Annu. Rev. Clin. Psychol. 2005. 1:31–65
doi: 10.1146/annurev.clinpsy.1.102803.144239

First published online as a Review in Advance on November 12, 2004

STRUCTURAL EQUATION MODELING: Strengths, Limitations, and Misconceptions

Andrew J. Tomarken[1] and Niels G. Waller[2]
Department of [1]Psychology and [2]Psychology and Human Development, Vanderbilt University, Nashville, Tennessee 37203; email: andrew.j.tomarke@vanderbilt.edu, niels.waller@vanderbilt.edu

Key Words covariance structure analysis, latent variables, causal models, statistical modeling, path analysis, confirmatory factor analysis

■ **Abstract** Because structural equation modeling (SEM) has become a very popular data-analytic technique, it is important for clinical scientists to have a balanced perception of its strengths and limitations. We review several strengths of SEM, with a particular focus on recent innovations (e.g., latent growth modeling, multilevel SEM models, and approaches for dealing with missing data and with violations of normality assumptions) that underscore how SEM has become a broad data-analytic framework with flexible and unique capabilities. We also consider several limitations of SEM and some misconceptions that it tends to elicit. Major themes emphasized are the problem of omitted variables, the importance of lower-order model components, potential limitations of models judged to be well fitting, the inaccuracy of some commonly used rules of thumb, and the importance of study design. Throughout, we offer recommendations for the conduct of SEM analyses and the reporting of results.

CONTENTS

INTRODUCTION

Writers of *Annual Review* chapters are confronted with difficult decisions concerning what features to highlight for an audience composed of both specialists and nonspecialists. Structural equation modeling (SEM) presents a particularly challenging topic for review because it is mathematically complex and constantly evolving. We do not provide a highly detailed summary of the many SEM applications by clinical scientists that have appeared during the recent past. Instead, we focus on two sets of issues: the strengths and limitations of SEM as an approach to model testing and important recent developments in modeling and software capabilities.

We begin by documenting the increasing popularity of SEM among clinical scientists. We then discuss several strengths of SEM as a data-analytic approach that likely account for this trend. We particularly emphasize several recent innovations that have extended the range of potential applications and exemplify the flexible and unique capabilities of SEM. In the final section, we consider the limitations of SEM and the misconceptions that it tends to elicit. This discussion is important due to a striking paradox: While the mathematical and computational complexity of SEM has increased over the past 20 years, it has become increasingly accessible to researchers who lack specialized statistical training (Steiger 2001). An important factor accounting for the latter trend is the development of new software programs (e.g., Arbuckle & Wothke 1999, Muthén & Muthén 2004, Neale et al. 2003, Steiger 1995) and updated versions of older programs (e.g., Bentler 1995, 2004; Jöreskog & Sörbom 1996) that combine powerful capabilities with relative ease of use. Although the increased accessibility of SEM to behavioral scientists is a positive development, it is associated with a potential downside: misconceptions about SEM and ignorance about its limitations and constraints.

Throughout the chapter, we focus on those features of SEM that appear particularly relevant to clinical scientists based on a review that we conducted of applications published in several relevant journals (e.g., *Journal of Abnormal Psychology, Journal of Consulting and Clinical Psychology*). At various points, we also offer recommendations to researchers for the conduct and reporting of SEM analyses.

INCREASING POPULARITY OF STRUCTURAL EQUATION MODELING

In the years since Peter Bentler's (1980) original review in the *Annual Review of Psychology*, SEM has become an increasingly popular data-analytic approach among psychologists (Hershberger 2003). Indeed, the growing interest in SEM is not limited to psychology but encompasses other social sciences (e.g., Hayduk 1996) and, increasingly, the biological sciences (e.g., Shipley 2000). In addition to the software developments already noted, there are a number of indicators of both the growth and increasing popularity of SEM. These include:

(*a*) the large number of published articles that have used SEM appearing in psychological journals (including common publication outlets for clinical scientists) (Hershberger 2003);

(*b*) evidence that SEM has become the multivariate technique most commonly used by psychologists (Hershberger 2003);

(*c*) evidence that SEM has elicited the greatest number of articles appearing in methodological journals whose primary audience is psychologists (Hershberger 2003);

(*d*) the publication of a variety of books that introduce readers to basic issues (e.g., Kline 1998, Loehlin 2004) or focus on more complex issues (e.g., Cudeck et al. 2001, Marcoulides & Schumaker 1996);

(*e*) the ready availability of workshops and short courses on SEM;

(*f*) the establishment of a journal (*Structural Equation Modeling*) devoted specifically to SEM.

Many of these developments are not simply responses to the growth of interest in SEM but are also contributors to its increasing popularity.

A perusal of journals that are primary publication outlets for clinical scientists indicates that SEM has been increasingly used to address a variety of questions of interest. For example, in recent years researchers have used this data-analytic technique to (*a*) assess the validity and other psychometric properties of measures of many constructs relevant to clinical science (e.g., Anthony et al. 1999, Brown et al. 1998); (*b*) assess the relative contribution of genetic and environmental factors to disorders and their comorbidity (e.g., Krueger et al. 2002, Slutske et al. 1998); (*c*) test etiological models that specify the direct and indirect (i.e., mediational) effects of hypothesized causes on psychiatric disorders or other outcomes of interest (e.g., Finn et al. 2000, Trull 2001); and (*d*) examine whether measurement or causal parameters are invariant across groups (e.g., King et al. 1995, Wills et al. 2002).

