.

In EQS, a key programming change is that the /EQUATIONS section now includes the intercept parameters, which are freely estimated in both groups (e.g., V1 = *V999 + F1 + E1, where *V999 is the freely estimated

TABLE 7.12. Computer Syntax (LISREL, Mplus, EQS, Amos) for Equal Factor Loadings and Equal Indicator Intercepts Multiple-Groups Model of Major Depression

LISREL 8.72

```
TITLE LISREL PROGRAM FOR EQUAL LOADINGS AND INTERCEPTS OF MDD (FEMALES)
DA NG=2 NI=9 NO=375 MA=CM                                    ! NOTE: NG = 2
LA
M1 M2 M3 M4 M5 M6 M7 M8 M9
KM
<Insert correlation matrix for Females from Figure 7.3>
ME
4.184  3.725  1.952  3.589  2.256  3.955  3.869  3.595  1.205
SD
1.717  2.015  2.096  2.212  2.132  2.005  2.062  2.156  1.791
MO NX=9 NK=1 PH=SY,FR LX=FU,FR TD=SY,FR TX=FR KA=FI
LK
DEPRESS
PA LX
0
1
1
1
1
1
1
1
1
VA 1.0 LX(1,1)                        ! SET THE METRIC OF THE LATENT VARIABLE
PA TD
1
1 1
0 0 1
0 0 0 1
0 0 0 0 1
0 0 0 0 0 1
0 0 0 0 0 0 1
0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 1
OU ME=ML RS MI SC AD=OFF IT=100 ND=4
EQUAL LOADINGS AND INTERCEPTS OF MDD (MALES)  ! SEPARATE TITLE FOR MALES
DA NI=9 NO=375 MA=CM
LA
M1 M2 M3 M4 M5 M6 M7 M8 M9
KM
<Insert correlation matrix for Males from Figure 7.3>
ME
4.171 3.685 1.739 3.357 2.235 3.661 3.421 3.517 1.259
SD
1.598 2.018 2.094 2.232 2.108 2.113 2.286 2.174 1.788
MO NX=9 NK=1 PH=PS LX=IN TD=PS TX=IN KA=FR    ! PROGRAMMING SHORTCUTS
OU ME=ML RS MI SC AD=OFF IT=100 ND=4
```

(cont.)

TABLE 7.12. *(cont.)*

Mplus 3.11

```
TITLE:       MPLUS PROGRAM FOR EQUAL LOADINGS AND INTERCEPTS OF MDD
DATA:        FILE IS "C:\MDDALL.DAT";
VARIABLE:    NAMES ARE ADIS SEX M1-M9;
             USEVAR ARE M1-M9;
             GROUPING IS SEX (0=FEMALE 1=MALE);
ANALYSIS:    ESTIMATOR=ML;
             TYPE=MEANSTRUCTURE;
MODEL:       DEPRESS BY M1-M9;
             M1 WITH M2;
MODEL MALE:  M1 WITH M2;  ! ALL MSMT PARAMETERS HELD EQUAL BY DEFAULT
OUTPUT:      SAMPSTAT MODINDICES(10.00) STAND RESIDUAL;
```

EQS 5.7b

```
/TITLE
 EQS SYNTAX FOR GENDER INVARIANCE OF MAJOR DEPRESSION (FEMALES)
/SPECIFICATIONS
 CASES=375; VAR=9; ME=ML; MA=COR; ANALYSIS=MOM; GROUPS=2;
/MATRIX
<Insert correlation matrix for Females from Figure 7.3>
/MEANS
4.184 3.725 1.952 3.589 2.256 3.955 3.869 3.595 1.205
/STANDARD DEVIATIONS
1.717 2.015 2.096 2.212 2.132 2.005 2.062 2.156 1.791
/LABELS
 V1=depmood; V2=anhedon; V3=weight; V4=sleep; V5=motor; V6=fatigue;
 V7=guilt; V8=concent; V9=suicide;
 F1 = DEPRESS;
/EQUATIONS
 V1 = *V999 + F1 + E1;
 V2 = *V999 + *F1 + E2;
 V3 = *V999 + *F1 + E3;
 V4 = *V999 + *F1 + E4;
 V5 = *V999 + *F1 + E5;
 V6 = *V999 + *F1 + E6;
 V7 = *V999 + *F1 + E7;
 V8 = *V999 + *F1 + E8;
 V9 = *V999 + *F1 + E9;
 F1 = 0.0V999 + D1;                         ! FEMALES ARE REFERENCE GROUP
/VARIANCES
 E1 TO E9= *;
 D1 = *;
/COVARIANCES
 E1,E2 = *;
/END
/TITLE
 EQS SYNTAX FOR GENDER INVARIANCE OF MAJOR DEPRESSION (MALES)
/SPECIFICATIONS
 CASES=375; VAR=9; ME=ML; MA=COR; ANALYSIS=MOM;
/MATRIX
<Insert correlation matrix for Males from Figure 7.3>
```

(cont.)

TABLE 7.12. (cont.)

```
/MEANS
4.171 3.685 1.739 3.357 2.235 3.661 3.421 3.517 1.259
/STANDARD DEVIATIONS
1.598 2.018 2.094 2.232 2.108 2.113 2.286 2.174 1.788
/LABELS
 V1=depmood; V2=anhedon; V3=weight; V4=sleep; V5=motor; V6=fatigue;
 V7=guilt; V8=concent; V9=suicide;
 F1 = DEPRESS;
/EQUATIONS
 V1 = *V999 + F1 + E1;
 V2 = *V999 + *F1 + E2;
 V3 = *V999 + *F1 + E3;
 V4 = *V999 + *F1 + E4;
 V5 = *V999 + *F1 + E5;
 V6 = *V999 + *F1 + E6;
 V7 = *V999 + *F1 + E7;
 V8 = *V999 + *F1 + E8;
 V9 = *V999 + *F1 + E9;
 F1 = *V999 + D1;               ! MALE LATENT MEAN FREELY ESTIMATED
/VARIANCES
 E1 TO E9= *;
 D1 = *;
/COVARIANCES
 E1,E2 = *;
/CONSTRAINTS
 (1,V2,F1)=(2,V2,F1);          ! FACTOR LOADING EQUALITY
 (1,V3,F1)=(2,V3,F1);
 (1,V4,F1)=(2,V4,F1);
 (1,V5,F1)=(2,V5,F1);
 (1,V6,F1)=(2,V6,F1);
 (1,V7,F1)=(2,V7,F1);
 (1,V8,F1)=(2,V8,F1);
 (1,V9,F1)=(2,V9,F1);
 (1,V1,V999)=(2,V1,V999);      ! INDICATOR INTERCEPT EQUALITY
 (1,V2,V999)=(2,V2,V999);
 (1,V3,V999)=(2,V3,V999);
 (1,V4,V999)=(2,V4,V999);
 (1,V5,V999)=(2,V5,V999);
 (1,V6,V999)=(2,V6,V999);
 (1,V7,V999)=(2,V7,V999);
 (1,V8,V999)=(2,V8,V999);
 (1,V9,V999)=(2,V9,V999);
/PRINT
 fit=all;
/LMTEST
/END
```

TABLE 7.12. *(cont.)*

Amos Basic 5.0

```
' Example of Equal Loadings and Intercepts in Amos 5.0

Sub Main ()
Dim sem As New AmosEngine

sem.TextOutput
sem.ModelMeansandIntercepts
sem.Standardized
sem.Smc

sem.BeginGroup "DepFEM.txt"
sem.GroupName "Females"
sem.Structure "m1 = (TX1) + (1) DEPRESS + (1) E1"
sem.Structure "m2 = (TX2) + (lam2) DEPRESS + (1) E2"
sem.Structure "m3 = (TX3) + (lam3) DEPRESS + (1) E3"
sem.Structure "m4 = (TX4) + (lam4) DEPRESS + (1) E4"
sem.Structure "m5 = (TX5) + (lam5) DEPRESS + (1) E5"
sem.Structure "m6 = (TX6) + (lam6) DEPRESS + (1) E6"
sem.Structure "m7 = (TX7) + (lam7) DEPRESS + (1) E7"
sem.Structure "m8 = (TX8) + (lam8) DEPRESS + (1) E8"
sem.Structure "m9 = (TX9) + (lam9) DEPRESS + (1) E9"
sem.Structure "E1 <-> E2"
sem.Mean "DEPRESS", "0"

sem.BeginGroup "DepMALE.txt"
sem.GroupName "Males"
sem.Structure "m1 = (TX1) + (1) DEPRESS + (1) E1"
sem.Structure "m2 = (TX2) + (lam2) DEPRESS + (1) E2"
sem.Structure "m3 = (TX3) + (lam3) DEPRESS + (1) E3"
sem.Structure "m4 = (TX4) + (lam4) DEPRESS + (1) E4"
sem.Structure "m5 = (TX5) + (lam5) DEPRESS + (1) E5"
sem.Structure "m6 = (TX6) + (lam6) DEPRESS + (1) E6"
sem.Structure "m7 = (TX7) + (lam7) DEPRESS + (1) E7"
sem.Structure "m8 = (TX8) + (lam8) DEPRESS + (1) E8"
sem.Structure "m9 = (TX9) + (lam9) DEPRESS + (1) E9"
sem.Structure "E1 <-> E2"
sem.Mean "DEPRESS", "Mn_DEP"

End Sub
```

indicator intercept; see Table 7.12). In females, the mean of the Depression factor (F1) is fixed to zero by the equation F1 = 0.0V999 + D1. "D1" reflects a residual variance (or *disturbance*, D) because, in fact, the analysis of mean structures requires the regression of the latent factors and indicators onto a constant (denoted in EQS as V999; Byrne, 1994). In the /VARIANCES section of the program, this disturbance variance is freely estimated in both groups (D1 = *), along with the indicator error variances (E1 TO E9 = *). In males, the only programming change is that the latent mean of Depression is freely estimated, F1 = *V999 + D1. The factor loadings and indicator intercepts are constrained to equality in the /CONSTRAINTS section of the syntax. For example, the command (1,V1,V999)= (2,V1,V999) informs EQS to hold the intercept of the V1 indicator (depressed mood) in the first group (women) equal to the intercept of the V1 indicator in the second group (men).

In Amos Basic, the mean structure component is brought into the solution with the "sem.ModelMeansandIntercepts" command. As in earlier examples (e.g., Table 7.5), parameters are held to equality by giving them the same names (e.g., TX1 for the intercept of the M1 indicator). The mean of the latent factor for females is fixed to zero by the statement, sem.Mean "DEPRESS", "0."

The equal measurement intercepts model is found to be good fitting and does not result in a significant degradation of fit relative to the equal factor loadings solution, $\chi^2_{diff}(8) = 12.47$, *ns* (see Table 7.9). The gain of 8 degrees of freedom (to a total of *df* = 68) is due to the additional 18 new elements of the input matrices (i.e., the 9 indicator means for men and women) minus the 10 mean structure parameters (9 intercepts held to equality, 1 freely estimated latent mean; for identification purposes, women served as the reference group by fixing their latent mean to zero). Because the factor loadings and indicator intercepts are invariant in men and women, comparison of the groups on the latent mean of Major Depression is interpretable. In men, the unstandardized parameter estimate for the latent mean is –.13 (not shown in the tables), indicating that, on average, men score .13 units below women on the dimension of Major Depression, a difference that is not statistically significant (z = 1.38). This lack of difference is upheld in a subsequent analysis of population heterogeneity (the final analysis in Table 7.9) that constrains the latent means to equality; that is, χ^2_{diff} = 1.92, which is roughly the same as $z = 1.38^2$ (cf. Wald test, Chapter 4).

As mentioned earlier, Figure 7.4 graphically displays different combinations of factor loading and intercept (non)invariance. Figure 7.4A is

graphically consistent with the results of the applied example presented thus far. In this graph, both the loading and the intercept of the indicator are equivalent between groups. Interpreted in isolation of factor loading invariance, the finding of intercept invariance suggests that both groups are expected to have the same observed value of the indicator (X2) when the latent factor (ξ) is zero. However, the result of factor loading invariance *and* intercept invariance can be interpreted as suggesting that, *for any given factor value*, the observed values of the indicator are expected to be statistically equivalent between groups (see dotted line in Figure 7.4A). This concept also upholds the statement made earlier in this chapter that group comparisons of the means of latent factors should be conducted only in the context of factor loading *and* indicator intercept invariance.

The remaining three graphs in Figure 7.4 show that if either the loading or intercept is noninvariant, the observed values of the indicator will differ between groups at a given level of the latent factor (see dotted lines).[2] For example, Figure 7.4B illustrates the situation of an equal factor loading but an unequal intercept. In this graph, although the indicator evidences the same relationship (regression slope) in both groups (i.e., a unit change in the factor is associated with the same amount of change in the indicator in both groups), the groups differ in the location parameter (origin) of the indicator, meaning that all predicted observed scores will differ at various levels of the latent factor. Group 1's predicted scores on the indicator will be higher than Group 2's across all levels of the "true score," suggesting that the indicator is biased. This is an example of *differential item functioning*, a term that is used to describe situations where an item yields a different mean response for the members of different groups with the same value of the underlying attribute (McDonald, 1999). An illustration of differential item functioning is presented in the "MIMIC Models" section of this chapter.

The final aspect of measurement invariance to be tested is the equality of the indicator error variances. As seen in Table 7.9, nested χ^2 evaluation indicates that the residual variances are equivalent in men and women patients, $\chi^2_{\text{diff}}(9) = 9.71$, *ns* (critical value of $\chi^2 = 16.92$, *df* = 9, α = .05). The gain of 9 degrees of freedom corresponds to the 9 residual variances held to equality. Although this example used a real data set, this outcome is rare in applied research data. In fact, most methodologists regard equality of error variances and covariances to be an overly restrictive test that is usually not important to the endeavor of measurement invariance evaluation (e.g., Bentler, 1995; Byrne, 1998). For instance, the illustration in Fig-

ure 7.4A shows that prediction of a group equivalent observed score (X2) by the latent variable model (i.e., factor loadings, indicator intercepts) does not rely on the condition of equal indicator error variances. However, the test of equal indicator errors is relevant in situations where the researcher is interested in determining whether the reliability of an assessment measure is invariant across groups. Nonetheless, in their review of the extant measurement invariance literature, Vandenberg and Lance (2000) observed that applied researchers often mistake the test of invariant indicator errors as a test of equivalent reliabilities. These authors remind us (cf. Cole & Maxwell, 1985) that if a test of invariant indicator error variances is conducted with the intent of testing the equality of reliabilities across groups, this evaluation should be preceded by a test that establishes that the variance of the factor on which the indicators load is invariant. This is because reliability is defined by the proportion of true score variance to total variance (total variance = true score variance + error variance), where true score variance is reflected by the factor variance. This concept is developed further in Chapter 8 where an alterative approach to scale reliability evaluation is described. Ordinarily, this test should be considered as an evaluation of indicator error invariance, *not* as a test of invariant reliability, an endeavor that has minimal substantive importance in most research situations.

The remaining analyses pertain to group comparisons on the structural parameters of the CFA model (i.e., tests of population heterogeneity). As noted earlier, the viability of these comparisons rests on the evaluation of measurement invariance. In other words, it is not useful to compare groups on aspects of the latent factors (factor variances, factor covariances, latent means) without first ascertaining that the latent factors measure the same constructs in the same fashion in each group. Specifically, group comparisons on factor variances are meaningful only if the factor loadings have been found to be invariant. Comparisons of the factor covariances (in CFA models with > 1 latent factor) are meaningful if both the factor loadings and factor variances are invariant. Finally, evaluation of group equality of latent factor means rests on the condition of invariant factor loadings and indicator intercepts. The conceptual logic of these statements should be apparent upon review of the equations related to these parameters (e.g., COV = $r_{1,2}SD_1SD_2$; $SD^2 = \sigma^2$), presented in this and earlier chapters (e.g., Chapter 3).

Evaluation of the equality of a factor variance examines whether the amount of within-group variability (dispersion) of the construct differs across groups. Although crucial to all aspects of invariance evaluation, the

test for equal factor variances best exemplifies why comparisons made by multiple-groups CFA rely on the unstandardized solution. The test of invariant factor variances would be meaningless if the metric of the factor was defined by fixing its variance to 1.0. Because the factor variance (ϕ) is strongly determined by the amount of variance in the marker indicator (i.e., ϕ = marker indicator variance multiplied by the marker indicator's squared completely standardized factor loading), the test of group equality of a factor variance can be regarded as an evaluation of whether the groups drew from similar ranges of the underlying construct to respond to the indicators of that construct. The question addressed by the test of factor variance equality often does not have clear substantive implications in applied research, although such evaluation is needed to establish the suitability of the potentially more interesting test of the invariance of factor covariances (i.e., are the latent variables more strongly related to each other in one group than another?). In the example of the measurement model of Major Depression, the factor variances were found to be equal in men and women, $\chi^2_{diff}(1) = 0.79$, *ns*. Because the example was based on a one-factor measurement model, the invariance evaluation of factor covariances was not relevant.

Table 7.13 presents the LISREL, Mplus, Amos, and EQS syntax for the fully invariant measurement model of Major Depression. In LISREL, all of the equality constraints except for the indicator residuals are made by commands on the Model line for males; that is, equal factor loadings: LX=IN; equal indicator intercepts: TX=IN; equal factor variances: PH=IN; equal latent means: KA=IN. The command TD=IN could also be employed to hold indicator error variances and covariances to equality. In this example, the covariance of the errors of M1 and M2 was allowed to vary freely in males and females to further illustrate the use of the EQ command. Because the indicator error variances, but not the M1–M2 error covariance, are to be held to equality, the EQ command is used. For instance, the statement EQ TD(1,1,1) TD(2,1,1) informs LISREL to hold the error variance of the M1 indicator (δ_{11}) in the first group (females, denoted by the first number "1" in parentheses after TD) equal to the error variance of the M1 indicator in the second group (males, denoted by the number "2" in parentheses after TD).

In Mplus, the equality constraints for indicator error variances and factor variance are placed on the solution using the numbers in parentheses that follow the names of the indicators and latent factor. This overrides the Mplus default of freely estimating these variances in both groups. As noted earlier, the factor loadings and indicator intercepts are held to equal-

TABLE 7.13. Computer Syntax (LISREL, Mplus, EQS, Amos) for Fully Invariant Multiple-Groups Model of Major Depression

LISREL 8.72

```
TITLE LISREL PROGRAM FOR INVARIANCE OF MAJOR DEPRESSION CRITERIA
    (FEMALES)
DA NG=2 NI=9 NO=375 MA=CM
LA
M1 M2 M3 M4 M5 M6 M7 M8 M9
KM
<Insert correlation matrix for Females from Figure 7.3>
ME
4.184 3.725 1.952 3.589 2.256 3.955 3.869 3.595 1.205
SD
1.717 2.015 2.096 2.212 2.132 2.005 2.062 2.156 1.791
MO NX=9 NK=1 PH=SY,FR LX=FU,FR TD=SY,FR TX=FR KA=FI
LK
DEPRESS
PA LX
0
1
1
1
1
1
1
1
1
VA 1.0 LX(1,1)                    ! SET THE METRIC OF THE LATENT VARIABLE
PA TD
1
1 1
0 0 1
0 0 0 1
0 0 0 0 1
0 0 0 0 0 1
0 0 0 0 0 0 1
0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 1
OU ME=ML RS MI SC AD=OFF IT=100 ND=4
INVARIANCE OF MAJOR DEPRESSION CRITERIA (MALES)
DA NI=9 NO=375 MA=CM
LA
M1 M2 M3 M4 M5 M6 M7 M8 M9
KM
<Insert correlation matrix for Males from Figure 7.3>
ME
4.171  3.685  1.739  3.357  2.235  3.661  3.421  3.517  1.259
SD
1.598  2.018  2.094  2.232  2.108  2.113  2.286  2.174  1.788
```

(cont.)

TABLE 7.13. *(cont.)*

```
MO NX=9 NK=1 PH=IN LX=IN TD=SY,FR TX=IN KA=IN ! IN = INVARIANT
PA TD
1
1 1
0 0 1
0 0 0 1
0 0 0 0 1
0 0 0 0 0 1
0 0 0 0 0 0 1
0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 1
EQ TD(1,1,1) TD(2,1,1)                   ! EQUALITY CONTRAINTS FOR
EQ TD(1,2,2) TD(2,2,2)                   ! RESIDUAL VARIANCES
EQ TD(1,3,3) TD(2,3,3)                   !
EQ TD(1,4,4) TD(2,4,4)                   !
EQ TD(1,5,5) TD(2,5,5)                   !
EQ TD(1,6,6) TD(2,6,6)                   !
EQ TD(1,7,7) TD(2,7,7)                   !
EQ TD(1,8,8) TD(2,8,8)                   !
EQ TD(1,9,9) TD(2,9,9)                   !
OU ME=ML RS MI SC AD=OFF IT=100 ND=4
```

Mplus 3.11

```
TITLE:    MPLUS PROGRAM FOR GENDER INVARIANCE OF MAJOR DEPRESSION
DATA:     FILE IS "C:\MDDALL.DAT";
VARIABLE: NAMES ARE SUBJ SEX M1-M9;
          USEVAR ARE M1-M9;
          GROUPING IS SEX (0=FEMALE 1=MALE);
ANALYSIS: ESTIMATOR=ML;
          TYPE=MEANSTRUCTURE;
MODEL:    DEPRESS BY M1-M9;
          M1 WITH M2;
          M1 (1); M2 (2); M3 (3); M4 (4); M5 (5); M6 (6); ! EQUAL ERRORS
          M7 (7); M8 (8); M9 (9);                          ! EQUAL ERRORS
          DEPRESS (10);                         ! EQUAL FACTOR VARIANCE
MODEL MALE: M1 WITH M2;             ! FREELY ESTIMATE ERROR COVARIANCE
          [DEPRESS@0];                          ! CONSTRAINT ON LATENT MEAN
OUTPUT:   SAMPSTAT MODINDICES(10.00) STAND RESIDUAL;
```

EQS 5.7b

```
/TITLE
 EQS SYNTAX FOR GENDER INVARIANCE OF MAJOR DEPRESSION (FEMALES)
/SPECIFICATIONS
 CASES=375; VAR=9; ME=ML; MA=COR; ANALYSIS=MOM; GROUPS=2;
/MATRIX
<Insert correlation matrix for Females from Figure 7.3>
/MEANS
4.184 3.725 1.952 3.589 2.256 3.955 3.869 3.595 1.205
```

(cont.)

TABLE 7.13. *(cont.)*

```
/STANDARD DEVIATIONS
1.717 2.015 2.096 2.212 2.132 2.005 2.062 2.156 1.791
/LABELS
 V1=depmood; V2=anhedon; V3=weight; V4=sleep; V5=motor; V6=fatigue;
 V7=guilt; V8=concent; V9=suicide;
 F1 = DEPRESS;
/EQUATIONS
 V1 = *V999 + F1 + E1;
 V2 = *V999 + *F1 + E2;
 V3 = *V999 + *F1 + E3;
 V4 = *V999 + *F1 + E4;
 V5 = *V999 + *F1 + E5;
 V6 = *V999 + *F1 + E6;
 V7 = *V999 + *F1 + E7;
 V8 = *V999 + *F1 + E8;
 V9 = *V999 + *F1 + E9;
 F1 = *V999 + D1;                   ! FACTOR MEAN FREED
/VARIANCES
 E1 TO E9= *;
 D1 = *;
/COVARIANCES
 E1,E2 = *;
/END
/TITLE
 EQS SYNTAX FOR GENDER INVARIANCE OF MAJOR DEPRESSION (MALES)
/SPECIFICATIONS
 CASES=375; VAR=9; ME=ML; MA=COR; ANALYSIS=MOM;
/MATRIX
<Insert correlation matrix for Males from Figure 7.3>
/MEANS
4.171 3.685 1.739 3.357 2.235 3.661 3.421 3.517 1.259
/STANDARD DEVIATIONS
1.598 2.018 2.094 2.232 2.108 2.113 2.286 2.174 1.788
/LABELS
 V1=depmood; V2=anhedon; V3=weight; V4=sleep; V5=motor; V6=fatigue;
 V7=guilt; V8=concent; V9=suicide;
 F1 = DEPRESS;
/EQUATIONS
 V1 = *V999 + F1 + E1;
 V2 = *V999 + *F1 + E2;
 V3 = *V999 + *F1 + E3;
 V4 = *V999 + *F1 + E4;
 V5 = *V999 + *F1 + E5;
 V6 = *V999 + *F1 + E6;
 V7 = *V999 + *F1 + E7;
 V8 = *V999 + *F1 + E8;
 V9 = *V999 + *F1 + E9;
 F1 = *V999 + D1;
```

(cont.)

TABLE 7.13. *(cont.)*

```
/VARIANCES
 E1 TO E9= *;
 D1 = *;
/COVARIANCES
 E1,E2 = *;
/CONSTRAINTS
 (1,V2,F1)=(2,V2,F1);              ! FACTOR LOADING EQUALITY
 (1,V3,F1)=(2,V3,F1);
 (1,V4,F1)=(2,V4,F1);
 (1,V5,F1)=(2,V5,F1);
 (1,V6,F1)=(2,V6,F1);
 (1,V7,F1)=(2,V7,F1);
 (1,V8,F1)=(2,V8,F1);
 (1,V9,F1)=(2,V9,F1);
 (1,V1,V999)=(2,V1,V999);          ! INDICATOR INTERCEPT EQUALITY
 (1,V2,V999)=(2,V2,V999);
 (1,V3,V999)=(2,V3,V999);
 (1,V4,V999)=(2,V4,V999);
 (1,V5,V999)=(2,V5,V999);
 (1,V6,V999)=(2,V6,V999);
 (1,V7,V999)=(2,V7,V999);
 (1,V8,V999)=(2,V8,V999);
 (1,V9,V999)=(2,V9,V999);
 (1,E1,E1) = (2,E1,E1);            ! EQUAL INDICATOR ERRORS
 (1,E2,E2) = (2,E2,E2);
 (1,E3,E3) = (2,E3,E3);
 (1,E4,E4) = (2,E4,E4);
 (1,E5,E5) = (2,E5,E5);
 (1,E6,E6) = (2,E6,E6);
 (1,E7,E7) = (2,E7,E7);
 (1,E8,E8) = (2,E8,E8);
 (1,E9,E9) = (2,E9,E9);
 (1,D1,D1) = (2,D1,D1);            ! EQUAL FACTOR VARIANCE
 (1,F1,V999) = (2,F1,V999);        ! EQUAL LATENT MEAN
/PRINT
 fit=all;
/LMTEST
/END
```

Amos Basic 5.0

```
' Example of Full Msmt. Invariance in Amos 5.0
Sub Main ()
Dim sem As New AmosEngine
sem.TextOutput
sem.ModelMeansAndIntercepts
sem.Standardized
sem.Smc
```

(cont.)

TABLE 7.13. *(cont.)*

```
sem.BeginGroup "DepFEM.txt"
sem.GroupName "Females"
sem.Structure "m1 = (TX1) + (1) DEPRESS + (1) E1"
sem.Structure "m2 = (TX2) + (lam2) DEPRESS + (1) E2"
sem.Structure "m3 = (TX3) + (lam3) DEPRESS + (1) E3"
sem.Structure "m4 = (TX4) + (lam4) DEPRESS + (1) E4"
sem.Structure "m5 = (TX5) + (lam5) DEPRESS + (1) E5"
sem.Structure "m6 = (TX6) + (lam6) DEPRESS + (1) E6"
sem.Structure "m7 = (TX7) + (lam7) DEPRESS + (1) E7"
sem.Structure "m8 = (TX8) + (lam8) DEPRESS + (1) E8"
sem.Structure "m9 = (TX9) + (lam9) DEPRESS + (1) E9"
sem.Structure "E1 (TD1)"
sem.Structure "E2 (TD2)"
sem.Structure "E3 (TD3)"
sem.Structure "E4 (TD4)"
sem.Structure "E5 (TD5)"
sem.Structure "E6 (TD6)"
sem.Structure "E7 (TD7)"
sem.Structure "E8 (TD8)"
sem.Structure "E9 (TD9)"
sem.Structure "E1 <-> E2"
sem.Structure "DEPRESS (DEP_PHI)"
sem.Mean "DEPRESS", "0"

sem.BeginGroup "DepMALE.txt"
sem.GroupName "Males"
<same syntax under sem.GroupName "Females" command is repeated here>
```

ity in Mplus by default, and thus no additional programming is required unless the user wishes to override these defaults. Finally, the latent mean of males is held equal to females by the command [DEPRESS@0]; that is, the deviation between males' and females' latent mean is constrained to not differ significantly from zero. In EQS, all of the equality constraints are made in the /CONSTRAINTS section of the syntax. In the sections of the program corresponding to males' and females' measurement model, all parameters other than the marker indicator are freely estimated; that is, the syntax is identical for men and women. In the /CONSTRAINTS section that follows, all of these parameters are then held to equality except for the factor loading of the marker indicator and the error covariance of V1 and V2.

The equality constraint on the factor means examines whether groups differ in their levels of the underlying construct. As shown in Table 7.9, this constraint did not significantly degrade the fit of the model, indicating

that men and women outpatients did not differ in their average levels of the underlying dimension of Major Depression, $\chi^2_{\text{diff}}(1) = 1.92$, *ns*. It was noted earlier in this section that this constraint was essentially redundant with the significance test of the freely estimated latent mean of male outpatients ($z = 1.38$, *ns*). This is because only two groups were used in the analysis. Hence, the omnibus test, which is the equality constraint on all factor means (in the current case, $\kappa = 2$), is approximately the same as the significance test of this single parameter; in both, *df* = 1.

In instances involving more than two groups, a significant omnibus test (i.e., a significant increase in χ^2 when the means of a given latent variable are held to equality across groups) is typically followed by post hoc evaluation to determine the nature of the overall effect, along the lines of simple effects testing in ANOVA. If the analysis used three groups, three possible follow-up analyses, in which the factor means of two groups are constrained to equality at a time, could be conducted to identify which groups differed on their latent means (perhaps with a control for experiment-wise error in the multiple comparisons, such as the modified Bonferroni procedure; Jaccard & Wan, 1996). Although the similarities to ANOVA and the independent *t*-test are apparent, it should be reemphasized that the CFA-based approach to group mean comparison has several advantages over these more traditional methods. A key strength of the CFA approach is that it establishes whether the group comparisons are appropriate. Although traditional analyses simply assume this is the case, the multiple-groups CFA may reveal considerable measurement noninvariance (e.g., unequal form, a preponderance of noninvariant factor loadings and indicator intercepts) that contraindicates groups comparison on the latent factor mean. As will be discussed shortly, in some instances it may possible to proceed with such comparisons in the context of partial measurement invariance. Moreover, group comparisons have more precision and statistical power in CFA because the structural parameters (factor means, variances, covariances) have been adjusted for measurement error. Traditional tests such as ANOVA assume perfect reliability.

A decade ago, it was rare to see a CFA analysis of mean structures in the applied research literature, primarily because the syntax specification of such models was very complex in early versions of latent variable software packages. At this writing, such analyses have become more common in the literature, although the technique continues to be underutilized somewhat; for instance, many psychometric studies continue to rely exclusively on the analysis of the covariance matrix in the evaluation of psychological testing instruments. As shown in the prior example, the

analysis of mean structures is straightforward in the latest releases of software programs such as LISREL, Mplus, Amos, and EQS. Given its advantages over traditional statistics (e.g., ANOVA) and its capability to evaluate other important substantive questions (e.g., psychometric issues such as differential item functioning), investigators are encouraged to incorporate mean structures in their applied CFA-based research.

Selected Issues in Single- and Multiple-Groups CFA Invariance Evaluation

In this section, three miscellaneous issues in CFA invariance evaluation are discussed. Although each of these issues can impact the conduct and interpretation of CFA invariance evaluation considerably, the extant SEM literature has provided minimal guidance on how these issues should be managed in applied research. Nevertheless, it is important that the researcher be aware of these issues, along with some tentative remedial strategies offered to date.

Partial Measurement Invariance

In their review of the applied CFA literature, Byrne et al. (1989) observed a widespread belief among researchers that when evidence of noninvariant measurement parameters (e.g., unequal factor loadings) is encountered, further testing of measurement invariance and population heterogeneity is not possible. As noted earlier, an omnibus test of invariance is conducted by placing equality constraints on a family of unstandardized parameters of the CFA model (e.g., factor loadings) and determining whether these restrictions produce a significant increase in model χ^2. If a significant increase in χ^2 is observed, then the null hypothesis is rejected (e.g., the factor loadings are not equal across groups). Fit diagnostics (e.g., modification indices, expected parameter change values) can assist the researcher in identifying the parameters that are noninvariant across groups, across time, and so on. Indeed, a significant omnibus χ^2 should not be interpreted as indicating that all parameters are noninvariant; for instance, this result might be obtained when there is a single noninvariant parameter within a complex measurement model.

Byrne et al. (1989) reminded researchers that, in many instances, invariance evaluation can proceed in the context of *partial measurement invariance*; that is, in CFA models where some but not all of the measurement parameters are equivalent. For ease of illustration, consider a

unidimensional model entailing five indicators (X1–X5, where X1 is specified as the marker indicator), which is tested for invariance between two groups. Given evidence of equal form, the researcher evaluates the between-group equality of factor loadings. The more restricted solution produces a significant increase in χ^2, suggesting that at least one of the factor loadings is noninvariant. Fit diagnostics suggest that the factor loading of X5 is noninvariant (e.g., associated with high modification index; the modification indices for X2, X3, and X4 are below 4.0). This is verified in a respecified multiple-groups solution where all factor loadings other than the loading of X5 (which is freely estimated in both groups) and X1 (which was previously fixed as the marker indicator) are constrained to equality between groups; that is, this respecified model does not produce a significant increase in χ^2 relative to the equal form solution. From a statistical perspective, the invariance evaluation may proceed to examine the equality of other measurement (e.g., indicator intercepts) and structural parameters (e.g., latent means) in context of this partial invariance. The researcher would freely estimate the factor loading of X5 in both groups in subsequent analyses. Indeed, Byrne et al. (1989) note that such analyses may proceed so long as there exists at least one noninvariant parameter other than the marker indicator; for instance, in the current example, invariance evaluation could continue if, say, only the factor loading of X2 was invariant in the two groups. The same logic would apply to other measurement parameters. For example, recall that comparison of latent means is meaningful only if factor loadings and indicator intercepts are invariant. In the present illustration, a between-group comparison of the latent means could be conducted, provided that there are at least partial factor loading and partial intercept invariance; for example, at least one indicator other than the marker indicator has a invariant factor loading and intercept.

The main advantage of this strategy is that it allows the invariance evaluation to proceed after some noninvariant parameters are encountered. This is very helpful in cases where the evaluation of structural parameters is of greatest substantive interest. For instance, consider the situation where the researcher wishes to evaluate a complex longitudinal structural equation modeling involving several latent variables. Of most interest are the structural parameters of this model; that is, the paths reflecting regressions within and among latent variables over time. Before these paths can be evaluated, the researcher must establish longitudinal measurement invariance (e.g., to rule out the spurious influence of temporal change in measurement). Without knowledge of the partial measure-

ment invariance strategy, the researcher might abandon the evaluation of structural parameters if measurement noninvariance is detected (e.g., the factor loadings of a few indicators differ over time). However, assuming a preponderance of invariant measurement parameters, the substantively more important analyses (e.g., cross-lagged effects among latent variables) can be conducted in the context of partial measurement invariance.

The primary disadvantages of partial invariance analysis are the fact that it is exploratory in nature and that it risks capitalizing on chance. These issues may be of greater concern when the purpose of the analysis is psychometric (e.g., evaluation of the measurement invariance of a psycho-logical questionnaire in demographic subgroups). As with other scenarios, psychometric invariance evaluation is often conducted in the absence of explicit hypothesis testing. In fact, although it is counter to the traditions of the scientific method, researchers typically conduct invariance evalua-tion with the hope of retaining the null hypothesis; that is, H_0: all parame-ters are the same across groups, thereby supporting the notion that the testing instrument is unbiased, has equivalent measurement properties, and so forth. Given the number of parameters that are constrained in invariance evaluations (especially when the measurement model contains a large number of indicators), it is possible that some parameters will differ by chance. In addition, the large sample sizes used in CFA have consider-able power to detect small differences as statistically significant, especially because invariance testing relies heavily on χ^2 (see the "Reliance on χ^2" section of this chapter). There are currently no methods of determining or adjusting for such issues (Byrne et al., 1989). Moreover, the guidelines and procedures for relaxing invariance constraints have not been fully devel-oped or studied by SEM methodologists. Vandenberg and Lance (2000) recommend that the partial invariance strategy should not be employed when a large number of indicators are found to be noninvariant. Indeed, although it is statistically possible to proceed when a single indicator other than the marker indicator is invariant, this outcome should prompt the researcher to question the suitability of the measurement model for further invariance testing. Although partial invariance evaluation is a post hoc procedure, this limitation can be allayed to some degree if the researcher can provide a substantively compelling account for the source of non-invariance (e.g., see the example of a MIMIC model later in this chapter). Finally, Byrne et al. (1989) underscore the importance of cross-validation using an independent sample. For example, if the overall sample is large enough, it can be randomly divided into two subsamples. In this strategy, the first sample is used to develop a good-fitting solution where some

parameters are freely estimated (e.g., across groups) when evidence of noninvariance arises (fostered by substantive arguments, it is hoped, for why these parameters differ across groups). The final model is then fit in the second sample to determine its replicability with independent data. Although cross-validation is a compelling method to address the limitations of partial invariance evaluation (and post hoc model revision in general), Byrne et al. (1989) highlighted some practical concerns, such as the feasibility of obtaining more than one sufficiently large sample and the fact that cross-validation is likely to be unsuccessful when multiple parameters are relaxed in the first sample.

Selection of Marker Indicator

Earlier it was suggested that selection of a latent variable's marker indicator should not be taken lightly (see Chapter 4). The selection of marker indicators can also greatly influence measurement and structural invariance evaluation. Difficulties will arise in multiple-groups CFA when the researcher inadvertently selects a marker indicator that is noninvariant across groups. First, the researcher may not detect this noninvariance because the unstandardized factor loadings of the marker indicator are fixed to 1.0 in all groups. Second, subsequent tests of partial invariance may be poor fitting because the unstandardized measurement parameters of the remaining indicators (e.g., factor loadings, intercepts) are impacted by the noninvariant marker indicator. The differences found in the remaining indicators may not reflect "true" differences among groups, but rather may be an artifact of scaling the metric of the latent factor with an indicator that has a different relationship to the factor in two or more groups. Although somewhat cumbersome, one approach to exploring this possibility is to re-run the multiple-groups CFA with different marker indicators. Chueng and Rensvold (1999) have also introduced procedures for addressing this issue, although more research is needed to determine the utility and viability of these approaches.

Reliance on χ^2

As the previous data-based examples illustrate, invariance evaluation relies strongly on the χ^2 statistic. For example, the omnibus test of equality of indicator intercepts across groups is conducted by determining whether the constrained solution (in which the intercepts are held to equal across groups) produces a significant increase in χ^2 relative to a less constrained

model (e.g., an equal factor loadings solution). When a significant degradation in model fit is encountered, procedures to identify noninvariant parameters rely in part on χ^2-based statistics (i.e., modification indices). As discussed in prior chapters, both model χ^2 and modification indices are sensitive to sample size. Researchers have noted that a double standard exists in the SEM literature (e.g., Chueng & Rensvold, 2000; Vandenberg & Lance, 2000). That is, given the limitations of χ^2, investigators are encouraged to use a variety of fit indices to evaluate the overall fit of a CFA solution (e.g., RMSEA, TLI, CFI, SRMR). However, in invariance evaluation, the χ^2 statistic is relied on exclusively to detect differences in more versus less constrained solutions. The reason why χ^2 is the only fit statistic used for this purpose is that its distributional properties are known and thus critical values can be determined at various degrees of freedom. This cannot be done for other fit indices; for example, a more constrained solution may produce an increase in the SRMR, but there is no way of determining at what magnitude this increase is statistically meaningful.

Researchers have begun to recognize and address this issue. Cheung and Rensvold (2000) conducted a large Monte Carlo simulation study to determine whether critical values of other goodness-of-fit statistics (e.g., the CFI) could be identified to reflect the presence/absence of measurement invariance in multiple-groups solutions; for example, does a .01- or .02-point reduction in the CFI reliably indicate the rejection of the null hypothesis that the measurement parameters are the same across groups? Although the authors proposed critical values for three fit statistics, the validity of these proposals awaits further research.

Another problem that the CFA researcher may encounter on occasion is that the omnibus test of invariance is statistically significant (i.e., the constraints result in a significant increase in model χ^2), but fit diagnostics reveal no salient strains with regard to any specific parameter; for instance, modification indices for all constrained parameters close to or below 4.0 (for an example of this outcome in the applied literature, see Campbell-Sills et al., 2004). Again, such a result may be indicative of χ^2's oversensitivity to sample size and the problem of relying exclusively on χ^2 in invariance evaluation. This outcome would suggest that the differences suggested by the statistically significant increase in model χ^2 are trivial and thus have no substantive importance. To verify this conclusion, Byrne et al. (1989) propose that a *sensitivity analysis* can be helpful. Sensitivity analysis is a post hoc method of determining whether the addition of minor parameters (e.g., relaxation of the constraints on parameters that had previously been held to equality) results in a clear change in the major

parameters of the model (e.g., the factor loadings). For example, if the major parameters do not change when minor parameters are included in the model, this could be considered evidence for the robustness of the initial solution. The less parsimonious model may now be statistically equivalent to the baseline solution, but these additional minor parameters are not needed because they do not appreciably impact the major parameters of the solution. However, if the additional parameters significantly alter the major parameters, this would suggest that exclusion of these post hoc parameters (e.g., factor loadings or intercepts that were freely estimated in some or all groups) would lead to biased estimates of the major parameters. However, sensitivity analysis is limited somewhat by its subjective nature (i.e., what constitutes "salient" change in the model's major parameters when post hoc parameters are included or excluded from the solution). To increase the objectivity of this approach, Byrne et al. (1989) suggest correlating the major parameters (e.g., factor loadings) in the initial model with those of the best-fitting post hoc solution. Although coefficients close to 1.00 might provide clear support for the more constrained model and argue against the importance of the additional freed parameters, this conclusion becomes more vague when the correlation is well below 1.00. Moreover, a high correlation may still mask the fact that one or two major parameters differ considerably in the context of a multiple-indicator model where all other parameters are roughly the same. The caveats of partial invariance evaluation should also be considered (i.e., exploratory nature, risk of chance effects, importance of cross-validation).

MIMIC Models (CFA with Covariates)

A less commonly used method of examining invariance in multiple groups entails regressing the latent factors and indicators onto covariates that represent group membership (e.g., Sex: 0 = female, 1 = male). This approach has been referred to as *CFA with covariates* or *MIMIC models* (MIMIC = multiple indicators, multiple causes; Jöreskog & Goldberger, 1975; Muthén, 1989). Unlike multiple-groups CFA, a single input matrix is used in the analysis. The input matrix contains the variances and covariances of both the latent variable indicators and the covariates that denote group membership. Indicator means are not included as input, as might be the case for multiple-groups CFA. The two basic steps of MIMIC modeling are (1) establish a viable CFA (or E/CFA; see Chapter 5) measurement model using the full sample (i.e., collapsing across groups); and (2) add covariate(s) to the model to examine their direct effects on the latent factors

and selected indicators. A significant direct effect of the covariate on the latent factor represents population heterogeneity; that is, the factor means are different at different levels of the covariate (akin to the test of equal latent means in multiple-groups CFA). A significant direct effect of the covariate on an indicator of a latent factor represents measurement noninvariance; that is, when the latent factor is held constant, the means of the indicator are different at different levels of the covariate, thereby pointing to differential item (indicator) functioning (akin to the test of equal indicator intercepts in multiple-groups CFA). Although both aspects of invariance evaluation in MIMIC modeling correspond to the mean structure component of multiple-groups CFA, indicator intercepts and factor means are not estimated in the MIMIC analysis and indicator means are not included in the input matrix. Rather, as in the analysis of covariance structure, these group mean differences are deviations conveyed by the parameter estimates of the direct effects (e.g., Sex $\rightarrow$ Latent Factor) where the means of the indicators and factors are assumed to be zero. For instance, if the unstandardized direct effect of Sex $\rightarrow$ Latent Factor is .75, then the latent means of males and females differ by .75 units, per the metric of the factor's marker indicator.

Unlike multiple-groups CFA, MIMIC models can test only the invariance of indicator intercepts and factor means. Thus, MIMIC models assume that all other measurement and structural parameters (i.e., factor loadings, error variances/covariances, factor variances/covariances) are the same across all levels of the covariates (groups). A primary advantage of MIMIC models is that they usually have smaller sample size requirements than multiple-groups CFA. Whereas multiple-groups CFA entails the simultaneous analysis of two or more measurement models, MIMIC involves a single measurement model and input matrix. For example, in the case where a researcher wishes to conduct an invariance analysis with three groups but has a total sample size of 150, the size of each group may not be sufficient for each within-group CFA using the multiple-groups approach; that is, the analysis would require three separate CFAs, each with an $n = 50$. However, the $N = 150$ may be suitable for a single CFA in which aspects of the measurement model are regressed onto covariates, although sample size considerations must also be brought to bear with regard to the statistical power of the direct effects of the covariates (see Chapter 10).[3]

Another potential strength of MIMIC over multiple-groups CFA arises when there are many groups involved in the comparison. Multiple-groups CFA with three or more groups can be very cumbersome (depending also

on the complexity of the measurement model) because of the number of parameters that must be estimated and held equal across all groups. In addition, post hoc testing is more complex when an omnibus test of the null hypothesis of a given aspect of invariance is rejected. In contrast, MIMIC models are more parsimonious because measurement model parameters are not estimated in each group.

In most applications of MIMIC, the covariate is a nominal variable that represents levels of known groups (e.g., Sex: 0 = female, 1 = male). When the levels of group (k) are three or more, group membership can be reflected as dummy codes, per the procedures found in multiple regression textbooks (e.g., Cohen et al., 2003). If $k = 3$, two ($k - 1$) binary codes are created that identify two of the three levels of the nominal variable, and the remaining level is treated as the reference group that does not receive its own code. Although categorical covariates are typically used in MIMIC models, a normal theory estimator such as ML may still be used if the indicators of the latent factors satisfy asymptotic theory assumptions (cf. Chapter 9). This is because the categorical variables are used as predictors instead of outcomes (cf. multiple regression vs. logistic regression). However, the covariates used in MIMIC models can also be dimensional (e.g., age). The ability to accommodate continuous predictors can be viewed as another potential advantage of MIMIC over multiple-groups CFA, because the latter would necessitate imposing categorical cutoffs on a dimensional variable to form "groups" for the analysis (cf. MacCallum, Zhang, & Preacher, 2002).

Ordinarily, the covariate is assumed to be free of measurement error (i.e., its error variance is fixed to zero). This assumption is reasonable when the covariate represents known groups (e.g., male vs. female). However, the desired amount of measurement error can be modeled in a dimensional covariate using the procedures described in Chapter 4; that is, the unstandardized error of the covariate can be fixed to some non-zero value on the basis of the sample variance estimate and known reliability information (see Eqs. 4.14 and 4.15, Chapter 4).

Unlike other examples of CFA presented in this book thus far, the MIMIC approach reflects a latent-Y (endogenous) variable specification (cf. Figure 3.4, Chapter 3). This is because the latent factors (and their indicators) are specified as outcomes predicted by the covariates (see Figure 7.5). This distinction has minimal impact on syntax programming in programs such as Mplus, Amos, and EQS, which accommodate the latent Y specification "behind the scenes," but does require several alterations if the analysis is conducted using the LISREL matrix programming language.

The MIMIC methodology is illustrated using the model presented in Figure 7.5. In this example, the researcher wishes to examine selected aspects of population heterogeneity (latent factor means) and measurement invariance (indicator intercepts) associated with an established questionnaire measure of social phobia (fear of social situations due to possibility of negative evaluation by others) and agoraphobia (fear of public situations due to possibility of experiencing unexpected panic attacks; this example is loosely based on a psychometric study conducted by Brown et al., 2005). Indicators of the Social Phobia and Agoraphobia constructs are provided in Figure 7.5 (items self-rated on 0–8 scales; higher scores = higher fear). In particular, the researcher expects to find sex differences with respect to the latent construct of Agoraphobia (population heterogeneity) based on the prior clinical literature indicates that women are more inclined than men to respond to unexpected panic attacks with situational fear and avoidance. Moreover, the researcher is concerned that an indicator of Agoraphobia (A3) functions differently for men and women. Specifically, it is anticipated that, regardless of the level of the underlying factor of Agoraphobia, women will score higher on the A3 indicator than men (A3: "walking alone in isolated areas") because of sex differences in the range of activities relating to personal safety; that is, irrespective of the presence/absence of the symptoms of agoraphobia, women are less likely than men to walk alone in isolated areas. To address this question, data were collected from 730 outpatients (365 men, 365 women) who presented for assessment and treatment of various anxiety disorders.

The path diagram and input matrix of this MIMIC model are presented in Figure 7.5. A path from the covariate to the Social Phobia factor has also been specified to test for sex differences on this construct. The covariate (Sex) is represented by a single dummy code (0 = female, 1 = male), and its standard deviation and correlations with the indicators of Agoraphobia and Social Phobia are included in the input matrix. As seen in Figure 7.5, the variances of Social Phobia and Agoraphobia are not estimated in the MIMIC model. Instead, because the model attempts to explain variance in these latent factors by the sex covariate, these parameters are residual variances (E; often referred to as *disturbances*). Also note that the residual variances of Social Phobia and Agoraphobia are specified to be correlated. This specification is justified by the argument that the constructs of Social Phobia and Agoraphobia are not completely orthogonal, and that this overlap can not be fully accounted for by the sex covariate; for example, sex is not a "third variable" responsible for the overlap between the dimensions of Social Phobia and Agoraphobia.
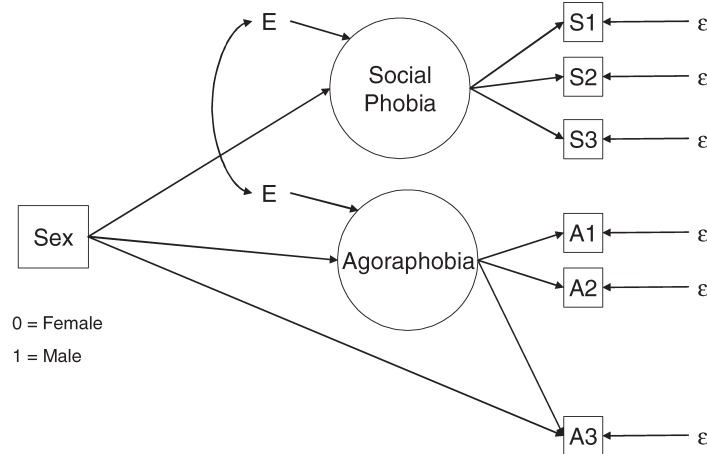
Sample Correlations and Standard Deviations (*SD*s); *N* = 730 (365 males, 365 females)

	S1	S2	S3	A1	A2	A3	Sex
S1	1.000						
S2	0.705	1.000					
S3	0.724	0.646	1.000				
A1	0.213	0.195	0.190	1.000			
A2	0.149	0.142	0.128	0.521	1.000		
A3	0.155	0.162	0.135	0.557	0.479	1.000	
Sex	−0.019	−0.024	−0.029	−0.110	−0.074	−0.291	1.000
SD:	2.260	2.730	2.110	2.320	2.610	2.440	0.500

FIGURE 7.5. MIMIC model of Social Phobia and Agoraphobia. S1, giving a speech; S2, meeting strangers; S3, talking to people; A1, going long distances from home; A2, entering a crowded mall; A3, walking alone in isolated areas (all questionnaire items rated on 0–8 scales, where 0 = no fear and 8 = extreme fear).

The first step is to ensure that the two-factor model of Social Phobia and Agoraphobia is reasonable and good fitting in the full sample (N = 730). In this step, the sex covariate is not included in the CFA and the variances (ϕ_{11}, ϕ_{22}) and covariance (ϕ_{21}) of the two factors are freely estimated (i.e., a typical CFA model is specified). This model provided a good fit to the data, $\chi^2(8)$ = 3.06, p = .93, SRMR = .012, RMSEA = 0.00 (90% CI = 0.00 to 0.01, CFit = 1.00), TLI = 1.005, CFI = 1.00. There are no salient areas of strain in the solution (e.g., no modification indices > 4.0), and all parameter estimates were reasonable and statistically significant (e.g., range of completely standardized factor loadings = .67 to .89; correlation between Social Phobia and Agoraphobia = .28).

Next, the sex covariate is added to the model (see Figure 7.5). Table 7.14 provides Mplus, LISREL, Amos, and EQS syntax for this model specification. As can be seen in Table 7.14, the programming is straightforward in Mplus. Regressing the latent factors of Social Phobia and Agoraphobia, as well as the A3 indicator, onto the sex covariate is accomplished with the "ON" keyword (e.g., "SOCIAL ON SEX"). The correlated residual of Social Phobia and Agoraphobia is specified by the "SOCIAL WITH AGORAPH" syntax. In EQS, the MIMIC portion of the model is represented in the last three lines of the /EQUATIONS command. The line V6 = *F2 + *V7 + E6 represents the fundamental equation that variance of A3 (V6) is to be accounted for by the Agoraphobia latent factor (F2), the sex covariate (V7), and residual variance (E6). The variance of the latent factors (F1, F2) is reproduced by variance explained by the sex covariate and residual variance (e.g., F1 = *V7 + D1). The /VARIANCE commands indicate that all residual variances (indicators: E1–E6; latent factors: D1, D2) and X-variable (V7) variances should be freely estimated. The /COVARIANCE command, D1,D2 = *, indicates that the correlated disturbance (residual) between Social Phobia and Agoraphobia should also be freely estimated. Amos Basic uses similar equation (line)-based programming.

LISREL matrix programming is more complicated (see Table 7.14). Although other programs (e.g., Mplus) rely on the same programming logic exemplified by the following LISREL programming, their software contains convenience features that allow the user to set up the analysis in a less complex manner. This is also true for the SIMPLIS sublanguage of LISREL. First, note that latent-Y programming is used (lambda-Y and theta-epsilon for factor loadings and indicator errors, respectively; psi for factor variances and covariances) because the CFA portion of the MIMIC model is endogenous (cf. Chapter 3). Although Sex is really an X variable, it is treated as a Y variable so that the overall model can be specified in LISREL. An "all Y" programming strategy is often used to accomplish other special analyses in the LISREL framework (e.g., see scale reliability estimation in Chapter 8). Another programming trick is that A3 is regarded as another factor (see LE line and pattern matrix for LY); thus, the model is programmed to have four latent factors (Social Phobia, Agoraphobia, A3, Sex). This programming (in tandem with PA BE matrix programming, discussed below) actually produces a solution equivalent to a typical CFA solution (see the "Equivalent CFA Solutions" section in Chapter 5). That is, although it will not be found in the estimates for lambda-Y, the solution will generate a factor loading for A3 onto Agoraphobia that is identical to what would be obtained in an ordinary CFA specification. For

TABLE 7.14. Computer Syntax (Mplus, LISREL, EQS, Amos) for MIMIC Model of Social Phobia and Agoraphobia (Regressing the Latent Factors and the A3 Indicator on the Sex Covariate)

Mplus 3.11

```
TITLE:     CFA MIMIC MODEL
DATA:      FILE IS MIMIC6.DAT;
           TYPE IS STD CORR;
           NOBS ARE 730;
VARIABLE: NAMES ARE S1 S2 S3 A1 A2 A3 SEX;
ANALYSIS: ESTIMATOR=ML;
MODEL:     SOCIAL BY S1-S3;
           AGORAPH BY A1-A3;
           SOCIAL ON SEX; AGORAPH ON SEX; A3 ON SEX; SOCIAL WITH AGORAPH;
OUTPUT:    SAMPSTAT MODINDICES(4.00) STAND RESIDUAL;
```

LISREL 8.72

```
TITLE LISREL PROGRAM FOR MIMIC MODEL
DA NI=7 NO=730 MA=CM
LA
S1 S2 S3 A1 A2 A3 SEX                    ! SEX: 0 = FEMALE, 1 = MALE
KM
1.000
 0.705 1.000
 0.724 0.646 1.000
 0.213 0.195 0.190 1.000
 0.149 0.142 0.128 0.521 1.000
 0.155 0.162 0.135 0.557 0.479 1.000
-0.019 -0.024 -0.029 -0.110 -0.074 -0.291 1.000
SD
2.26 2.73 2.11 2.32 2.61 2.44 0.50
MO NY=7 NE=4 LY=FU,FR TE=SY,FR PS=SY,FR BE=FU,FR
LE
SOCIAL AGORAPH A3 SEX
PA LY
0 0 0 0
1 0 0 0
1 0 0 0
0 0 0 0
0 1 0 0
0 0 0 0                                  ! A3 IS A PSEUDOFACTOR
0 0 0 0
VA 1.0 LY(1,1) LY(4,2) LY(6,3) LY(7,4)  ! SET METRIC OF VARIABLES
PA TE
1
0 1
0 0 1
0 0 0 1
0 0 0 0 1
0 0 0 0 0 0
0 0 0 0 0 0 0    ! FIX ERROR OF SEX COVARIATE TO ZERO
```

                                                                    (cont.)

TABLE 7.14. *(cont.)*

```
PA PS
1
1 1             ! CORRELATED DISTURBANCE BETWEEN SOCIAL AND AGORAPH
0 0 1           ! ERROR VARIANCE OF A3
0 0 0 1         ! VARIANCE OF SEX
PA BE
0 0 0 1         ! PATH FROM SEX TO SOCIAL LATENT VARIABLE
0 0 0 1         ! PATH FROM SEX TO AGORAPH LATENT VARIABLE
0 1 0 1         ! PATH (FACTOR LOADING) FROM AGORAPH TO A3 INDICATOR
0 0 0 0         ! AND PATH FROM SEX TO A3 INDICATOR
OU ME=ML RS MI SC AD=OFF IT=100 ND=4
```

EQS 5.7b

```
/TITLE
 EQS SYNTAX FOR MIMIC MODEL
/SPECIFICATIONS
 CASES=730; VAR=7; ME=ML; MA=COR;
/MATRIX
1.000
 0.705 1.000
 0.724 0.646 1.000
 0.213 0.195 0.190 1.000
 0.149 0.142 0.128 0.521 1.000
 0.155 0.162 0.135 0.557 0.479 1.000
-0.019 -0.024 -0.029 -0.110 -0.074 -0.291 1.000
/STANDARD DEVIATIONS
2.26 2.73 2.11 2.32 2.61 2.44 0.50
/LABELS
 V1=s1; V2=s2; V3=s3; V4=a1; V5=a2; V6=a3; V7=sex;
 F1 = SOCIAL;
 F2 = AGORAPH;
/EQUATIONS
 V1 = F1 + E1;
 V2 = *F1 + E2;
 V3 = *F1 + E3;
 V4 = F2 + E4;
 V5 = *F2 + E5;
 V6 = *F2 + *V7 + E6;
 F1 = *V7 + D1;
 F2 = *V7 + D2;
/VARIANCES
 E1 TO E6 = *;
 D1,D2 = *;
 V7 = *;
/COVARIANCES
 D1,D2 = *;
/PRINT
 fit=all;
/LMTEST
/END
```

(cont.)

TABLE 7.14. *(cont.)*

Amos Basic 5.0

```
' Example of MIMIC Model in Amos 5.0

Sub Main ()
Dim sem As New AmosEngine

sem.TextOutput
sem.Standardized
sem.Smc
sem.Mods 10

sem.BeginGroup "mimic.txt"
sem.Structure "S1 = (1) SOCIAL + (1) E1"
sem.Structure "S2 = SOCIAL + (1) E2"
sem.Structure "S3 = SOCIAL + (1) E3"
sem.Structure "A1 = (1) AGORAPH + (1) E4"
sem.Structure "A2 = AGORAPH + (1) E5"
sem.Structure "A3 = AGORAPH + SEX + (1) E6"
sem.Structure "SOCIAL = SEX + (1) D1"
sem.Structure "AGORAPH = SEX + (1) D2"
sem.Structure "D1 <-> D2"

End Sub
```

the pattern matrix programming of lambda-Y (PA LX), the model specification for Social Phobia is typical (S1 is used as the marker indicator, the factor loadings of S2 and S3 are freely estimated). For Agoraphobia, A1 is used as the marker variable, and the factor loading of A2 is freely estimated. In the remaining part of PA LY (and its associated VA commands), the metric of the A3 indicator is passed onto the A3 "pseudofactor"; the metric of the sex covariate (0/1) is passed onto the Sex "factor." In the theta-epsilon pattern matrix (PA TE), the error variances of the first five indicators (S1, S2, S3, A1, A2) are freely estimated. The measurement errors of the A3 and Sex indicators are fixed to zero. Although the Sex variable will be assumed to have no measurement error in the analysis (i.e., it corresponds to known groups), the error variance of the A3 indicator will be estimated in another portion of the solution.

Because an "all Y" specification is employed, variances and covariances (not including the error variances of indicators S1 through A2) must be estimated by the psi (PS) matrix (Ψ). The first two diagonal elements of PS (i.e., ψ_{11}, ψ_{22}) freely estimate the residual variances of Social Phobia and Agoraphobia. The first off-diagonal element of PS (ψ_{21}) esti-

mates the correlated disturbance of these latent factors. The third (ψ_{33}) diagonal element of PS estimates the residual variance of the A3 pseudo-factor; this parameter estimate will in fact represent the measurement error variance of the A3 indicator. The fourth (ψ_{44}) diagonal element of PS freely estimates the variance of Sex.

The MIMIC portion of the LISREL specification resides in the pattern matrix programming of beta (BE). The beta (B) matrix has not been employed in prior examples in this book because it represents a structural component of an SEM model; in this context, "structural" refers to directional relationships among latent variables. Specifically, the beta matrix focuses on directional relationships among endogenous (Y) variables. Alternatively, the gamma matrix (Γ) focuses on the directional relationships between exogenous and endogenous variables. Because four latent Y variables have been specified, B will be a 4×4 full matrix; the row and column order of variables is SOCIAL AGORAPH A3 SEX, as specified by earlier programming. Thus, the β_{14} and β_{24} parameters in PA BE inform LISREL to freely estimate the regressive path of Sex to Social Phobia and Sex to Agoraphobia, respectively (population heterogeneity). The β_{34} element corresponds to the regressive path of Sex to A3 (measurement invariance). Finally, the β_{32} element informs LISREL to freely estimate a path from Agoraphobia to A3. Although this estimate is obtained in a structural matrix (B), it should be interpreted as the factor loading of the A3 indicator on the Agoraphobia latent factor; in fact, this estimate would equal the LY estimate if the analysis was not conducted as a MIMIC model.

The MIMIC model provides a good fit to the data, χ^2 (11) = 3.80, p = .98, SRMR = .011, RMSEA = 0.00 (90% CI = 0.00 to 0.00, CFit = 1.00), TLI = 1.008, CFI = 1.00. Inclusion of the sex covariate did not alter the factor structure or produce salient areas of strain in the solution (e.g., all modification indices < 4.0). Table 7.15 provides selected results of this solution. Of particular interest are the regressive paths linking Sex to the latent factors and the A3 indicator. As predicted, the path of Sex $\rightarrow$ Agoraphobia was statistically significant (z = 2.97, p < .01). Given how the sex covariate was coded (0 = females, 1 = males) and the negative sign of this parameter estimate (e.g., unstandardized estimate = –0.475), it can be concluded that males have a lower mean than females on the Agoraphobia factor; more specifically, the mean of females is .475 units higher than the mean of males. The completely standardized estimate of this parameter (StdYX = –0.13) is not readily interpretable because of the binary predictor (i.e., the sex covariate). In other words, it is not meaningful to discuss this relationship in terms of a standardized score change in Sex when the level of this

variable is either male or female. But the standardized estimate (Std = –0.261) can convey useful information about this effect. In a standardized solution, only the latent variable is standardized. Thus, the estimate of –0.261 can be interpreted as indicating that a unit increase in Sex is associated with a .261 standardized score decrease in the latent factor of Agoraphobia; or more directly, women are .261 standardized scores higher than men on the latent dimension of Agoraphobia. This value can be interpreted akin to Cohen's *d* (Cohen, 1988, 1992). Following Cohen's guidelines (*d* = .20, .50, and .80 for small, medium, and large effects, respectively; cf. Cohen, 1992), the sex difference for Agoraphobia is a small effect. The results in Table 7.15 also reveal that men and women do not differ with respect to Social Phobia (*z* = 0.69, *ns*).

Consistent with the researcher's predictions, the results of the MIMIC model show that the A3 indicator is not invariant for males and females (akin to intercept noninvariance in multiple-groups CFA). This is reflected by the significant direct effect of Sex on the A3 indicator (*z* = 6.65, *p* < .001) that is not mediated by Agoraphobia. In other words, holding the latent factor of Agoraphobia constant, there is a significant direct effect of Sex on the A3 indicator. Thus, at any given value of the latent factor, women score significantly higher on the A3 indicator than men (by .985 units, or nearly a full point on the 0–8 scale). This is evidence of *differential item functioning*; that is, the item behaves differently as an indicator of Agoraphobia in men and women. For the substantive reasons noted earlier (sex differences in the range of activities relating to personal safety), the A3 item is biased against females. Even when their level of the underlying dimension of Agoraphobia is the same as in men, women will have higher scores on the A3 indicator (cf. Figure 7.4B) because women's responses to this item ("walking alone in isolated areas") are affected by other influences that are less relevant to men (i.e., perceptions of personal safety, in addition to the underlying construct of Agoraphobia).

Although the current example had a specific hypothesis about the gender noninvariance of the social phobia-agoraphobia questionnaire (re: item A3), measurement invariance is frequently evaluated in an exploratory fashion. In the context of a MIMIC model, this can be done by fixing all direct effects between the covariate and the indicators to zero and then inspecting modification indices (and associated expected parameter change values) to determine whether salient direct effects may be present in the data. The researcher should not pursue the alternative of freely estimating all of these direct effects, because the model would be under-

TABLE 7.15. Mplus Results of MIMIC Model of Social Phobia and Agoraphobia

```
MODEL RESULTS
            Estimates    S.E.    Est./S.E.      Std    StdYX
SOCIAL BY
   S1          1.000     0.000      0.000     2.007    0.889
   S2          1.079     0.045     23.967     2.166    0.794
   S3          0.855     0.035     24.534     1.716    0.814

AGORAPH BY
   A1          1.000     0.000      0.000     1.820    0.785
   A2          0.956     0.066     14.388     1.739    0.667
   A3          0.917     0.063     14.495     1.669    0.684

SOCIAL ON
   SEX        -0.109     0.158     -0.690    -0.054   -0.027

AGORAPH ON
   SEX        -0.475     0.160     -2.973    -0.261   -0.130

A3 ON
   SEX        -0.985     0.148     -6.653    -0.985   -0.202

SOCIAL WITH
   AGORAPH     0.999     0.171      5.857     0.273    0.273

Residual Variances
   S1          1.072     0.126      8.533     1.072    0.210
   S2          2.750     0.195     14.087     2.750    0.370
   S3          1.501     0.114     13.169     1.501    0.338
   A1          2.062     0.217      9.498     2.062    0.384
   A2          3.777     0.264     14.301     3.777    0.555
   A3          2.705     0.214     12.642     2.705    0.455
   SOCIAL      4.026     0.284     14.175     0.999    0.999
   AGORAPH     3.257     0.317     10.269     0.983    0.983
```

identified. This could be accomplished by modifying the Mplus syntax in Table 7.14 with the following MODEL commands (the bolded syntax fixes the direct effects of the sex covariate on the six indicators to zero):

```
MODEL: SOCIAL BY S1-S3;
       AGORAPH BY A1-A3;
       SOCIAL ON SEX; AGORAPH ON SEX; S1-A3 ON SEX@0; SOCIAL
           WITH AGORAPH;
```

This analysis produces the following modification indices in regard to covariate–indicator direct effects:

```
                 M.I.    E.P.C.    Std E.P.C.   StdYX E.P.C.
ON Statements
A1   ON SEX     13.302   0.559        0.559        0.120
A2   ON SEX     10.102   0.549        0.549        0.105
A3   ON SEX     42.748  -1.046       -1.046       -0.214
```

As in the planned analysis, this result clearly suggests the salience of the direct effect of the sex covariate on the A3 indicator (e.g., modification index = 42.78). Because a substantive argument could be made in regard to the source of this noninvariance, this parameter could be freed in a subsequent analysis (i.e., the model depicted in the path diagram in Figure 7.5). Although the initial analysis suggests that direct effects may exist between Sex and the A1 and A2 indicators (e.g., modification indices = 13.30 and 10.10, respectively), these relationships are less interpretable and are not salient (i.e., modification indices drop below 4.0) after the direct effect of Sex → A3 is added to the model (cf. Jöreskog, 1993, and guidelines for model revision in Chapter 5).

SUMMARY

This chapter introduced the reader to the methods of evaluating the equivalence of CFA parameters within and across groups. These procedures provide a sophisticated approach to examining the measurement invariance and population heterogeneity of CFA models. Invariance evaluation may be the primary objective of an investigation (e.g., determine the generalizability of a test instrument) or may establish the suitability of subsequent analyses (e.g., verify longitudinal measurement invariance to justify group comparisons on structural parameters such as the equivalence of regres-

sive paths among latent variables). The chapter also discussed the methodology for the analysis of mean structures. Incorporating means into CFA allows for other important analyses such as the between-group equivalence of indicator intercepts (cf. differential item functioning) and between-group comparisons on latent means. While the latter is analogous to ANOVA, CFA provides a much stronger approach because the group comparisons are conducted in context of a measurement model (e.g., adjusting for measurement error and an error theory).

In Chapter 8, three new specific applications of CFA are discussed: higher-order factor models, CFA evaluation of scale reliability, and models with formative indicators. Although the specific applications and issues covered in the second half of this book are presented in separate chapters (i.e., Chapter 6 through Chapter 10), it is important to underscore the fact that these procedures can be integrated into the same analysis. For instance, the multiple-groups approach presented in this chapter can be employed to examine the between-group equivalence of a higher-order CFA, a CFA evaluation of scale reliability, or a model containing indicators that cause the latent construct (formative indicators). Missing and non-normal data are common to all types of CFA and must be dealt with appropriately (Chapter 9). These topics are covered in the next two chapters.

NOTES

1. In addition, the term "strictly parallel" has been used for indicators that, in addition to possessing invariant factor loadings and error variances, have equivalent intercepts.

2. In Figure 7.4C, the predicted value of X2 would be the same for Groups 1 and 2 at $\xi = 0$, because the intercepts are invariant. However, predicted values of X2 would be expected to differ between groups at all non-zero values of ξ.

3. The level of statistical power is the same for between-groups comparisons of parameters tested by both the MIMIC and multiple-groups approach (e.g., group differences in factor means). The difference lies in the fact that MIMIC assumes that other measurement and structural parameters (e.g., factor loadings) are equivalent across groups, and thus the measurement model is not first tested separately in each group (where sample size may not be sufficient to ensure adequate power and precision of parameter estimates within each group).

Appendix 7.1

Reproduction of the Observed Variance–Covariance Matrix with Tau-Equivalent Indicators of Auditory Memory

Running the model specified by the syntax listed in Table 7.2 produces the following model-estimated variance–covariance matrix (obtained from LISREL output):

```
        X1       X2       X3       X4       X5       X6
X1    6.8011
X2    4.4302   6.8467
X3    4.4302   4.4302   6.9595
X4    1.4125   1.4125   1.4125   3.7636
X5    1.4445   1.4445   1.4445   3.1681   4.1209
X6    1.4449   1.4449   1.4449   3.1689   3.2407   4.2025
```

and the following residual (unstandardized) matrix:

```
        X1       X2       X3       X4       X5       X6
X1     0.0110
X2     0.1588   0.2289
X3    -0.1715  -0.0004  -0.2514
X4    -0.0454   0.1356  -0.0659   0.0000
X5     0.1291  -0.0136  -0.2615   0.0022   0.0000
X6     0.1068   0.1201  -0.1281  -0.0032   0.0011   0.0000
```

A few observations will be made using the X1 indicator as an example (sample SD = 2.61, factor loading = 2.1048, error variance = 2.3708):

1. Unlike other CFA examples discussed in previous chapters of this book, the model-estimated variances of the indicators loading on Auditory Memory (whose factor loadings are held to equality) may not be reproduced perfectly. For instance, the sample variance of the X1 indicator is 6.8121 (SD^2 = 2.61^2 = 6.8121), but the model-estimated variance of this measure is 6.8011 (resulting in a residual of 0.011). Thus, these model-implied variances, as well as the model-implied covariances, will count toward any discrepancies between the observed and predicted matrices (cf. Appendix 3.3 in Chapter 3 on the calculation of F_{ML}). Note that the model perfectly reproduces the variances

318

of the Visual Memory indicators because the factor loadings of these measures have not been constrained to equality in this solution.

2. In the present case where the metric of the latent factor has been defined by fixing its variance to 1.0, the model estimate of the variances of the indicators whose factor loadings have been constrained to equality is the sum of the indicator's squared unstandardized factor loading and model estimate of the indicator's error variance. For instance, the model estimate of the variance of X1 = $\lambda_{11}^2 + \delta_{11}$ = 2.1048^2 + 2.3708 = 6.801 (same as X1 variance estimate in the predicted matrix presented above). The model estimate of the covariances of the indicators loading on Auditory Memory is simply the unstandardized factor loading squared (i.e., 2.1048^2 = 4.4302). Thus, note that the estimated covariances of X1–X2, X1–X3, and X2–X3 are the same (4.4302) because the factor loadings of these indicators have been constrained to be the same (see also the predicted covariances of these indicators with indicators that are specified to load on Visual Memory).

3. Inspection of the selected Mplus output presented in the text of Chapter 7 indicates that while the unstandardized loadings for X1, X2, and X3 are the same (all λs = 2.10, because this portion of the unstandardized solution has been constrained to equality), the completely standardized estimates of these loadings differ (λs range from .798 to .807). This is because the error variances of these indicators are permitted to vary (unlike in the model that tests for parallel indicators). Recall from Chapter 4 that unstandardized factor loadings can be readily converted to completely standardized coefficients using a formula from multiple regression, $b^* = (bSD_x)/(SD_y)$, where b = the unstandardized coefficient (in this case the unstandardized factor loading), SD_x = the SD of the predictor (in this case, the SD of the latent factor), and SD_y = the SD of the outcome variable (in this case, the predicted SD of the indicator). Using the X1 indicator and model estimates from the model of tau equivalence of Auditory Memory, b = 2.1048, SD_x = 1.00 (because the variance of Auditory Memory was fixed to 1.0), and SD_y = 2.6079 (the square root of the model estimated variance of X1, 6.8011). Thus, the completely standardized factor loading of X1 is calculated as 2.1048 (1.0) / 2.6079 = .807 (the same as the completely standardized factor loading for X1 presented in Chapter 7). The completely standardized loadings for the remaining two indicators of Auditory Memory differ because the error variances (and hence model-implied variances) are allowed to differ; for example, completely standardized loading for X3 = 2.1048 (1.0) / 2.6381 = .798.

4. In models that test for parallel indicators, all completely standardized estimates for indicators loading on a given factor will be the same (i.e., factor loadings, errors, communalities). This is because the error variances (and hence model-implied variances of the indicators) are the same. Thus, all values of the equation $b^* = (bSD_x)/(SD_y)$ are identical (unlike in the tau-equivalent model, where SD_y may differ for each indicator).

8

Other Types of CFA Models

Higher-Order Factor Analysis,
Scale Reliability Evaluation,
and Formative Indicators

CFA provides a unifying analytic framework for addressing a wide range of questions commonly asked by social and behavioral scientists. This notion is further evidenced in the present chapter on other types of CFA models. Three different applications of CFA are presented: higher-order factor analysis, scale reliability estimation, and constructs defined by formative indicators. Higher-order factor analysis is a theory-driven procedure in which the researcher imposes a more parsimonious structure to account for the interrelationships among factors established by CFA. CFA can also be employed to evaluate the reliability of a testing instrument in a manner that overcomes limitations of traditional methods (e.g., Cronbach's alpha). The concluding section describes models in which it is more reasonable to believe that a set of indicators is the cause of a construct rather than vice versa. The presentation of these models will underscore several concepts introduced in earlier chapters (e.g., model identification, equivalent models).

HIGHER-ORDER FACTOR ANALYSIS

Empirically, it may be of interest to examine the *higher-order structure* of a CFA measurement model. Hierarchical factor analysis is often used for the-

ory testing. For example, this analytic procedure is popular in intelligence research where it is believed that more specialized facets of ability (e.g., verbal comprehension, perceptual organization, memory) are influenced by a broader dimension of general intelligence (*g*). Another major reason why higher-order factor analysis is used in applied research is to rescue a construct. It is often the case that for a construct initially predicted to be unidimensional, the research evidence reveals that multiple factors are required to explain the covariation among the set of indicators of the construct. In these instances, a single higher-order factor can often be specified to account for the covariation among the multiple factors; that is, the construct consists of a single broader dimension and several subdimensions (for an applied example in the clinical psychology literature, see factor analytic research on the construct of anxiety sensitivity; e.g., Taylor, 1999; Zinbarg, Barlow, & Brown, 1997).

All CFA examples presented in this book thus far have been *first-order* measurement models. These analyses entailed specification of the number of factors, the pattern of indicator–factor relationships (i.e., factor loadings), and a measurement error theory (random or correlated indicator error variances). In multiple-factor CFA models, the factors have been specified to be intercorrelated (oblique); in other words, the factors are presumed to be interrelated, but the nature of these relationships is unanalyzed; that is, the researcher makes no substantive claims about the directions or patterns of factor interrelationships. In higher-order factor analysis (in both EFA and CFA), the focus is on the intercorrelations among the factors. In essence, these factor correlations represent the input matrix for the higher-order factor analysis. A goal of higher-order factor analysis is to provide a more parsimonious account for the correlations among lower-order factors. Higher-order factors account for the correlations among lower-order factors, and the number of higher-order factors and higher-order factor loadings is less than the number of factor correlations. Accordingly, the rules of identification used in first-order CFA apply to the higher-order component of a hierarchical solution. For instance, the number of higher-order factors that can be specified is dictated by the number of lower-order factors (discussed below). Unlike first-order CFA, higher-order CFA tests a theory-based account for the patterns of relationships among the first-order factors. These specifications assert that higher-order factors have direct effects on lower-order factors; these direct effects and the correlations among higher-order factors are responsible for the covariation of the lower-order factors.

Second-Order Factor Analysis

An example of a higher-order CFA model is presented in Figure 8.1. In this example, the researcher is evaluating the latent structure of a questionnaire measure of coping styles (the illustration is loosely based on the stress and coping literature, e.g., Folkman & Lazarus, 1980; Tobin, Holroyd, Reynolds, & Wigal, 1989). The questionnaire was administered
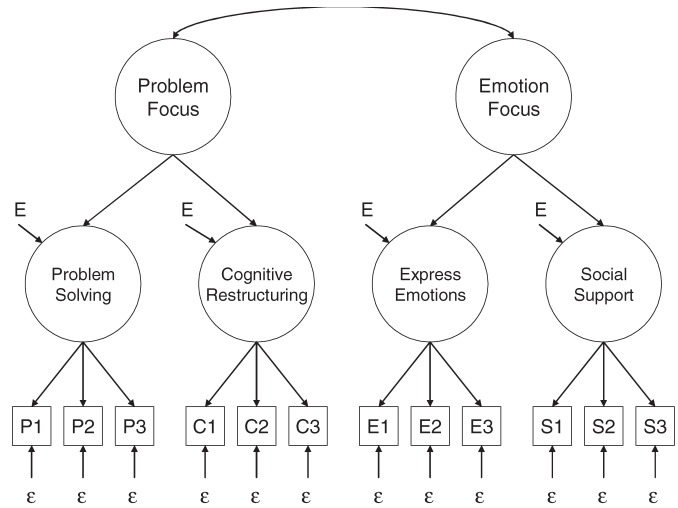
Sample Correlations and Standard Deviations (*SD*s); *N* = 275

```
      P1    P2    P3    C1    C2    C3    E1    E2    E3    S1    S2    S3
P1   1.00
P2   0.78  1.00
P3   0.80  0.77  1.00
C1   0.56  0.51  0.48  1.00
C2   0.52  0.51  0.46  0.78  1.00
C3   0.59  0.51  0.51  0.80  0.79  1.00
E1   0.16  0.15  0.17  0.14  0.18  0.16  1.00
E2   0.19  0.13  0.18  0.14  0.16  0.16  0.81  1.00
E3   0.12  0.17  0.17  0.17  0.20  0.16  0.75  0.80  1.00
S1   0.16  0.13  0.17  0.15  0.16  0.18  0.56  0.52  0.50  1.00
S2   0.16  0.14  0.18  0.15  0.16  0.18  0.51  0.58  0.51  0.81  1.00
S3   0.16  0.15  0.14  0.16  0.16  0.14  0.52  0.57  0.52  0.80  0.79  1.00

SD:  1.40  2.10  1.30  2.30  2.40  1.90  2.00  2.90  2.30  3.10  2.20  1.20
```

FIGURE 8.1. Higher-order factor model of a questionnaire measure of coping styles. P1–P3 = problem-solving items; C1–C3 = cognitive restructuring items; E1–E3 = express emotion items; S1–S3 = social support items.

to 275 college undergraduates. The latent structure of this questionnaire is predicted to be characterized by four first-order factors that represent four distinctive ways of coping with stressful events: Problem Solving, Cognitive Restructuring, Express Emotions, and Social Support. The four factors are presumed to be intercorrelated. If no relationships were observed among the first-order order factors, there would be no justification to pursue higher-order factor analysis. Two higher-order factors are predicted to account for the six correlations among the first-order factors: Problem-Focused Coping and Emotion-Focused Coping. The conceptual basis for this specification is that problem solving and cognitive restructuring are believed to represent styles of coping aimed directly at resolving the stressor (problem-focused). Conversely, expressing emotions and seeking social support are coping strategies not directed toward the stressor, but instead are aimed at managing emotional reactions to the stressor (emotion-focused). The empirical feasibility of the higher-order model should be evidenced by the patterning of correlations among factors in the first-order model; for example, the latent factors of Problem Solving and Cognitive Restructuring should be more strongly correlated with Express Emotions and Social Support, and vice versa.

On a terminological note, Problem-Focused Coping and Emotion-Focused Coping could be more specifically referred to as *second-order factors* (i.e., a second level of factors that account for the correlations among first-order factors), as opposed to the more general term, *higher-order factors*. In context of substantive theory and statistical identification (i.e., a sufficient number of correlated, second-order factors), higher-order factor analysis can proceed to the third order and beyond, although such analyses are rare in the applied literature. It is also important to note that higher-order factors are not defined by observed measures; this aspect requires some alterations to typical CFA syntax programming.

The general sequence of CFA-based higher-order factor analysis is as follows: (1) develop a well-behaved (e.g., good-fitting, conceptually valid) first-order CFA solution; (2) examine the magnitude and pattern of correlations among factors in the first-order solution; and (3) fit the second-order factor model, as justified on conceptual and empirical grounds. Thus, using the data set presented in Figure 8.1, the first step is to fit a four-factor CFA model, allowing the correlations among the factors to be freely estimated. The four-factor solution provides a good fit to the data, $\chi^2(48) = 82.668$, $p < .01$, SRMR = .017, RMSEA = 0.049 (90% CI = 0.03 to 0.07, CFit = .51), TLI = 0.99, CFI =.99 (based on LISREL output). The

completely standardized parameter estimates of this solution are presented in Table 8.1. As seen in this table, all 12 items are reasonable indicators of their respective latent factors (range of factor loadings = .86 to .93).

The results from the phi matrix in Table 8.1 provide the correlations among the factors. Consistent with prediction, all factors are significantly interrelated (range of zs = 2.92–9.54, not shown in Table 8.1). It is of importance that the pattern of correlations speaks to the viability of the posited second-order model. The model depicted in Figure 8.1 asserts that

TABLE 8.1. Selected LISREL Results of the Four-Factor (First-Order) Measurement Model of Coping

```
Completely Standardized Solution
```

LAMBDA-X

	PROBSLV	COGRES	EXPREMOT	SOCSPT
P1	0.9108	– –	– –	– –
P2	0.8643	– –	– –	– –
P3	0.8787	– –	– –	– –
C1	– –	0.8876	– –	– –
C2	– –	0.8739	– –	– –
C3	– –	0.9045	– –	– –
E1	– –	– –	0.8733	– –
E2	– –	– –	0.9285	– –
E3	– –	– –	0.8600	– –
S1	– –	– –	– –	0.9008
S2	– –	– –	– –	0.8965
S3	– –	– –	– –	0.8863

PHI

	PROBSLV	COGRES	EXPREMOT	SOCSPT
PROBSLV	1.0000			
COGRES	0.6624	1.0000		
EXPREMOT	0.2041	0.2005	1.0000	
SOCSPT	0.1959	0.2018	0.6699	1.0000

THETA-DELTA

P1	P2	P3	C1	C2	C3
0.1705	0.2529	0.2279	0.2122	0.2362	0.1819

THETA-DELTA

E1	E2	E3	S1	S2	S3
0.2374	0.1379	0.2604	0.1885	0.1963	0.2145

Problem Solving and Cognitive Restructuring, and Express Emotions and Social Support, are more specific dimensions of Problem-Focused Coping and Emotion-Focused Coping, respectively. If this is true in the data, the magnitude of factor correlations should show a clear pattern; that is, Problem Solving and Cognitive Restructuring will be more strongly correlated with each other than with Express Emotions and Social Support, and Express Emotions and Social Support will be more highly interrelated than either factor is with Problem Solving and Cognitive Restructuring. As shown in Table 8.1, the factor correlations follow this pattern. For example, the magnitude of the correlation between Problem Solving and Cognitive Restructuring is considerably higher (ϕ_{21} = .66) than the correlations between Problem Solving and Cognitive Restructuring and Express Emotions and Social Support (e.g., ϕ_{31} = .20).

Other patterns of factor correlations would contradict the posited higher-order model. For instance, if all factor correlations were roughly the same (e.g., $\phi s \approx .60$), this would favor a single second-order factor. Although a single second-order factor would adequately account for these factor correlations, it should not be specified in the absence of conceptual justification. As with CFA in general, analysis of a higher-order solution should be fully confirmatory. This is particularly important, considering the fact that higher-order factors are specified without indicators. In addition to specifying the structure of higher-order factors, the researcher must name these factors in a manner that can be defended by theory.

Moreover, the higher-order portion of the solution must be statistically identified. The rules of identification used for a typical CFA model readily generalize to higher-order solutions (see Chapter 3, Figures 3.6 and 3.7). Like lower-order factors, higher-order factors must have a metric. Because interest is primarily in the completely standardized solution, the metric of higher-order factors is usually defined by standardizing the entire higher-order portion of the solution. Although less often used, it is possible to generate an unstandardized higher-order solution by specifying lower-order factors as "marker indicators" for the higher-order factors (both approaches are illustrated below). This approach would be favored in some situations, such as when the researcher is interested in evaluating the measurement invariance of the higher-order solution in multiple groups (e.g., Chen, Sousa, & West, 2005). Both strategies produce an identical completely standardized solution and fit to the data. In addition, the number of freely estimated parameters in the higher-order portion of the model must not exceed the total number of factor variances and covariances in the first-order solution. For this aspect of identification, it

can be helpful to view the lower-order factors as "indicators" of the higher-order factors. For example, a single higher-order factor cannot be specified to account for the factor correlation from a first-order CFA model with two factors because it would be underidentified (cf. Figure 3.6A, Chapter 3), unless other (potentially unreasonable) constraints were placed on the solution (e.g., constraining the higher-order factor loadings to equality would just-identify the solution). If the first-order model has three factors, a solution that specifies a single higher-order factor would be just-identified (cf. Figure 3.6B, Chapter 3); that is, the higher-order solution would produce the same goodness of fit as the first-order model in which the three factors are allowed to freely covary. Despite its just-identified nature, it may nonetheless be substantively meaningful to evaluate such a solution in order to examine the magnitude (and statistical significance) of the higher-order factor loadings and relationships of the higher-order factors to the observed measures (discussed in more detail below; e.g., Schmid–Leiman transformation).

The higher-order portion of the model presented in Figure 8.1 is overidentified by a single degree of freedom. Specifically, in the first-order model there are four factor variances and six factor covariances. In contrast, the higher-order portion of the solution contains nine freely estimated parameters: four higher-order factor loadings, four residual variances (also called disturbances, depicted as "E"s in Figure 8.1), and the correlation between Problem-Focused Coping and Emotion-Focused Coping. The variances of Problem-Focused Coping and Emotion-Focused Coping are not included in this tally because they will be fixed to 1.0 to define the metric of these higher-order factors. Accordingly, the higher-order solution has one degree of freedom (10 – 9; cf. Figure 3.7B, Chapter 3).

However, the researcher must be mindful of the possibility of empirical underidentification in the higher-order portion of a solution. For instance, empirical underidentification would occur in the Figure 8.1 model if the correlations of the Problem Solving and Cognitive Restructuring factors with the Express Emotions and Social Support factors were (close to) zero (cf. Figure 3.7D, Chapter 3). In this case, the correlation between the higher-order factors of Problem-Focused Coping and Emotion-Focused Coping would be zero, and there would be infinite pairs of higher-order factor loadings that would reproduce the correlations between Problem Solving and Cognitive Restructuring, and between Express Emotions and Social Support, if these parameters were freely estimated (see Chapter 3, "CFA Model Identification" section).

Table 8.2 presents computer syntax from various software packages to program the high-order model in Figure 8.1. As in MIMIC models (Chapter 7), the factors that were developed in the initial CFA (Problem Solving, Cognitive Restructuring, etc.) are now latent-Y variables (endogenous) because they are being predicted by the higher-order factors. This is reflected in several aspects of the LISREL programming; that is, the 12 items from the coping questionnaire are treated as Y indicators (NY = 12), four eta (endogenous) variables are specified (NE = 4), and the lambda-Y (LY) and theta-epsilon (TE) matrices are used to program the first-order measurement model. Because these factors are now endogenous variables, their residual variances (disturbances) are estimated using the psi (PS) matrix (as opposed to estimation of factor variances, using the phi matrix, PH). On the model line, the command PS=DI accomplishes this by informing LISREL that psi is a diagonal (DI) matrix; that is, freely estimate residual variances and fix all off-diagonal relationships—residual covariances—to zero.

In LISREL matrix programming, the two second-order factors are ksi (exogenous) variables (NK = 2) that have no observed measures. However, these factors must be provided a metric. As noted earlier, because the tradition of factor analysis is to focus on the completely standardized solution (higher-order factor analysis was first developed in EFA), the most common approach to setting the metric of higher-order factors is to fix their variances to 1.0. In the LISREL syntax presented in Table 8.2, this is done using the PH=ST command on the model line (i.e., standardize the phi matrix).[1] The gamma (GA) matrix is needed to specify the directional relationships between the second-order and first-order factors.

Table 8.2 also provides LISREL syntax for occasions when the user wishes to generate an unstandardized solution for the higher-order component of the model. In this case, the phi matrix should not be standardized (i.e., PH=SY,FR replaces PH=ST). Rather, the scale of the higher-order factors is set by passing the metric of a lower-order factor up to them. In this example, Problem Solving and Express Emotion are selected to serve as "marker indicators" for Problem-Focused Coping and Emotion-Focused Coping, respectively. The same logic for specifying marker indicators in first-order CFA is used for higher-order factors. However, instead of the lambda-X or lambda-Y matrices, the gamma matrix is used for this purpose. Table 8.2 shows that in the pattern matrix for gamma, the paths of Problem Focus → Problem Solving and Emotion Focus → Express Emotions are fixed to zero; this is then overridden by the value (VA) statement, which fixes these unstandardized paths to 1.0.

TABLE 8.2. Computer Syntax (LISREL, Mplus, EQS, Amos, CALIS) for Specification of a Higher-Order CFA Factor Model of Coping

LISREL 8.72 (Higher-Order Factors Standardized)

```
TITLE LISREL PROGRAM FOR HIGHER-ORDER FACTOR MODEL OF COPING
DA NI=12 NO=275 MA=CM
LA
P1 P2 P3 C1 C2 C3 E1 E2 E3 S1 S2 S3
KM
<Insert correlation matrix from Figure 8.1 here>
SD
1.40 2.10 1.30 2.30 2.40 1.90 2.00 2.90 2.30 3.10 2.20 1.20
MO NY=12 NK=2 NE=4 LY=FU,FR TE=DI GA=FU,FR PS=DI PH=ST
LK
PROBFOC EMOTFOC
LE
PROBSLV COGRES EXPREMOT SOCSPT
PA LY
0 0 0 0
1 0 0 0
1 0 0 0
0 0 0 0
0 1 0 0
0 1 0 0
0 0 0 0
0 0 1 0
0 0 1 0
0 0 0 0
0 0 0 1
0 0 0 1
VA 1.0 LY(1,1) LY(4,2) LY(7,3) LY(10,4)
PA GA
1 0
1 0
0 1
0 1
OU ME=ML RS MI SC AD=OFF IT=100 ND=4
```

LISREL 8.72 (Higher-Order Factors Unstandardized)

Model line

```
MO NY=12 NK=2 NE=4 LY=FU,FR TE=DI GA=FU,FR PS=DI PH=SY,FR
```

Gamma matrix:

```
PA GA
0 0
1 0
0 0
0 1
VA 1.0 GA(1,1) GA(3,2)
```

(cont.)

TABLE 8.2. *(cont.)*

Mplus 3.11 (Higher-Order Factors Unstandardized)

```
TITLE:     HIGHER ORDER CFA MODEL OF COPING
DATA:      FILE IS "C:\HIGHORD.DAT";
           TYPE IS STD CORR;
           NOBS ARE 275;
VARIABLE:  NAMES ARE P1-P3 C1-C3 E1-E3 S1-S3;
ANALYSIS:  ESTIMATOR=ML;
MODEL:     PROBSLV BY P1-P3;
           COGRES BY C1-C3;
           EXPREMOT BY E1-E3;
           SOCSPT BY S1-S3;
           PROBFOC BY PROBSLV COGRES;
           EMOTFOC BY EXPREMOT SOCSPT;
OUTPUT:    SAMPSTAT MODINDICES(10.00) STAND RESIDUAL;
```

EQS 5.7b (Higher-Order Factors Standardized)

```
/TITLE
 HIGHER-ORDER MODEL OF COPING: Chapter 8
/SPECIFICATIONS
 CASES=275; VARIABLES=12; METHODS=ML; MATRIX=COR; ANALYSIS=COV;
/LABELS
 v1=item1; v2= item2; v3= item3; v4= item4; v5= item5; v6= item6;
 v7= item7; v8= item8; v9= item9; v10= item10; v11= item11; v12= item12;
 f1 = PROBSLV; f2 = COGRES; f3 = EXPREMOT; f4 = SOCSPT; f5 = PROBFOC;
 f6 = EMOTFOC;
/EQUATIONS
 V1 = F1+E1;
 V2 = *F1+E2;
 V3 = *F1+E3;
 V4 = F2+E4;
 V5 = *F2+E5;
 V6 = *F2+E6;
 V7 = F3+E7;
 V8 = *F3+E8;
 V9 = *F3+E9;
 V10 = F4+E10;
 V11 = *F4+E11;
 V12 = *F4+E12;
 F1 = *F5+D1;
 F2 = *F5+D2;
 F3 = *F6+D3;
 F4 = *F6+D4;
/VARIANCES
 F5 = 1;
 F6 = 1;
 D1 TO D4 = *;
 E1 TO E12 = *;
/COVARIANCES
 F5, F6 = *;
/MATRIX
<Insert correlation matrix from Figure 8.1 here>
```

TABLE 8.2. *(cont.)*

```
/STANDARD DEVIATIONS
1.40 2.10 1.30 2.30 2.40 1.90 2.00 2.90 2.30 3.10 2.20 1.20
/PRINT
 fit=all;
/LMTEST
/WTEST
/END
```

SAS 8.2 PROC CALIS (Higher-Order Factors Standardized)

```
Title "Higher-Order CFA Model of Coping";
Data COPE (type=CORR);
 input _TYPE_ $ _NAME_ $ V1-V12;
 label V1 = 'item1'
 V2 = 'item2'
 V3 = 'item3'
 V4 = 'item4'
 V5 = 'item5'
 V6 = 'item6'
 V7 = 'item7'
 V8 = 'item8'
 V9 = 'item9'
 V10 = 'item10'
 V11 = 'item11'
 V12 = 'item12';
cards;
mean . 0 0 0 0 0 0 0 0 0 0 0 0
 std .    1.40 2.10 1.30 2.30 2.40 1.90 2.00 2.90 2.30 3.10 2.20 1.20
 N   .     275  275  275  275  275  275  275  275  275  275  275  275
corr V1  1.00  .    .    .    .    .    .    .    .    .    .    .
corr V2  0.78 1.00  .    .    .    .    .    .    .    .    .    .
corr V3  0.80 0.77 1.00  .    .    .    .    .    .    .    .    .
corr V4  0.56 0.51 0.48 1.00  .    .    .    .    .    .    .    .
corr V5  0.52 0.51 0.46 0.78 1.00  .    .    .    .    .    .    .
corr V6  0.59 0.51 0.51 0.80 0.79 1.00  .    .    .    .    .    .
corr V7  0.16 0.15 0.17 0.14 0.18 0.16 1.00  .    .    .    .    .
corr V8  0.19 0.13 0.18 0.14 0.16 0.16 0.81 1.00  .    .    .    .
corr V9  0.12 0.17 0.17 0.17 0.20 0.16 0.75 0.80 1.00  .    .    .
corr V10 0.16 0.13 0.17 0.15 0.16 0.18 0.56 0.52 0.50 1.00  .    .
corr V11 0.16 0.14 0.18 0.15 0.16 0.18 0.51 0.58 0.51 0.81 1.00  .
corr V12 0.16 0.15 0.14 0.16 0.16 0.14 0.52 0.57 0.52 0.80 0.79 1.00
;
run;
proc calis data=COPE cov method=ml pall pcoves;
var = V1-V12;
lineqs
 V1 = 1.0 f1 + e1,
 V2 = lam2 f1 + e2,
 V3 = lam3 f1 + e3,
 V4 = 1.0 f2 + e4,
 V5 = lam5 f2 + e5,
 V6 = lam6 f2 + e6,
```

(cont.)

330

TABLE 8.2. *(cont.)*

```
 V7 = 1.0 f3 + e7,
 V8 = lam8 f3 + e8,
 V9 = lam9 f3 + e9,
 V10 = 1.0 f4 + e10,
 V11 = lam11 f4 + e11,
 V12 = lam12 f4 + e12,
 f1 = ga1 f5 + D1,
 f2 = ga2 f5 + D2,
 f3 = ga3 f6 + D3,
 f4 = ga4 f6 + D4;
std
 f5 = 1.0,
 f6 = 1.0,
 d1-d4 = ps1-ps4,
 e1-e12 = te1-te12;
cov
 f5-f6 = ph1;
run;
```

Amos Basic 5.0 (Higher-Order Factors Standardized)

```
' Example of Second-Order CFA in Amos 5.0
'
Sub Main ()
Dim sem As New AmosEngine

sem.TextOutput
sem.Standardized
sem.Smc
sem.Mods 10

sem.BeginGroup "coping.txt"

sem.Structure "p1 = (1) PROBSLV + (1) th1"
sem.Structure "p2 =     PROBSLV + (1) th2"
sem.Structure "p3 =     PROBSLV + (1) th3"
sem.Structure "c1 = (1) COGRES + (1) th4"
sem.Structure "c2 =     COGRES + (1) th5"
sem.Structure "c3 =     COGRES + (1) th6"
sem.Structure "e1 = (1) EXPREMOT + (1) th7"
sem.Structure "e2 =     EXPREMOT + (1) th8"
sem.Structure "e3 =     EXPREMOT + (1) th9"
sem.Structure "s1 = (1) SOCSPT + (1) th10"
sem.Structure "s2 =     SOCSPT + (1) th11"
sem.Structure "s3 =     SOCSPT + (1) th12"

sem.Structure "PROBSLV = PROBFOC + (1) D1"
sem.Structure "COGRES = PROBFOC + (1) D2"
sem.Structure "EXPREMOT = EMOTFOC + (1) D3"
sem.Structure "SOCSPT = EMOTFOC + (1) D4"

sem.Structure "PROBFOC (1)"
sem.Structure "EMOTFOC (1)"
sem.Structure "PROBFOC <—> EMOTFOC"

End Sub
```

If the reader has read the preceding chapters of this book, the programming logic for the Mplus, EQS, Amos, and CALIS syntax should be self-explanatory. As with LISREL, note that the EQS, Amos, and CALIS programming differentiates disturbance variance (D1–D4 in all examples) and measurement error variance. In the EQS, Amos, and CALIS examples, the higher-order solution is standardized (e.g., F5 = 1.0, F6 = 1.0). The Mplus example will also produce an unstandardized solution, although this syntax could be readily modified to generate a completely standardized solution only (i.e., fix the variances of the higher-order factors to 1.0). As noted earlier, regardless of whether an unstandardized solution is generated, the specifications will produce identical results in terms of overall fit, modification indices, completely standardized parameter estimates, and so forth.

The fit of the higher-order solution is as follows: $\chi^2(49) = 82.703$, $p < .01$, SRMR = .017, RMSEA = 0.047 (90% CI = 0.03 to 0.07, CFit = .55), TLI = 0.99, CFI =.99 (LISREL output). A higher-order solution cannot improve goodness of fit relative to the first-order solution where the factors are freely intercorrelated. If the higher-order aspect of the model is overidentified, it is attempting to reproduce these factor correlations with fewer number of freely estimated parameters. When the higher-order model is overidentified, the nested χ^2 test can be used to determine whether the specification produces a significant degradation in fit relative to the first-order solution. In this example, the higher-order solution is found to be equally good fitting, $\chi^2_{\text{diff}}(1) = 0.035$, *ns* (i.e., 82.703 – 82.668, *df* = 49 – 48). The gain of one degree of freedom in this particular higher-order model is caused by the model's attempting to account for the six correlations among the lower-order factors with one less freely estimated parameter (i.e., 4 higher-order factor loadings, 1 correlation between higher-order factors = 5).

In addition to goodness of fit, the acceptability of the higher-order model must be evaluated with regard to the magnitude of the higher-order parameters; that is, size of higher-order factor loadings and higher-order factor correlations. The completely standardized estimates from this solution are presented in Table 8.3. Each of the first-order factors loads strongly onto the second-order factors (range of loadings = .81–.83). The estimates provided in the psi matrix indicate the proportion of variance in the lower-order factors that is not explained by the second-order factors (i.e., completely standardized disturbances). Using these estimates, it can be seen that the higher-order factors account for 66%–68% of the variance in the first-order factors; for example, Express Emotion, 1 – .3185 = .6815.

TABLE 8.3. Selected Output (LISREL) of Higher-Order CFA Model of Coping

```
Completely Standardized Solution
LAMBDA-Y

          PROBSLV    COGRES   EXPREMOT    SOCSPT
          ____       ____      ____       ____
   P1     0.9108     - -       - -        - -
   P2     0.8644     - -       - -        - -
   P3     0.8787     - -       - -        - -
   C1     - -        0.8876    - -        - -
   C2     - -        0.8740    - -        - -
   C3     - -        0.9044    - -        - -
   E1     - -        - -       0.8733     - -
   E2     - -        - -       0.9284     - -
   E3     - -        - -       0.8601     - -
   S1     - -        - -       - -        0.9008
   S2     - -        - -       - -        0.8965
   S3     - -        - -       - -        0.8863
```

```
GAMMA

           PROBFOC    EMOTFOC
           ____       ____
 PROBSLV   0.8118      - -
  COGRES   0.8159      - -
 EXPREMOT   - -       0.8255
  SOCSPT    - -       0.8116
```

```
Correlation Matrix of ETA and KSI

          PROBSLV    COGRES   EXPREMOT    SOCSPT    PROBFOC   EMOTFOC
          ____       ____      ____       ____      ____      ____
 PROBSLV  1.0000
  COGRES  0.6624     1.0000
 EXPREMOT 0.2018     0.2028    1.0000
  SOCSPT  0.1984     0.1994    0.6699     1.0000
 PROBFOC  0.8118     0.8159    0.2485     0.2443    1.0000
 EMOTFOC  0.2444     0.2457    0.8255     0.8116    0.3011    1.0000
```

```
PSI

          PROBSLV    COGRES   EXPREMOT    SOCSPT
          ____       ____      ____       ____
          0.3410     0.3343    0.3185     0.3414
```

```
          THETA-EPS
              P1        P2        P3        C1        C2        C3
            ____      ____      ____      ____      ____      ____
           0.1705    0.2529    0.2279    0.2122    0.2361    0.1820
```

```
          THETA-EPS
              E1        E2        E3        S1        S2        S3
            ____      ____      ____      ____      ____      ____
           0.2374    0.1380    0.2602    0.1886    0.1963    0.2145
```

The correlation between the higher-order factors is estimated to be .301 (seen in "Correlation Matrix of ETA and KSI" section of output).

Because the higher-order solution did not result in a significant decrease in model fit, it can be concluded that the model provided a good account for the correlations among the first-order factors. This can be demonstrated using the tracing rules presented in earlier chapters (e.g., Chapter 3). For example, in the first-order CFA model, the correlation between Problem Solving and Cognitive Restructuring was .66 (see Table 8.1). Multiplying the higher-order factor loadings of Problem Focus → Problem Solving and Problem Focus → Cognitive Restructuring perfectly reproduces this correlation; that is, .812(.816) = .66. Similarly, in the initial CFA solution, the correlation between Problem Solving and Express Emotions was estimated to be .196. This relationship is accounted for multiplying the following three parameters: Problem Focus → Problem Solving (.812), Emotion Focus → Express Emotion (.823), correlation between Problem Focus and Emotion Focus (.301); that is, .812(.823)(.301) = .20.

Schmid-Leiman Transformation

The researcher can use the estimates provided by the completely standardized solution to estimate the relationship of the observed measures to the higher-order factors. This is calculated by multiplying the appropriate higher-order and first-order factor loadings. For instance, the completely standardized effect of Problem Focus Coping on indicator P1 is .74; that is, .812(.911). This can be interpreted as the indirect effect of the higher-order factor on the indicator; for example, the effect of Problem Focused Coping on P1 is mediated by Problem Solving. However, because the higher-order factors have no direct effects on the indicators, these values can also be regarded as total effects; that is, the completely standardized relationship between Problem Focused Coping and P1 is .74. Squaring this result provides the proportion of the variance in P1 explained by the second-order factor, Problem Focused Coping; that is, $.74^2 = .55$.

Using a few additional simple calculations, the hierarchical solution can be transformed to further elucidate the relationships of the first- and second-order factors with the observed measures. This procedure was introduced by Schmid and Leiman (1957). Although developed for use within the EFA framework, the procedure readily generalizes to CFA. The transformation treats first-order factors as residualized factors; that is, second-order factors explain as much variance as possible, and the vari-

ance in the indicators that cannot be accounted for by second-order factors is ascribed to the first-order factors. As Loehlin (2004) notes, whereas the explanatory power of first-order factors is lessened by the transformation, the substantive interpretation of first-order factors may be fostered. This is because the residualized factor loadings of the first-order factors represent the unique contribution of the first-order factors to the prediction of the indicators. In addition, as a result of the transformation, the first-order factors are not correlated with second-order factors.

To illustrate this procedure, a Schmid–Leiman transformation was conducted on the higher-order solution of coping. The results of this analysis are presented in Table 8.4 (a spreadsheet prepared in Microsoft Excel). The two columns following the item names contain the first-order factor loadings and second-order factor loadings that were obtained from the hierarchical CFA (see Table 8.3). In the fourth column, each item's loading on the second-order factor is calculated by the product of its loading on the first-order factor (column 2) and the first-order factor's loading on the second-order factor (column 3); for example, loading of indicator P1 on Problem Focused Coping = .9108(.8118) = .7394 (the same calculation is presented in the first paragraph of this section). The residualized first-order loading (column 6) is computed by multiplying the first-order loading (column 2) by the square root of the uniqueness of the higher-order factor. Recall that in a completely standardized solution, a uniqueness represents the proportion of variance in an outcome (in this case, a first-order factor) that is not explained by a factor (in this case, the second-order factor). The terms "disturbance" and "residual variance" are synonymous in this context (cf. "E," Figure 8.1). For example, the uniqueness of Problem Focused Coping is estimated to be $1 - .8118^2 = .341$ (cf. psi matrix, Table 8.3). The square roots of the uniquenesses are presented in the fifth column of Table 8.4. Thus, the residualized first-order loading of P1 is computed: .9108(.5839) = .532.

The second row of calculations indicates how much variance in each indicator is explained by the second-order factors ("Higher-Order Rsq" column) and first-order factors ("Residualized Loading Rsq" column). These results are obtained by squaring the indicator's transformed loadings on the first- and second-order factors. For example, for indicator P1, 55% of its variance is explained by the second-order factor, Problem Focused Coping ($.739^2 = .55$), and 28% of its variance is explained by the first-order factor, Problem Solving ($.531^2 = .28$). Based on this transformation, it could be concluded that whereas slightly more than half (55%) of the variance in the P1 indicator is accounted for by a broader trait reflecting

TABLE 8.4. Spreadsheet of Schmid–Leiman Transformation of the Higher-Order Model of Coping

Item	Factor Loading	Higher Order Loading	Item: High-Order Factor	SQRT(Uniq)	Residualized Primary Loading
P1	0.9108	0.8118	0.73938744	0.583935579	0.531848525
P2	0.8644	0.8118	0.70171992	0.583935579	0.504753914
P3	0.8787	0.8118	0.71332866	0.583935579	0.513104193
C1	0.8876	0.8159	0.72419284	0.578193039	0.513204141
C2	0.874	0.8159	0.7130966	0.578193039	0.505340716
C3	0.9044	0.8159	0.73789996	0.578193039	0.522917784
E1	0.8733	0.8255	0.72090915	0.564402117	0.492892369
E2	0.9284	0.8255	0.7663942	0.564402117	0.523990926
E3	0.8601	0.8255	0.71001255	0.564402117	0.485442261
S1	0.9008	0.8116	0.73108928	0.584213523	0.526259541
S2	0.8965	0.8116	0.7275994	0.584213523	0.523747423
S3	0.8863	0.8116	0.71932108	0.584213523	0.517788445

	Loading Rsq	High-Order Rsq		Residual Loading Rsq	Sum
P1	0.82955664	0.546693786	+	0.282862854	0.82955664
P2	0.74718736	0.492410846	+	0.254776514	0.74718736
P3	0.77211369	0.508837777	+	0.263275913	0.77211369
C1	0.78783376	0.52445527	+	0.26337849	0.78783376
C2	0.763876	0.508506761	+	0.255369239	0.763876
C3	0.81793936	0.544496351	+	0.273443009	0.81793936
E1	0.76265289	0.519710003	+	0.242942887	0.76265289
E2	0.86192656	0.58736007	+	0.27456649	0.86192656
E3	0.73977201	0.504117821	+	0.235654189	0.73977201
S1	0.81144064	0.534491535	+	0.276949105	0.81144064
S2	0.80371225	0.529400887	+	0.274311363	0.80371225
S3	0.78552769	0.517422816	+	0.268104874	0.78552769

the tendency to engage in coping efforts aimed directly at resolving the stressor (Problem Focused Coping), slightly more than a quarter of its variance (28%) is explained by the influence of a more specific problem-focused coping strategy (Problem Solving). The remaining 17% of P1's variance (i.e., $1 - .55 - .28 = .17$) is not accounted for by either the first- or second-order factors (i.e., unique variance or measurement error; cf. theta-epsilon matrix in Table 8.3).

It is important to note that this transformation does not alter the explanatory power of the original CFA solution; for example, the proportion of unexplained variance in P1 is .17 before and after the transformation. This is also demonstrated in Table 8.4 where it is shown that the sum of variance explained by the transformed first- and second-order factor loadings (e.g., P1: $.547 + .283 = .83$; see "Sum" column) is equal to the communalities of the first-order factor loadings (communality = squared factor loading) obtained in the initial hierarchical solution (e.g., P1 communality = $.9108^2 = .83$; see "Loading Rsq" column). Rather, the Schmid–Leiman procedure is simply a method of calculating the contribution of lower- and higher-order factors to the prediction of observed measures (with explanatory preference afforded to the higher-order factors). Applied examples of this procedure can be found in Campbell-Sills et al. (2004) and Brown et al. (2004).

SCALE RELIABILITY ESTIMATION

Point Estimation of Scale Reliability

Multiple-item measures (e.g., questionnaires) are commonly used in the social and behavioral sciences. A key aspect of the psychometric development of such measures is the evaluation of *reliability*. Reliability refers to the precision or consistency of measurement; that is, the overall proportion of true score variance to total observed variance of the measure. Early in the process of scale development, researchers often submit the measure's items to factor analysis (EFA, CFA). The resulting latent structure is then used to guide the scoring of the measure for further psychometric analyses (e.g., convergent and discriminant validity evaluation) and, ultimately, applied use; for example, if factor analyses reveal two factors, the measure is deemed to have two subscales. Most often, such scoring entails unrefined composites where items found to load on a given factor are simply summed or averaged (cf. Grice, 2001). The reliability of these composite

scores is usually estimated by Cronbach's (1951) coefficient alpha (α). Despite its widespread popularity, researchers have long known that α is a misestimator of scale reliability. If the scale contains no correlated measurement errors, α will underestimate scale reliability unless the condition of tau equivalence holds (Lord & Novick, 1968; Zimmerman, 1972; see Chapter 7 for definition and evaluation of tau equivalence). However, the condition of tau equivalence is frequently not realized in actual data sets, in part because the units of measurement are often arbitrary (Raykov, 2001a, 2004). Further, if the measure contains correlated measurement errors, α can either underestimate or overestimate scale reliability, depending on the underlying measurement parameters (Raykov, 2001b; Zimmerman, 1972). Thus, this and other research (e.g., Green & Hershberger, 2000; Komaroff, 1997; Raykov, 1997) has shown that Cronbach's coefficient α does not provide a dependable estimate of scale reliability of multiple-item measures.

Raykov (2001a, 2004) has developed a CFA-based method of estimating scale reliability that reconciles the problems with Cronbach's coefficient α. As shown in the next example, an appealing feature of this approach is that it provides an estimate of scale reliability directly in the context of the CFA measurement model. CFA estimates of factor loadings, error variances, and error covariances are used to calculate the scale's true and error variance. As noted earlier, scale reliability represents the proportion of true score variance to total observed variance in the measure. This is expressed by the following equation (Lord & Novick, 1968):

$$\rho_Y = Var(T) \ / \ Var(Y) \qquad (8.1)$$

where ρ_Y is the scale reliability coefficient, $Var(T)$ is the true score variance of the measure, and $Var(Y)$ is the total variance of the measure. $Var(Y)$ is the sum of the true score variance and error variance of the measure. In the case of a congeneric measurement model (Jöreskog, 1971a) without correlated measurement errors, this equation can be re-expressed using CFA measurement parameters as

$$\rho = (\Sigma\lambda_i)^2 \ / \ [(\Sigma\lambda_i)^2 + \Sigma\theta_{ii}] \qquad (8.2)$$

where $(\Sigma\lambda_i)^2$ is the squared sum of unstandardized factor loadings, and $\Sigma\theta_{ii}$ is the sum of unstandardized measurement error variances. In instances where the measurement model contains correlated measurements errors, scale reliability would be calculated as:

$$\rho = (\Sigma\lambda_i)^2 / [(\Sigma\lambda_i)^2 + \Sigma\theta_{ii} + 2\Sigma\theta_{ij}] \qquad (8.3)$$

where $2\Sigma\theta_{ij}$ is the sum of nonzero error covariances multiplied by 2. In latent software programs that accommodate nonlinear constraints (e.g., LISREL), Raykov's (2001a) procedure entails specifying three dummy latent variables whose variances are constrained to equal the numerator (true score variance), denominator (total variance), and corresponding ratio of true score variance to total score variance (ρ), per the classic formula for scale reliability estimation (Lord & Novick, 1968). Beginning in version 6.0 of EQS, estimates of scale reliability (and Cronbach's α) can be requested using the /RELIABILITY command; for example,

> /RELIABILITY
> SCALE = V1,V2,V3,V4;

This procedure is now illustrated using the measurement model and data provided in Figure 8.2. In this example, the researcher has developed a multi-item questionnaire designed to assess subjective distress following exposure to traumatic stressors such as natural disasters, sexual assault, and motor vehicle accidents (the example is based loosely on a measure developed by Horowitz, Wilner, & Alvarez, 1979). The questionnaire is purported to assess two related dimensions of psychological reactions to extreme stress: (1) Intrusions—intrusive thoughts, images, dreams, or flashbacks of a traumatic event; and (2) Avoidance—overt and covert avoidance of the traumatic event (e.g., avoidance of situations, thoughts, feelings, reminders of the trauma). A correlated error is predicted to exist between items Y7 and Y8 owing to a method effect (e.g., reverse wording). The measure was administered to 500 participants.

A CFA revealed that the two-factor model provided a good fit to the data, $\chi^2(18) = 20.48$, $p = .306$, SRMR = .021, RMSEA = 0.017 (90% CI = 0.00 to 0.036, CFit = .98), TLI = 1.00, CFI = 1.00 (LISREL output). The unstandardized parameter estimates of the factor loadings, error variances, and error covariance from this solution are provided in Table 8.5 (standard errors and test statistics have been deleted). These estimates will be used shortly to illustrate the calculation of point estimates of scale reliability.

Table 8.6 presents the LISREL syntax used to estimate the scale reliability of the dimensions of Intrusions and Avoidance. This matrix programming requires that a latent-Y variable specification be used, owing to the use of nonlinear constraints (dummy latent variables). Note that while there are only two substantive latent factors (Intrusions, Avoidance), eight
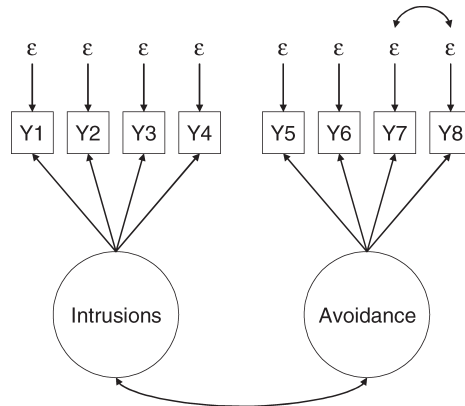
Sample Correlations and Standard Deviations (*SD*s); *N* = 500

```
        Y1      Y2      Y3      Y4      Y5      Y6      Y7      Y8
Y1    1.000
Y2    0.594   1.000
Y3    0.607   0.613   1.000
Y4    0.736   0.765   0.717   1.000
Y5    0.378   0.321   0.360   0.414   1.000
Y6    0.314   0.301   0.345   0.363   0.732   1.000
Y7    0.310   0.262   0.323   0.337   0.665   0.583   1.000
Y8    0.317   0.235   0.276   0.302   0.632   0.557   0.796   1.000

SD:   1.150   1.200   1.570   2.820   1.310   1.240   1.330   1.290
```

FIGURE 8.2. Latent structure of a questionnaire measure of distress following exposure to traumatic stress.

latent variables are specified (see LE and MO lines, Table 8.6). The remaining six latent variables represent the dummy latent variables needed to convey the true score variance, total variance, and ratio of true to total variance for the two latent dimensions of substantive interest. Accordingly, lambda-Y is an 8 (indicator) $\times$ 8 (latent variable) matrix, but only eight elements of this matrix are freely estimated in order to relate the observed measures (Y1–Y8) to the substantive latent factors (see PA LY). The theta-epsilon (TE) matrix is specified in the typical fashion (indicator measurement errors are freely estimated; note error covariance of Y7 and Y8). Because eight latent variables have been specified, psi (PS) is an 8×8 symmetric matrix. Elements PS(1,1) and PS(5,5) are fixed to 1.0 to set the metric of the latent factors, Intrusions and Avoidance. The off-diagonal

element, PS(5,1) is freely estimated to obtain the covariance of Intrusions and Avoidance. The values of the remaining diagonal elements of the psi matrix will be determined by the constraints (CO) that follow (see Table 8.6).

The first constraint holds the value of PS(2,2) equal to the true score variance of Intrusions. Although the mathematical operations to the right of the equals sign do not readily convey this fact, the constrained value of PS(2,2) is simply the squared sum of the unstandardized factor loadings of indicators of Intrusions, that is, $(\Sigma\lambda_i)^2$. Because of syntax programming

TABLE 8.5. Selected Unstandardized Parameter Estimates of Measurement Model of the Reactions to Traumatic Stress Questionnaire

```
LISREL  Estimates  (Maximum Likelihood)

     LAMBDA-Y
      INTRUS      AVOID
      _____      _____

Y1    0.8912       - -
Y2    0.9595       - -
Y3    1.1968       - -
Y4    2.6757       - -
Y5      - -      1.1942
Y6      - -      0.9954
Y7      - -      0.9701
Y8      - -      0.8938

     THETA-EPS
          Y1        Y2        Y3        Y4        Y5        Y6
          ____      ____      ____      ____      ____      ____

Y1    0.5284
Y2      - -      0.5194
Y3      - -        - -     1.0325
Y4      - -        - -       - -     0.7931
Y5      - -        - -       - -       - -     0.2901
Y6      - -        - -       - -       - -       - -     0.5468
Y7      - -        - -       - -       - -       - -       - -
Y8      - -        - -       - -       - -       - -       - -

     THETA-EPS
          Y7        Y8
          ____      ____

Y7    0.8277
Y8    0.4986    0.8652
```

TABLE 8.6. LISREL Syntax and Selected Results for Estimating Scale Reliability of the Reactions to Traumatic Stress Questionnaire

```
TITLE CFA SCALE RELIABILITY
DA NI=8 NO=500 MA=CM
LA
Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8
KM
1.000
0.594 1.000
0.607 0.613 1.000
0.736 0.765 0.717 1.000
0.378 0.321 0.360 0.414 1.000
0.314 0.301 0.345 0.363 0.732 1.000
0.310 0.262 0.323 0.337 0.665 0.583 1.000
0.317 0.235 0.276 0.302 0.632 0.557 0.796 1.000
SD
1.15 1.20 1.57 2.82 1.31 1.24 1.33 1.29
MO NY=8 NE=8 LY=FU,FR TE=SY,FR PS=SY,FR
LE
INTRUS INTNUM INTDEN INTREL AVOID AVNUM AVDEN AVOIDREL
PA LY
1 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0
0 0 0 0 1 0 0 0
0 0 0 0 1 0 0 0
0 0 0 0 1 0 0 0
0 0 0 0 1 0 0 0
PA TE
1
0 1
0 0 1
0 0 0 1
0 0 0 0 1
0 0 0 0 0 1
0 0 0 0 0 0 1
0 0 0 0 0 0 1 1
PA PS
0
0 1
0 0 1
0 0 0 1
1 0 0 0 0
0 0 0 0 0 1
0 0 0 0 0 0 1
0 0 0 0 0 0 0 1
VA 1.0 PS(1,1) PS(5,5)
CO PS(2,2) = LY(1,1)**2+LY(2,1)**2+LY(3,1)**2+LY(4,1)**2+C
```

(cont.)

TABLE 8.6. *(cont.)*

```
2*LY(1,1)*LY(2,1)+2*LY(1,1)*LY(3,1)+2*LY(1,1)*LY(4,1)+2*LY(2,1)*LY(3,1)+C
2*LY(2,1)*LY(4,1)+2*LY(3,1)*LY(4,1)
CO PS(3,3) = LY(1,1)**2+LY(2,1)**2+LY(3,1)**2+LY(4,1)**2+C
2*LY(1,1)*LY(2,1)+2*LY(1,1)*LY(3,1)+2*LY(1,1)*LY(4,1)+2*LY(2,1)*LY(3,1)+C
2*LY(2,1)*LY(4,1)+2*LY(3,1)*LY(4,1)+TE(1,1)+TE(2,2)+TE(3,3)+TE(4,4)
CO PS(4,4) = PS(2,2)*PS(3,3)**-1
CO PS(6,6) = LY(5,5)**2+LY(6,5)**2+LY(7,5)**2+LY(8,5)**2+C
2*LY(5,5)*LY(6,5)+2*LY(5,5)*LY(7,5)+2*LY(5,5)*LY(8,5)+2*LY(6,5)*LY(7,5)+C
2*LY(6,5)*LY(8,5)+2*LY(7,5)*LY(8,5)
CO PS(7,7) = LY(5,5)**2+LY(6,5)**2+LY(7,5)**2+LY(8,5)**2+C
2*LY(5,5)*LY(6,5)+2*LY(5,5)*LY(7,5)+2*LY(5,5)*LY(8,5)+2*LY(6,5)*LY(7,5)+C
2*LY(6,5)*LY(8,5)+2*LY(7,5)*LY(8,5)+ TE(5,5)+TE(6,6)+TE(7,7)+TE(8,8)+C
2*TE(8,7)
CO PS(8,8) = PS(6,6)*PS(7,7)**-1
OU ME=ML SC AD=OFF ND=4
```

Selected Results

```
LISREL Estimates (Maximum Likelihood)
 PSI
```

	INTRUS	INTNUM	INTDEN	INTREL	AVOID	AVNUM	
	____	____	____	____	____	____	____
INTRUS	1.0000						
INTNUM	- -	32.7543					
		(2.2588)					
		14.5009					
INTDEN	- -	- -	35.6277				
			(2.2555)				
			15.7962				
INTREL	- -	- -	- -	0.9193			
AVOID	0.4857	- -	- -	- -	1.0000		
	(0.0389)						
	12.5013						
AVNUM	- -	- -	- -	- -	- -	16.4307	
						(1.2664)	
						12.9740	
AVDEN	- -	- -	- -	- -	- -	- -	
AVOIDREL	- -	- -	- -	- -	- -	- -	

```
     PSI
```

	AVDEN	AVOIDREL
	____	____
AVDEN	19.9577	
	(1.2631)	
	15.8008	
AVOIDREL	- -	0.8233

restrictions in LISREL (i.e., parentheses are not permitted in CO commands except in matrix elements), the squared sum of factor loadings must be calculated by summing the squared factor loadings, for example, LY(1,1)**2, with the sum of the product of each factor loading pair multiplied by 2, for example, 2*LY(1,1)*LY(2,1). For instance, if $x_1 = 4$, $x_2 = 3$:

$$(\Sigma x_i)^2 = 4^2 + 3^2 + 2(4 * 3) = 49$$

The "C" at the end of this and other constraint commands is used to inform LISREL that the statement continues on the next line.

The next constraint, CO PS(3,3), represents the denominator of the scale reliability formula; in this case, the total variance of Intrusions. This is calculated by adding the sum of indicator error variances (unstandardized theta-epsilon estimates for Y1, Y2, Y3, and Y4) to the true score variance of Intrusions. The constraint on PS(4,4) provides the proportion of true score variance to the total variance of Intrusion; that is, the point estimate of the scale reliability of Intrusions. Division is not permitted in LISREL (or in matrix algebra), so this estimate is calculated using a negative exponent; that is, CO PS(4,4) = PS(2,2) * PS(3,3)**−1.

The calculation of the true score variance, total variance, and proportion of true to total variance (scale reliability) is performed in the same fashion for the dimension of Avoidance; see constraints for PS(6,6), PS(7,7), and PS(8,8), respectively (Table 8.6). The only difference is that the error covariance of indicators Y7 and Y8 must be included in the total variance of Avoidance. This is reflected by the addition of "2*TE(8,7)" at the end of the CO PS(7,7) command.

Selected results of this analysis are presented in Table 8.6. Although not shown in the table, this specification provides the same fit to the data as the CFA without nonlinear constraints, for example, χ^2 (18) = 20.48. Estimates provided in the psi matrix include the true score variance, total variance, and proportion of true to total variance (scale reliability) of Intrusions and Avoidance. For example, the true score and total variance of Intrusions is 32.75 and 35.63, respectively. The point estimate of the scale reliability (ρ) of Intrusions is .919; that is, 32.75 / 35.63 (92% of the total variance of Intrusions is true score variance). The scale reliability (ρ) estimate for Avoidance is .823; that is, 16.43 / 19.96.

Because the computation of ρ and the current CFA model are not complex, the scale reliability of Intrusions and Avoidance could be readily calculated by hand. Using the estimates provided in Table 8.5, the true score variance of Intrusions would be

$$(0.8912 + 0.9595 + 1.1968 + 2.6757)^2 = (5.723)^2 = 32.75$$

$$(8.4)$$

(cf. INTNUM, Table 8.6). The total variance of Intrusions would be computed as

$$32.75 + 0.5284 + 0.5194 + 1.0325 + 0.7931 = 35.63 \quad (8.5)$$

(cf. INTDEN, Table 8.6).

In addition to estimating scale reliability within complex CFA models (e.g., multiple latent factors, indicators, and error covariances), the aforementioned LISREL parameterization can be adapted for various extensions of this methodology such as testing for differences in scale reliability across groups (Raykov, 2002a).

In applied research, the reliability of the dimensions ("subscales") of a questionnaire is usually estimated by Cronbach's α. In the current data set, Cronbach's α for the Intrusions indicators is .840. For Avoidance, Cronbach's α is .845. Although the population values of the measurement parameters in this example are not known, it can be presumed that Cronbach's α underestimated the reliability of Intrusions (cf. ρ = .919) because the condition of tau equivalence was not present (see the *SD* and factor loading of Y4). However, Cronbach's α overestimated the reliability of Avoidance (cf. ρ = .823) owing to the presence of an error covariance involving items Y7 and Y8.

Standard Error and Interval Estimation of Scale Reliability

The preceding section illustrated how one may obtain a point estimate of scale reliability within the CFA framework. Although representing the "optimal" estimate, this point estimate should not be regarded as the "true" (population) value of a scale's reliability coefficient. Indeed, a point estimate contains no information about how closely (or distantly) it approximates the population value. Thus, as with other parameter estimates (e.g., means, factor loadings), it is helpful to calculate confidence intervals (CI) for point estimates of scale reliability. Interval estimation provides the range of plausible values that the population value is within. The degree of confidence to which the population value is contained within the CI is stated in terms of a probability or percentage; the CI most often used in the social and behavioral sciences is the 95% CI.

Confidence intervals are typically calculated by adding and subtracting from the point estimate the product of an appropriate z value and the standard error (SE) of the estimate. For instance, in Chapter 4, it was shown that the 95% CI of a factor loading could be calculated as

$$95\% \text{ CI } (\lambda) = \lambda \pm 1.96 SE(\lambda) \qquad (8.6)$$

If $\lambda = 1.05$, and $SE_\lambda = .05$, the 95% CI is .952–1.148 (the value of 1.96 is taken from the z distribution; i.e., 95% of all scores in a normal distribution occur between –1.96 and +1.96 SDs from the mean). Thus, with a high degree of confidence (95%), we can assert that the true population value of λ lies between .952 and 1.148.

These calculations are straightforward because latent variable software programs provide SEs in addition to the point estimates of the various parameters. To compute the CI of a point estimate of scale reliability, its SE must be obtained. Raykov (2002b) has developed a method for the estimation of SE and CI of scale reliability within the CFA framework. This method furnishes an approximate SE of ρ in the form

$$SE(\rho) = \text{SQRT}[D_1{}^2 Var(u) + D_2{}^2 Var(v) + 2D_1 D_2 Cov(u,v)] \quad (8.7)$$

where u is the sum of estimated unstandardized factor loadings, v is the sum of estimated error variances, $Cov(u,v)$ is the covariance of u and v (if any), and D_1 and D_2 are the partial derivatives of the scale reliability coefficient (ρ) with respect to u and v. D_1 and D_2 are obtained by the following formulas:

$$D_1 = 2uv / (u^2 + v)^2 \qquad (8.8)$$
$$D_2 = u^2 / (u^2 + v)^2 \qquad (8.9)$$

Once the SE is obtained, the 95% CI of ρ is easily computed using the formula

$$95\% \text{ CI } (\rho) = \rho \pm 1.96 SE(\rho) \qquad (8.10)$$

Other CIs can be estimated by substituting the appropriate z value; for example, 90% CI $(\rho) = \rho \pm 1.645 SE(\rho)$.

The Raykov (2002b) procedure for interval estimation requires two analytic steps. In the first step, a latent variable software program that accommodates nonlinear constraints (e.g., LISREL) is used to obtain the

estimates of u and v and their variances and covariance. LISREL syntax for this step is shown in Table 8.7, again using the two-factor measurement model of the reactions to the traumatic stress questionnaire. Using a latent-Y specification, six latent variables are specified: Intrusions and Avoidance, and four dummy latent variables of their corresponding us and vs. The programming of the LY, TE, and PS matrices follows along the same lines as the approach used to estimate scale reliability (cf. Table 8.6). In this case, certain diagonal elements of the psi (PS) matrix are constrained to equal the values of u and v for Intrusions and Avoidance. For example, PS(2,2) equals the sum of the unstandardized factor loadings of the Intrusion indicators (u, labeled "UINT" in the syntax). PS(3,3) equals the sum of the error variances of these indicators (v, labeled "VINT"). The remaining two constraints produce u and v for Avoidance ("UAVD" and "VAVD," respectively); note that the error covariance of Y7 and Y8 is included in the calculation of v. On the output (OU) line, the keyword "ALL" is included (Table 8.7). When "ALL" is requested, the LISREL output includes the variance–covariance matrix of the parameter estimates, which are also needed in the calculation of $SE(\rho)$.

Table 8.7 presents selected output of this analysis. Although not shown in Table 8.7, this model specification provides the same fit to the data as the initial CFA without nonlinear constraints; for example, $\chi^2(18)$ = 20.48. In the psi matrix, u and v values for Intrusions and Avoidance are on the diagonal (*SE*s and test statistics have been deleted from the output); for example, u for Intrusions = 5.723 (the same value that was hand calculated in the preceding section of this chapter, using unstandardized factor loadings in Table 8.5). Below this output is the portion of the covariance matrix of parameter estimates needed for computation of $SE(\rho)$. The first two diagonal elements are the variances of u and v (i.e., 0.0389 and 0.0223, respectively) for Intrusions (cf. $Var(u)$ and $Var(v)$) in the formula for $SE(\rho)$. The remaining two diagonal elements are the variances of u and v (0.0244 and 0.0425, respectively) for Avoidance. The immediate off-diagonal elements of these values are the covariances of u and v (cf. $Cov(u,v)$) in $SE(\rho)$ formula. The covariances are –0.0016 and –0.0031 for Intrusions and Avoidance, respectively.

In the second step, these values are used in a set of simple calculations that, for convenience, can be performed in a major software program such as SPSS or SAS. Table 8.8 provides SPSS syntax for calculating D_1, D_2, scale reliability (SR), the standard error (SE), and the 95% confidence interval (CI95LO, CI95HI) using the values of u, v, $Var(u)$, $Var(v)$, and $Cov(u,v)$ obtained in the prior LISREL analysis (and the formulas presented earlier

TABLE 8.7. LISREL Syntax and Selected Results for Computing Standard Errors and Confidence Intervals for Point Estimates of Scale Reliabilities of the Reactions to Traumatic Stress Questionnaire (Step 1)

```
TITLE CFA SCALE RELIABILITY: STD ERRORS AND CONFIDENCE INTERVALS
    (STEP ONE)
DA NI=8 NO=500 MA=CM
LA
Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8
KM
1.000
0.594 1.000
0.607 0.613 1.000
0.736 0.765 0.717 1.000
0.378 0.321 0.360 0.414 1.000
0.314 0.301 0.345 0.363 0.732 1.000
0.310 0.262 0.323 0.337 0.665 0.583 1.000
0.317 0.235 0.276 0.302 0.632 0.557 0.796 1.000
SD
1.15 1.20 1.57 2.82 1.31 1.24 1.33 1.29
MO NY=8 NE=6 LY=FU,FR TE=SY,FR PS=SY,FR
LE
INTRUS UINT VINT AVOID UAVD VAVD
PA LY
1 0 0 0 0 0
1 0 0 0 0 0
1 0 0 0 0 0
1 0 0 0 0 0
0 0 0 1 0 0
0 0 0 1 0 0
0 0 0 1 0 0
0 0 0 1 0 0
PA TE
1
0 1
0 0 1
0 0 0 1
0 0 0 0 1
0 0 0 0 0 1
0 0 0 0 0 0 1
0 0 0 0 0 0 1 1
PA PS
0
0 1
0 0 1
1 0 0 0
0 0 0 0 1
0 0 0 0 0 1
VA 1.0 PS(1,1) PS(4,4)
CO PS(2,2) = LY(1,1)+LY(2,1)+LY(3,1)+LY(4,1)
CO PS(3,3) = TE(1,1)+TE(2,2)+TE(3,3)+TE(4,4)
```

(cont.)

TABLE 8.7. *(cont.)*

```
CO PS(5,5) = LY(5,4)+LY(6,4)+LY(7,4)+LY(8,4)
CO PS(6,6) = TE(5,5)+TE(6,6)+TE(7,7)+TE(8,8)+2*TE(8,7)
OU ME=ML ALL ND=4
```

Selected Results

```
LISREL  Estimates  (Maximum  Likelihood)
         PSI
           INTRUS       UINT       VINT      AVOID       UAVD       VAVD
           _____      _____     _____     _____     _____     _____

INTRUS     1.0000
   UINT      - -       5.7231
   VINT      - -        - -       2.8734
  AVOID     0.4857      - -        - -       1.0000
   UAVD      - -        - -        - -        - -       4.0535
   VAVD      - -        - -        - -        - -        - -       3.5270


         Covariance  Matrix of  Parameter  Estimates
           PS 2,2     PS 3,3     PS 5,5     PS 6,6
           _____     _____     _____     _____

PS 2,2     0.0389
PS 3,3    -0.0016     0.0223
PS 5,5     0.0055     0.0000     0.0244
PS 6,6     0.0000     0.0000    -0.0031     0.0425
```

in this section). As shown in this table, the SPSS calculations reproduce the point estimates of scale reliability obtained previously in LISREL; ρs = .919 and .823 for Intrusions and Avoidance, respectively. For Intrusions, the standard error of ρ is 0.0066. Using this value, the 95% CI for scale reliability point estimate of .919 is found to be .9065–.9322; that is, 95% CI = .919 ± 1.96(0.0066). Thus, while the population reliability of Intrusions is unknown, with a considerable degree of confidence we can assert that it lies between the values of .9065 and .9322, with .919 being the optimal estimate. It is noteworthy that the Cronbach's α estimate of reliability (.84) falls well below the lower range of this CI. For Avoidance, the standard error of ρ is 0.0147, and thus the 95% CI of this point estimate (.8233) is .7944–.8521. In this case, the Cronbach's α estimate of .845 falls within the upper range of this CI.

These CFA-based procedures (Raykov, 2001a, 2002b) represent a very useful and more dependable method of calculating point and interval estimates of scale reliability. Raykov has developed a number of important extensions to this work, including the estimation of scale reliability in context of noncongeneric scales (Raykov & Shrout, 2002), weighted scales

TABLE 8.8. SPSS Syntax for Computing Standard Errors and Confidence Intervals for Point Estimates of Scale Reliabilities of the Reactions to Traumatic Stress Questionnaire (Step 2)

```
TITLE SPSS SYNTAX FOR SE AND CI OF SR POINT ESTIMATES:
    INTRUSION SUBSCALE.

COMPUTE U = 5.7231.
COMPUTE V = 2.8734.
COMPUTE D1 = (2*u*v) / (U**2 + V)**2.
COMPUTE D2 = -U**2 / (U**2 + V)**2.
COMPUTE VARU = 0.0389.
COMPUTE VARV = 0.0223.
COMPUTE COVUV = -0.0016.
COMPUTE SR = U**2 / (U**2 + V).
COMPUTE SE = SQRT(D1**2*VARU+D2**2*VARV+2*D1*D2*COVUV).
COMPUTE CI95LO = SR-1.96*SE.
COMPUTE CI95HI = SR+1.96*SE.

LIST SR SE CI95LO CI95HI.


SR          SE          CI95LO      CI95HI

0.9193      0.0066      0.9065      0.9322




TITLE SPSS SYNTAX FOR SE AND CI OF SR POINT ESTIMATES:
    AVOIDANCE SUBSCALE.

COMPUTE U = 4.0535.
COMPUTE V = 3.527.
COMPUTE D1 = (2*u*v) / (U**2 + V)**2.
COMPUTE D2 = -U**2 / (U**2 + V)**2.
COMPUTE VARU = 0.0244.
COMPUTE VARV = 0.0425.
COMPUTE COVUV = -0.0031.
COMPUTE SR = U**2 / (U**2 + V).
COMPUTE SE = SQRT(D1**2*VARU+D2**2*VARV+2*D1*D2*COVUV).
COMPUTE CI95LO = SR-1.96*SE.
COMPUTE CI95HI = SR+1.96*SE.

LIST SR SE CI95LO CI95HI.


SR          SE          CI95LO      CI95HI

0.8233      0.0147      0.7944      0.8521
```

(Raykov, 2001a), and scales with higher-order factor structure (Raykov & Shrout, 2002). In addition, these procedures can be adapted to test for significant change in composite reliability as the result of scale refinement (e.g., addition or deletion of items; Raykov & Grayson, 2003) and to test whether scale reliabilities significantly differ across groups (Raykov, 2002a). Use of some of these extensions in applied research can be found in Brown et al. (2004) and Campbell-Sills et al. (2004).

MODELS WITH FORMATIVE INDICATORS

All CFA examples discussed thus far in this book have entailed models with *reflective indicators* (also referred to as *effects indicators*). In a model with reflective indicators, the paths relating the indicators to the factor (i.e., factor loadings) emanate from the latent variable to the indicator (e.g., Figure 8.2). In accord with classical test theory (Lord & Novick, 1968), the direction of causality is from the latent construct to the observed measure. For indicators specified to load on a given factor, the latent variable should explain most of the covariation among the observed measures; for example, in Figure 8.2, the only reason that indicators Y1–Y4 are intercorrelated is that they share a common cause—the underlying construct of Intrusions. Related to this is the fact that measurement error in reflective models is taken into account at the indicator level (i.e., the εs in Figure 8.2). Moreover, given that the direction of causality is from the factor to the observed measures, a change in the latent variable will result in a change in all constituent indicators. In such models, the indicators of a given latent factor are considered to be interchangeable. For instance, eliminating the Y4 indicator from the Figure 8.2 model would not alter the meaning of the Intrusions latent factor.

In many scenarios, it may be more natural to define a construct as being influenced by a set of indicators. In *formative indicator* models (also referred to as *composite cause* or *cause indicators*), the direction of causality is from the observed measures to the construct (Bollen, 1989; Bollen & Lennox, 1991; Edwards & Bagozzi, 2000; MacCallum & Browne, 1993). An oft-used example is that of socioeconomic status (SES), whose causal indicators might be income, education level, and occupational status. In this case, it seems more reasonable to assert that these three variables are the cause of one's SES than the converse, which would claim income, education level, and occupational status are interrelated because they share

the common underlying cause of SES. The term *composite cause* is sometimes used in reference to such models because the latent variable is posited to be a construct that is a weighted, linear combination of its observed measures, plus error. Although rarely considered in some substantive domains (e.g., mental health research), formative indicator models are very common in many fields, such as marketing, economics, and consumer research (e.g., Diamantopoulos & Winklhofer, 2001; Jarvis, MacKenzie, & Podsakoff, 2003). Path diagrams of models containing formative indicators are presented in Figure 8.3. In these examples, Chronic Life Stress is the composite latent variable influenced by a variety of stressful events (e.g., financial difficulties, interpersonal conflict). In practice, causal measures are usually single indicators, although higher-order factor analysis can be extended to specify first-order factors as causes of second-order constructs (cf. Jarvis et al., 2003).

In addition to the direction of construct–indicator causality, the assumptions of formative indicator models differ from those of reflective indicator models in other important ways. First, in the formative model, although the cause indicators are specified to be intercorrelated (see Figure 8.3), the correlations among the indicators are not relevant to the goodness of fit and conceptual viability of the model (except for the issue of multicollinearity, which would indicate undue redundancy in the indicator set used to form the composite latent variable). In other words, the formative indicators may influence the composite construct independently of one another. For example, SES will increase if income increases even when education and occupation status remain the same. Similarly, financial difficulties should increase chronic life stress when job stress and interpersonal conflicts are static (Figure 8.3). Thus, causal indicators of a concept can be positively or negatively correlated or have no association whatsoever. Second, as is not the case with reflective models, eliminating a formative indicator from the measurement model is apt to change the meaning of the composite construct; that is, because the construct is a weighted, linear combination of all its observed measures. Formative indicators are not interchangeable. If the income indicator is omitted from the SES composite, part of the construct of SES is not represented. Elimination of the financial difficulties indicator removes an important variable that contributes to chronic life stress. Finally, although identification issues must be considered (discussed below), measurement error in the formative indicator model is taken into account at the latent construct level (i.e., the disturbance of the composite variable, E; see Figure 8.3), not at the indicator level (i.e., εs in Figure 8.2).

Model A: Single Construct Defined Solely by Formative Indicators (Unidentified)
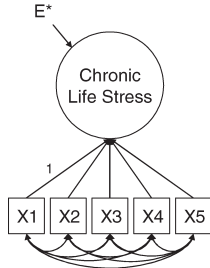
Model B: Single Construct Defined by Formative and Reflective Indicators (Identified)
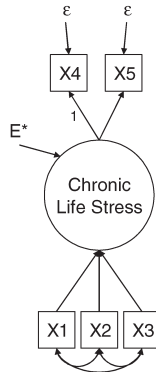
Model C: Formative Indicator Construct Predicting a Reflective Indicator Construct (Unidentified)
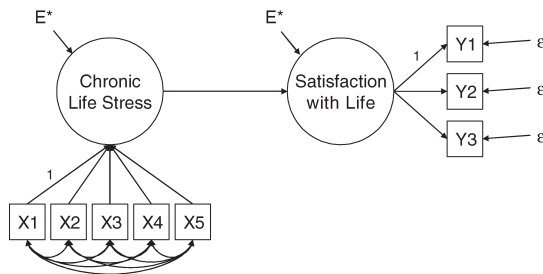
(cont.)

FIGURE 8.3. Examples of unidentified and identified models with formative indicators. X1, financial difficulties; X2, job stress; X3, interpersonal/familial conflict; X4, health problems (self); X5, health problems (significant others); Y1–Y3, indicators of life satisfaction; Y4–Y6, indicators of optimism with life; E*, disturbance parameter freely estimated; E = 0, disturbance parameter fixed to zero.

Model D: Formative Indicator Construct Predicting a Reflective Indicator Construct
(Identified by fixing disturbance of composite latent variable to zero)
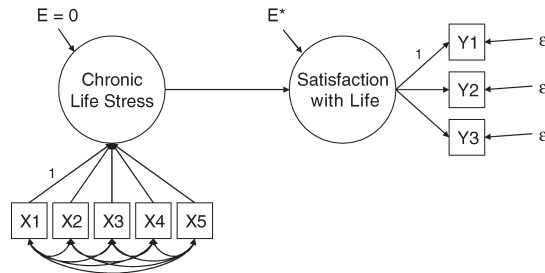
Model E: Formative Indicator Construct Predicting a Reflective Indicator Construct
(Identified by including reflective indicators on the latent composite variable)
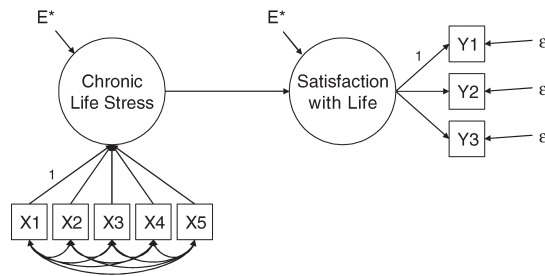
Model F: Formative Indicator Construct Predicting More Than One Reflective Indicator
Construct (Identified)
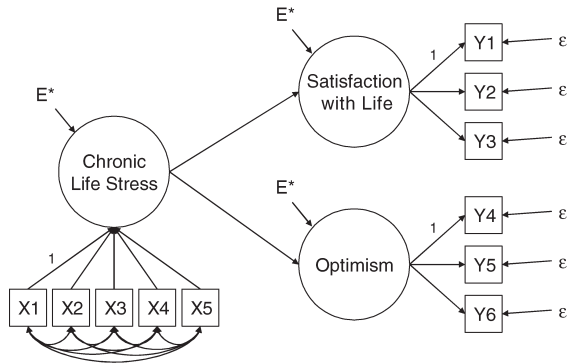
FIGURE 8.3. (*cont.*)

Model G: Formative Indicator Construct Predicting More Than One Reflective Indicator
Construct (Unidentified)
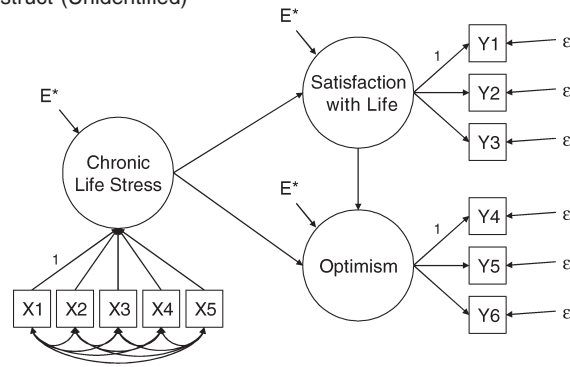
FIGURE 8.3. *(cont.)*

Identification problems are an issue in models with formative indicators (MacCallum & Browne, 1993). The core aspects of model identification discussed in Chapter 3 apply to formative indicator models; for example, the degrees of freedom of the model must be at least zero, and the latent variable must be provided a scale. The metric of the latent composite variable can be defined either by fixing a formative indicator path to one (e.g., see Figure 8.3A) or by fixing the factor variance to unity. Although necessary, these conditions are not sufficient to identify a model containing formative indicators. Indeed, as in some CFAs of multitrait–multimethod matrices (e.g., correlated methods models, Chapter 6), identification problems are common in some types of models with formative indicators.

Figure 8.3 presents several examples of identified and unidentified models that contain formative indicators. Measurement models that consist solely of formative indicators are not identified (Figure 8.3A). MacCallum and Browne (1993) note that many identification problems of formative indicator constructs stem from indeterminacies associated with the construct-level error term (i.e., the "E" in Figure 8.3A). Depending on the nature of the model, these indeterminacies can be resolved in various ways. As shown in Figure 8.3B, the identification problem can be addressed by adding two reflective indicators to the formative construct. This strategy should not be employed simply to identify the model. Indeed, this approach may not be appropriate if it cannot be justified on substantive grounds. In the current example (Figure 8.3B), note that the X4 and X5

indicators (which really should be labeled as Y indicators in Figure 8.3B because they are predicted by the formative construct) are measures of health problems in the respondent and his or her significant others. A compelling conceptual argument for treating X4 and X5 as reflective indicators could perhaps be forwarded in this instance; that is, although health problems are often construed as stressors, they can be caused by exposure to chronic life stress (e.g., job and family stress; cf. Figure 8.3). If the identification issue can be remedied in this manner, the formative construct is identified on its own and can reside anywhere in the measurement or structural model (e.g., as a latent-X or latent-Y variable, as a factor in a CFA model). This strategy carries the advantage of modeling flexibility (e.g., Figure 8.3E). For instance, if the formative construct is included in a broader CFA model, its discriminant validity and measurement parameters can be fully evaluated (Jarvis et al., 2003).

Models are also not identified if the formative construct predicts a single reflective construct (Figure 8.3C). One way the indeterminacy of the construct error term could be addressed is by fixing the disturbance of the composite latent variable to zero (Figure 8.3D).[2] However, this approach is often unreasonable because it assumes that the formative indicators represent the composite latent construct perfectly (cf. principal components analysis, Chapter 2; canonical correlation, Thompson, 1984). It was noted in the preceding paragraph that the error term of the formative construct could be freely estimated if the latent composite contains some reflective indicators. In fact, unlike the case in the Figure 8.3B model, only one reflective indicator is needed to identify the model depicted in Figure 8.3E. But if either strategy shown in Figure 8.3D or Figure 8.3E is employed, the issue of equivalent models must be considered (discussed below).

The residual variances of composite latent variables may be identified if the formative construct emits paths to two or more latent constructs defined by reflective indicators. An identified model is shown in Figure 8.3F where the formative construct of Chronic Life Stress predicts the reflective constructs of Satisfaction with Life and Optimism. In order for the identification problem to be resolved, there must be no relationship specified between the two latent-Y reflective constructs. Figure 8.3G depicts an unidentified model where both Chronic Life Stress and Satisfaction with Life have direct effects on Optimism. The same problem would arise if a correlated disturbance was specified between Satisfaction with Life and Optimism. Thus, models along the lines of Figure 8.3F, while more apt to be empirically identified, may also be substantively unreasonable. For instance, the model specification in Figure 8.3F indicates that the

relationship between Satisfaction with Life and Optimism is spurious (i.e., fully accounted for by Chronic Life Stress), an assertion that may be at odds with theory (e.g., a conceptual model that posits life satisfaction to be a predictor of optimism). The identification issue could also be addressed by including reflective indicators on the composite latent variable, if justified.

In addition to identification, the issue of equivalent models (see Chapter 5) must be considered in solutions involving formative indicators. Models with formative indicators may be statistically equivalent to MIMIC models (Chapter 7), where reflective latent constructs are predicted by single indicators but no direct effects of the covariates on the reflective indicators are specified (cf. Figure 7.5, Chapter 7). Figure 8.4 provides two

MIMIC Model A (Equivalent to Model D in Figure 8.3)
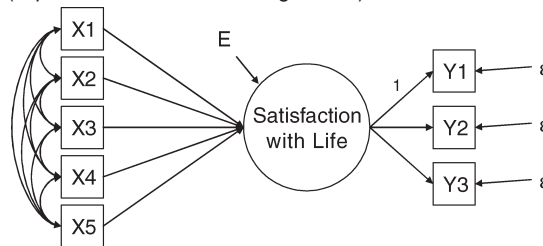
MIMIC Model B (Equivalent to Model E in Figure 8.3)
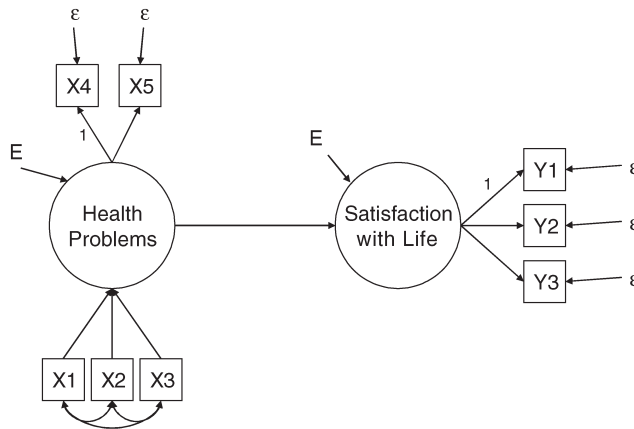
FIGURE 8.4. Equivalence of formative indicator models to MIMIC models. X1, financial difficulties; X2, job stress; X3, interpersonal/familial conflict; X4, health problems (self); X5, health problems (significant others); Y1–Y3, indicators of life satisfaction.

MIMIC models equivalent to formative construct models presented in Figure 8.3. Figure 8.4A depicts a model where the Satisfaction with Life latent variable is regressed onto the five single indicators of different forms of life stress. Although a different set of regressive parameter estimates will be obtained (i.e., five direct effects between the single indicators and Satisfaction with Life), the Figure 8.4A model will provide exactly the same predicted covariance matrix and goodness-of-fit statistics as the model shown in Figure 8.3D. Similarly, the model shown in Figure 8.3E could be alternatively construed as a model with two reflective constructs (Health Problems, Satisfaction with Life; see Figure 8.4B) in which one of the factors (Health Problems) is predicted by three background variables (X1, X2, X3). These single indicators may have indirect effects on Satisfaction with Life, as fully mediated by Health Problems. Although quite different conceptually, the models depicted in Figure 8.3E and Figure 8.4B will produce the same fitted covariance matrix. As noted in Chapter 5, the procedures of CFA/SEM cannot inform the researcher on the comparative validity of these solutions.

To foster the reader's understanding of how models with formative constructs are specified, a data-based example is provided in Figure 8.5. This model is identical to the model presented in Figure 8.3F, except that the formative construct of Chronic Life Stress is caused by three indicators instead of five. Mplus syntax is provided in Table 8.9. The latent endogenous factors of Satisfaction with Life and Optimism are programmed in the usual way (e.g., the Mplus default is to automatically set the first indicator listed—in this case Y1 and Y4—as the marker indicator). In order for Mplus to recognize chronic stress (Stress) as a latent variable, the programming indicates that Stress is "measured by" (i.e., the BY keyword) Life Satisfaction (Satis) and Optimism (Optim). Although the BY keyword is typically used in reference to factor loadings, the actual parameters involved are paths (beta, β) of the regressions of Satisfaction with Life and Optimism on the composite latent variable, Chronic Life Stress. In addition, a "*" is used to override the Mplus default of fixing the unstandardized parameter of Stress $\rightarrow$ Satis to 1.0; instead, this regressive path is freely estimated (*). The statement STRESS ON X1@1 X2 X3 regresses the formative construct of Chronic Life Stress onto its three causal indicators. The parameter relating X1 to Chronic Life Stress is fixed to 1.0 to define the metric of the composite latent variable. All three disturbances (E) are freely estimated by Mplus default. It is also an Mplus default to freely intercorrelate exogenous variables; in this case, the relationships among
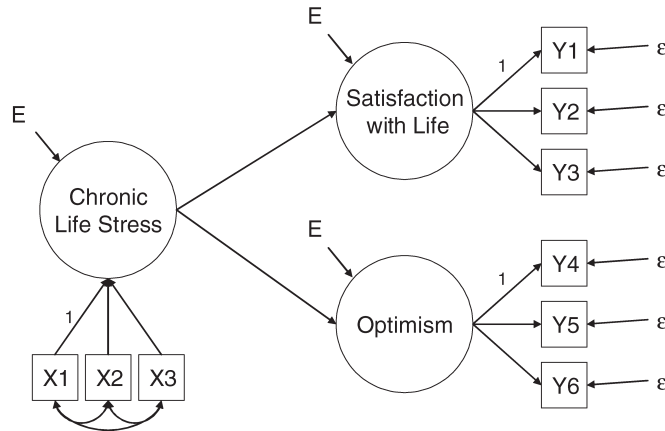
Correlations (and *SD*s, *N* = 500):

```
        Y1     Y2     Y3     Y4     Y5     Y6     X1     X2     X3
Y1   1.000
Y2   0.700 1.000
Y3   0.713 0.636 1.000
Y4   0.079 0.066 0.076 1.000
Y5   0.088 0.058 0.070 0.681 1.000
Y6   0.084 0.056 0.074 0.712 0.633 1.000
X1   0.279 0.248 0.240 0.177 0.155 0.170 1.000
X2   0.250 0.214 0.222 0.157 0.143 0.152 0.373 1.000
X3   0.280 0.236 0.251 0.173 0.178 0.171 0.448 0.344 1.000

SD:   2.5    2.1    3.0    4.1    3.9    4.4    1.2    1.0    1.2
```

FIGURE 8.5. Path diagram and input data for example of a model containing a formative construct. X1, financial difficulties; X2, job stress; X3, interpersonal/familial conflict; Y1–Y3, indicators of life satisfaction; Y4–Y6, indicators of optimism.

X1, X2, and X3. But to make the details of formative construct specification more explicit, these interrelationships are written out in the Mplus syntax example.

The Figure 8.5 model fits the data well; for example, $\chi^2(22) = 2.17$, $p = 71.00$. The parameter estimates provided in Table 8.9 should be self-explanatory if the reader is comfortable interpreting prior examples in this book (e.g., MIMIC models). Although the measurement model holds up well and the composite latent variable is significantly related to Satisfaction with Life and Optimism (i.e., zs = 3.805 and 3.357, respectively), the

TABLE 8.9. Mplus Syntax and Selected Results for Model
with a Formative Construct

Syntax File

```
TITLE:      MPLUS PROGRAM FOR MODEL WITH A FORMATIVE CONSTRUCT
DATA:       FILE IS STRESS.DAT;
            TYPE IS STD CORR;
            NOBS ARE 500;
VARIABLE:   NAMES ARE Y1 Y2 Y3 Y4 Y5 Y6 X1 X2 X3;
ANALYSIS:   ESTIMATOR=ML;
MODEL:      SATIS BY Y1 Y2 Y3;
            OPTIM BY Y4 Y5 Y6;
            STRESS BY SATIS* OPTIM;
            STRESS ON X1@1 X2 X3;
            X1 WITH X2 X3; X2 WITH X3;
OUTPUT:     SAMPSTAT MODINDICES(4.0) STAND FSDETERMINACY RESIDUAL;
```

Selected Results

MODEL RESULTS

	Estimates	S.E.	Est./S.E.	Std	StdYX
SATIS BY					
Y1	1.000	0.000	0.000	2.217	0.888
Y2	0.746	0.038	19.571	1.655	0.789
Y3	1.086	0.054	19.931	2.409	0.804
OPTIM BY					
Y4	1.000	0.000	0.000	3.580	0.874
Y5	0.848	0.045	18.733	3.035	0.779
Y6	1.000	0.051	19.441	3.579	0.814
STRESS BY					
SATIS	0.317	0.083	3.805	0.402	0.402
OPTIM	0.338	0.101	3.357	0.266	0.266
STRESS ON					
X1	1.000	0.000	0.000	0.355	0.425
X2	1.054	0.445	2.369	0.374	0.374
X3	1.073	0.435	2.469	0.381	0.456
X1 WITH					
X2	0.447	0.057	7.815	0.447	0.373
X3	0.644	0.070	9.142	0.644	0.448
X2 WITH					
X3	0.412	0.057	7.274	0.412	0.344
Variances					
X1	1.437	0.091	15.811	1.437	1.000
X2	0.998	0.063	15.811	0.998	1.000
X3	1.437	0.091	15.811	1.437	1.000

(cont.)

TABLE 8.9. *(cont.)*

Residual Variances

Y1	1.321	0.186	7.082	1.321	0.212
Y2	1.661	0.142	11.735	1.661	0.378
Y3	3.181	0.284	11.214	3.181	0.354
Y4	3.964	0.528	7.509	3.964	0.236
Y5	5.971	0.506	11.795	5.971	0.393
Y6	6.510	0.622	10.459	6.510	0.337
SATIS	4.121	0.498	8.279	0.838	0.838
OPTIM	11.905	1.130	10.534	0.929	0.929
STRESS	0.488	3.435	0.142	0.061	0.061

R-SQUARE

Observed Variable	R-Square
Y1	0.788
Y2	0.622
Y3	0.646
Y4	0.764
Y5	0.607
Y6	0.663

Latent Variable	R-Square
SATIS	0.162
OPTIM	0.071
STRESS	0.939

substantive significance of the results might be questioned on the basis of the magnitude of these structural relations (e.g., Chronic Life Stress accounts for 7.1% of the variance in Optimism). Because the error term of the formative construct was freely estimated, the paths linking Chronic Life Stress to Satisfaction with Life and Optimism are adjusted for measurement error. The paths linking the X indicators to Chronic Life Stress are interpreted as regression coefficients; for example, holding X2 and X3 constant, a one standardized unit increase in X1 is associated with a .425 standardized unit increase in Chronic Life Stress. Collectively, the three formative indicators in this example have considerable strength in predicting Chronic Life Stress (R^2 = .939).

SUMMARY

This chapter dealt with three specialized applications of CFA: higher-order factor analysis, scale reliability evaluation, and models with formative indicators. Each analysis provides a compelling method to address substantively important research questions in a manner unparalleled by alternative approaches (e.g., CFA-based scale reliability vs. Cronbach's alpha, formative latent constructs vs. coarse composites). As with the other types of models discussed throughout this book (e.g., MTMM models, Chapter 6), higher-order factor models and formative indicator models can be embedded in broader SEM analyses (e.g., Figure 8.5), as determined by the research context.

Now that the major types of CFA models have been presented, this book turns to consider analytic issues relevant to any form of CFA. Chapter 9 addresses the complications of missing, non-normal, and categorical data. The final chapter describes methods for determining the appropriate sample size of a CFA study.

NOTES

1. In earlier versions of LISREL, it was often necessary to provide starting values when this form of specification was done. In my experience, this is not necessary in later versions.

2. A related strategy is to impose equality constraints on the formative construct and reflective construct error terms. Often, this approach is not conceptually well defended.

9

Data Issues in CFA

Missing, Non-Normal, and Categorical Data

This chapter presents several data issues frequently encountered in CFA with applied data sets and ways to deal with them (e.g., missing data, non-normality, parceling). In addition, procedures for conducting CFA with categorical indicators are presented. In context of this discussion, the parallels between CFA and item response theory are illustrated. The topics discussed in this chapter are equally germane to broader applications of SEM.

CFA WITH MISSING DATA

Rarely will an applied researcher have the luxury of conducting a statistical analysis (CFA or otherwise) using a data set in which all variables are present for all cases. Indeed, missing data are the norm in research data sets. However, applied research reports often fail to mention how missing data were handled. In addition, the missing data strategies most often used by researchers (e.g., listwise or pairwise deletion) are inappropriate and have deleterious effects, such as loss of statistical power, and bias in parameter estimates, standard errors, and test statistics (Allison, 2002, 2003; Little & Rubin, 2002; Schafer & Graham, 2002).

Mechanisms of Missing Data

Missing data may occur for a variety of reasons. For example, data may be missing by design ("planned missingness") as would occur in cohort-sequential research designs (a longitudinal design frequently used in developmental research), matrix sampling (used in educational testing where all students are not administered identical item sets), or other designs where participants in the sample are randomly selected to complete various subsets of the full assessment battery (possibly due to practical considerations such as financial or time constraints). In addition, data may be missing by chance for reasons other than the research design (e.g., accidental omission of a questionnaire in some assessment packets). In such conditions, data are assumed to be *missing completely at random* (MCAR). The assumption of MCAR holds if the probability of missing data on Y is unrelated to Y or to the values of any other variable in the data set. Using a two-variable (X = gender, Y = life satisfaction) illustration similar to one presented by Allison (2003), the MCAR assumption would hold if the probability of missing data on life satisfaction (Y) is not related to either life satisfaction or gender (X). The condition of MCAR can be readily tested in the research data; for example, do males and females differ in the rate of missing data on life satisfaction? If the assumption holds for all variables in the data set, the cases in the sample with complete data can be considered as a random subsample of the total sample (as would occur if missing data were planned as part of the research design).

The probability of missing data may be related to other variables in the data set. The assumption of *missing at random* (MAR) is met when the probability that data are missing on Y may depend on the value of X, but is not related to the value of Y when holding X constant. Although a weaker assumption, MAR is more likely than MCAR to hold in applied data sets when missing data were not planned by research design. Again using the two-variable example, MAR would hold if males and females differ in their rates of missing data on life satisfaction, but within both levels of gender (i.e., for males and females separately), the likelihood of missing data on life satisfaction does not depend on the level of life satisfaction. It is impossible to test the condition of MAR in research data because the values of missing data are unknown. As Allison (2003) notes, "In essence, MAR allows missingness to depend on things that are observed, but not on things that are not observed" (p. 545). This would be evident in a research study that used a sampling procedure of planned missingness by selective design (e.g., participants are selected for additional testing on the basis of

initial testing scores). MAR is closely related (and sometimes used interchangeably) with the missing data mechanism referred to as *ignorable* (for further discussion, see Allison, 2002, 2003).

Missing data are said to be *nonignorable* if the missingness is related to values that would have been observed (i.e., the condition of MAR does not hold). Missing data are nonignorable if cases with missing data on a given variable would have higher or lower values on that variable than cases with data present, controlling for all other variables in the data set. For instance, using the example of a longitudinal research design (e.g., a treatment outcome study where participants are assessed at pretreatment, posttreatment, and various follow-up points), the data would be MAR if dropouts were related to variables collected at testing occasions prior to the dropout (Schafer & Graham, 2002). The data would be nonignorable if dropouts are related to unseen responses on variables that would have been collected after participants drop out (e.g., missing posttreatment data from participants who failed to respond to treatment; treatment nonresponse is unrelated to variables obtained at pretreatment).

Conventional Approaches to Missing Data

Over the past several years, sophisticated methods for handling missing data have been developed when MCAR or MAR is true (e.g., direct maximum likelihood, multiple imputation; Allison, 1987; Little & Rubin, 2002; Schafer, 1997; Schafer & Graham, 2002). Before discussing these state-of-the-art methods, more common approaches to missing data management are reviewed.

One such approach is *listwise deletion*, in which cases with missing data on any variable used in the analysis are removed from the sample. Listwise deletion has several advantages, including its ease of implementation, the fact that it can be used for any form of statistical analysis, and that the same sample is used for all analyses. In fact, when data are MCAR, listwise deletion will produce *consistent* (i.e., unbiased) parameter estimates, standard errors, and test statistics (e.g., model χ^2). As noted earlier, when MCAR holds, the listwise complete sample can be regarded as a random subsample of the original sample. However, estimates produced using listwise deletion are usually not *efficient*. This is because listwise deletion often results in loss of a considerable proportion of the original sample. Thus, standard errors will frequently be larger when listwise deletion is used, as compared with alternative methods (e.g., multiple imputation) that use all of the available data. The inflation in standard errors will thus

decrease statistical power and lower the precision of the parameter estimates (wider confidence intervals). Indeed, Schafer and Graham (2002) note that listwise deletion can be a very effective (and straightforward) missing data strategy only when the MCAR assumption holds and only a very small part of the sample is removed. When data are MAR, listwise deletion may produce results that are neither consistent nor efficient. Returning to the two-variable example, if males are more likely to have missing data on the life satisfaction variable than females, and if males report less life satisfaction than females, listwise deletion will result in a positively biased estimate of average life satisfaction (although Allison, 2002, discusses scenarios where listwise deletion may outperform other missing data methods when the assumptions of MAR are violated).

Another common missing data strategy is *pairwise deletion*. A variety of statistical analyses (e.g., EFA, CFA, multiple regression, ANOVA) can be performed using means, variances, and covariances as input data (i.e., do not require raw data as input). When pairwise deletion is used, the input vectors (means) and matrices (variances–covariances) are estimated using all cases that have data present for each variable (means, variances) or each pair of variables (covariances). For example, in a sample with 300 total participants who were administered variables X1, X2, and X3, the frequencies with which data are present for the three pairs of indicators is: X1,X2 = 226, X1,X3 = 56, and X2,X3 = 55. Using pairwise deletion, correlations (or covariances) would be estimated for the input matrix using three different Ns (226, 56, 55). This approach could be considered superior to listwise deletion because more data are preserved. When the data are MCAR, pairwise deletion typically produces consistent parameter estimates (fairly unbiased in large samples). Although one might expect that pairwise deletion is more efficient than listwise deletion (because fewer data are lost), this is not always the case (see Allison, 2002). Pairwise deletion has a few other more serious problems. For one, pairwise deletion produces biased standard errors. A related issue is the decision of what sample size to specify when the input matrix is created under pairwise deletion. In the example provided above, N might be specified as 300 (the total sample), 226 (the highest number of cases that have data present on two indicators), or 55 (the smallest number of cases that have data present on two indicators). No single number is satisfactory: the largest N is too liberal (negatively biased standard errors), but the smallest N is too conservative (inflated standard errors). Second, if the data are MAR, the parameter estimates (as well as the standard errors) are often severely biased. Finally, covariance matrices prepared using pairwise deletion may not be positive definite, and thus the statistical model cannot be estimated

(Wothke, 1993). Nonpositive definite matrices were discussed in Chapters 4 and 5. For instance, an input matrix will be nonpositive definite when one or more of its elements have out-of-bounds values. The range of possible values that a correlation (or covariance) may possess is dependent on all other relationships in the input matrix. To illustrate: for three variables, X, Y, and Z, the correlation between X and Y ($r_{x,y}$) must be within a certain range, as determined by the following equation:

$$r_{x,z}r_{y,z} \pm \text{SQRT}[(1 - r_{x,z}^2)(1 - r_{y,z}^2)] \tag{9.1}$$

For example, if $r_{x,z} = .70$ and $r_{y,z} = .80$, then the value of $r_{x,y}$ must be with the range of .13 to .99 (i.e., $.56 \pm .43$). If $r_{x,y} < .13$, then the input matrix would not be positive definite. Thus, out-of-range values may occur when correlations (covariances) are estimated using pairwise deletion.

A less frequently used conventional method of missing data management is *simple (or single) imputation* (mean and regression imputation). For example, regression imputation (referred to by Allison, 2002, as "conditional mean imputation") entails regressing the variable with missing data on other variables in the data set for cases with complete data. The resulting regression equation is used to predict scores for cases with missing data. However, this approach frequently produces underestimates of variances, overestimates of correlations, and underestimates of standard errors. As discussed later in this chapter, the method of *multiple imputation* reconciles these problems in part by introducing random variation into the imputations and by repeating the imputation process several times.

Recommended Missing Data Strategies

More sophisticated methods for handling missing data have been developed in recent years. Maximum likelihood (direct ML) and multiple imputation are the most widely preferred methods for handling missing data in SEM and other data analytic contexts (Allison, 2003; Schafer & Graham, 2002). Both approaches use all the available data; that is, N = the total sample size, including cases with missing data. When either MCAR or MAR is true (and the data have a multivariate normal distribution), maximum likelihood and multiple imputation produce parameter estimates, standard errors, and test statistics (e.g., model χ^2) that are consistent and efficient. These two approaches are now discussed.

The ML estimator can be readily used for the estimation of CFA and SEM models with missing data. When ML is used in context of missing data, it is often referred to as *full information maximum likelihood*, or FIML.

However, this term is not descriptive because ML is a full information estimator regardless of whether the data set is (in)complete. Thus, some methodologists (e.g., Allison, 2003) prefer the term *direct ML* or *raw ML*, because ML estimation with missing data requires that raw data be input to the analysis rather than a variance–covariance matrix (and means).

Before direct ML became widely available, some researchers relied on the *EM algorithm* (Little & Rubin, 2002) for handling missing data in CFA/SEM models (under the assumptions of MAR and multivariate normality). The EM algorithm is a computational device for obtaining ML estimates of the means and the covariance matrix (for details on its computation, see Allison, 2002). These estimates (e.g., variance–covariance matrix) are then used as the input matrix in the CFA/SEM analysis. An advantage of the EM algorithm is that it can be easily implemented in a variety of popular statistical software packages (e.g., SPSS, SAS, SYSTAT, PRELIS). However, a key limitation of using the EM algorithm to calculate input matrices for CFA/SEM is that the resulting standard errors of the parameter estimates are not consistent. Thus, confidence intervals and significance tests may be compromised. As with pairwise deletion, this is due in part to the problem of specifying the proper sample size (Allison, 2003).

Direct ML

Direct ML is free of the problems associated with using the EM algorithm (Allison, 1987; Arbuckle, 1996; Muthén, Kaplan, & Hollis, 1987). Methodologists generally regard direct ML to be the best method for handling missing data in most CFA and SEM applications (e.g., Allison, 2003; Duncan, Duncan, & Li, 1998). Direct ML is available in most current latent variable software packages (e.g., Amos, Mplus, LISREL, Mx) and is very straightforward to use. For example, in LISREL, direct ML is invoked by the "MI" keyword on the Data (DA) line (in tandem with programming LISREL to read a raw data file; e.g., RA=FILE.DAT); for example, MI = 9 would inform LISREL to treat the value of 9 in the raw data file as missing data. Table 9.1 presents Mplus syntax for estimating a simple measurement model of a questionnaire (single factor, four items, one error covariance) with missing data. In this example, an actual research data file was used ($N = 650$). Because the original data set was complete, a number of cases were randomly selected to create missingness on each of the four items (mostly item S4, as might occur if an item was illegible or accidentally omitted on several questionnaires). Thus, in this illustration, it would be reasonable to assume the data are MCAR.

TABLE 9.1. Computer Syntax (Mplus) and Selected Output for Estimation of CFA Model with Missing Data Using Direct Maximum Likelihood

```
TITLE:  CFA WITH MISSING ITEMS (DIRECT ML)
DATA:
  FILE IS CFAMISS.DAT;
VARIABLE:
  NAMES ARE SUBJ S1 S2 S3 S4;
  USEV ARE S1 S2 S3 S4;
  MISSING ARE ALL (9);
ANALYSIS: ESTIMATOR IS ML;
        TYPE=MISSING H1;
MODEL:
    ESTEEM BY S1 S2 S3 S4; S2 WITH S4;
OUTPUT:  STANDARDIZED MODINDICES(4.00) SAMPSTAT PATTERNS;



SUMMARY OF DATA
     Number of patterns            5


SUMMARY OF MISSING DATA PATTERNS

      MISSING DATA PATTERNS

           1  2  3  4  5
  S1       x  x  x  x
  S2       x  x  x     x
  S3       x  x     x  x
  S4       x     x  x  x

      MISSING DATA PATTERN FREQUENCIES

   Pattern   Frequency    Pattern   Frequency    Pattern   Frequency
         1         385          3          25          5          25
         2         190          4          25

COVARIANCE COVERAGE OF DATA

Minimum covariance coverage value   0.100

      PROPORTION OF DATA PRESENT


          Covariance Coverage
            S1          S2          S3          S4
          _____    _____    _____    _____
  S1        0.962
  S2        0.923       0.962
  S3        0.923       0.923       0.962
  S4        0.669       0.669       0.669       0.708
```

Only a couple additions are needed to Mplus syntax for a typical CFA model. First, the statement "MISSING ARE ALL (9)" informs Mplus that, for all variables, "9" indicates a missing value. Second, under the ANALYSIS command, the statement "TYPE = MISSING" informs Mplus to conduct direct ML. The H1 option allows the estimation of an unrestricted mean and covariance model under direct ML. On the OUTPUT line, the PATTERNS option is requested to obtain a summary of missing data patterns. Otherwise, the programming and output of CFA results (e.g., fit statistics, parameter estimates, standard errors) are the same as those of a CFA analysis without missing data.

Table 9.1 presents output of the analysis produced by the PATTERNS option. It is important to examine the missing data patterns in order to determine how well the model can be estimated. The results show that there are five missing data patterns ("x" = data are present, a blank = data are missing). In the first pattern, data are present for all four items (n = 385). When data are missing, the second pattern is most common—that is, all items except S4 are present (n = 190). In the remaining three patterns (ns = 25), a single item other than S4 is missing. The next section of the output presents the covariance coverage in the data. Covariance coverage indicates the proportion of data that were available for each indicator and pair of indicators. For example, as only 25 cases were missing item S1 (missing data pattern 5), coverage of this item is .962 (625/650 = .962). Sixty-seven percent (.669) of the sample had data present for both S1 and S4. Muthén and Muthén (1998–2004) note that coverage of about 50% usually causes no problems with direct ML estimation (and H1) in Mplus. By default, Mplus stops the analysis if coverage is less than 10% for any covariance (although this default can be overridden by an option on the ANALYSIS command; e.g., COVERAGE = .05).

Direct ML assumes that the data are MCAR or MAR and multivariate normal. However, when data are non-normal, direct ML can be implemented to provide standard errors and test statistics that are robust to nonnormality using the MLR estimator (robust ML; Yuan & Bentler, 2000). This feature has been added to recent releases of many programs such as Mplus and EQS.

Multiple Imputation

In most cases, direct ML represents the best and easiest way to manage missing data in CFA and SEM analyses. However, the procedure of *multiple imputation* (Rubin, 1987) is a valuable alternative to direct ML. Multiple

imputation possesses statistical properties that closely approximate ML (Allison, 2002, 2003; Schafer, 1997). As Allison (2003) notes, multiple imputation is a useful approach to missing data when the researcher does not have access to a program capable of direct ML or wishes to estimate a CFA/SEM model with a fitting function other than ML.

As discussed earlier, simple imputation procedures such as mean or regression imputation are problematic because they produce underestimates of variances and overestimates of correlations among the variables with imputed data. For instance, if regression imputation was used to supply values on variable Y from data that are available on variable X (i.e., $\hat{Y} = a + bX$), the correlation between X and Y would be overestimated (i.e., for cases with imputed Y values, X is perfectly correlated with Y). Multiple imputation reconciles this problem by introducing random variation into the process. In other words, missing values for each case are imputed on the basis of observed values (as in regression imputation), but random noise is incorporated to preserve the proper degree of variability in the imputed data. In the bivariate example, inclusion of this random variation is reflected by the equation

$$\hat{Y} = a + bX + S_{x,y}E \qquad (9.2)$$

where $S_{x,y}$ is the estimated standard deviation of the regression's error term (root mean squared error), and E is a random draw (with replacement) from a standard normal distribution. However, if this procedure is conducted a single time, the resulting standard errors would be too small because the standard error calculation would not account adequately for the additional variability in the imputed data. This problem is solved by repeating the imputation process multiple times. Because random variation is included in the process, the resulting data sets will vary slightly. The variability across imputed data sets is used to upwardly adjust the standard errors.[1]

Thus, there are three basic steps of multiple imputation. The first step is to impute multiple data sets. When the amount of missing data is no more than moderate, generating five imputed data sets ($M = 5$) is usually sufficient (Allison, 2003) and is the default in some software packages (e.g., SAS PROC MI; Yuan, 2000).[2] In the second step, the M data sets are analyzed using standard analytic procedures (i.e., CFA, ANOVA, etc.). In the third step, the results from the M analyses are combined into a single set of parameter estimates, standard errors, and test statistics. Parameter estimates are combined by simply averaging the estimates across the M

analyses. Standard errors are combined using the average of the standard errors over the set of analyses and the between-analysis parameter estimate variation (Schafer, 1997); for the exact formula of this calculation, see either Allison (2003) or Rubin (1987).

A few programs for multiple imputation are now available. The first widely used program for multiple imputation, NORM (a free, downloadable program for Windows) was developed by Schafer (1997). Using the algorithms contained in NORM, SAS has introduced a relatively new procedure (PROC MI; Yuan, 2000) for multiple imputation. MI procedures are also available in recent releases of LISREL (e.g., version 8.72). Recent releases of Mplus (e.g., version 3.1) have the capability of combining parameter estimates, standard errors, and test statistics of CFA/SEM analyses conducted on data sets that were created by multiple imputation in another software program.

Table 9.2 presents the SAS syntax for estimating a CFA model with missing data using multiple imputation. The same data set used in Table 9.1 (direct ML using Mplus) is used in this example. The syntax listed in Step 1 generates multiple imputations using SAS PROC MI for this example. The PROC MI syntax is very straightforward. The names of the input and output data files are specified, along with the list of variables to be used in the imputation process. A couple of PROC MI options are also illustrated in the syntax. The NIMPU option specifies the number of imputations to be conducted (in this example, the number of imputations specified is redundant with the SAS default). The SEED option specifies a positive integer that is used to start the pseudorandom number generator (the SAS default is a value generated from reading the time of day from the computer's clock). In this example, a seed value is specified so that the results can be duplicated in separate computer runs. This option touches on a minor drawback in multiple imputation: because random variation is introduced, multiple imputation will yield slightly different results each time it is conducted.[3] There are a variety of other options available in PROC MI, including options to make the imputed values consistent with the observed variable values (i.e., set minimum and maximum values, round imputed values to desired units) and to specify the method of imputation (the SAS default is Markov Chain Monte Carlo, MCMC, which in essence is a method of improving the estimates of imputed data in the context of complex missing data patterns by using the imputed data to produce optimal estimates of the regression coefficients). See Yuan (2000) and Allison (2003) for a discussion of other options and complications with multiple imputation.

TABLE 9.2. Computer Syntax (SAS) and Selected Output for Estimation of CFA Model with Missing Data Using Multiple Imputation

---

Step 1: Creation of Data Sets with Imputed Data

```
proc mi data=sedata nimpu=5 seed=44176 out=seimp;
  var s1 s2 s3 s4;
run;
```

Step 2: CFA of Data Sets Created by Multiple Imputation

```
libname sys 'c:\missing\';
data esteem;
 set sys.seimp;

proc calis data=esteem method=ml cov privec pcoves;
var = s1-s4;
lineqs
  s1 = 1.0 f1 + e1,
  s2 = lam2 f1 + e2,
  s3 = lam3 f1 + e3,
  s4 = lam4 f1 + e4;
std
  f1 = ph1,
  e1-e4 = th1 th2 th3 th4;
cov
  e2 e4 = th5;
by _imputation_;
ods output Estimates=a covmat=b;
run;
```

Step 3: Combining Parameter Estimates and Standard Errors

```
proc mianalyze parms=a covb=b;
var lam2 lam3 lam4 ph1 th1 th2 th3 th4 th5;
run;
```

Selected Output

```
            Multiple  Imputation  Parameter  Estimates
```

Parameter	Estimate	Std Error	95% Confidence	Limits	DF
lam2	1.375094	0.062479	1.252596	1.497593	3486.7
lam3	1.354973	0.057990	1.241247	1.468698	2076
lam4	1.265969	0.061006	1.145971	1.385966	339.46
ph1	0.834902	0.078748	0.680498	0.989305	3095.1
th1	0.690190	0.044470	0.602668	0.777712	292.53
th2	0.288787	0.044836	0.200662	0.376912	429.23
th3	0.451633	0.041280	0.370563	0.532703	602.7
th4	0.298328	0.050084	0.195041	0.401615	24.359
th5	0.254208	0.043327	0.164350	0.344070	21.985

TABLE 9.2. *(cont.)*

Multiple Imputation Parameter Estimates

Parameter	Minimum	Maximum	t for H0: Parameter=Theta0	Pr > |t|
lam2	1.364608	1.387795	22.01	<.0001
lam3	1.345343	1.371108	23.37	<.0001
lam4	1.246292	1.288201	20.75	<.0001
ph1	0.814763	0.852187	10.60	<.0001
th1	0.679365	0.712501	15.52	<.0001
th2	0.266607	0.298306	6.44	<.0001
th3	0.435737	0.463256	10.94	<.0001
th4	0.256843	0.330129	5.96	<.0001
th5	0.233794	0.298554	5.87	<.0001

At a minimum, all variables that are to be used in the substantive analysis (e.g., CFA) should be included in the multiple imputation process. Although not shown in the Table 9.2 example, multiple imputation is often fostered by making use of additional variables that are not used in the substantive analysis. This is also true for direct ML. Such "extra" variables would entail variables that are correlated with the variables of interest that have missing data, and variables that are associated with the probability of missingness for the variables of interest. In SAS PROC MI, these extra variables would simply be added to the variable list.

In the second step of Table 9.2, SAS PROC CALIS is used to conduct a CFA for each of the five completed data sets produced by PROC MI. PROC MI writes the five data sets to a single file, adding the variable _IMPUTATION_ (with range of values = 1–5) to distinguish each data set. As can be seen in Table 9.2, the CALIS syntax looks typical except for the final two lines. The "by _imputation_" line requests a separate CFA for each value of the variable, _IMPUTATION_ (i.e., five separate CFAs). The final line (ods = output delivery system) informs SAS to write the parameter estimates ("Estimates") and covariance matrix ("covmat") to SAS data sets (named "a" and "b") for further analysis.

The third step of Table 9.2 demonstrates the use of SAS PROC MIANALYZE, a SAS procedure that combines the results of the multiple analyses and generates valid statistical inferences about each parameter. The keyword "PARMS" is used to indicate a SAS data set that contains parameter estimates from imputed data sets; the keyword "COVB" names a SAS data set that contains covariance matrices of the parameter estimates

from imputed data sets (if the COVB option is used, so must PARMS, and vice versa). The variable (VAR) statement is required to list the names of the parameters to be combined (in this example, the freely estimated factor loadings, factor variance, measurement error variances, and error covariance).

In addition, Table 9.2 presents selected output generated by PROC MIANALYZE. Included in this output are the averaged unstandardized parameter estimates, their standard errors, their 95% confidence intervals, range of values across imputations, and test statistics (see "t for H0:" column; that is, parameter estimate divided by its standard error) and associated p values. These results would be interpreted in the same fashion as a standard CFA of a single, complete data set (e.g., unstandardized factor loading of S2 = 1.375, SE = .0625, z = 22.01, p < .001). Through simple manipulations (e.g., see Chapter 4), a completely standardized solution could be readily calculated.

Table 9.3 presents Mplus syntax for analyzing and combining the results of data sets generated by multiple imputation.[4] In Mplus, the multiple imputations must be produced by another program (e.g., SAS PROC MI, NORM). In this example, the five data sets created by PROC MI (see Table 9.2) were exported to a text file. The text file was then split into five files corresponding to the five separate data sets generated by PROC MI (named data1.dat, data2.dat . . . data5.dat). In Mplus, the FILE option of the DATA command is used to provide the name of the file that contains the names of the multiple imputation data sets to be analyzed. In this case, the file is named "IMPUTE.DAT." IMPUTE.DAT is a text file created by the user that contains the names of the multiple imputation data sets (e.g., data1.dat, data2.dat . . . data5.dat; each file name is on a separate line). Specifying "TYPE=IMPUTATION" on the DATA command instructs Mplus to run an analysis for each data set named in IMPUTE.DAT. The results of the analyses are then combined using the same methods used by PROC MIANALYZE (although Mplus also provides averaged fit statistics such as model χ^2, RMSEA, CFI, etc.; Table 9.3 presents the average model χ^2). As can be seen in Table 9.3, the combined Mplus results are roughly the same (within rounding error) as the combined results produced by PROC MIANALYZE (cf. Table 9.2).

Table 9.4 presents the unstandardized parameter estimates, standard errors, and model χ^2 generated by the various methods of handling missing data discussed in this chapter. Recall that a problem with pairwise deletion is determining the appropriate sample size to specify in the analysis. For this illustration, two Ns were used: (1) the number of nonmissing

TABLE 9.3. Computer Syntax (Mplus) and Selected Output for Combining CFA Results from Multiple Imputation Data Sets

```
TITLE:  CFA WITH MISSING ITEMS (MULTIPLE IMPUTATION)
DATA:
  FILE IS "C:\IMPUTE.DAT";
  TYPE = IMPUTATION;
VARIABLE:
  NAMES ARE IMP SUBJ S1 S2 S3 S4;
  USEV ARE S1 S2 S3 S4;
ANALYSIS: ESTIMATOR IS ML;
MODEL:
    ESTEEM BY S1 S2 S3 S4; S2 WITH S4;



TESTS OF MODEL FIT

Number of Free Parameters                        9

Chi-Square Test of Model Fit

        Degrees of freedom                       1

        Mean                                 2.057
        Std Dev                              1.232
        Number of successful computations        5


MODEL RESULTS

                    Estimates    S.E.   Est./S.E.
 ESTEEM    BY
    S1                 1.000    0.000      0.000
    S2                 1.375    0.062     22.087
    S3                 1.355    0.058     23.467
    S4                 1.266    0.060     20.956

 S2        WITH
    S4                 0.254    0.042      6.090

 Variances
    ESTEEM             0.834    0.078     10.641

 Residual Variances
    S1                 0.689    0.044     15.685
    S2                 0.288    0.044      6.498
    S3                 0.451    0.041     11.023
    S4                 0.298    0.048      6.171
```

TABLE 9.4. CFA Unstandardized Parameter Estimates and Model χ^2 for Various Methods of Missing Data Management

	$\lambda_{2,1}$	$\lambda_{3,1}$	$\lambda_{4,1}$	$\theta_{4,2}$	$\phi_{1,1}$	χ^2
No Missing ($N = 650$)						
Parameter Estimate	1.360	1.351	1.230	0.238	0.843	1.32
Standard Error	0.060	0.056	0.056	0.032	0.077	
Listwise Deletion ($N = 385$)						
Parameter Estimate	1.411	1.361	1.273	0.285	0.844	0.11
Standard Error	0.081	0.074	0.075	0.043	0.102	
Pairwise Deletion						
Parameter Estimate	1.383	1.366	1.268	0.235	0.827	
Standard Error ($N = 460$)	0.073	0.068	0.068	0.039	0.091	0.89
Standard Error ($N = 650$)	0.062	0.057	0.058	0.033	0.077	1.26
Direct ML ($N = 650$)						
Parameter Estimate	1.383	1.360	1.275	0.258	0.826	1.27
Standard Error	0.064	0.059	0.063	0.039	0.078	
Multiple Imputation ($N = 650$)						
Parameter Estimate	1.375	1.355	1.266	0.254	0.834	2.06
Standard Error	0.062	0.058	0.060	0.042	0.078	

Note. $\lambda_{2,1}$–$\lambda_{4,1}$ = unstandardized factor loadings for items S2, S3, and S4; $\theta_{4,2}$ = error covariance of items S2 and S4; $\phi_{1,1}$ = Self-Esteem factor variance; ML, maximum likelihood.

cases for the variable with the most missing data ($N = 460$); and (2) the size of the full sample ($N = 650$; cf. Allison, 2003). In this example, the results produced by the various missing data methods are similar. This can be attributed in large part to the size of the sample (elimination of 41% of the sample with at least one missing observation still results in an $N = 385$) and the fact that the data are MCAR. The biggest difference is seen in the standard errors produced in the analysis using listwise deletion. Specifically, these standard errors are somewhat larger than those produced by the other methods, which use more of the available data. Note the similarity of the parameter estimates and standard errors produced by direct ML and multiple imputation. If one were to round to the second decimal point, virtually all of these values would be the same.

In summary, direct ML and multiple imputation are strong methodologies for handling missing data when the data are either MCAR or MAR. If missing data are nonignorable (i.e., the MAR assumption does not hold), these procedures will yield misleading results (unfortunately, although there are often reasons to believe that data are not MAR, there is no way to test this assumption). Methodologists have developed procedures of estimation with nonignorable missing data (e.g., pattern-mixture models). However, such procedures require firm knowledge of the missing data mechanism. In addition, such models are prone to underidentification and are very difficult to estimate properly (Allison, 2002, 2003; Little & Rubin, 2002). Allison (2002, 2003) urges that methods for nonignorable missing data should be interpreted with great caution and should be accompanied by sensitivity analyses to explore the effects of different modeling assumptions (e.g., mechanisms of missingness).

CFA WITH NON-NORMAL OR CATEGORICAL DATA

In previous examples in this book, the maximum likelihood (ML) estimator was used. The vast majority of CFA and SEM models reported in the applied research literature use ML. However, an alternative to ML for normal, continuous data is generalized least squares (GLS). GLS is a computationally simpler fitting function and produces approximately the same goodness of fit as ML (i.e., $F_{ML} = F_{GLS}$), especially when sample size is large. Nevertheless, ML (and GLS) are appropriate only for multivariate normal, interval-type data (i.e., the joint distribution of the continuous variables is distributed normally). When continuous data depart markedly from normality (i.e., marked skewness or kurtosis), or when some of the

indicators are not interval level (i.e., binary, polytomous, ordinal), an estimator other than ML should be used.[5] In this section, alternatives to ML are discussed for such situations.

Non-Normal, Continuous Data

Research has shown that ML (and GLS) is robust to minor departures in normality (e.g., Chou & Bentler, 1995). However, when non-normality is more pronounced, an estimator other than ML should be used to obtain reliable statistical results (i.e., accurate goodness-of-fit statistics and standard errors of parameter estimates). ML is particularly sensitive to excessive kurtosis. The consequences of using ML under conditions of severe non-normality include (1) spuriously inflated model χ^2 values (i.e., overrejection of solutions); (2) modest underestimation of fit indices such as the TLI and CFI; and (3) moderate to severe underestimation of the standard errors of the parameter estimates (inflating the risk of Type I error—i.e., concluding that a parameter is significantly different from zero when that is not the case in the population) (West, Finch, & Curran, 1995). These deleterious effects are exacerbated as sample size decreases (i.e., the risk for nonconverging or improper solutions increases). The two most commonly used estimators for non-normal continuous data are (1) robust ML (Bentler, 1995; Satorra & Bentler, 1994); and (2) weighted least squares (WLS; Browne, 1984b). For reasons discussed in the next section (e.g., requirement of extremely large samples), WLS is not recommended. In contrast, research has shown that robust ML is a very well-behaved estimator across different levels of non-normality (except in instances of severe floor or ceiling effects), model complexity, and sample size (e.g., Chou & Bentler, 1995; Curran et al., 1996).

The robust ML estimator (hereafter abbreviated, MLM) provides ML parameter estimates with standard errors and a mean-adjusted χ^2 test statistic that are robust to non-normality. The mean-adjusted χ^2 test statistic is often referred to as the Satorra–Bentler scaled χ^2 (SB χ^2; Satorra & Bentler, 1994). Although MLM was first introduced in the EQS program, it is now available in some other latent variable software packages (e.g., Mplus, LISREL). To use MLM in EQS or Mplus, raw data must be inputted. In LISREL, its preprocessor companion, PRELIS, must be used to generate a covariance matrix and asymptotic covariance matrix for use as input in the subsequent CFA analysis. The asymptotic covariance matrix is used to compute a weight matrix (*W*, to adjust fit statistics and standard errors for non-normality) in subsequent LISREL analyses that rely on a non-normal theory estimator such as MLM (Jöreskog & Sörbom, 1996b).

Table 9.5 presents Mplus, EQS, and PRELIS/LISREL syntax for conducting a simple CFA model (one factor, five indicators, $N = 870$) using the MLM estimator. The table shows how raw data can be read into Mplus and EQS (see DATA and /SPECIFICATIONS lines, respectively). The MLM is requested in Mplus and EQS by the "ESTIMATOR IS MLM" and "METHODS=ML, ROBUST" syntax, respectively. Otherwise, the syntax is like that of a typical CFA analysis. In this first example, a congeneric model is specified (one factor, no correlated measurement errors). A later example entails a respecified CFA model of this data set (i.e., correlated error of X1 and X3) to illustrate nested χ^2 evaluation when the MLM estimator is employed.

The LISREL preprocessing program, PRELIS, serves a variety of functions, including tests of normality, multiple imputation, bootstrapping, Monte Carlo studies, and generating various types of matrices (e.g., covariance, tetrachoric correlations, polychoric correlations) from raw text-file data or data imported from other software programs such as SPSS. In this example, PRELIS is used to test for normality and create matrices to be used as input for LISREL (e.g., covariance and asymptotic covariance matrices). The line CO ALL is included to declare all five indicators as continuous. PRELIS regards any variable with less than 16 distinct values as ordinal by default. Thus, this command is necessary in instances where indicators with a sample range of 15 or less are desired to be treated as interval level. Syntax on the PRELIS output line (OU) saves the covariance and asymptotic covariance matrices to external files (in addition, a file containing indicator means is also saved). The LISREL programming is the same as for a typical CFA except that the covariance matrix (CM) and asymptotic covariance matrix (AC) are used as input for the analysis. Note that ML is still requested as the estimator on the output (OU) line (i.e., ME=ML). However, because an asymptotic covariance matrix has been inputted, LISREL will provide the SB χ^2 (along with the usual model χ^2) and standard errors that are robust to non-normality.

Table 9.6 provides selected results from EQS and PRELIS with regard to the univariate normality of the five indicators (normality tests are automatically produced by EQS when MLM is requested). As shown in this table, EQS and PRELIS produce very similar results. Some of the indicators evidence considerable non-normality (e.g., kurtosis of X5 = 9.4), and thus the assumption of multivariate normality does not hold (although univariate normality does not ensure multivariate normality, univariate non-normality does ensure multivariate non-normality).

TABLE 9.5. Mplus, EQS, and LISREL/PRELIS Syntax for Conducting CFA
with Non-Normal, Continuous Data Using Robust Maximum Likelihood

Mplus

```
TITLE:  CFA WITH NON-NORMAL, CONTINUOUS DATA (ROBUST ML)
DATA:
    FILE IS NONML.DAT;
VARIABLE:
    NAMES ARE x1 x2 x3 x4 x5;
ANALYSIS: ESTIMATOR IS MLM;
MODEL:
    F1 BY x1 x2 x3 x4 x5;
                !without x1, x3 correlated residual (ADDED IN 2ND RUN);
OUTPUT:  STANDARDIZED MODINDICES(10.00) SAMPSTAT;
```

EQS

```
/TITLE
 one-factor measurement model: non-normal, continuous data
/SPECIFICATIONS
 CASES=870; VARIABLES=5; METHODS=ML, ROBUST; MATRIX=RAW; ANALYSIS=COV;
 DATA = 'NONML.DAT'; FO = '(5F2.0)';
/LABELS
 v1=item1; v2= item2; v3= item3; v4= item4; v5= item5;
 f1 = factor1;
/EQUATIONS
 V1 =  F1+E1;
 V2 = *F1+E2;
 V3 = *F1+E3;
 V4 = *F1+E4;
 V5 = *F1+E5;
/VARIANCES
 F1 = *;
 E1 TO E5= *;
 !/COVARIANCES           ! ADDED IN SECOND RUN
 ! E1, E3 = *;
/PRINT
 fit=all;
/LMTEST
/END
```

LISREL/PRELIS

```
PRELIS PROGRAM: CFA WITH NON-NORMAL CONTINUOUS DATA
DA NI=5 NO=870
RA FI=NONML.DAT
CO ALL
LA
X1 X2 X3 X4 X5
OU MA=CM ME=NNML.ME CM=NNML.CM AC=NNML.ACC
```

(cont.)

TABLE 9.5. *(cont.)*

```
TITLE LISREL CFA WITH NON-NORMAL, CONTINUOUS DATA
DA NI=5 NO=870 MA=CM
CM=NNML.CM
AC=NNML.ACC
ME=NNML.ME    ! NOT NEEDED UNLESS CONDUCTING MEANSTRUCTURE ANALYSIS
LA
X1 X2 X3 X4 X5
MO NX=5 NK=1 PH=SY,FR LX=FU,FR TD=SY,FR
LK
FACTOR1
PA LX
0
1
1
1
1
VA 1.0 LX(1,1)
PA TD
1
0 1
0 0 1        ! X1, X3 CORRELATED ERROR ADDED ON SECOND RUN, TD(3,1)
0 0 0 1
0 0 0 0 1
OU ME=ML RS MI SC AD=OFF IT=100 ND=4
```

When MLM is requested, EQS will output both the ML χ^2 and SB χ^2, as well as both the ML standard errors and robust standard errors of the unstandardized parameter estimates (although LISREL will also provide the ML χ^2, Mplus and LISREL provide only robust standard errors when MLM is requested). In Table 9.7, χ^2 values and the unstandardized factor loadings (and their standard errors) are presented from EQS output. These results demonstrate the typical consequences of using ML in the context of non-normal data. Specifically, the ML χ^2 (87.48) is considerably larger than the SB χ^2 (33.13), reflecting the tendency for ML to produce inflated χ^2 values when data are non-normal. In addition, the standard errors of the ML estimates are noticeably smaller than those based on MLM (e.g., .027 vs. .051 for the V2 indicator), illustrating the propensity for ML to underestimate standard errors in this context. The underestimation of standard errors results in inflated test statistics (e.g., zs for V2 = 22.60 and 12.15 for ML and MLM, respectively), thus increasing the risk of Type I error. Note that the parameter estimates are not affected (e.g., λ_{21} = .618) by the type of estimator used (e.g., λ_{21} = .618 in both ML and MLM). Although not

TABLE 9.6. EQS and PRELIS Selected Results of Tests of Normality

EQS

UNIVARIATE STATISTICS

VARIABLE	ITEM1	ITEM2	ITEM3	ITEM4	ITEM5
MEAN	1.4701	0.8230	1.2655	1.0264	0.6069
SKEWNESS (G1)	1.5090	2.4025	1.8010	2.1612	3.1093
KURTOSIS (G2)	1.2624	5.6901	2.3555	3.9936	9.4023
STANDARD DEV.	2.1728	1.6015	2.0700	1.9280	1.5192

MULTIVARIATE KURTOSIS

MARDIA'S COEFFICIENT (G2,P) = 62.5153
NORMALIZED ESTIMATE = 110.1963

PRELIS

Univariate Summary Statistics for Continuous Variables

Variable	Mean	St. Dev.	T-Value	Skewness	Kurtosis	Minimum	Freq.	Maximum	Freq.
X1	1.470	2.173	19.957	1.512	1.277	0.000	472	8.000	24
X2	0.823	1.601	15.158	2.407	5.730	0.000	591	8.000	8
X3	1.266	2.070	18.032	1.804	2.376	0.000	509	8.000	24
X4	1.026	1.928	15.703	2.165	4.024	0.000	575	8.000	22
X5	0.607	1.519	11.783	3.115	9.464	0.000	669	8.000	7

Test of Univariate Normality for Continuous Variables

	Skewness		Kurtosis		Skewness and Kurtosis	
Variable	Z-Score	P-Value	Z-Score	P-Value	Chi-Square	P-Value
X1	13.682	0.000	5.020	0.000	212.393	0.000
X2	17.956	0.000	10.666	0.000	436.185	0.000
X3	15.269	0.000	7.222	0.000	285.303	0.000
X4	16.957	0.000	9.275	0.000	373.562	0.000
X5	20.434	0.000	12.571	0.000	575.587	0.000

TABLE 9.7. Selected EQS Results of CFA with Robust Maximum Likelihood Estimation

```
GOODNESS OF FIT SUMMARY
INDEPENDENCE MODEL CHI-SQUARE =        2523.941 ON    10 DEGREES OF FREEDOM

CHI-SQUARE =        87.478 BASED ON     5 DEGREES OF FREEDOM
PROBABILITY VALUE FOR THE CHI-SQUARE STATISTIC IS LESS THAN 0.001

THE NORMAL THEORY RLS CHI-SQUARE FOR THIS ML SOLUTION IS         90.680.

SATORRA-BENTLER SCALED CHI-SQUARE =     33.1281
PROBABILITY VALUE FOR THE CHI-SQUARE STATISTIC IS     0.00000


MEASUREMENT EQUATIONS WITH STANDARD ERRORS AND TEST STATISTICS
(ROBUST STATISTICS IN PARENTHESES)

ITEM1   =V1  =   1.000 F1    + 1.000 E1




 ITEM2   =V2  =    .703*F1    + 1.000 E2
                   .035
                  20.121
                (   .062)
                ( 11.339)

 ITEM3   =V3  =   1.068*F1    + 1.000 E3
                   .034
                  31.712
                (   .044)
                ( 24.119)

 ITEM4   =V4  =    .918*F1    + 1.000 E4
                   .042
                  21.762
                (   .063)
                ( 14.613)

 ITEM5   =V5  =    .748*F1    + 1.000 E5
                   .033
                  22.404
                (   .055)
                ( 13.588)
```

Note. Unstandardized parameter estimates are derived from a revised solution that adds the parameter of an error covariance between items 1 and 3 (see Table 9.5).

shown in Table 9.7, the values of other commonly used fit indices (e.g., TFI, CFI, RMSEA, SRMR) also differ to varying degrees (e.g., RMSEA = 0.138 and .082 in ML and MLM, respectively).

Chi-square difference testing can be conducted using the SB χ^2 statistic. However, unlike ML-based analysis, this test cannot be conducted by simply calculating the difference in χ^2 values produced by the nested and comparison models. This is a common mistake in the applied research literature. The reason is that a difference between two SB χ^2 values for nested models is not distributed as χ^2. Thus, a *scaled difference in χ^2s* (SDCS) test should be used (Satorra & Bentler, 1994). The SDCS test statistic, T_S, is computed as

$$T_S = (T_0 - T_1) \ / \ c_d \tag{9.3}$$

where T_0 is the regular ML χ^2 for the nested model, T_1 is the regular ML χ^2 for the comparison (less restricted) model, and c_d is the difference test scaling correction. c_d is defined as

$$c_d = [(d_0 * c_0) - (d_1 * c_1)] \ / \ (d_0 - d_1) \tag{9.4}$$

where d_0 is the degrees of freedom of the nested model, d_1 is the degrees of freedom of the comparison model, c_0 is the scaling correction factor for the nested model, and c_1 is the scaling correction factor for the comparison model. In Mplus, scaling correction factors are automatically provided when the MLM estimator is used. If a program other than Mplus is being used, scaling correction factors can be readily computed by dividing the regular ML χ^2 by the SB χ^2:

$$c_0 = T_0 \ / \ T_0{}^* \tag{9.5}$$

where $T_0{}^*$ is the SB χ^2 value.

The process of computing the SCDS test statistic (T_S) is illustrated using the previous one-factor CFA model. After fitting the initial CFA model, the results indicate that the fit of the solution could be improved by specifying a correlated error between indicators X1 and X3. Because this modification can be defended on substantive grounds, a revised solution is pursued. Table 9.8 breaks down the process of computing T_S into three steps. First, the χ^2s must be obtained from ML and MLM for the nested and comparison models (the initial CFA model is nested within the CFA

TABLE 9.8. Computing the Scaled Difference in Chi-Squares (SDCS) Test for Nested Models Estimated with Robust Maximum Likelihood (MLM)

Step 1. Obtain T_0 and T_1 Values and Scaling Correction Factors for Nested and Comparison Models

Nested Model (One-Factor, No Correlated Measurement Errors)

	T_0	$T_0{}^*$	d_0	c_0
ML	87.478		5	
MLM		33.128	5	2.641[*]

$$^*c_0 = T_0 \, / \, T_0{}^* = 87.478 \, / \, 33.128 = 2.641$$

Comparison Model (One-Factor, One Correlated Measurement Error)

	T_1	$T_1{}^*$	d_1	c_1
ML	25.833		4	
MLM		10.137	4	2.502[*]

$$^*c_1 = T_1 \, / \, T_1{}^* = 25.833 \, / \, 10.137 = 2.548$$

Step 2. Compute the Difference Test Scaling Correction (c_d)
$$c_d = [(d_0 * c_0) - (d_1 * c_1)] \, / \, (d_0 - d_1)$$
$$c_d = [(5 * 2.641) - (4 * 2.548)] \, / \, (5 - 4) = 3.013$$

Step 3. Compute the Satorra–Bentler SDCS Test
$$T_S = (T_0 - T_1) \, / \, c_d$$
$$T_S = (87.478 - 25.833) \, / \, 3.013 = 20.46$$
$$T_S = 20.46, \, df = 1, \, p < .001$$

Note. T_0 = ML χ^2 for nested model; $T_0{}^*$ = Satorra–Bentler scaled χ^2 for nested model; T_1 = ML χ^2 for comparison model; $T_1{}^*$ = Satorra–Bentler scaled χ^2 for comparison model; d_0 = degrees of freedom of nested model; d_1 = degrees of freedom of comparison model; c_0 = scaling correction factor for nested model; and c_1 = scaling correction factor for comparison model.

model with a correlated error with a single degree of freedom; $d_0 - d_1 = 5 - 4 = 1$). These four χ^2 values are then used to calculate the scaling correction factors for the nested and comparison models; for example, $c_0 = T_0 \, / \, T_0{}^* = 87.478 \, / \, 33.128 = 2.641$ (see Table 9.8). In the second step, the difference test scaling correction (c_d) is computed using the scaling correction factors and degrees of freedom from the nested and comparison models. As shown in Table 9.8, c_d in this example equals 3.013. In the third and final step, T_S is obtained by dividing the difference between the ML χ^2 values of the nested and comparison models by c_d; that is, $T_S = (87.478 - 25.833) \, / \, 3.013 = 20.46$. T_S is interpreted in the same fashion as the regular χ^2 difference test. Because the T_S value is statistically significant ($df = 1$,

$p < .001$), it can be concluded that the revised one-factor model provides a significantly better fit to the data than the original one-factor solution. In this example, the traditional χ^2 difference test would have yielded the same conclusion (i.e., $T_0 - T_1 = 87.478 - 25.833 = 61.645$, $p < .001$), although the differential magnitude of the χ^2 differences is appreciable (20.46 vs. 61.645). However, there are many situations where using the standard χ^2 difference test to compare nested models estimated by MLM will yield misleading results (i.e., provide false evidence of the statistical equivalence or difference between the fit of the nested and comparison model solutions). Thus, the SCDS test should always be employed when comparing nested solutions estimated by MLM.

Categorical Data

When at least one factor indicator is categorical (i.e., dichotomous, polytomous, ordinal), ordinary ML should not be used to estimate CFA models. The potential consequences of treating categorical variables as continuous variables in CFA are multifold, including that it can (1) produce attenuated estimates of the relationships (correlations) among indicators, especially when there are floor or ceiling effects; (2) lead to "pseudofactors" that are artifacts of item difficulty or extremeness; and (3) produce incorrect test statistics and standard errors. ML can also produce incorrect parameter estimates, such as in cases where marked floor or ceiling effects exist in purportedly interval-level measurement scales (i.e., because the assumption of linear relationships does not hold). Thus, it is important that an estimator other than ML be used with categorical outcomes or severely non-normal data.[6]

There are a handful of estimators that can be used with categorical indicators; for example, weighted least squares (WLS), robust weighted least squares (WLSMV), and unweighted least squares (ULS). Historically, WLS (Browne, 1984b) has been the most frequently used estimator for categorical outcomes (despite the fact that it was originally intended for use with non-normal, continuous data). WLS is available in all of the major latent variable software programs (in Amos, WLS is referred to as asymptotically distribution-free, ADF; in EQS, WLS is referred to as arbitrary generalized least squares, AGLS). WLS is closely related to the GLS estimator. Like ML, GLS minimizes the discrepancy between the observed (S) and predicted (Σ) covariance matrices. However, GLS uses a weight matrix (W) for the residuals. In GLS, W is typically the inverse of S. WLS uses a different W; specifically, one that is based on estimates of the variances and

covariances of each element of S, and fourth-order moments based on multivariate kurtosis (Jöreskog & Sörbom, 1996b; Kaplan, 2000). Thus, unlike GLS, the WLS fit function is weighted by variances/covariances and kurtosis to adjust for violations in multivariate normality; that is, if there is no kurtosis, WLS and GLS will produce the same minimum fit function value, $F_{WLS} = F_{GLS}$.

Because W in WLS is based on the variances and covariances of each element of S (in other words, the "covariances of the covariances"), it can be extremely large, especially when there are many indicators in the model. Consider a three-factor CFA model in which each latent factor is defined by 6 indicators ($p = 18$). Thus, there are 171 elements of S; that is, $b = 18(19) / 2$ (see Eq. 3.14, Chapter 3). In this example, W is of the order $b \times b$ (171×171) and has 14,706 distinct elements; $b(b + 1) / 2 = 171(172) / 2 = 14,706$. The need to store and invert such large matrices in the itera- tive model-estimation process may pose serious difficulties in practical applications (e.g., quite demanding of computer resources). Moreover, WLS requires very large samples in order to accurately estimate the matrix of fourth-order moments (Jöreskog & Sörbom, 1996b). In addition, WLS requires that sample size exceeds $b + p$ (number of elements of S plus number of indicators) to ensure that W is nonsingular (EQS will not per- form WLS if $N < b$). Unless the sample size is quite large, very skewed items can make W not invertible; W will frequently be nonpositive definite in small to moderate samples with variables that evidence floor or ceiling effects. Consequently, WLS behaves very poorly in small or moderately sized samples. Moreover, in Monte Carlo studies that evaluated the perfor- mance of various estimators of non-normally distributed continuous data, results have suggested that WLS does not perform as well as MLM (e.g., Chou & Bentler, 1995). Evidence also suggests that the performance of the WLS estimator with categorical outcomes is not favorable (e.g., over- sensitivity of χ^2 and considerable negative bias in standard errors as model complexity increases; Muthén & Kaplan, 1992). Thus, as with non-normal continuous data, WLS is not a good estimator choice with categorical out- comes, especially in small to moderate samples (Flora & Curran, 2004).

Currently, the Mplus program appears to provide the best options for CFA modeling with categorical data. This is due in part to the WLSMV estimator, which is currently available only in Mplus. The WLSMV estima- tor provides weighted least square parameter estimates using a *diagonal* weight matrix (W) and robust standard errors and a mean- and variance- adjusted χ^2 test statistic (Muthén & Muthén, 1998-2004). Unlike WLS, WLSMV does not require W to be positive definite, because W is not

inverted as part of the estimation procedure. In WLSMV, the number of elements in the diagonal W equals the number of sample correlations in S, but this matrix is not inverted during estimation. Nevertheless, WLSMV estimation is fostered by N being larger than the number of rows in W. In the computation of the χ^2 test statistic and standard errors, the full W is used but not inverted. Thus, although CFA models using categorical indicators necessitate larger samples than comparably sized models using continuous indicators, the sample size requirements of WLSMV are far less restrictive than those of WLS. For example, Muthén has conducted unpublished simulation studies and has found that Ns of 150 to 200 may be sufficient for medium-sized models (e.g., 10 to 15 indicators). Flora and Curran (2004) confirmed these results by showing WLSMV produced accurate test statistics, parameter estimates, and standard errors of CFA models under a variety of conditions (e.g., sample sizes ranging from 100 to 1,000, varying degrees of non-normality and model complexity). In addition, preliminary simulation research has shown that WLSMV performs well with samples as small as 200 for variables with floor or ceiling effects (although, as with continuous indicators, very skewed categorical indicators call for larger Ns). More studies are needed to more fully establish the performance of WLSMV with various sample sizes and under other conditions (e.g., skewness, model complexity, size of indicator relationships). Because requisite sample size is closely tied to the specific model and data of a given study, general rules of thumb are of limited utility. Instead, researchers are encouraged to use the Monte Carlo routines available in the recent versions of some latent variable programs (e.g., Mplus) to determine the necessary N for their particular CFA investigations (see Chapter 10).

The framework and procedures of CFA differ considerably from normal theory CFA when categorical indicators are used. For instance, S is a correlation matrix rather than a covariance matrix; for example, a tetrachoric correlation matrix is used for binary indicators; a polychoric correlation matrix is used for polytomous indicators. Within the Mplus framework, various response models for categorical indicators are placed into a unifying framework by the use of latent continuous response variables, y^* (e.g., Muthén & Asparouhov, 2002). In the latent response variable framework, y^* reflects the amount of an underlying continuous and normally distributed characteristic (e.g., intelligence, attitude, personality trait) that is required to respond in a certain category of an observed categorical variable. For example, in the case of an IQ test, y^* would express the level of underlying intelligence needed to provide the correct response on a binary

test item. This framework assumes that the latent variable could have been measured in a finer-grained fashion; for example, individual differences in intelligence could be more precisely measured by items that render more units of discrimination than dichotomous or polytomous indicators. Correlations of the underlying y^* variables are used for S rather than the correlations of observed variables. The underlying y^* variables are related to observed categorical variables by means of threshold parameters (τ).[7] In the case of a binary indicator ($y = 0$ or 1), the threshold is the point on y^* where $y = 1$ if the threshold is exceeded (and where $y = 0$ if the threshold is not exceeded). Polytomous items have more than one threshold parameter. Specifically, the number of thresholds is equal to the number of categories minus one; for example, an ordinal item with three categories ($y = 0$, 1, or 2) has two thresholds (i.e., point on y^* where $y = 1$, point on y^* where $y = 2$). Thresholds are part of the mean structure of a CFA model and can be used in multiple-group comparisons (Chapter 7, cf. intercept invariance)[8] or in instances where the researcher wishes to convert the parameters of the CFA model into item response theory (IRT) parameters (described later in this chapter).

Moreover, because a correlation matrix for the y^*s is used as S, the observed variances of the indicators are not analyzed. Thus, as is not the case in CFA with continuous indicators, the residual variances of categorical indicators are not identified and are not estimated (because y^* is a latent variable, it does not have a metric). Therefore, the measurement errors (θ) of the CFA model with categorical indicators are not free parameters but instead reflect the remainder of 1 minus the product of the squared factor loading and factor variance; that is,

$$\theta = 1 - \lambda^2\phi \qquad (9.6)$$

(or simply 1 minus the squared completely standardized factor loading).

Estimation of a CFA model with categorical indicators is now illustrated using an applied example. In this example, the researcher wishes to verify a unifactorial model of alcohol dependence in a sample of 750 outpatients. Indicators of alcoholism are binary items reflecting the presence/absence of six diagnostic criteria for alcoholism (0 = criterion not met, 1 = criterion met). Mplus syntax for this one-factor CFA is presented in Table 9.9. As with CFA with non-normal, continuous indicators, a raw data file must be used as input. The third line of the VARIABLE command informs Mplus that all six indicators are categorical. Mplus determines the number of categories for each indicator (in this case, all indicators are dichoto-

TABLE 9.9. Mplus Syntax for Conducting CFA with Categorical Indicators

```
TITLE:     UNIFACTORIAL MODEL OF ALCOHOLISM
DATA:      FILE IS BINARY.DAT;
           FORMAT IS F4,F2,5F1;
VARIABLE: NAMES ARE ID Y1-Y6;
           USEV = Y1-Y6;
           CATEGORICAL ARE Y1-Y6;
ANALYSIS: ESTIMATOR=WLSMV;
           TYPE = MEANSTRUCTURE;  ! this command not necessary for
               standard CFA
MODEL:     ETOH BY Y1-Y6;
OUTPUT:    SAMPSTAT MODINDICES(10.00) STAND RESIDUAL TECH2;
SAVEDATA: DIFFTEST = DERIV.DAT;  ! this command is to be used for
               nested chi2
```

mous) and then calculates the appropriate type of correlation matrix (in this case, tetrachoric). On the ANALYSIS command, the estimator is specified as WLSMV (although this is the Mplus default with categorical outcomes). A mean structure analysis is also requested, although this is not necessary in a typical single-group CFA. If mean structure is requested, Mplus will output indicator thresholds, which are used in another illustration in this chapter. The *df*, overall fit, and other parameter estimates (e.g., factor loadings, factor variances) of the single-group CFA will not differ regardless of whether TYPE=MEANSTRUCTURE is included in the command syntax. The MODEL and OUTPUT commands are written in the same fashion as in CFA models with continuous indicators; for example, as before, the unstandardized loading of the first indicator, Y1, is fixed to 1.0 by Mplus default to define the metric of the latent factor. Beginning with Version 3.1, Mplus provides modification indices (and nested χ^2 evaluation; see below) when the WLSMV estimator is employed. WLSMV-based modification indices and associated values (e.g., expected parameter change) are interpreted in the same manner as CFA with continuous outcomes (see Chapter 5). The final command in Table 9.9 (SAVEDATA) is not necessary for the current CFA, but is to be used in a forthcoming illustration of nested model evaluation in context of the WLSMV estimator.

Selected Mplus output of the one-factor CFA results appear in Table 9.10. The first section of the output provides the proportions of the sample that endorsed each diagnostic criterion of alcohol dependence (e.g., 77.6% of the sample met the criterion assessed by the Y1 indicator). As noted earlier, the estimation process benefits from an absence of highly skewed indicators (e.g., to ensure that univariate and bivariate distributions contain

TABLE 9.10. Selected Mplus Output of One-Factor CFA of Alcohol Dependence Using Binary Indicators

```
SUMMARY OF CATEGORICAL DATA PROPORTIONS
    Y1
      Category 1    0.224
      Category 2    0.776
    Y2
      Category 1    0.345
      Category 2    0.655
    Y3
      Category 1    0.107
      Category 2    0.893
    Y4
      Category 1    0.213
      Category 2    0.787
    Y5
      Category 1    0.351
      Category 2    0.649
    Y6
      Category 1    0.207
      Category 2    0.793


SAMPLE STATISTICS

    ESTIMATED SAMPLE STATISTICS

        SAMPLE TETRACHORIC CORRELATIONS
            Y1          Y2          Y3          Y4          Y5
          _____    _____    _____    _____    _____
Y1
Y2          0.494
Y3          0.336       0.336
Y4          0.610       0.513       0.479
Y5          0.632       0.515       0.362       0.572
Y6          0.437       0.317       0.277       0.521       0.468

TESTS OF MODEL FIT

Chi-Square Test of Model Fit
        Value                         9.540*
        Degrees of Freedom              9**
        P-Value                       0.3889

*   The chi-square value for MLM, MLMV, MLR, WLSM and WLSMV cannot
    be used for chi-square difference tests.    MLM, MLR and WLSM
    chi-square difference testing is described in the Mplus Technical
    Appendices at www.statmodel.com.
    See chi-square difference testing in the index of the Mplus
    User's Guide.

**  The degrees of freedom for MLMV and WLSMV are estimated
    according to a formula given in the Mplus Technical Appendices at
    www.statmodel.com.
    See degrees of freedom in the index of the Mplus User's Guide.
```

TABLE 9.10. *(cont.)*

```
Chi-Square Test of Model Fit for the Baseline Model

        Value                                694.433
        Degrees of Freedom                         9
        P-Value                              0.0000


CFI/TLI
        CFI    0.999
        TLI    0.999


Number of Free Parameters                         12

RMSEA (Root Mean Square Error Of Approximation)

        Estimate                              0.009

WRMR (Weighted Root Mean Square Residual)

        Value                                 0.519


MODEL RESULTS
              Estimates        S.E.     Est./S.E.         Std        StdYX
ETOH    BY
   Y1           1.000        0.000        0.000         0.775        0.775
   Y2           0.822        0.072       11.407         0.637        0.637
   Y3           0.653        0.092        7.107         0.506        0.506
   Y4           1.031        0.075       13.721         0.799        0.799
   Y5           1.002        0.072       13.880         0.776        0.776
   Y6           0.759        0.076       10.024         0.589        0.589

Thresholds
   Y1$1        -0.759        0.051      -14.910        -0.759       -0.759
   Y2$1        -0.398        0.047       -8.448        -0.398       -0.398
   Y3$1        -1.244        0.061      -20.305        -1.244       -1.244
   Y4$1        -0.795        0.051      -15.457        -0.795       -0.795
   Y5$1        -0.384        0.047       -8.158        -0.384       -0.384
   Y6$1        -0.818        0.052      -15.796        -0.818       -0.818

Variances
   ETOH         0.601        0.063        9.609         1.000        1.000

R-SQUARE


   Observed   Residual
   Variable   Variance   R-Square


   Y1            0.399      0.601
   Y2            0.594      0.406
   Y3            0.744      0.256
   Y4            0.361      0.639
   Y5            0.397      0.603
   Y6            0.653      0.347
```

several observations per cell). In the next section, the sample tetrachoric correlations reflect the zero-order relationships among the six y^* variables. Because these coefficients are based on the latent variable underlying the binary indicators, they differ in value from phi correlations, which are based on observed measures. The results indicate that the one-factor model fits the data well, χ^2 (9) = 9.54, p = .39, RMSEA = 0.009, TLI = 0.999, CFI =.999. Note that Mplus does not provide an SRMR fit statistic. Presumably, this is due to recent evidence from simulation research (Yu, 2002) that the SRMR does not perform well with binary indicators.[9]

The unstandardized factor loadings are probit coefficients (because a linear equation can be written for the y^*s), which can be converted to probabilities if desired (see below). In the CFA, the y^* variances are standardized to 1.0, and thus parameter estimates should be interpreted accordingly. Squaring the completely standardized factor loadings yields the proportion of variance in y^* that is explained by the latent factor (e.g., Y1 = .775² = .601), not the proportion of variance explained in the observed measure (e.g., Y1), as in the interpretation of CFA with continuous indicators. Similarly, the residual variances convey the proportion of y^* variance that is not accounted for by the latent factor; for example, for Y1: 1 − .775² = .399.

As with SB χ^2, the difference in χ^2 values for nested models estimated with WLSMV is not distributed as χ^2. In addition, the calculation of model *df* is not the same as in CFA solutions estimated with ML (in fact, computation of WLSMV model *df* is not straightforward; see Muthén & Muthén, 1998–2004). To obtain the correct χ^2 difference test with the WLSMV estimator, a two-step procedure is required. In the first step, the less restricted model is estimated and the DIFFTEST option of the SAVEDATA command is used to save the derivatives needed for the nested χ^2 test. In the second step, the more constrained model is fitted to the data and the χ^2 difference test is calculated using the derivatives from both analyses.

To illustrate this procedure, a second model is estimated in which the five previously freely estimated factor loadings are constrained to equality (tau equivalence; see Chapter 7). The syntax for this model is provided in Table 9.11. In addition to this constraint, the Mplus syntax is modified such that the DIFFTEST option of the ANALYSIS command is used to inform the program of the name of the data file that contains the derivatives from the less restricted model. As shown in Table 9.11, Mplus provides the model χ^2 of the new solution as well as the χ^2 difference test for the two solutions. These results are interpreted in the same fashion as χ^2

TABLE 9.11. Nested Model Comparison with WLSMV: One-Factor CFA of Alcohol Dependence with Binary Indicators (Factor Loadings Constrained to Equality)

Syntax File

```
TITLE:     UNIFACTORIAL MODEL OF ALCOHOLISM (EQUAL FACTOR LOADINGS)
DATA:      FILE IS BINARY.DAT;
           FORMAT IS F4,F2,5F1;
VARIABLE: NAMES ARE ID Y1-Y6;
           USEV = Y1-Y6;
           CATEGORICAL ARE Y1-Y6;
ANALYSIS: ESTIMATOR=WLSMV;
           TYPE = MEANSTRUCTURE;
           DIFFTEST = DERIV.DAT;         ! line added for chisq diff test
MODEL:     ETOH BY Y1@1 Y2-Y6 (1);      ! equal Y2-Y6 lambdas
OUTPUT:    SAMPSTAT MODINDICES(10.00) STAND RESIDUAL TECH2;
```

Selected Output

```
TESTS OF MODEL FIT

Chi-Square Test of Model Fit

           Value                         40.033*
           Degrees of Freedom               10**
           P-Value                       0.0000


Chi-Square Test for Difference Testing

           Value                         27.958
           Degrees of Freedom                4**
           P-Value                       0.0000


*   The chi-square value for MLM, MLMV, MLR, WLSM and WLSMV cannot
    be used for chi-square difference tests.  MLM, MLR and WLSM
    chi-square difference testing is described in the Mplus Technical
    Appendices at www.statmodel.com.
    See chi-square difference testing in the index of the Mplus
    User's Guide.

**  The degrees of freedom for MLMV and WLSMV are estimated according
    to a formula given in the Mplus Technical Appendices at
    www.statmodel.com.
    See degrees of freedom in the index of the Mplus User's Guide.
```

difference testing with the ML estimator; that is, χ^2_{diff} (4) = 27.96, $p < .001$ indicates that the restriction of equal factor loadings significantly degrades the fit of the model.

Comparison with Item Response Theory (IRT) Models

It is well known that factor analysis with binary outcomes is equivalent to a two-parameter normal ogive *item response theory* model (e.g., Ferrando & Lorenza-Sevo, 2005; Glöckner-Rist & Hoijtink, 2003; MacIntosh & Hashim, 2003; Moustaki, Jöreskog, & Mavridis, 2004; Muthén, Kao, & Burstein, 1991; Reise, Widaman, & Pugh, 1993; Takane & de Leeuw, 1987). Although a detailed review of IRT is beyond scope of this book, a few fundamental aspects are noted to demonstrate the comparability of IRT and CFA.

IRT, which has also been referred to as *latent trait theory*, relates characteristics of items (item parameters) and characteristics of individuals (latent traits) to the probability of endorsing a particular response category (Bock, 1997; Lord, 1980). Whereas CFA aims to explain the correlations among test items (or the y^*s underlying the test items), IRT models account for participants' item responses. In other words, an IRT model specifies how both the level of the latent trait and the item properties are related to a person's item responses (which can be measured by either binary or polytomous items). The probability of answering correctly or endorsing a particular response category is graphically depicted by an *item response function* (IRF, also referred to as an *item characteristic curve*, ICC). IRFs reflect the nonlinear (logit) regression of a response probability on the latent trait. Most often, applied IRT models entail a single latent trait, although multidimensional IRT models can be analyzed (Bock, Gibbons, & Muraki, 1988; Embretson & Reise, 2000). Many objectives of IRT are similar to those of CFA. For instance, IRT can be used to explore the latent dimensionality of categorical outcomes, to evaluate the psychometric properties of a test, and to conduct differential item functioning analysis (DIF; see Chapter 7). The results of an IRT model can be used to assign sample participants a latent trait level estimate (akin to a factor score) on a standard z score metric (test scoring). IRT is frequently used in the domains of computerized and educational testing (e.g., Scholastic Aptitude Test, Graduate Record Examination) for item parameter estimation, test calibration, and test equating and scoring (e.g., identification of items that yield the highest measure-

ment precision/information about the examinee in a given trait domain, development of parallel tests or shorter test forms, equating scores across different subsets of items).

In addition to the latent trait level (denoted θ in the IRT literature), either one, two, or three item parameters can be estimated in an IRT model. The choice of IRT model should be based on substantive and empirical considerations (e.g., model fit, although IRT currently provides limited information in regard to goodness of model fit). Examples with binary items are discussed below, although these models can be readily extended to polytomous outcomes. The simplest model is the *one-parameter logistic model* (1PL), also known as the *Rasch model* (Rasch, 1960). In the 1PL model, the probability of responding positively on an item (e.g., correct response on an ability test, meeting diagnostic criterion in the alcoholism example) is predicted by the latent trait (θ) and a single item parameter, *item difficulty* (denoted either as b or β in the IRT literature). This is represented by the logistic function

$$P(y_{is} = 1 \mid \theta_s, b_i) = \exp(\theta_s - b_i) / [1 + \exp(\theta_s - b_i)] \quad (9.7)$$

in other words, the probability (P) that y equals 1 for a specific item (i) and participant (s), given (|) the participant's trait level (θ_s) and the item's difficulty (b_i), is the exponent (exp) of the difference of θ_s and b_i, divided by 1 plus the exponent (exp) of the difference of θ_s and b_i.

An item difficulty conveys the level of the latent trait (θ) where there is a 50% chance of a positive response on the item; for example, if $b = .75$, there is a .50 probability that a person with a trait level of .75 will respond positively to the item. The item difficulty parameter (b) represents the location of the curve along the horizontal axis of the IRF; for example, relatively "easier" items have lower b values and are represented by curves closer to the horizontal axis. Accordingly, b is inversely related to a proportion-correct score (p; or proportion of items endorsed if the presence/absence of a symptom, trait, attitude, and so forth). Item difficulties have been alternatively referred to in the IRT literature as *item threshold* or *item location* parameters. In fact, item difficulty parameters are analogous to item thresholds (τ) in CFA with categorical outcomes (cf. Muthén et al., 1991).

In a *two-parameter logistic model* (2PL), an *item discrimination* parameter is included (denoted either as a or α in the IRT literature). Thus, the probability of a positive response is predicted by the logistic function

$$P(y_{is} = 1 \mid \theta_s, b_i, a_i) = \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta_s - b_i)]} \qquad (9.8)$$

Item discrimination parameters are analogous to factor loadings in CFA and EFA because they represent the relationship between the latent trait and the item responses. Thus, discrimination parameters influence the steepness of the *slope* of the IRF curves. For instance, items with relatively high *a* parameter values are more strongly related to the latent variable (θ) and have steeper IRF curves. As can be seen in the above equation, item discrimination (*a*) is a multiplier of the difference between trait level (θ) and item difficulty (*b*). This reflects the fact that the impact of the difference between θ and *b* on the probability of a positive response (*P*) depends on the discriminating power of the item (Embretson & Reise, 2000).

A *three-parameter logistic model* (3PL) can also be estimated in IRT, which includes a "guessing" parameter (denoted either as *c* or γ). This additional parameter (also referred to as an "asymptote") is used to represent IRF curves that do not fall to zero on the vertical axis (i.e., > .00 probability of a positive response for persons with very low θ levels). In other words, if an item can be correctly answered by guessing (as in true/false or multiple-choice items on an aptitude test), the probability of a positive response is greater than zero even for persons with low levels of the latent trait characteristic. Accordingly, 3PL models are most germane to IRT analyses of aptitude or ability tests (e.g., intelligence and educational testing).

Unlike the case with 1PL and 2PL models, there currently does not exist a CFA counterpart to the 3PL IRT model (i.e., there is no analogous CFA parameter for the IRT "guessing" parameter). As will be seen shortly, the one-factor CFA model of alcoholism (Tables 9.9 and 9.10) is analogous to a 2PL IRT model; cf. factor loadings and item thresholds to item discrimination and item difficulty parameters, respectively. A CFA counterpart to a 1PL IRT model could be estimated by holding the factor loadings equal across items.

To illustrate these concepts, a unidimensional 2PL IRT model is fit to the alcohol dependence data using the BILOG-MG software program (Version 3.0; Zimowski, Muraki, Mislevy, & Bock, 2003). Item parameters are estimated by a marginal maximum likelihood method (MML), the most commonly used estimator in IRT analyses (Embretson & Reise, 2000). The syntax and selected output of this analysis are presented in Table 9.12. The first section of the selected output provides descriptive and classical test statistics. #TRIED indicates that there were no missing responses in the

TABLE 9.12. BILOG-MG 3.0 Estimation of a Unidimensional Two-Parameter Logistic IRT Model of Alcohol Dependence

Syntax File

```
IRT MODEL OF
ALCOHOLISM CRITERIA
>GLOBAL DFName = 'C:\BINARY.DAT',
        NPArm = 2;
>LENGTH NITems = (6);
>INPUT NTOtal = 6,
       NIDchar = 4;
>ITEMS ;
>TEST1 TNAme = 'ETOHDEP',
       INUmber = (1(1)6);
(4A1, 1X, 9A1)
>CALIB ACCel = 1.0000;
>SCORE ;
```

Selected Output

ITEM STATISTICS FOR SUBTEST ETOHDEP

						ITEM*TEST	CORRELATION
ITEM	NAME	#TRIED	#RIGHT	PCT	LOGIT/1.7	PEARSON	BISERIAL
1	ITEM0001	750.0	582.0	77.6	-0.73	0.486	0.678
2	ITEM0002	750.0	491.0	65.5	-0.38	0.409	0.527
3	ITEM0003	750.0	670.0	89.3	-1.25	0.267	0.449
4	ITEM0004	750.0	590.0	78.7	-0.77	0.514	0.724
5	ITEM0005	750.0	487.0	64.9	-0.36	0.500	0.644
6	ITEM0006	750.0	595.0	79.3	-0.79	0.358	0.508

SUBTEST ETOHDEP; ITEM PARAMETERS

ITEM	INTERCEPT	SLOPE	THRESHOLD	LOADING	ASYMPTOTE
ITEM0001	1.213	1.218	-0.996	0.773	0.000
	0.123*	0.155*	0.085*	0.099*	0.000*
ITEM0002	0.517	0.831	-0.622	0.639	0.000
	0.067*	0.097*	0.085*	0.075*	0.000*
ITEM0003	1.515	0.656	-2.309	0.549	0.000
	0.104*	0.099*	0.276*	0.083*	0.000*
ITEM0004	1.319	1.292	-1.021	0.791	0.000
	0.136*	0.171*	0.086*	0.105*	0.000*
ITEM0005	0.608	1.211	-0.502	0.771	0.000
	0.088*	0.159*	0.068*	0.101*	0.000*
ITEM0006	1.020	0.736	-1.386	0.593	0.000
	0.080*	0.092*	0.144*	0.074*	0.000*

* STANDARD ERROR

sample of 750 outpatients. #RIGHT provides the frequency with which the item (criterion) was endorsed in the sample. PCT provides the proportion of individuals who responded positively to the item (cf. "Summary of Categorical Data Proportions" in Mplus, Table 9.10). The last two columns present item-total correlations, both Pearson coefficients and point-biserial correlations that have been adjusted for the base rate of responses.

The second portion of the selected output presents the item calibrations (i.e., IRT parameter estimates). The item discrimination (*a*) and item difficulty (*b*) parameter estimates are provided under the "Slope" and "Threshold" columns, respectively. All asymptote parameters (*c*, "guessing") equal zero because a 2PL model was specified. As in CFA, the item loadings are interpreted as the correlations between the items and the latent traits (θ). Loadings can be calculated by the equation

$$a \; / \; SQRT(1 \; + \; a^2) \qquad\qquad (9.9)$$

for example, loading for Y1 = 1.218 / SQRT(1 + 1.218²) = .773. Item intercepts can be computed by the equation

$$-ab \qquad\qquad (9.10)$$

for example, intercept for Y1 = –(1.218 × –0.996) = 1.213. Item thresholds (*b*) always have the opposite sign of item intercepts. On the basis of these results, it appears that items Y1, Y4, and Y5 have the highest discrimination; that is, are more strongly related to θ, such that the probability of a positive response changes most rapidly with a change in θ. The Y3 indicator has the lowest difficulty (b_3 = -2.309); that is, lower levels of the latent dimension of Alcohol Dependence (θ) are required for a positive response on the Y3 criterion (e.g., when θ = –2.309, there is a 50% likelihood that this criterion will be endorsed). As noted earlier, *b* is related to the rate of item endorsement in the sample (e.g., p_3 = .893, Table 9.12) and the type of sample (e.g., larger *b* values would be likely if a community sample rather than an outpatient sample was used).

Using the 2PL equation presented earlier, an IRF curve can be plotted for each item. IRFs for three items are displayed in Figure 9.1. The form of the curves depict how change in the latent trait (θ, horizontal axis) relate to change in the probability of a positive response (*P*, vertical axis). The three curves have some similarities. For example, the steepest section of each is the middle of the curve, where small changes in θ are associated with the greatest increase in probability of a positive response. The end of
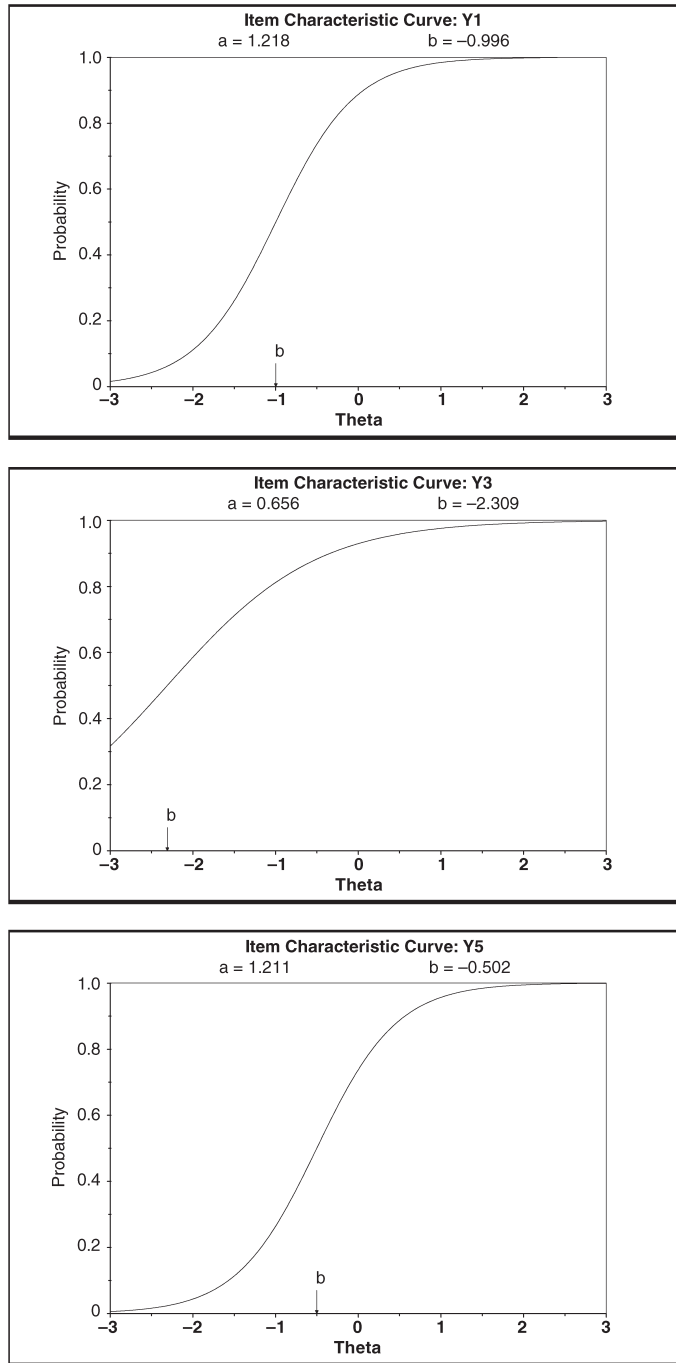
FIGURE 9.1. Item response functions for three selected alcohol dependence criteria (Y1, Y3, Y5).

each curve indicates that once a certain trait level is reached (i.e., $\theta \approx 1.5$), increases in the trait are not associated with appreciable change in item endorsement. Several differences are also evident. For instance, although items Y1 and Y5 have similar discrimination parameters (a = 1.218 and 1.211, respectively), Y5 is a "less difficult" item (b = –0.996 and –0.502, respectively). Item difficulty describes the extent to which items differ in probabilities across trait levels. Thus, although items Y1 and Y5 have similar a parameters, the shape of their curves differs somewhat because lower levels of θ are needed for a positive response to item Y5 (its IRF curve is a bit closer to the horizontal axis than is the IRF curve of Y1). Item Y3 has the relatively weakest relationship with θ (a = .656); hence, the probability of a positive response on Y3 changes most slowly with a change in θ. Item 3 also has the lowest threshold (b = –2.309).

With this background, the parallels of CFA and IRT should become clear. Comparing the CFA estimates provided in Table 9.10 and the IRT estimates provided in Table 9.12, it can be seen that the item loadings are quite similar in value (e.g., Y1 = .775 and .773 for CFA and IRT, respectively). Because factor loadings are closely linked to item discrimination parameters, the CFA factor loadings can be roughly interpreted in IRT terms (i.e., item Y4 has the highest factor loading and thus evidences the highest discrimination). However, it is possible to directly convert CFA parameters into IRT parameters (cf. Muthén et al., 1991; Muthén & Asparouhov, 2002). Using CFA parameterization (and symbols), an IRT discrimination parameter can be calculated as

$$a = \lambda / \text{SQRT}(\theta) \qquad (9.11)$$

where λ is the factor loading, and θ is the residual variance (*not* the trait level, as in IRT parameterization). For example, based on the CFA parameter estimates presented in Table 9.10, the item discrimination of Y1 would be calculated as a_1 = .775 / SQRT(.399) = 1.23 (cf. a_1 = 1.22 in Table 9.12).

Similarly, the item thresholds in Table 9.10 correspond to the item difficulty parameters that were estimated in the IRT analysis; for example, as the rank order of the item thresholds in the CFA is the same as in IRT, the results of the CFA also convey that item Y3 has the lowest difficulty. However, again using CFA parameterization (and symbols), an IRT difficulty parameter can be directly calculated as

$$b = \tau / \lambda \qquad (9.12)$$

where τ is the CFA item threshold, and λ is the CFA factor loading. Thus, the item difficulty parameter of item Y1 is computed $b_1 = -0.759 / .775 = -0.979$ (cf. $b_1 = -0.996$ in Table 9.12).

Table 9.13 presents a summary of all CFA-converted IRT parameters along with the item parameters estimated in BILOG-MG. Standard errors of IRT parameter estimates can also be obtained from CFA using the delta method (see MacIntosh & Hashim, 2003, for details). In Table 9.13 it can be seen that, although both approaches yield very similar results, the CFA-converted estimates and IRT estimates do not correspond exactly. This can be attributed mainly to the use of differing estimation methods (WLSMV in CFA, MML in IRT).

As noted earlier, IRT is frequently used to assess differential item functioning (DIF) of test items. Muthén (1988; Muthén et al., 1991) has shown that MIMIC models (see Chapter 7) with categorical indicators are equivalent to DIF analysis in the IRT framework (see also Meade & Lautenschlager, 2004). Recently, MacIntosh and Hashim (2003) provided an applied illustration of converting MIMIC model parameters (and standard errors) to IRT in the context of a DIF analysis. The MIMIC parameter estimates pertaining to covariate → item direct effects can be used to calculate item conditional probabilities (i.e., likelihood of item endorsement, given various levels of the covariate and latent factor). Moreover, Muthén (1998; Muthén et al., 1991) notes that the MIMIC framework offers several potential advantages over IRT. These include the ability to (1) use either continuous covariates (e.g., age) or categorical background variables (e.g., gender); (2) model a direct effect of the covariate on the latent

TABLE 9.13. Conversion of CFA Parameters to IRT Parameters: Two-Parameter Logistic IRT Model of Alcohol Dependence

Item	CFA to IRT		IRT	
	a	*b*	*a*	*b*
Y1	1.23	−0.98	1.22	−1.00
Y2	0.83	−0.62	0.83	−0.62
Y3	0.59	−2.46	0.66	−2.31
Y4	1.33	−1.00	1.29	−1.02
Y5	1.23	−0.49	1.21	−0.50
Y6	0.73	−1.39	0.74	−1.39

Note. a = item discrimination parameter; *b* = item difficulty parameter; CFA, confirmatory factor analysis; IRT, item response theory.

factor (in addition to direct effects of the covariate on test items); (3) readily evaluate multidimensional models (i.e., measurement models with > one latent factor); and (4) incorporate an error theory (e.g., measurement error covariances). Indeed, a general advantage of the covariance structure analysis approach is that the IRT model can be embedded in a larger structural equation model (e.g., Lu, Thomas, & Zumbo, 2005).

Other Potential Remedies for Indicator Non-Normality

Three other remedial strategies for non-normality are briefly presented: bootstrapping, item parceling, and data transformation. The advantages and disadvantages of each method are summarized. These strategies are reviewed primarily to increase the reader's familiarity with alternative approaches that have been used for dealing with non-normal data. However, given the advent of other full information estimators (e.g., MLM, WLSMV) and the limitations associated with each method, utilization of these alternative strategies is becoming less common.

Bootstrapping

Bootstrapping is a resampling procedure in which the original sample serves as the population (cf. Efron & Tibshirani, 1993; Mooney & Duval, 1993). Multiple samples (with the same N as the original sample) are randomly drawn from the original sample *with replacement* (i.e., a given case may be randomly selected more than once in any given bootstrapped data set), the CFA model is estimated in each data set, and the results are averaged over the data sets. The number of bootstrapped samples can be specified by the researcher, but should be sufficiently large to foster the quality of the averaged estimates (e.g., 500 samples is common). Using the one-factor CFA model presented earlier in this chapter (Table 9.5, $N = 870$), bootstrapping might entail generating 500 random samples of $N = 870$ from the original data set, fitting the one-factor CFA model in each sample (i.e., 500 times), and then averaging the results (e.g., parameter estimates, standard errors) across the 500 analyses. The procedure is most appropriate for models with non-normal, continuous indicators. Bootstrapping should not be confused with Monte Carlo simulation, although the two procedures have some similarities (see Chapter 10 for a discussion of Monte Carlo simulation). In Monte Carlo simulation, multiple samples (e.g., > 500) are randomly generated on the basis of population parameter

values and other data aspects (e.g., sample size, amount of non-normality) that are prespecified by the researcher (i.e., unlike the situation in boot-strapping, the researcher has population parameters on hand, but not a sample from the population). As in bootstrapping, the results of models fitted in the simulated data sets are averaged to examine the behavior of the estimates (e.g., stability and precision of parameter estimates and test statistics).

Bootstrapping is based on the notion that when the distributional assumptions of normal-theory statistics are violated, an *empirical sampling distribution* can be relied upon to describe the actual distribution of the population on which the parameter estimates are based. Unlike theoretical sampling distributions, the bootstrapped sampling distribution is *concrete* because it is based on the multiple samples spawned from the original data set. Accordingly, the bootstrapped averaged estimates and their standard errors (and possibly fit statistics, depending on the software program used) can be compared against the results from the original sample to eval-uate the stability of model parameters.

The bootstrapping procedure is straightforward in most latent vari-able software programs. Table 9.14 presents Amos Basic syntax for gener-ating bootstrapped estimates of the one-factor CFA presented earlier in Table 9.5 (1 factor, $N = 870$, 5 non-normal, continuous indicators). A raw data file must be used as input (in this example, the data are read from an SPSS .sav file). The command "sem.Bootstrap 500" instructs Amos to gen-erate 500 bootstrap samples. The command "sem.ConfidenceBC 90" is used to obtain bootstrapped bias-corrected 90% confidence intervals for freely estimated parameters (a value other than 90 can be specified, if desired). The remaining syntax specifies the CFA model to be tested.

The first portion of the selected output in Table 9.14 provides the normal-theory, ML parameter estimates and standard errors for the one-factor solution. For the sake of brevity, only unstandardized factor load-ings are presented. These results are identical to the EQS output in Table 9.7. The next section of the output provides results from the bootstrapped samples. The first column, "SE," is the bootstrap estimate of the standard errors of the factor loadings (specifically, these values represent the stan-dard deviations of the parameter estimates across the 500 bootstrapped samples). As in the MLM illustration (Table 9.7), the bootstrapped stan-dard errors are considerably larger than the maximum likelihood esti-mates; for example, X2: the bootstrapped standard error of .0606 is 74% larger than the ML estimate of .0349. It is noteworthy that the boot-

TABLE 9.14. Amos Basic Syntax and Selected Output for Bootstrapping a One-Factor CFA with Non-Normal, Continuous Indicators

Syntax File

```
' Example of Bootrapping in Amos 5.0
' One-factor model
'
Sub Main ()
Dim sem As New AmosEngine

sem.Bootstrap 500
sem.TextOutput
sem.Standardized
sem.ConfidenceBC 90

sem.BeginGroup "nonml.SAV"
sem.Structure "X1 = (1) FACTOR1 + (1) th1"
sem.Structure "X2 =     FACTOR1 + (1) th2"
sem.Structure "X3 =     FACTOR1 + (1) th3"
sem.Structure "X4 =     FACTOR1 + (1) th4"
sem.Structure "X5 =     FACTOR1 + (1) th5"
sem.Structure "th1 <-> th3"

End Sub
```

Selected Output

```
Maximum Likelihood Estimates
```

Regression Weights: (Group number 1 - Model 1)

	Estimate	S.E.	C.R.	P	Label
X1 <-- FACTOR1	1.0000				
X2 <-- FACTOR1	.7027	.0349	20.1211	***	
X3 <-- FACTOR1	1.0682	.0337	31.7119	***	
X4 <-- FACTOR1	.9181	.0422	21.7623	***	
X5 <-- FACTOR1	.7483	.0334	22.4031	***	

Bootstrap (Group number 1 - Model 1)

Parameter	SE	SE-SE	Mean	Bias	SE-Bias
X1 <-- FACTOR1	.0000	.0000	1.0000	.0000	.0000
X2 <-- FACTOR1	.0606	.0019	.7076	.0049	.0027
X3 <-- FACTOR1	.0429	.0014	1.0709	.0027	.0019
X4 <-- FACTOR1	.0627	.0020	.9172	-.0009	.0028
X5 <-- FACTOR1	.0581	.0018	.7489	.0006	.0026

90% confidence intervals (bias-corrected percentile method)

Parameter	Estimate	Lower	Upper	P
X1 <-- FACTOR1	1.0000	1.0000	1.0000	...
X2 <-- FACTOR1	.7027	.5904	.7962	.0072
X3 <-- FACTOR1	1.0682	1.0023	1.1426	.0043
X4 <-- FACTOR1	.9181	.8217	1.0326	.0027
X5 <-- FACTOR1	.7483	.6625	.8568	.0030

strapped standard errors are very similar in magnitude to the standard errors produced by MLM; for example, X2 = .062 and .061 for MLM and bootstrap, respectively (see Table 9.7). Indeed, the primary objective of bootstrapping is often to obtain better standard errors for the purpose of significance testing, calculation of confidence intervals, and so forth. The second column, "SE-SE," presents standard errors of the bootstrapped standard error estimates. These should be low in magnitude, given the number of bootstrapped samples and original sample size. Values in the "Mean" column are the average unstandardized factor loadings across the 500 bootstrap samples (virtually the same as the original sample estimates, although this is not always the case; cf. Arbuckle & Wothke, 1999). Values in the "Bias" column represent the difference between the original estimates and the averaged bootstrapped estimates (e.g., X2 = .7076 – .7027 = .0049). The last column, "SE-Bias" presents the standard errors of these bias estimates. The final section of the output lists the bias-corrected 90% confidence intervals of the unstandardized factor loadings; that is, confidence intervals of the original sample parameter estimates using standard errors that have been adjusted on the basis of bootstrapped results. These biased-correct intervals are interpreted in the same fashion as ordinary confidence intervals; for example, it is 90% likely that the true population value of the factor loading of X2 falls between .590 and .796 (because this interval does not include zero, the parameter estimate is significantly different from zero). Results in the "P" column indicate how small the confidence interval must be to include the value of zero. For example, the *p* value of .0072 indicates that the confidence interval for X2 must be at the 99.3% level in order for the lower-bound value to be zero (Byrne, 2001).

Yung and Bentler (1996) warn researchers to avoid believing that bootstrapping will always yield more accurate and reliable results. The success of bootstrapping depends on a number of aspects, including (1) the representativeness of the original sample to the population (if not representative, bootstrapping will produce misleading results); and (2) "the sampling behavior of a statistic being the same when the samples are drawn from the empirical distribution and when they are taken from the original population" (p. 113, Bollen & Stine, 1993). Bollen and Stine (1993) have also demonstrated that the bootstrap distribution will follow a noncentral χ^2 distribution rather than a central χ^2 distribution in accord with statistical theory (due to sampling fluctuation in the original sample). These authors have introduced transformation procedures that seem to work reasonably well to minimize this problem. Moreover, bootstrapping is not a remedy for small sample size (Yung & Bentler, 1996). As these

authors note, "the success of the bootstrap depends on the accuracy of the estimation of the parent distribution (and/or under a particular model) by the observed sample distribution" (p. 223). Thus, a very small sample will not render an acceptable level of accuracy. Finally, Yung and Bentler (1996) assert that while bootstrapping may provide more accurate standard error estimates, this approach is not necessarily the best method. The authors acknowledge that other approaches (such as using a non-normal theory estimator such as MLM) may be preferred to acquire certain statistical properties such as efficiency, robustness, and so forth.

Item Parceling

Another remedial approach that has been used to address non-normality is *item parceling*. A parcel (also referred to as a "testlet") is a sum or average of several items that presumably measure the same construct. For guidelines on constructing parcels, see Yuan, Bentler, and Kano (1997). The primary potential advantages of using parcels are that (1) parcels may be more apt to approximate normality than individual items (thereby, the assumptions of ML are more likely to be met); (2) it offers improved reliability and relationships with other variables (Kishton & Widaman, 1994); and (3) models based on parcels may be considerably less complex (i.e., smaller input matrix, fewer estimated parameters) than models based on individual items (it has been assumed that less model complexity will foster the stability of parameter estimates, but see below).

However, there are a variety of potential problems with using item parcels as indicators in CFA models. The most serious problem occurs when the underlying structure of the items in a parcel is not unidimensional. If such instances (i.e., one or more parcels is multifactorial), the use of parcels will obscure rather than clarify the latent structure of the data (West et al., 1995). Latent structures may also be obscured when the uniquenesses (measurement error variances) of items within a given parcel correlate with the unique or common factors of items in other parcels (Bandalos & Finney, 2001; Hall, Snell, & Foust, 1999). In many situations, the use of parcels may not be feasible, such as in instances where there are too few items to form a sufficient number of parcels. For example, in the case of binary indicators, it may take 10 or more items to create a parcel that approximates normality. If the entire scale contains only 20–30 items, then only two or three parcels can be created. Particularly when the intent of the CFA is to validate a testing instrument (or other multicomponent

measure), the use of a small handful of indicators (parcels) will render an overliberal and incomplete test of latent structure (West et al., 1995). In addition, simulation research has shown that the likelihood of improper or nonconvergent solutions increases as the number of parcels decreases (Nasser & Wisenbaker, 2003). Other research has shown that, under many different conditions (e.g., sample size), parcels do not outperform models based on individual items (Hau & Marsh, 2004; Marsh et al., 1998). For example, in the past parceling has been considered a good method of addressing non-normality when sample size does not permit the use of a non-normal theory estimator such as WLS (Browne, 1984b). With the advent of non-normal theory estimators with less restrictive sample size requirements (e.g., WLSMV in Mplus), item-level analysis is now more feasible.

Data Transformation

A final potential remedial strategy for non-normality is to transform raw scores of a variable so they more closely approximate a normal distribution. Different transformation procedures are available for different forms of non-normality (e.g., logarithmic transformation for substantial negative skewness; for an overview, see Tabachnick & Fidell, 2001). However, this approach has a few possible drawbacks. First, transformation is not always successful at reducing the skewness or kurtosis of a variable (e.g., data with severe floor or ceiling effects). Transformed variables must be reassessed in order to verify the success of the transformation in approximating normality at the uni- and multivariate levels. Second, in addition to altering the distribution of variables, nonlinear transformation often changes the relationships a variable has with other variables in the analysis. Thus, the resulting fit statistics, parameter estimates, and standard errors of a CFA based on transformed indicators may differ markedly from an analysis based on the original variables (West et al., 1995). This may be problematic for at least two reasons: (1) it "strains" reality if the true population distribution is not normally distributed; and (2) it makes the interpretability of the parameter estimates more complex. For example, if the variable "years of education" was log transformed to approximate a normal distribution, all resulting parameter estimates involving this variable should be interpreted with respect to log years, not calendar years. The latter issue is less salient when the original metric of a variable is arbitrary (e.g., Likert response scales of questionnaire items).

SUMMARY

This chapter dealt with several situations often encountered in the analysis of applied data sets (missing data, non-normal data, categorical data). As noted at the outset of this chapter, these data situations, and the methods of addressing them, are pertinent to any form of latent variable analysis (e.g., structural regression models, latent growth models). The field of latent variable statistical analysis has evolved tremendously over the past several years. As shown in this chapter, latent variable software programs are now well equipped to accommodate these data difficulties (e.g., direct ML for missing data, estimators other than ML for non-normal or categorical indicators). The final chapter of this book addresses another issue that has received scant attention in SEM sourcebooks to date: determination of the sample size needed to obtain sufficient power and precision of estimates in CFA. As shown in Chapter 10, current latent variable software programs offer elegant procedures for addressing this important question (e.g., Monte Carlo evaluation).

NOTES

1. In addition, a single imputation (with added random variation) will produce parameter estimates that are not fully efficient (i.e., estimated variances of the parameter estimates will be biased toward zero) (Allison, 2002; Schafer & Graham, 2002). This problem stems from the fact that the analysis would assume that the missing data are predicted on the basis of equations containing true population values. In other words, single imputation does not incorporate the uncertainty about the predictions of the unknown missing values. Within a multiple imputation framework, uncertainty about these parameters is reflected by taking random draws from the Bayesian posterior distribution of the parameters. The form of this posterior distribution is described by Iversen (1985); procedures for random draws from this distribution (e.g., data augmentation) are discussed by Schafer (1997).

2. However, in view of the ease of conducting multiple imputation in current software (e.g., SAS PROC MI), Allison (2003) recommends that more than five imputations be computed to improve the estimates of standard errors and stability of parameter estimates.

3. As Allison (2002) notes, the slight variability in multiply-imputed data sets could lead to the awkward situation where different researchers obtain different results from the same data and standard analyses. In addition, such variability could promote unscrupulous uses of the procedure (e.g., repeat the multiple imputation process until a "borderline" effect becomes statistically significant).

4. John Graham (e.g., Schafer & Graham, 2002) has written downloadable software routines for automating multiple imputation using NORM for LISREL and EQS (mcgee.hhdev.psu.edu/missing/index.html).

5. In CFA, if one or more indicators of a latent factor are not continuous, then an estimator other than ML must be used. However, this does not apply to exogenous predictors (covariates) of latent factors. For example, in MIMIC models (where the covariate may be categorical dummy codes reflecting levels of group; Chapter 7), ML is appropriate, assuming that indicators of the latent factors are normal, continuous.

6. However, in the version 3.0 release of Mplus, ML can be used as an estimator for CFA models with categorical outcomes when used in conjunction with *numerical integration* (a computationally complex and time-consuming algorithm that is needed when the posterior distribution for the latent variables does not have a closed form expression). The primary advantage of this approach is the ability to bring all the full information capabilities of ML into a categorical model framework; for example, one can model missing data with categorical indicators using direct ML. This method is also useful for modeling interaction effects between continuous latent variables (cf. Klein & Moosbrugger, 2000).

7. Note that τ is also used to denote indicator intercepts in mean structure analysis of continuous indicators (Chapter 7).

8. In addition, multiple-groups CFA with categorical outcomes also requires the inclusion of scale factors (symbolized as Δ). Scale factors contain information about the residual variance of y^*, the factor loadings, and the factor variances (see Muthén & Asparouhov, 2002, for more details).

9. As an alternative to SRMR, the results of Yu (2002) suggest that a cutoff of ≤ 1.0 on a relatively new fit statistic, the weighted root mean square residual (WRMR), can be used for models with binary outcomes when $Ns \geq 250$.

10

Statistical Power
and Sample Size

In this closing chapter, appropriate strategies are presented for deter-mining the sample size required to achieve adequate statistical power and precision of parameter estimates in a CFA study. This topic is often neglected in SEM sourcebooks and widely misunderstood in the applied literature. For instance, applied researchers often cite a gen-eral guideline (e.g., a subject: indicator ratio) in support of study sam-ple size. As discussed in this chapter, these rules of thumb are very crude and usually do not generalize to the researcher's data set and model. Thus, sample size requirements should be evaluated in the con-text of the particular data set and model at hand. As with many of the topics presented in Chapter 9, the concepts and strategies presented in this chapter are relevant to SEM models of any type (e.g., CFA, structural regression models). The book concludes with an overview of new horizons involving CFA methodology (multilevel factor models, factor mixture models).

OVERVIEW

In designing a CFA investigation, the researcher must address the critical question of how many cases (participants) should be collected to obtain an acceptable level of precision and statistical power of the model's parameter estimates, as well as reliable indices of overall model fit. The extant litera-ture provides little guidance on this issue. Some SEM sourcebooks provide general rules of thumb based on a small set of Monte Carlo studies. Many rules of thumb have been offered, including minimum sample size

(e.g., $N \geq 100$ to 200), minimum number of cases per each freed parameter (e.g., at least 5 to 10 cases per parameter), and minimum number of cases per indicator in the model (cf. Bentler & Chou, 1987; Boomsma, 1983; Ding, Velicer, & Harlow, 1995; Tanaka, 1987). Such guidelines are limited by their poor generalizability to any given research data set. That is, the models and assumptions used in Monte Carlo studies to provide sample size guidelines are often dissimilar to the types of models and data used by the applied researcher. Indeed, requisite sample size depends on a variety of aspects such as the study design (e.g., cross-sectional vs. longitudinal), the size of the relationships among the indicators, the reliability of the indicators, the scaling (e.g., categorical, continuous) and distribution of the indicators, estimator type (e.g., ML, robust ML, WLSMV), the amount and patterns of missing data, and the size of the model (model complexity). These features will vary widely from data set to data set. Thus, the goal of determining required sample size is fostered by the extent to which the actual model and data can be approximated in the power analysis.

Sample size affects the *statistical power* and *precision* of the model's parameter estimates. Statistical power is defined as one minus the probability of Type II error. Based on the work of Cohen (1988, 1992), the cutoff most frequently used to define acceptable power is .80; that is, an 80% likelihood of rejecting a false null hypothesis (thus, risk of Type II error is 20%). In SEM and CFA, power pertains to both the test of the model (e.g., sensitivity of χ^2 to detect model misspecifications) and the model parameter estimates (i.e., probability of detecting a parameter estimate as significantly different from zero). The closely related concept of precision pertains to the ability of the model's parameter estimates to capture true population values; for example, in the population, the correlation between variables X and Y is .30—do the sample data and model reasonably approximate this population value? Precision can be gauged in part by the amount of bias in the parameter estimates and their standard errors. Traditionally, model-based approaches to determining appropriate sample size have focused on the issue of statistical power. Some latent software packages allow researchers to also consider the precision of model estimates.

SATORRA–SARIS METHOD

The most widely known approach for conducting power analysis for multiple-indicator SEM models was introduced by Satorra and Saris (1985). This approach focuses on the power of the χ^2 difference test to

detect specification errors associated with a single parameter. The researcher specifies a model associated with a null hypothesis (H_0), which is evaluated in comparison with a model that represents the alternative hypothesis (H_1). In simplest terms, H_1 reflects the "true" model (i.e., contains the true population values for all parameters) and H_0 is identical to H_1 except for the parameter(s) to be tested. Thus, a population covariance matrix is generated from the parameter estimates of H_1. This covariance matrix is used as input in the test of H_0, which contains the misspecified parameter and the sample size of interest. In practice, the misspecified parameter is most often fixed to zero, and thus the analysis focuses on the power of model χ^2 to detect that a parameter is different from zero. The test of H_0 produces a nonzero model χ^2 value resulting from the misspecified parameter. This value represents a noncentrality parameter (NCP, often symbolized as λ, but not to be confused with a factor loading) of the noncentral χ^2 distribution. The noncentral χ^2 distribution reflects the χ^2 distribution when the null hypothesis is false. Using the resulting NCP value, the power of the test can be determined from tabled values (e.g., Saris & Stronkhorst, 1984) or simple routines in commercial software packages such as SPSS and SAS. Various sample sizes are considered to identify the required N for power = .80 (α = .05).

To illustrate the logic and procedures of this approach, a simple example is given. The model (H_1) containing the population values of the various parameters is presented in Figure 10.1. Ideally, these values
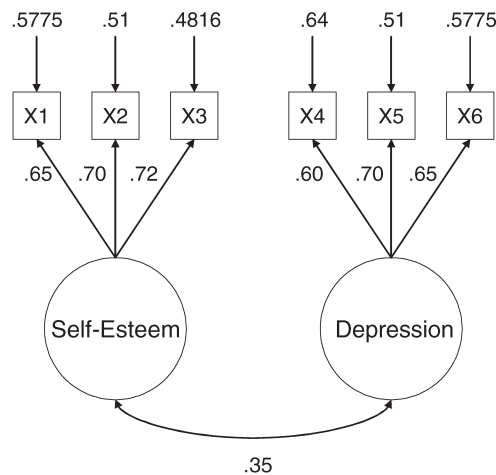
FIGURE 10.1. Hypothesized model of self-esteem and depression.

should be based on prior research (e.g., pilot data). The solution of interest is a two-factor CFA model of Self-Esteem and Depression (measured by three indicators each). The analysis will focus on the sample size needed to have adequate power (.80) to detect the factor covariance of Self-Esteem and Depression (ϕ_{21} = .35) as significantly different from zero (α = .05).

The first step is to generate a population covariance matrix from the parameter estimates of the H_1 model. In this example, standardized values are used for ease of illustration and interpretation. To accomplish this, a covariance matrix of null relationships (i.e., zero values in the off-diagonal, variances on the diagonal) is used as input, and the various model parameters (i.e., factor loadings, error variances and covariances, factor covariances) are fixed to equal the conjectured population values. Of course, this specification will produce a poor fit to the data because all relationships are null in the input matrix, but it is not the purpose of this step to obtain goodness-of-fit information. Rather, this analysis will produce a fitted covariance matrix that can be used as the population covariance matrix in subsequent steps of the procedure.

Table 10.1 provides LISREL syntax for conducting this initial step using the parameter estimates from Figure 10.1. Again, note that the covariance matrix (CM) used as input contains zeros in all off-diagonal elements (i.e., no relationships among the six indicators) and the indicator variances on the diagonal. There are no freely estimated parameters in this analysis because all parameters are fixed to population values using the value (VA) command. A sample size of 500 is specified. In this step, the size of the sample is arbitrary. In addition, to reduce rounding error, the analysis should be conducted with a sufficient number of decimal points (in this case, ND = 4 on the Output line). This syntax produces the following fitted covariance matrix that will be used as the population matrix in the next two steps. Although more time-consuming, this matrix could also be calculated by hand using the equations presented in Chapter 3; for example, covariance of X1 and X4 = .65(.35)(.60) = .1365 (cf. Eq. 3.8).

```
        Fitted Covariance Matrix
        X1        X2        X3        X4        X5        X6
        ____      ____      ____      ____      ____      ____

   X1   1.0000
   X2   0.4550    1.0000
   X3   0.4680    0.5040    1.0000
   X4   0.1365    0.1470    0.1512    1.0000
   X5   0.1592    0.1715    0.1764    0.4200    1.0000
   X6   0.1479    0.1592    0.1638    0.3900    0.4550    1.0000
```

TABLE 10.1. LISREL Syntax for Satorra–Saris Method of Determining Power to Detect That Factor Covariance Is Significantly Different from Zero

Step 1: Generate Population Covariance Matrix from H_1 Model

```
TITLE SATORRA-SARIS METHOD OF POWER CALCULATION: STEP ONE
DA NI=6 NO=500 MA=CM
LA
X1 X2 X3 X4 X5 X6
CM
1.0
0.0 1.0
0.0 0.0 1.0
0.0 0.0 0.0 1.0
0.0 0.0 0.0 0.0 1.0
0.0 0.0 0.0 0.0 0.0 1.0
MO NX=6 NK=2 PH=SY,FI LX=FU,FI TD=SY,FI
LK
ESTEEM DEPRESS
VA .65 LX(1,1)
VA .70 LX(2,1)
VA .72 LX(3,1)
VA .60 LX(4,2)
VA .70 LX(5,2)
VA .65 LX(6,2)
VA .5775 TD(1,1)
VA .51   TD(2,2)
VA .4816 TD(3,3)
VA .64   TD(4,4)
VA .51   TD(5,5)
VA .5775 TD(6,6)
VA 1.0 PH(1,1) PH(2,2)
VA .35 PH(2,1)
OU ME=ML RS AD=OFF IT=100 ND=4
```

Step 2: Analyze Residual Covariance Matrix to Ensure That Population Values Are Recovered

```
TITLE SATORRA-SARIS METHOD OF POWER CALCULATION: STEP TWO
DA NI=6 NO=500 MA=CM
LA
X1 X2 X3 X4 X5 X6
CM
1.0000
0.4550  1.0000
0.4680  0.5040  1.0000
0.1365  0.1470  0.1512  1.0000
0.1592  0.1715  0.1764  0.4200  1.0000
0.1479  0.1592  0.1638  0.3900  0.4550  1.0000
MO NX=6 NK=2 PH=SY,FR LX=FU,FR TD=DI
LK
ESTEEM DEPRESS
```

(cont.)

TABLE 10.1. *(cont.)*

```
PA LX
1 0
1 0
1 0
0 1
0 1
0 1
PA PH
0
1 0
VA 1.0 PH(1,1) PH(2,2)
OU ME=ML RS AD=OFF IT=100 ND=4
```

Step 3: Fit H_0 Model That Contains the Misspecified Parameter
(i.e., factor covariance fixed to zero) and Target Sample Size (e.g., N = 100)

```
TITLE SATORRA-SARIS METHOD OF POWER CALCULATION: STEP THREE
DA NI=6 NO=100 MA=CM                 ! SPECIFY SAMPLE SIZE OF INTEREST
LA
X1 X2 X3 X4 X5 X6
CM
1.0000
0.4550  1.0000
0.4680  0.5040  1.0000
0.1365  0.1470  0.1512  1.0000
0.1592  0.1715  0.1764  0.4200  1.0000
0.1479  0.1592  0.1638  0.3900  0.4550  1.0000
MO NX=6 NK=2 PH=SY,FR LX=FU,FR TD=DI
LK
ESTEEM DEPRESS
PA LX
1 0
1 0
1 0
0 1
0 1
0 1
PA PH
0
0 0                                  ! FACTOR COVARIANCE FIXED TO ZERO
VA 1.0 PH(1,1) PH(2,2)
OU ME=ML RS MI SC AD=OFF IT=100 ND=4
```

Step 4: Use χ^2 from Step 3 as Noncentrality Parameter
to Estimate Power at Targeted Sample Sizes (e.g., N = 100)

SAS Syntax:

```
DATA POWER;
DF=1; CRIT = 3.841459;
LAMBDA = 6.3724;
POWER = (1-(PROBCHI(CRIT,DF,LAMBDA)));
RUN;                                                    (cont.)
```

TABLE 10.1. *(cont.)*

SPSS Syntax:

```
DATA LIST FREE /X DF NCP.
COMPUTE POWER = 1 - NCDF.CHISQ(X,DF,NCP).
BEGIN DATA.
3.841459 1 6.3724
END DATA.
LIST POWER.
```

The second step is simply an accuracy check. The H_1 model is freely estimated using the fitted covariance matrix from the first step as input data (see Table 10.1). If the first two steps are conducted properly, this analysis will fit the data perfectly and the parameter estimates will be identical within rounding error to the original population values (e.g., λ_{11} = .65, ϕ_{21} = .35).

In the third step, the prior analysis is repeated except for two key alterations. First, the parameter of interest is misspecified. In this step, the null hypothesis is specified. Specifically, we are interested in determining the power to detect that the factor correlation of Self-Esteem and Depression is significantly different from zero. Thus, this parameter should be fixed to zero in this analysis (see PA PH commands in Table 10.1). Second, the sample size (N) for which power is desired must be specified. In Table 10.1, N is specified as 100. Therefore, the current analysis will estimate the power to detect a factor covariance of .35 as statistically significant when N = 100.

This analysis produces a model χ^2 value of 6.3724 [remember that Mplus users will obtain a slightly different χ^2 value because χ^2 is calculated as $F_{ML}(N)$ instead of the more widely used $F_{ML}(N-1)$]. Because all other aspects of the solution are properly specified, this nonzero χ^2 value is due to the misspecification of the factor covariance parameter. As noted earlier, this χ^2 value is the noncentrality parameter (NCP, λ) of the noncentral χ^2 distribution; that is, distribution of χ^2 when the null hypothesis is false.

In the fourth step, the χ^2 value from the preceding analysis (e.g., 6.3724) is used as the NCP to calculate the power to detect the model misspecification. Although tabled values are available (e.g., Saris & Stronkhorst, 1984), this step can be performed using simple routines in commercial software packages. These routines require three pieces of information: the NCP, degrees of freedom (df), and the critical χ^2 value. Because this particular power analysis focuses on a single parameter, df is 1

(i.e., this *df* should not be confused with the *df* of the H_1 or H_0 model) and the critical χ^2 value will be 3.84 (i.e., when α is set at .05). In Table 10.1, the critical χ^2 value is written out to six decimal points to reduce rounding error. Table 10.1 provides SAS and SPSS syntax for these routines.

The analysis yields a power value of .714. Thus, when $N = 100$, there is a 71.4% likelihood of rejecting a false null hypothesis (i.e., detecting the factor covariance of Self-Esteem and Depression as significantly different from zero at the .05 alpha level). Because .714 is below conventional cut-offs (i.e., power = .80), the results indicate that sample size should be larger to obtain adequate statistical power. NCPs (χ^2) for other sample sizes can be calculated in a few ways (all produce identical results). The Step 3 analysis could be re-run, changing the N of 100 to a larger value. Alternatively, most programs (e.g., LISREL) provide the minimized fit function value (F_{ML}), which can be used to calculate χ^2 at various sample sizes; that is, $\chi^2 = F_{ML}(N - 1)$. In programs that do not provide this value, F_{ML} can be readily computed as $\chi^2 / (N - 1)$. The NCPs for other sample sizes are then used as input in the power calculations (e.g., SAS or SPSS routine presented in Table 10.1).

In the current example, increasing sample size to 125 produces an NCP value of 7.9816. Using this NCP in the SAS or SPSS routine produces a power estimate of .807. Thus, an $N = 125$ appears more suitable in terms of the power to detect the factor covariance as statistically different from zero. For the reader's information, NCPs and power estimates for a few other sample sizes are as follows:

N	NCP	Power
100	6.3724	.713
125	7.9816	.807
150	9.5908	.872
200	12.8091	.947

The Satorra–Saris (1985) method is superior to general rules of thumb (e.g., ratio of cases to freed parameters) because it is a model-based, quantitative approach to determining requisite sample size. However, this approach has several disadvantages. As noted by Kaplan (1995), the method requires that the researcher specify the alternative value(s) to be tested (e.g., $\phi_{21} = .35$). Satorra (1989) later demonstrated that modification indices can be used as approximate NCP values for each restriction (e.g., fixed parameter) in the model (recall the relationship of the modification index and χ^2 at $df = 1$, Chapter 5). Thus, alternative values for the parame-

ters of interest do not need to be specified, but instead the modification indices are used as NCP values to determine the power to detect mis-specification of the restricted parameter. In addition, the aforementioned approach focuses on a single parameter at a time. Although Saris and Satorra (1993) subsequently developed a procedure for assessing power for multiple simultaneous mis-specifications ("isopower contours"), these procedures are not readily implemented in current latent variable software packages.

Some have criticized model-based approaches to power analysis because the procedures require the researcher to make exact estimates of population values for each parameter in the model. For instance, although the prior example focused on the factor covariance, all the factor loadings, factor variances, and indicator residuals had to be properly specified. This is often difficult to carry out in practice. A few misestimates of the parameter population values may undermine the power analysis. Accordingly, some methodologists (e.g., Jaccard & Wan, 1996) have suggested that OLS-based power analysis (e.g., Cohen, 1988) may be useful to determine power/requisite sample size for salient structural parameters of the model (i.e., effects involving relationships among latent variables). Bootstrapping (e.g., Bollen & Stine, 1993; see also Chapter 9) has been proposed as another approach to power analysis, although this method has limited utility in research planning (e.g., determining target N) because it requires a sufficiently large existing data set to generate "bootstrapped" samples (for an overview of bootstrapping as a power analysis procedure, see Jaccard & Wan, 1996).

MONTE CARLO APPROACH

In addition to the aforementioned drawbacks, the Satorra–Saris and OLS methods are limited by the fact that they do not address the precision of the model parameter estimates or directly take into account other salient aspects of the research data, such as degree of non-normality, type of indicators (e.g., binary, continuous), and the amount and patterns of missing data. Recent developments in latent variable software packages permit researchers to use the Monte Carlo methodology to determine the power and precision of model parameters in context of a given model, sample size, and data set. Typically, Monte Carlo studies are used in SEM research to study the behavior of statistical estimators and test statistics under various conditions manipulated by the researcher, such as sample size, degree

of model misspecification, and degree of data non-normality (cf. Paxton, Curran, Bollen, Kirby, & Chen, 2001). However, Muthén and Muthén (2002) have shown how the Monte Carlo method can be conveniently used to determine sample size and power in the design of SEM studies.

As in the Satorra–Saris (1985) approach, the Monte Carlo method requires specification of an H_1 model containing the population values of all parameters (e.g., Figure 10.1). Ideally, these values should be guided by existing data (e.g., pilot studies). Numerous samples are randomly generated on the basis of the population values of the model. Appendix 10.1 provides an illustration of the data generation process used in Monte Carlo research (for further details, the reader is referred to Mooney, 1997, Fan & Fan, 2005, and Muthén & Muthén, 1998–2004). In the absence of additional programming, the data sets generated by the Monte Carlo utility will consist of normally distributed, continuous data (with no missing data). However, programs such as Mplus and EQS offer programming features to produce the desired amount of non-normality and missing data, as well as the capability of estimating in the context of noncontinuous indicators (and categorical estimators such as WLSMV; see Chapter 9).[1] The specified SEM model is estimated in each sample, and the results of these analyses (e.g., parameter values, standard errors, fit statistics) are averaged across the samples. These averages are used to determine the precision (e.g., bias and coverage) and power of the estimates (i.e., proportion of samples in which the parameter is significantly different from zero). As in the Satorra–Saris approach, various sample sizes are studied to determine the N needed to obtain parameter estimates with sufficient power and precision.

The CFA model in Figure 10.1 is used to illustrate this approach. This CFA model is first examined under the condition of normally distributed indicators without missing data (for examples of Monte Carlo evaluations with non-normal indicators, see Muthén & Muthén, 2002). As in the Satorra–Saris example, the initial analysis focuses on the suitability of a sample size of 100. Mplus syntax for this analysis is presented in Table 10.2. The MONTECARLO command describes the details of the Monte Carlo study, such as the number and names of indicators (X1–X6) and the sample size (NOBSERVATIONS = 100) used for data generation and analysis. The NREPS statement specifies how many samples (replications) are to be generated from the parameters of the population model. This number can be regarded as the sample size of the Monte Carlo study. Current computer technology (e.g., fast CPU speeds) allows the user to specify a large number of replications (10,000, in this example) to ensure the stability of

TABLE 10.2. Mplus Syntax for Monte Carlo Approach to Determining Power and Precision of CFA Model Parameter Estimates (Normal and No Missing Data, *N* = 100)

```
TITLE:          CFA TWO-FACTOR, NORMAL DATA, NO MISSING
MONTECARLO:
                NAMES ARE X1-X6;
                NOBSERVATIONS = 100;            ! SAMPLE SIZE OF INTEREST
                NREPS = 10000;
                SEED = 53567;
MODEL POPULATION:
                ESTEEM BY X1*.65 X2*.70 X3*.72;
                DEPRESS BY X4*.60 X5*.70 X6*.65;
                ESTEEM@1; DEPRESS@1;
                X1*.5775; X2*.51; X3*.4816; X4*.64; X5*.51; X6*.5775;
                ESTEEM WITH DEPRESS*.35;
MODEL:
                ESTEEM BY X1*.65 X2*.70 X3*.72;
                DEPRESS BY X4*.60 X5*.70 X6*.65;
                ESTEEM@1; DEPRESS@1;
                X1*.5775; X2*.51; X3*.4816; X4*.64; X5*.51; X6*.5775;
                ESTEEM WITH DEPRESS*.35;
ANALYSIS:       ESTIMATOR = ML;
OUTPUT:         TECH9;
```

the simulation results (e.g., average of parameter estimates across replications). The SEED statement specifies the starting point used for the random draws from the population. Muthén and Muthén (2002) recommend the use of multiple seeds to verify the consistency of results across different seed values.

The MODEL POPULATION command describes the data generation model for the Monte Carlo study. On this command, the true population parameter values used for data generation are provided. In the current example, the parameter values listed in Figure 10.1 are specified as population values (and factor variances of Self-Esteem and Depression are fixed to 1.0). The MODEL command is used to indicate the analysis model (i.e., the CFA). In some applications, the analysis model will be different from the data generation model; for example, the data are generated as categorical, but analyzed as continuous to study the impact on various statistical estimators. In the typical power analysis, the MODEL command will be identical to the data generation model (MODEL POPULATION). Accordingly, these commands are used to provide both population values and starting values in the estimation of the model (for computing coverage, described below). In this example, ML is specified as the statistical estima-

tor on the ANALYSIS command. TECH9 is requested on the OUTPUT line to obtain any errors that are encountered in regard to the convergence of each Monte Carlo replication.

Selected output of the analysis is presented in Table 10.3. Muthén and Muthén (2002) forwarded the following criteria for determining sample size: (1) bias of the parameters and their standard errors do not exceed 10% for any parameter in the model; (2) for parameters that are the specific focus of the power analysis (e.g., the factor covariance of Self-Esteem and Depression), bias of their standard errors does not exceed 5%; and (3) coverage is between .91 and .98. In addition to these criteria, appropriate sample size is determined when the power of salient model parameters is .80 or above (cf. Cohen, 1988). In order (by column), the Mplus summary of the analysis provides the true population value for each parameter (Pop-

TABLE 10.3. Selected Mplus Output for Monte Carlo Approach to Determining Power and Precision of CFA Model Parameter Estimates (Normal and No Missing Data, $N = 100$)

```
MODEL  RESULTS


                    ESTIMATES            S. E.  M. S. E.   95%    % Sig
        Population Average Std. Dev.  Average           Cover   Coeff
ESTEEM    BY
 X1           0.650  0.6439    0.1100   0.1067    0.0121  0.941  1.000
 X2           0.700  0.6955    0.1127   0.1077    0.0127  0.938  1.000
 X3           0.720  0.7154    0.1129   0.1081    0.0128  0.939  1.000


DEPRESS   BY
 X4           0.600  0.5949    0.1189   0.1134    0.0142  0.936  0.999
 X5           0.700  0.6953    0.1223   0.1178    0.0150  0.944  0.999
 X6           0.650  0.6468    0.1199   0.1154    0.0144  0.939  0.999


ESTEEM   WITH
 DEPRESS      0.350  0.3495    0.1334   0.1247    0.0178  0.928  0.761


Variances
 ESTEEM       1.000  1.0000    0.0000   0.0000    0.0000  1.000  0.000
 DEPRESS      1.000  1.0000    0.0000   0.0000    0.0000  1.000  0.000


Residual  Variances
 X1           0.577  0.5628    0.1165   0.1127    0.0138  0.936  0.995
 X2           0.510  0.4926    0.1195   0.1156    0.0146  0.946  0.972
 X3           0.482  0.4625    0.1232   0.1178    0.0155  0.945  0.946
 X4           0.640  0.6209    0.1302   0.1231    0.0173  0.937  0.994
 X5           0.510  0.4893    0.1442   0.1353    0.0212  0.957  0.915
 X6           0.577  0.5567    0.1330   0.1275    0.0181  0.943  0.969
```

ulation), the average of the parameter estimates across replications (Average), the standard deviation of the parameter estimates across replications (Std. Dev.), the average of the standard errors across replications (S.E. Average), the mean square error of each parameter (M.S.E., calculated as the variance of the estimate across replications plus the square of the bias), 95% coverage (95% Cover, the proportion of replications for which the 95% confidence interval contains the true population parameter value), and the proportion of replications in which the parameter is significantly different from zero at the .05 alpha level (% Sig Coeff).

The percentage of parameter bias can be calculated by subtracting the population parameter value from the average parameter value, dividing this difference by the population value, and then multiplying the result by 100. For example, using the results in Table 10.3, the percent bias of the factor covariance parameter is

$$\text{Bias}(\phi_{21}) = [(.3495 - .35) / .35](100) = -.14\% \qquad (10.1)$$

The bias of this estimate is less than a percentage point, which is negligible. Indeed, bias is well below 10% for each parameter in the model, which satisfies the first criterion.

Bias in the standard errors of parameters is calculated in a similar fashion. Muthén and Muthén (2002) indicate that the standard deviation of the parameter estimate over replications can be treated as the population standard error when the number of replications is large (as in the current example, NREPS = 10,000). The percentage of standard error bias is calculated by subtracting the average of the estimated standard errors across replications (S.E. Average) from the standard deviation of the parameter estimate (Std. Dev.), dividing this difference by the standard deviation of the parameter estimate (Std. Dev.), and multiplying the result by 100. Again using the factor covariance (Table 10.3), the calculation is:

$$SE\ \text{Bias}(\phi_{21}) = [(.1247 - .1334) / .1334](100) = -6.52\%$$
$$(10.2)$$

Because this parameter is a focus of the power analysis, the bias would exceed the guideline provided by Muthén and Muthén (2002) of 5%, suggesting that the $N = 100$ may not be sufficient to obtain precise parameter estimates. Although coverage values are satisfactory (except for the fixed estimates, all are close to the correct value of .95), further evidence against the suitability of an $N = 100$ is provided by the proportion of replications

for which the factor covariance parameter is significantly different from zero (% Sig Coeff = .761; Table 10.3). Specifically, the value of .761 indicates that the factor covariance of Self-Esteem and Depression was statistically significant (α = .05) in 76.1% of the replications. Note that this result is close to the theoretical value of .714 obtained previously in the Satorra–Saris approach. This can be regarded as the power estimate of this parameter; that is, the probability of rejecting the null hypothesis when it is false. Alternatively, if a parameter's population value was specified to be 0.0, this would provide an estimate of Type I error—the likelihood of rejecting a true null hypothesis.

The analysis would proceed by examining larger sample sizes (and other seed values, to ensure stability once a suitable N has been tentatively identified). This is accomplished by changing the number of observations in the syntax provided in Table 10.2. Table 10.4 provides selected results for a Monte Carlo analysis conducted with a sample size of 125. Consistent with the results of the Satorra–Saris method, this analysis suggests that an N = 125 has sufficient power (.836) to reject a false null hypothesis in regard to the factor covariance. However, the results indicate that the standard error of this estimate continues to be somewhat biased (–5.95%; cf. Muthén & Muthén, 2002, guidelines), suggesting that a sample size larger than 125 may be warranted (although this conclusion would be bolstered by re-running the Monte Carlo analyses with different seed values).

Beginning with version 3.0, Mplus summarizes data pertaining to the behavior of selected fit indices (e.g., χ^2, RMSEA, SRMR) for a target model, data, and sample size. In Table 10.4, the mean and standard deviation of model χ^2 across the 10,000 replications of the Monte Carlo analysis is provided. The Proportions Expected and Percentiles Expected columns convey aspects of the χ^2 distribution. For instance, with a Proportions Expected value of .05 (third row from the bottom), there is a 5% probability of obtaining a χ^2 value that exceeds the Percentiles Expected value of 15.507 at df = 8 (the df of the model presented in Figure 10.1 and analyzed in this Monte Carlo study). In other words, the value of 15.507 is the critical value of χ^2 at df = 8, α = .05 (i.e., the same information could be obtained from tabled values of the χ^2 distribution in an appendix of a statistics textbook). However, the Proportions Observed and Percentiles Expected provide the values obtained in the Monte Carlo replications; for example, in the preceding example, the proportion of replications for which the critical value was exceeded is .058. The Percentiles Observed value of 15.937 is the χ^2 at this percentile from the Monte Carlo analysis that has 5% of the values in replications above it. Because it is very close to

TABLE 10.4. Selected Mplus Output for Monte Carlo Approach to Determining Power and Precision of CFA Model Parameter Estimates (Normal and No Missing Data, *N* = 125)

```
MODEL RESULTS
                        ESTIMATES            S. E.   M. S. E.    95%    % Sig
          Population  Average  Std. Dev.  Average            Cover   Coeff
ESTEEM B
 X1         0.650    0.6445    0.0971    0.0955     0.0095  0.944   1.000
 X2         0.700    0.6968    0.0996    0.0963     0.0099  0.940   1.000
 X3         0.720    0.7155    0.0998    0.0966     0.0100  0.941   1.000

DEPRESS BY
 X4         0.600    0.5959    0.1053    0.1013     0.0111  0.939   1.000
 X5         0.700    0.6966    0.1089    0.1051     0.0119  0.944   1.000
 X6         0.650    0.6476    0.1049    0.1030     0.0110  0.946   1.000

ESTEEM WITH
 DEPRESS    0.350    0.3491    0.1194    0.1123     0.0143  0.929   0.836

Variances
 ESTEEM     1.000    1.0000    0.0000    0.0000     0.0000  1.000   0.000
 DEPRESS    1.000    1.0000    0.0000    0.0000     0.0000  1.000   0.000

Residual Variances
 X1         0.577    0.5657    0.1030    0.1007     0.0107  0.944   0.998
 X2         0.510    0.4957    0.1050    0.1032     0.0112  0.947   0.989
 X3         0.482    0.4678    0.1071    0.1047     0.0117  0.949   0.976
 X4         0.640    0.6252    0.1138    0.1095     0.0132  0.936   0.997
 X5         0.510    0.4928    0.1309    0.1202     0.0174  0.954   0.949
 X6         0.577    0.5613    0.1156    0.1132     0.0136  0.947   0.989

TESTS OF MODEL FIT

Number of Free Parameters                        13

Chi-Square Test of Model Fit

        Degrees of freedom                        8

        Mean                                   8.213
        Std Dev                                4.084
        Number of successful computations    9998

           Proportions            Percentiles
         Expected   Observed   Expected   Observed
          0.990      0.991      1.646      1.844
          0.980      0.983      2.032      2.138
          0.950      0.958      2.733      2.844
          0.900      0.909      3.490      3.583
          0.800      0.814      4.594      4.725
          0.700      0.718      5.527      5.695
          0.500      0.522      7.344      7.509
          0.300      0.316      9.524      9.746
          0.200      0.218     11.030     11.344
          0.100      0.109     13.362     13.633
          0.050      0.058     15.507     15.937
          0.020      0.023     18.168     18.619
          0.010      0.012     20.090     20.405
```

the theoretical value of 15.507 (in tandem with the similarity of the observed and expected proportions values, .058 vs. .05), this supports the notion that the χ^2 distribution was well approximated in this instance (e.g., $N = 125$); that is, bias of less than 3%. Similarly, the average χ^2 across replications (8.213) is roughly the same as the model df (8); the variance of χ^2 across replications ($4.084^2 = 16.68$) is near the value of $2df$ (16). This also provides evidence that the χ^2 distribution was approximated.

Mplus has extensive Monte Carlo facilities beyond those illustrated in the prior example. For instance, a variety of data types (e.g., categorical, non-normal continuous, clustered) and models (e.g., multiple groups, mixtures, multilevel) can be studied in this framework. In addition, the Monte Carlo routine has several other useful features, such as the option to save all or a portion of the generated data sets and the ability to import estimates from an analysis of real data as population or coverage values in the Monte Carlo study. Another important feature is the ability to determine sample size and power in the context of missing data (see also Dolan, der Sluis, & Grasman, 2005). As noted in Chapter 9, missing data is a reality in most applied research (especially in longitudinal designs). In addition, missing data can be planned as part of the research design (e.g., as in the cohort sequential design to study life span developmental processes in a more compressed time frame). However, the effects of missing data are rarely considered in power analysis.

Table 10.5 presents Mplus syntax for a study with planned missingness in which half of the sample is administered all six indicators, a quarter is administered the Depression indicators only, and the remaining quarter is administered the Self-Esteem indicators only (based on the model presented in Figure 10.1). In the MONTECARLO command, the PATMISS option is used to specify the missing data patterns and the proportion missing for each indicator. In the Table 10.5 example, there are three missing data patterns (the patterns are separated by "|"). In the first pattern, all six indicators are present for all cases (the numbers in parentheses indicate the proportion of data missing for a given indicator; in this pattern, all proportions are 0). In the second pattern, all of the Self-Esteem indicators are missing (indicated by the value "1" in parentheses following the X1, X2, and X3 indicators—100% missing), and all of the Depression indicators are present. In the third pattern, the Depression indicators are missing, and the Self-Esteem indicators are present. The PATPROBS statement specifies the proportion of cases with each missing data pattern. In the current example, 50% of cases have data for all six indicators, 25% of data for Depression indicators only (X4–X6), and 25% have data for Self-Esteem

TABLE 10.5. Mplus Syntax for Monte Carlo Approach to Determining Power
and Precision of CFA Model Parameter Estimates (Normal and Planned Missing Data,
N = 200)

```
TITLE:      CFA TWO-FACTOR, NORMAL DATA, PLANNED MISSING
MONTECARLO:
            NAMES ARE X1-X6;
            NOBSERVATIONS = 200;
            NREPS = 10000;
            SEED = 53567;
            PATMISS =   X1 (0) X2 (0) X3 (0) X4 (0) X5 (0) X6 (0) |
                        X1 (1) X2 (1) X3 (1) X4 (0) X5 (0) X6 (0) |
                        X1 (0) X2 (0) X3 (0) X4 (1) X5 (1) X6 (1);
            PATPROB = .5 | .25 | .25;
ANALYSIS:   TYPE = MISSING;
            ESTIMATOR = ML;
MODEL POPULATION:
            ESTEEM BY X1*.65 X2*.70 X3*.72;
            DEPRESS BY X4*.60 X5*.70 X6*.65;
            ESTEEM@1; DEPRESS@1;
            X1*.5775; X2*.51; X3*.4816; X4*.64; X5*.51; X6*.5775;
            ESTEEM WITH DEPRESS*.35;
MODEL:
            ESTEEM BY X1*.65 X2*.70 X3*.72;
            DEPRESS BY X4*.60 X5*.70 X6*.65;
            ESTEEM@1; DEPRESS@1;
            X1*.5775; X2*.51; X3*.4816; X4*.64; X5*.51; X6*.5775;
            ESTEEM WITH DEPRESS*.35;
OUTPUT:     TECH9;
```

indicators only (X1–X3). The proportions in this statement must sum to
one. The only other addition to the Mplus syntax is the "TYPE =
MISSING" option on the ANALYSIS command (see Table 10.5), which
specifies use of the full information ML estimator (cf. Allison, 2003) to
accommodate missing data.

This example is based on a cross-sectional analysis in which some
data are missing by design (e.g., due to practical or financial constraints
that preclude administration of a test battery to the full sample). More
often, data are lost owing to participant attrition, incomplete responding,
noncompliance, and so on. For example, if the researcher is interested in
evaluating the temporal stability of a latent construct (defined by three
indicators, A, B, and C), he or she might specify the following pattern of
missingness for a two-wave study (in which a quarter of the sample is
expected to be lost at the second testing occasion):

```
PATMISS = A1 (0) B1 (0) C1 (0) A2 (.25) B2 (.25) C2 (.25);
PATPROB = 1;
```

In fact, the A1, B1, and C1 indicators could be omitted from the PATMISS statement, because indicators omitted from this statement are assumed to have no missing data by Mplus default. Ideally, the specification of 25% at the Time 2 assessment would be evidence based (e.g., pilot data, extant literature). Unlike the prior example, this example has a single pattern of missing data (all cases have complete data at Time 1, 25% of the sample is lost at Time 2). Accordingly, the probability of this missing pattern is specified as 1.0 (PATPROB = 1).

SUMMARY AND FUTURE DIRECTIONS IN CFA

This book draws to a close on the important, yet often overlooked, topic of sample size determination in planning a CFA investigation. As this chapter has shown, current latent variable software programs provide elegant methods of determining the sample size required for obtaining adequate statistical power and sufficient precision of parameter estimates. Two procedures were described in detail: the Satorra–Saris method and the Monte Carlo approach. It is hoped that these illustrations have convinced readers not to rely on general rules of thumb that seem to persist in the applied research literature (e.g., minimum sample size). Rather, sample size and power determinations should be based on models and data that mirror the actual empirical context. Recent innovations in software packages allow the researcher to also consider the impact of other common real-world complications (e.g., non-normality, missing data) on sample size and power.

A theme echoed in this and all preceding chapters is that methodological developments in CFA and SEM have advanced rapidly in recent years. The aim of this book has been to provide a practical discussion and guide to CFA, including the very latest developments in these methods most germane to the applied researcher (e.g., IRT models, Monte Carlo study). In closing, it seems appropriate to mention two additional CFA modeling possibilities that have developed rapidly during the course of writing this book: *factor mixture models* and *multilevel factor models*. In addition to their infancy, the reason that these models are not discussed in detail here is that, while closely related to CFA, they could be viewed as falling more

within the realm of SEM. However, readers should be aware of these emerging latent variable methodologies.

Factor mixture models are a combination of latent class models (cf. Lazarsfeld & Henry, 1968) and common factor models (e.g., CFA). Latent class models are used to explore unobserved population heterogeneity. The sources of population heterogeneity can be known or unobserved. For instance, in Chapter 7, multiple-groups CFA was described as a method of examining group differences on the parameters of a measurement model (e.g., gender may be a source of heterogeneity in the measurement model). In this case, heterogeneity is observed because it is possible to define subpopulations based on an observed variable (e.g., gender). Unobserved heterogeneity exists when it is not possible to identify the sources of heterogeneity beforehand; that is, the sample may be heterogeneous because it is comprised of cases that belong to different subpopulations, but it is not possible to identify these subgroups a priori. Factor mixture models can be conducted to identify salient *latent classes*; that is, homogeneous clusters with a heterogeneous sample. The term "latent class" is used because subpopulation membership is not observed but must be inferred from the data. Unlike all other latent variable models discussed in this book, the latent variable in the latent class model is categorical, and the number of categories (i.e., latent classes) reflects the number of classes in the sample. These classes may differ qualitatively or quantitatively. For example, using the Figure 10.1 model, a factor mixture model might identify three classes: an "unaffected" class (i.e., cases with high self-esteem, and no depression), a so-called "vulnerability" class (i.e., cases with no depression but with low self-esteem), and a "disordered" class (i.e., cases with depression and low self-esteem). Various aspects of the model results (e.g., goodness-of-fit statistics such as the Bayesian information criterion, BIC, and interpretability/quality of classification and posterior probabilities) are used to evaluate the factor mixture model. After an appropriate number of classes are identified (keeping substantive issues firmly in mind), the estimates (posterior class probabilities) of the factor mixture model can be used to divide the sample into subpopulations (e.g., Case #1 is a member of the unaffected class). Covariates and distal outcomes can be brought into the analysis in effort to validate the factor mixture model. For instance, an observed background variable (e.g., family history of depression) might be strongly predictive of class membership. These models may be more informative and interpretable than a single-group model that simply regresses the latent factors on a background variable (e.g., the background variable

may be differentially predictive of classes). At this writing, Mplus is the only software program available for fitting factor mixture models. For further information on this emerging methodology, the reader is referred to Lubke and Muthén (2005).

Another relatively new methodology is the multilevel factor model. Although multilevel models have been around for years (cf. hierarchical linear models; Raudenbush & Bryk, 2002), only recently has this methodology merged with CFA factor models in a manner readily accessible to applied researchers. Multilevel models should be employed when data have been obtained by cluster or unequal probability sampling. In these instances, the data are said to have a hierarchical or nested structure (in this context, the term "hierarchical" should not be confused with higher-order factor analysis, Chapter 8). For example, consider a study of grade school scholastic achievement where the student data have been collected statewide. The data have a multilevel hierarchical structure; that is, students are nested within classrooms, classrooms are nested within schools, and schools are nested within school districts. Other common examples of clustered data structures are family data (i.e., cases nested within families) and repeated measures data (i.e., cases nested by time). The observations may not be independent within clusters. For example, one would not employ an independent sample *t*-test to compare the means of a sample assessed at two time points, because the observations are not independent. In cross-sectional data, the observations also may not be independent because of the clustered data structure. For instance, the scholastic achievement of students within a given classroom of a given school may be homogeneous because students share the same teacher and classroom dynamics and come from similar family/socioeconomic backgrounds. In data sets of this nature, multilevel modeling is employed to avoid biases in parameter estimates, standard errors, and tests of model fit. In other words, if the hierarchical structure is ignored, so is the non-independence of observations. Consequently, for example, the standard errors of parameter estimates may be underestimated, resulting in positively biased statistical significance testing.

Moreover, the multilevel model can be estimated to learn more about within- and between-cluster relationships. Multilevel models are also referred to as *random coefficient models*. Random coefficients are parameters in a model that vary across clusters. Covariates can be included in the multilevel model to account for variability between and within clusters. Using the example of scholastic achievement, a multilevel regression

model might find that gender of student is a significant predictor of reading achievement; that is, that girls are better at reading than boys. In this example, this is a within-level effect (also referred to as a Level 1 effect); that is, the gender covariate accounts for variation in reading achievement among individuals. However, the multilevel analysis reveals that the effect varies across classrooms (i.e., a between-cluster or Level 2 effect). Thus, the Gender → Reading Achievement effect is a *random slope* (the slope varies across classrooms). Covariates can be brought into the model to explain the variability of this coefficient across clusters (classrooms). For instance, it may be found that the level of teacher training/experience (a cluster-level covariate) is inversely related to this random slope (a between-level effect); that is, the effect of gender on reading decreases as teacher experience increases. This cross-level interaction might be interpreted substantively as indicating that girls do better at reading than boys because, in the classroom, boys are less attentive, more recalcitrant, and so forth (the within-level effect). However, experienced teachers are more adept at implementing strategies to direct boys' attention to classroom learning (the cross-level interaction). As noted earlier, this two-level analysis could be expanded to incorporate additional levels of the data structure (e.g., classrooms within schools, schools within school districts).

Multilevel models can now be employed to analyze within- and between-cluster latent factors. In fact, the number of factors can differ at the within and between levels. Indeed, early evidence in applied data sets suggests fewer factors are obtained at the between level owing to lack of variability across clusters. As in the single indicator example described in the preceding paragraph, covariates can be brought into the model to explain variability within and between clusters. Any parameter of the CFA solution (e.g., a factor loading) can be treated as a random coefficient, if justified on substantive and empirical grounds. Most latent variable software programs (e.g., Mplus, EQS, LISREL) now have multilevel factor modeling capabilities. Examples of this methodology are still rather sparse in the applied literature, but illustrations can be found in Muthén (1991), Kaplan and Kreisman (2000), Heck (2001), and Dyer, Hanges, and Hall (2005).

Factor mixture models and multilevel factor models offer exciting new modeling possibilities to the researcher working within the realm of CFA and SEM. These developments perpetuate the hallmarks of CFA/SEM as an exceptionally versatile analytic framework for addressing a wide array of questions in the social and behavioral sciences.

NOTE

1. However, no current program can generate data with the desired amount of multivariate non-normality. For instance, the Mplus program uses a mixture modeling approach to generate non-normal data; that is, by varying the amount of overlap of the normal distributions of two subpopulations and then combining the data, the desired marginal skewness and kurtosis is obtained on a trial-and-error basis. Nonetheless, the procedures in Mplus and other programs with extensive data generation facilities (e.g., EQS) do not address joint skewness and kurtosis (multivariate non-normality), distribution shape, and so forth. For more information on these issues and their impact on power analyses based on simulations, see Yuan and Hayashi (2003).

Appendix 10.1

Monte Carlo Simulation in Greater Depth: Data Generation

As noted in this chapter, the Monte Carlo approach to power analysis involves several steps. First, the hypothesized model is established, which contains the population values of the model parameters (see Figure 10.1). Population values are often derived from the research literature and pilot studies. Second, multiple random samples are generated from the population values of the hypothesized model, based on the conditions specified by the researcher (e.g., sample size, patterns/amount of missing data). The model is then analyzed in each sample and the results are averaged across all simulated data sets. The averaged results provide information about the precision and statistical power of the model; for example, the proportion of replications for which a given parameter is statistically significant indicates the power of the estimate under the specified conditions (e.g., N). Although the rationale of this procedure is relatively straightforward, perhaps the most mysterious aspect of the Monte Carlo approach is how samples are generated from the population values of the model. There are various ways data can be simulated. To foster the reader's understanding of this process, a matrix decomposition procedure (Kaiser & Dickman, 1962) is illustrated using the Figure 10.1 model. To avoid undue complexity, only the indicators of the Self-Esteem factor (X1, X2, X3) will be used, under the assumptions of multivariate normality and no missing data. A more extensive illustration of this approach under other conditions (e.g., non-normality) can be found in Fan and Fan (2005).

Step 1. Calculate correlation matrix (R) from the population values of the hypothesized model.

Because the values in Figure 10.1 are completely standardized, it is straightforward to compute correlations among the X1, X2, and X3 indicators using basic equations presented in earlier chapters, for example, population r of X1,X2 = .65(.70) = .455 (cf. Eq. 3.7, Chapter 3). Thus, the population correlation matrix for these three indicators is as follows:

```
            X1        X2        X3
X1      1.000
X2       .455     1.000
X3       .468      .504     1.000
```

Step 2. Obtain a factor pattern matrix (F) by conducting a principal components analysis (PCA) on the population correlation matrix (R), where the requested number of components equals the number of indicators.

Because the number of components is the same as the number of indicators, the input correlation matrix is reproduced perfectly. SAS PROC FACTOR syntax for this step is provided below:

```
Title "Principal components analysis of population
correlation matrix";
Data SE (type=CORR);
input _TYPE_ $ _NAME_ $ x1-x3;
cards;
corr x1  1.000   .     .
corr x2  0.455  1.000  .
corr x3  0.468  0.504  1.000
;
run;
proc factor data=SE method=p nfactors=3;
run;
```

The resulting pattern matrix (F) is:

```
                  Factor Pattern

              Factor1        Factor2        Factor3
      x1      0.79017        0.60799        0.07738
      x2      0.81110       -0.36403        0.45781
      x3      0.81816       -0.22629       -0.52859
```

Step 3. Generate p uncorrelated random normal variables ($M = 0$, $SD = 1$), each with N observations (X).

In the current example, three variables will be generated (i.e., X1, X2, X3). To obtain a closer correspondence to the original population values, a large sample size will be requested ($N = 2,000$). The line of SAS syntax below accomplishes this step.

```
X = RANNOR(J(2000,3,0));
```

Step 4. Premultiply the random data matrix (X) with the pattern matrix (F) to create the correlated data matrix (Z).

The fundamental equation of the Kaiser and Dickman (1962) matrix decomposition procedure is

$$Z_{(p \times N)} = F_{(p \times p)} X_{(p \times N)} \tag{9.3}$$

where Z is the correlated data matrix with N cases and p variables; F is a $p \times p$ pattern matrix from PCA (see Step 2); and X is an uncorrelated random normal data matrix with p variables, N cases (see Step 3).

This step imposes the population correlations (R) among the variables on the sample data (X), as if the data were sampled from a population with the intercorrelations represented by the imposed correlation matrix (reflected by F, which reproduces R exactly).

The following SAS PROC IML syntax performs Steps 3 and 4:

```
PROC IML;
F = {0.79017   0.60799   0.07738,      *  1
     0.81110  -0.36403   0.45781,
     0.81816  -0.22629  -0.52859};
X  = RANNOR(J(2000,3,0));              *  2
X  = X';                              *  3
Z  = F*X;                             *  4
Z  = Z';                              *  5
```

*1 = factor pattern matrix (F; Step 2); *2 = generate three uncorrelated random normal variables (X), each with $N = 2,000$ (Step 3); *3 = transpose X to a $3 \times 2,000$ matrix in preparation for multiplication with F; *4 = Eq. 9.3; *5 = transpose Z back into a $2,000 \times 3$ matrix.

If desired, the SAS syntax above can be extended to assign means and standard deviations to the sample data. The SAS syntax below executes a linear transformation on X1 and X2 (target $M = 5$, $SD = 1$); no such transformation is done on X1 so it remains standardized ($M = 0$, $SD = 1$).

```
X1  =Z[,1];
X2  =Z[,2]* 5 + 1;
X3  =Z[,3]* 5 + 1;
Z = X1||X2||X3;
CREATE EST FROM Z[COLNAME={X1 X2 X3}];
     *output data file is EST
APPEND FROM Z;
```

Step 5. Obtain descriptive statistics and correlations from the simulated data. These statistics can be obtained from the SAS syntax below:

```
proc means data=EST n mean std skewness kurtosis;
   var x1 x2 x3;
proc corr data=EST;
   var x1 x2 x3;
run;
```

The results are as follows:

```
The MEANS  Procedure
Variable   N         Mean      Std Dev      Skewness     Kurtosis
X1       2000   0.0073703   1.0087635   -0.0441861   -0.0042880
X2       2000   0.9776742   5.0160837   -0.0363566    0.0504317
X3       2000   0.8757570   5.0265543   -0.1164463    0.0314612


The  CORR  Procedure
       Pearson Correlation Coefficients, N = 2000
              Prob > |r|  under H0: Rho=0
                      X1           X2           X3
         X1       1.00000      0.44867      0.48283
                               <.0001       <.0001
         X2       0.44867      1.00000      0.49222
                               <.0001       <.0001
         X3       0.48283      0.49222      1.00000
                               <.0001       <.0001
```

The sample data obtained in this initial simulation approximate the target population characteristics in terms of their univariate distributions (e.g., X2: M = 0.98, SD = 5.02) and their intercorrelations (e.g., r of X1, X2 = .45).

In context of the Mplus illustration provided in this chapter, this process would be repeated multiple times (e.g., number of replications = 10,000; Table 10.2) for the entire set of indicators (X1–X6). The hypothesized measurement model is fitted to each generated data set, and the fit statistics (e.g., χ^2) and parameter estimates are averaged across replications (cf. Table 10.4).

References

[Brackets indicate the chapter(s) in which the reference is cited.]

Ahmavaara, Y. (1954). Transformation analysis of factorial data. *Annales Academiae Scientiarium Fennicae*, *88*, (Ser. B), 150. [3]

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317–322. [5]

Allison, P. D. (1987). Estimation of linear models with incomplete data. In C. Clogg (Ed.), *Sociological methodology 1987* (pp. 71–103). Washington, DC: American Sociological Association. [9]

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage. [9]

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*, 545–557. [9, 10]

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author. [6, 7]

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, *49*, 155–173. [5]

Arbuckle, J. L. (1996). Full information estimation in the presence of missing data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–277). Mahwah, NJ: Erlbaum. [9]

Arbuckle, J. L., & Wothke, W. (1999). *Amos 4.0 user's guide*. Chicago: Smallwaters. [9]

Bagozzi, R. P. (1993). Assessing construct validity in personality research: Applications to measures of self-esteem. *Journal of Research in Personality*, *27*, 49–87. [6]

Bagozzi, R. P., & Yi, Y. (1990). Assessing method variance in multitrait–multimethod matrices: The case of self-reported affect and perceptions at work. *Journal of Applied Psychology*, *75*, 547–560. [6]

Bandalos, D., & Finney, J. S. (2001). Item parceling issues in structural equation modeling: In G. A. Marcoulides & R. E. Schumacker (Eds.), *New develop-*

ments and techniques in structural equation modeling (pp. 269–295). Mahwah, NJ: Erlbaum. [9]

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182. [3]

Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology*, *28*, 97–104. [2]

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin*, *107*, 238–246. [3]

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software. [3, 7, 9]

Bentler, P. M., & Chou, C. (1987). Practical issues in structural modeling. *Sociological Methods and Research*, *16*, 78–117. [10]

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, *16*, 21–33. [9]

Bock, R. D., Gibbons, R., & Muraki, E. J. (1988). Full information item factor analysis. *Applied Psychological Measurement*, *12*, 261–280. [9]

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley. [1, 3, 4, 8]

Bollen, K. A. (1990). Outlier screening and a distribution-free test for vanishing tetrads. *Sociological Methodology and Research*, *19*, 80–92. [5]

Bollen, K. A., & Curran, P. J. (2004). Autoregressive latent trajectory (ALT) models: A synthesis of two traditions. *Sociological Methods and Research*, *32*, 336–383. [7]

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*, 305–314. [8]

Bollen, K. A., & Stine, R. A. (1993). Bootstrapping goodness-of-fit measures in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 111–135). Newbury Park, CA: Sage. [9, 10]

Bollen, K. A., & Ting, K. (1993). Confirmatory tetrad analysis. In P. M. Marsden (Ed.), *Sociological methodology* (pp. 147–175). Washington, DC: American Sociological Association. [5]

Bollen, K. A., & Ting, K. (2000). A tetrad test for causal indicators. *Psychological Methods*, *5*, 3–22. [5]

Boomsma, A. (1983). *On the robustness of LISREL against small sample size and nonnormality*. Amsterdam: Sociometric Research Foundation. [10]

Brown, T. A. (2003). Confirmatory factor analysis of the Penn State Worry Questionnaire: Multiple factors or method effects? *Behaviour Research and Therapy*, *41*, 1411–1426. [3, 5]

Brown, T. A., Chorpita, B. F., & Barlow, D. H. (1998). Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. *Journal of Abnormal Psychology*, *107*, 179–192. [2]

Brown, T. A., White, K. S., & Barlow, D. H. (2005). A psychometric reanalysis of

the Albany Panic and Phobia Questionnaire. *Behaviour Research and Therapy*, *43*, 337–355. [5, 7]

Brown, T. A., White, K. S., Forsyth, J. P., & Barlow, D. H. (2004). The structure of perceived emotional control: Psychometric properties of a revised Anxiety Control Questionnaire. *Behavior Therapy*, *35*, 75–99. [5, 8]

Browne, M. W. (1984a). The decomposition of multitrait–multimethod matrices. *British Journal of Mathematical and Statistical Psychology*, *37*, 1–21. [6]

Browne, M. W. (1984b). Asymptotic distribution free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62–83. [9]

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research*, *24*, 445–455. [5]

Browne, M. W., & Cudek, R. (1993). Alternate ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage. [3]

Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum. [7]

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum. [4, 7]

Byrne, B. M. (2001). *Structural equation modeling with Amos: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum. [9]

Byrne, B., & Goffin, R. (1993). Modeling multitrait–multimethod data from additive and multiplicative covariance structures: An audit of construct validity concordance. *Multivariate Behavioral Research*, *28*, 67–96. [6]

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. [3, 7]

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, *56*, 81–105. [6]

Campbell, D. T., & O'Connell, E. J. (1967). Method factors in multitrait–multimethod matrices: Multiplicative rather than additive? *Multivariate Behavioral Research*, *2*, 409–426. [6]

Campbell-Sills, L. A., Liverant, G., & Brown, T. A. (2004). Psychometric evaluation of the Behavioral Inhibition/Behavioral Activation Scales (BIS/BAS) in large clinical samples. *Psychological Assessment*, *16*, 244–254. [5, 7, 8]

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276. [2]

Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM). *Organizational Research Methods*, *1*, 421–483. [7]

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models. *Sociological Methods and Research*, *29*, 468–508. [5]

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, *12*, 471–492. [8]

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*, 1–27. [7]

Cheung, G. W., & Rensvold, R. B. (2000). Evaluating goodness-of-fit indices for testing measurement invariance. *Structural Equation Modeling*, *9*, 233–255. [7]

Chou, C. P., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 37–55). Thousand Oaks, CA: Sage. [9]

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum. [4, 7, 10]

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. [7, 10]

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum. [2, 5, 7]

Cole, D. A. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology*, *55*, 584–594. [6]

Cole, D. A., & Maxwell, S. E. (1985). Multitrait–multimethod comparisons across populations: A confirmatory factor analysis approach. *Multivariate Behavioral Research*, *20*, 389–417. [7]

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Hillsdale, NJ: Erlbaum. [2]

Cooper, M. L. (1994). Motivations for alcohol use among adolescents: Development and validation of a four-factor model. *Psychological Assessment*, *6*, 117–128. [5]

Coovert, M. D., Craiger, J. P., & Teachout, M. S. (1997). Effectiveness of the direct product versus confirmatory factor model for reflecting the structure of multimethod–multirater job performance data. *Journal of Applied Psychology*, *82*, 271–280. [6]

Cox, B. J., Walker, J. R., Enns, M. W., & Karpinski, D. (2002). Self-criticism in generalized social phobia and response to cognitive-behavioral treatment. *Behavior Therapy*, *33*, 479–491. [2]

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of a test. *Psychometrika*, *16*, 297–334. [8]

Cudeck, R. (1988). Multiplicative models and MTMM matrices. *Journal of Educational Statistics*, *13*, 131–147. [6]

Curran, P. J., & Hussong, A. M. (2003). The use of latent trajectory models in psychopathology research. *Journal of Abnormal Psychology*, *112*, 526–544. [7]

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*, 16–29. [3, 9]

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with forma-

tive indicators: An alternative to scale development. *Journal of Marketing Research*, *38*, 269–277. [8]

Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions in structural equation modeling fit indices. *Structural Equation Modeling*, *2*, 119–143. [10]

Dolan, C., der Sluis, S., & Grasman, R. (2005). A note on normal theory power calculation in SEM with data missing completely at random. *Structural Equation Modeling*, *12*, 245–262. [10]

Duncan, T. E., Duncan, S. C., & Li, F. (1998). A comparison of model- and multiple imputation-based approaches to longitudinal analysis with partial missingness. *Structural Equation Modeling*, *5*, 1–21. [9]

Duncan, T. E., Duncan, S. C., Strycker, L. A., Li, F., & Alpert, A. (1999). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Mahwah, NJ: Erlbaum. [7]

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *Leadership Quarterly*, *16*, 149–167. [10]

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*, 155–174. [8]

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall. [9]

Eid, M. (2000). A multitrait–multimethod model with minimal assumptions. *Psychometrika*, *65*, 241–261. [6]

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait–multimethod models: A multiple-indicator CT-C(*M*-1) model. *Psychological Methods*, *8*, 38–60. [6]

Eliason, S. R. (1993). *Maximum likelihood estimation: Logic and practice*. Newbury Park, CA: Sage. [3]

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum. [9]

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. [2]

Fan, X., & Fan, X. (2005). Using SAS for Monte Carlo simulation research in SEM. *Structural Equation Modeling*, *12*, 299–333. [10]

Ferrando, P. J., & Lorenza-Sevo, U. (2005). IRT-related factor analytic procedures for testing the equivalence of paper-and-pencil and Internet-administered questionnaires. *Psychological Methods*, *10*, 193–205. [9]

Flamer, S. (1983). Assessment of the multitrait–multimethod matrix validity of Likert scales via confirmatory factor analysis. *Multivariate Behavioral Research*, *18*, 275–308. [6]

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative meth-

ods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466–491. [9]

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7, 286–299. [2]

Folkman, S., & Lazarus, R. S. (1980). An analysis of coping in a middle-aged community sample. *Journal of Health and Social Behavior*, *21*, 219–239. [8]

Fresco, D. M., Heimberg, R. G., Mennin, D. S., & Turk, C. L. (2002). Confirmatory factor analysis of the Penn State Worry Questionnaire. *Behaviour Research and Therapy*, *40*, 313–323. [5]

Glöckner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, 10, 544–565. [9]

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. [2]

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science*, *12*, 133–157. [7]

Gonzalez, R., & Griffin, D. (2001). Testing parameters in structural equation modeling: Every "one" matters. *Psychological Methods*, *6*, 258–269. [4]

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum. [2]

Gorsuch, R. L. (1997). New procedures for extension analysis in exploratory factor analysis. *Educational and Psychological Measurement*, 57, 725–740. [3]

Green, D. P., Goldman, S. L., & Salovey, P. (1993). Measurement error masks bipolarity in affect ratings. *Journal of Personality and Social Psychology*, *64*, 1029–1041. [6]

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. [8]

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*, 430–450. [2, 8]

Hall, R. J., Snell, A. F., & Foust, M. S. (1999). Item parceling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. *Organizational Research Methods*, 2, 233–256. [9]

Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press. [2]

Hau, K. T., & Marsh, H. W. (2004). The use of item parcels in structural equation modeling: Non-normal data and small sample sizes. *British Journal of Mathematical and Statistical Psychology*, 57, 327–351. [9]

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*, 191–205. [2]

Hazlett-Stevens, H., Ullman, J. B., & Craske, M. G. (2004). Factor structure of the

Penn State Worry Questionnaire: Examination of a method factor. *Assessment*, *11*, 361–370. [5]

Heck, R. H. (2001). Multilevel modeling with SEM. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 89–127). Mahwah, NJ: Erlbaum. [10]

Hershberger, S. L. (1994). The specification of equivalent models before the collection of data. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis* (pp. 68–105). Thousand Oaks, CA: Sage. [5]

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185. [2]

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117–144. [7]

Horowitz, M. J., Wilner, N., & Alvarez, W. (1979). The Impact of Event Scale: A measure of subjective stress. *Psychosomatic Medicine*, *41*, 209–218. [8]

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 158–176). Thousand Oaks, CA: Sage. [4]

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Thousand Oaks, CA: Sage. [3]

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424–453. [3]

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. [3, 4, 5]

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*, 351–362. [3]

Humphreys, L. G., & Montanelli, R. G. (1975). An investigation of parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, *10*, 191–205. [2]

Iversen, G. (1985). *Bayesian statistical inference*. Thousand Oaks, CA: Sage. [9]

Jaccard, J., & Wan, C. K. (1996). *LISREL approaches to interaction effects in multiple regression*. Thousand Oaks, CA: Sage. [3, 4, 7, 10]

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, *30*, 199–218. [8]

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183–202. [2, 5]

Jöreskog, K. G. (1971a). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*, 109–133. [2, 3, 8]

Jöreskog, K. G. (1971b). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409–426. [7]

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage. [4, 5, 7]

Jöreskog, K. G. (1999). *How large can a standardized coefficient be* [webpage]? Chicago: Scientific Software International. www.ssicentral.com/lisrel/column2.htm [4]

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. [7]

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models* (J. Magidson, Ed.). Cambridge, MA: Abt Books. [5]

Jöreskog, K. G., & Sörbom, D. (1996a). *LISREL 8: User's reference guide*. Chicago: Scientific Software International. [3]

Jöreskog, K. G., & Sörbom, D. (1996b). *PRELIS 2: User's reference guide*. Chicago: Scientific Software International. [4, 9]

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika*, *27*, 179–182. [10]

Kaplan, D. (1989). Model modification in covariance structure analysis: Application of the expected parameter change statistic. *Multivariate Behavioral Research*, *24*, 285–305. [4]

Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation. *Multivariate Behavioral Research*, *25*, 137–155. [4]

Kaplan, D. (1995). Statistical power in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 100–117). Thousand Oaks, CA: Sage. [10]

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage. [4, 5, 9]

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. [10]

Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley-Interscience. [3, 6]

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait–multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, *112*, 165–172. [6]

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. T. Gilbert & S. T. Fiske (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 233–265). New York: McGraw-Hill. [5]

Kishton, J. M., & Widaman, K. F. (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement*, *54*, 757–765. [9]

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, *65*, 457–474. [9]

Kline, R. B. (2005). *Principles and practices of structural equation modeling* (2nd ed.). New York: Guilford Press. [4]

Komaroff, E. (1997). Effect of simultaneous violations of essential tau-equivalence and correlated errors on coefficient alpha. *Applied Psychological Measurement*, *21*, 337–348. [8]

Kumar, A., & Dillon, W. R. (1992). An integrative look at the use of additive and multiplicative covariance models in the analysis of MTMM data. *Journal of Marketing Research*, *29*, 51–64. [6]

Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait–multimethod data. *Psychological Methods*, *7*, 228–244. [6]

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin. [10]

Lee, S., & Hershberger, S. (1990). A simple rule for generating equivalent models in structural equation modeling. *Multivariate Behavioral Research*, *25*, 313–334. [5]

Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait–multimethod matrices. *Journal of Applied Psychology*, *86*, 1202–1222. [6]

Lischetzke, T., & Eid, M. (2003). Is attention to feelings beneficial or detrimental to affective well-being?: Mood regulation as a moderator variable. *Emotion*, *3*, 361–377. [6]

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley. [9]

Loehlin, J. C. (2004). *Latent variable models* (4th ed.). Mahwah, NJ: Erlbaum. [3, 8]

Lord, F. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum. [9]

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley. [8]

Lu, I. R. R., Thomas, D. R., & Zumbo, B. D. (2005). Embedding IRT in structural equation models: A comparison with regression based on IRT scores. *Structural Equation Modeling*, *12*, 263–277 [9].

Lubke, G. H., & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*, 21–39. [10]

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, *100*, 107–120. [4, 5]

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*, 201–226. [4]

MacCallum, R. C., & Browne, M. W. (1993). The use of causal indicators in covariance structure models: Some practical issues. *Psychological Bulletin*, *114*, 533–541. [8]

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130–149. [3]

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490–504. [4]

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. [5]

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

MacIntosh, R., & Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Applied Psychological Measurement*, *27*, 372–379. [9]

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait–multimethod data: Many problems and a few solutions. *Applied Psychological Measurement*, *13*, 335–361. [6]

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*, 810–819. [3, 5, 6]

Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analysis of multitrait–multimethod data: A comparison of the behavior of alternative models. *Applied Psychological Measurement*, *15*, 47–70. [6]

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*, 391–410. [3]

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait–multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 177–198). Thousand Oaks, CA: Sage. [6]

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much?: The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*, 181–220. [3, 9]

Marsh, H. W., & Hocevar, D. (1983). Confirmatory factor analysis of multitrait–multimethod matrices. *Journal of Educational Measurement*, *20*, 231–248. [6]

McDonald, R. P. (1981). Constrained least squres estimators of oblique common factors. *Psychometrika*, *46*, 337–341. [2]

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum. [7]

McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*, 64–82. [4]

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*, 361–388. [9]

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543. [7]

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development

and validation of the Penn State Worry Questionnaire. *Behaviour Research and Therapy*, 28, 487–495. [5]

Mooney, C. Z. (1997). *Monte Carlo simulation*. Thousand Oaks, CA: Sage. [10]

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Thousand Oaks, CA: Sage. [9]

Moustaki, I., Jöreskog, K. G., & Mavridis, D. (2004). Factor models for ordinal variables with covariate effects on the manifest and latent variables: A comparison of LISREL and IRT approaches. *Structural Equation Modeling*, 11, 487–513. [9]

Muthén, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 213–238). Hillsdale, NJ: Erlbaum. [9]

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, 54, 557–585. [7]

Muthén, B. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, 28, 338–354. [10]

Muthén, B., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus*. Unpublished manuscript. (Available at www.statmodel.com/mplus/examples/webnote.html#web4). [9]

Muthén, B. O., Kao, C., & Burstein, L. (1991). Instructional sensitivity in mathematics achievement test items: Applications of a new IRT-based detection technique. *Journal of Educational Measurement*, 28, 1–22. [9]

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology*, 45, 19–30. [9]

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52, 431–462. [9]

Muthén, L. K., & Muthén, B. O. (1998–2004). *Mplus user's guide* (3rd ed). Los Angeles: Author. [5, 9, 10]

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, 4, 599–620. [10]

Nasser, F., & Wisenbaker, J. (2003). A Monte Carlo study investigating the impact of item parceling on measures of fit in confirmatory factor analysis. *Educational and Psychological Measurement*, 63, 729–757. [9]

Neale, M. C., & Miller, M. B. (1997). The use of likelihood-based confidence intervals in genetic models. *Behavior Genetics*, 27, 113–120. [4]

O'Connor, B. P. (2001). Extension: SAS, SPSS, and MATLAB programs for extension analysis. *Applied Psychological Measurement*, 25, 88. [2]

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, 8, 287–312. [10]

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879–903. [5, 6]

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, 2, 13–43. [2]

Price, L. R., Tulsky, D., Millis, S., & Weiss, L. (2002). Redefining the factor structure of the Wechsler Memory Scale-III: Confirmatory factor analysis with cross-validation. *Journal of Clinical and Experimental Neuropsychology*, *24*, 574–585. [7]

Rasch, G. (1960). *Probablistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press. [9]

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage. [10]

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, *32*, 329–353. [8]

Raykov, T. (2001a). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology*, *54*, 315–323. [7, 8]

Raykov, T. (2001b). Bias of Cronbach's alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, *25*, 69–76. [8]

Raykov, T. (2002a). Examining group differences in reliability of multiple-component instruments. *British Journal of Mathematical and Statistical Psychology*, *55*, 145–158. [8]

Raykov, T. (2002b). Analytic estimation of standard error and confidence interval for scale reliability. *Multivariate Behavioral Research*, *37*, 89–103. [8]

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy*, *35*, 299–331. [8]

Raykov, T., & Grayson, D. A. (2003). A test for change of composite reliability in scale development. *Multivariate Behavioral Research*, *38*, 143–159. [8]

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling*, *9*, 195–212. [8]

Raykov, T., Tomer, A., & Nesselroade, J. R. (1991). Reporting structural equation modeling results in *Psychology and Aging*: Some proposed guidelines. *Psychology and Aging*, *6*, 499–503. [4]

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*, 552–566. [9]

Rosenberg, M. (1965). *Society and the adolescent child*. Princeton, NJ: Princeton University Press. [3, 5]

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley. [9]

Russell, J. A. (1979). Affective space is bipolar. *Journal of Personality and Social Psychology*, *37*, 1161–1178. [6]

Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 181–204). Newbury Park, CA: Sage. [10]

Saris, W. E., & Stronkhorst, H. (1984). *Causal modeling in nonexperimental research*. Amsterdam: Sociometric Research Foundation. [10]

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika*, *54*, 131–151. [10]

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage. [9]

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*, 83–90. [10]

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall. [9]

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177. [9]

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, 22, 53–61. [8]

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait–multimethod matrices. *Applied Psychological Measurement*, *10*, 1–22. [6]

Self, S. G., & Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*, 605–610. [4]

Silvia, E. S., & MacCallum, R. C. (1988). Some factors affecting the success of specification searches in covariance structure modeling. *Multivariate Behavioral Research*, *23*, 297–326. [4]

Sörbom, D. (1989). Model modification. *Psychometrika*, *54*, 371–384. [4]

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, *15*, 201–293. [2]

Spearman, C. (1927). *The abilities of man*. New York: Macmillan. [2]

Steiger, J. H., & Lind, J. M. (1980, June). *Statistically based tests for the number of common factors*. Paper presented at the meeting of the Psychometric Society, Iowa City, IA. [2, 3]

Stelzl, I. (1986). Changing the causal hypothesis without changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research*, *21*, 309–331. [5]

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston: Allyn & Bacon. [2, 4, 5, 9]

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52, 393–408. [9]

Tanaka, J. S. (1987). "How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development*, *58*, 134–146. [10]

Taylor, S. (1999). (Ed.). *Anxiety sensitivity: Theory, research, and treatment of the fear of anxiety*. Mahwah, NJ: Erlbaum. [8]

Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation*. Newbury Park, CA: Sage. [8]

Thurstone, L. L. (1935). *The vectors of mind*. Chicago: University of Chicago Press. [2]

Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago: University of Chicago Press. [2]

Tobin, D., Holroyd, K., Reynolds, R., & Wigal, J. (1989). The hierarchical factor structure of the Coping Strategies Inventory. *Cognitive Therapy and Research*, *13*, 343–361. [8]

Tomás, J. M., Hontangas, P. M., & Oliver, A. (2000). Linear confirmatory factor models to evaluate multitrait–multimethod matrices: The effects of number of indicators and correlation among methods. *Multivariate Behavioral Research*, *35*, 469–499. [6]

Tomás, J. M., & Oliver, A. (1999). Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling*, *6*, 84–98. [5, 6]

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10. [3]

Tulsky, D. S., & Price, L. R. (2003). The joint WAIS-III and WMS-III factor structure: Development and cross-validation of a six-factor model of cognitive functioning. *Psychological Assessment*, *15*, 149–162. [7]

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–69. [7]

Wang, J., Siegal, H. A., Falck, R. S., & Carlson, R. G. (2001). Factorial structure of Rosenberg's self-esteem scale among crack-cocaine drug users. *Structural Equation Modeling*, *8*, 275–286. [5]

Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item Short-Form Health Survey (SF-36). I: Conceptual framework and item selection. *Medical Care*, *30*, 473–483. [4]

Wechsler, D. (1997). *Wechsler Memory Scale—Third Edition*. San Antonio, TX: Psychological Corporation. [7]

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA: Sage. [9]

Wheaton, B. (1987). Assessment of fit in overidentified models with latent variables. *Sociological Methods and Research*, *16*, 118–154. [3]

Wheaton, B., Muthén, B., Alwin, D., & Summers, G. (1977). Assessing reliability and stability in panel models. In D. R. Heise (Ed.), *Sociological methodology* (pp. 84–136). San Francisco: Jossey-Bass. [3]

Widaman, K. F. (1985). Hierachically nested covariance structure models for multitrait–multimethod data. *Applied Psychological Measurement*, *9*, 1–26. [6]

Widaman, K. F. (1993). Common factor analysis versus principal components

analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research*, *28*, 263–311. [2]

Wiggins, J. S. (1996). *The five-factor model of personality: Theoretical perspectives.* New York: Guilford Press. [4]

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology*, *10*, 395–426. [3, 4]

Wothke, W. A. (1993). Nonpositive definite matrices in structural modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 256–293). Newbury Park, CA: Sage. [3, 4, 5, 9]

Wothke, W. A. (1996). Models for multitrait–multimethod matrix analysis. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 7–56). Mahwah, NJ: Erlbaum. [6]

Wothke, W. A., & Browne, M. W. (1990). The direct product model for the MTMM matrix parameterized as a second order factor analysis model. *Psychometrika*, *55*, 255–262. [6]

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, *5*, 161–215. [3]

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Unpublished doctoral dissertation, University of California, Los Angeles. [3, 9]

Yuan, K. H. (2005). Fit statistics versus test statistics. *Multivariate Behavioral Research*, *40*, 115–148. [3]

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. In M. E. Sobel & M. P. Becker (Eds.), *Sociological methodology 2000* (pp. 165–200). Washington, DC: American Sociological Association. [9]

Yuan, K. H., Bentler, P. M., & Kano, Y. (1997). On averaging variables in a CFA model. *Behaviormetrika*, *24*, 71–83. [9]

Yuan, K. H., & Hayashi, K. (2003). Bootstrap approach to inference and power analysis based on three statistics for covariance structure models. *British Journal of Mathematical and Statistical Psychology, 56*, 93–110. [10]

Yuan, K. H., Marshall, L. L., & Bentler, P. M. (2003). Assessing the effect of model misspecifications on parameter estimates in structural equation models. *Sociological Methodology, 33*, 241–265. [5]

Yuan, Y. C. (2000). *Multiple imputation for missing data: Concepts and new development* (SAS Tech. Rep. No. P267-25). Rockville, MD: SAS Institute. [9]

Yung, Y. F., & Bentler, P. M. (1996). Bootstrapping techniques in analysis of mean and covariance structures. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 195–226). Mahwah, NJ: Erlbaum. [9]

Zimmerman, D. W. (1972). Test reliability and the Kuder–Richarson formulas: Derivation from probability theory. *Educational and Psychological Measurement*, *32*, 939–954. [8]

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2003). *BILOG-MG, Version 3.0.* [Computer software]. Chicago: Scientific Software International. [9]

Zinbarg, R. E., Barlow, D. H., & Brown, T. A. (1997). The hierarchical structure and general factor saturation of the Anxiety Sensitivity Index: Evidence and implications. *Psychological Assessment*, *9*, 277–284. [8]

Zwick, W. R., & Velicer, W. F. (1986). Factors influencing five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432–442. [2]

Author Index

Subject Index

Page numbers in *italics* indicate pages where a subject is defined or illustrated

459