Over the past ten years, one of the major methodological developments in the SEM domain has been the formulation of new approaches to the analysis of longitudinal or other forms of repeated measures data (e.g., Curran & Hussong 2003, Duncan et al. 1999, Kenny & Zautra 1995, McArdle 2001, Muthén & Curran

1997, Willett & Sayer 1994). Clinical scientists have taken advantage of these developments to address a variety of questions. For example, researchers have specified and tested (*a*) cross-lagged panel models to test hypotheses concerning the causal relations among constructs (e.g., Sher et al. 1996); (*b*) trait-state models to estimate components of variance due to traits, states, and random error and to assess the degree to which trait versus state factors contribute to comorbidity (e.g., Dumenici & Windle 1998, Jackson et al. 2000); and (*c*) latent growth models to assess the developmental trajectory of constructs over time (e.g., Curran & Hussong 2003, Garber et al. 2002).

In addition, although the traditional province of SEM has been research that relies on self-report or behavioral measures, a broader set of applications has appeared in recent years. For example, SEM has been used in neuroimaging studies to test hypotheses concerning patterns of functional connectivity among brain regions (e.g., Büchel & Friston 1997, Bullmore et al. 2000, Burgess et al. 2000) and has been applied in psychophysiological studies to model properties of the autonomic nervous system (e.g., Llabre et al. 2004). Given the diversity of applications, it is not surprising that special sections on SEM have appeared in two journals, the *Journal of Abnormal Psychology* (Tomarken & Baker 2003) and the *Journal of Consulting and Clinical Psychology* (Hoyle 1994), whose primary audience consists of clinical psychologists.

STRENGTHS OF STRUCTURAL EQUATION MODELING AS A DATA-ANALYTIC APPROACH

One obvious reason why SEM has become such an increasingly popular data-analytic option is that it has a number of strengths. One well-known feature is the ability to specify latent variable models that provide separate estimates of relations among latent constructs and their manifest indicators (the measurement model) and of the relations among constructs (the structural model). By these means, it is commonly argued, researchers can assess the psychometric properties of measures and estimate relations among constructs that are corrected for biases attributable to random error and construct-irrelevant variance (Bollen 1989). However, we should caution that the psychometric benefits of latent variable modeling can be overstated and are conditional upon various methodological factors operative in a given study (e.g., DeShon 1998, Little et al. 1999).

Another commonly acknowledged strength is the availability of measures of global fit that can provide a summary evaluation of even complex models that involve a large number of linear equations. Most alternative procedures that might be used in place of SEM (e.g., multiple regression) to test such models would provide only separate "mini-tests" of model components that are conducted on an equation-by-equation basis. In addition, via nested chi-square tests and other means, users can comparatively evaluate the fit of alternative models that differ in complexity. In this regard, SEM supports the model comparison approach to data analysis (e.g., Judd et al. 1995).

SEM also allows researchers to directly test the model of interest rather than a straw-man alternative. In most statistical contexts encountered by behavioral scientists, the researcher's theoretical hypothesis is aligned with the alternative hypothesis rather than the null hypothesis (e.g., Meehl 1978, Steiger & Fouladi 1997). For instance, although the null hypothesis tested by a typical between-groups *t* test is that two population means are equal, the researcher is typically conducting the study because he or she believes that the two means actually differ. In contrast, in SEM the theoretical hypothesis is often aligned with the null hypothesis, which specifies that the model fits exactly or at least approximately (e.g., MacCallum et al. 1996). We should note, however, that some types of SEM analyses represent exceptions to this conclusion (e.g., between-group comparisons of factor means).

SEM is also an exceedingly broad data-analytic framework that is associated with unique capabilities relative to the statistical procedures traditionally used by clinical scientists. We focus on these features in the sections below, with an emphasis on recent innovations that have further increased the scope and capacities of SEM.

STRUCTURAL EQUATION MODELING AS AN INCREASINGLY BROAD ANALYTIC FRAMEWORK: RECENT MODELING AND SOFTWARE INNOVATIONS

SEM is a quite general analytic framework with many types of models as special cases. For example, multiple regression, path analysis, and confirmatory factor analysis are special cases of the SEM model that, historically, have been heavily utilized by behavioral scientists (for reviews, see, e.g., Bentler 1980, MacCallum & Austin 2000). Causal modeling with latent variables conceptually represents the union of the latter two techniques by combining the psychometric perspective characteristic of factor analysis and the emphasis on causal modeling characteristic of path analysis. In recent years, there has been a growing recognition on the part of both methodological specialists and applied users that SEM can be applied to an even wider array of data analytic problems.

Latent Growth Modeling and Other Multilevel Modeling Capabilities

LATENT GROWTH MODELING In recent years, latent growth modeling (LGM) (e.g., Curran & Hussong 2003, Duncan et al. 1999, Muthén & Curran 1997, Willett & Sayer 1994) and related SEM approaches to the analysis of repeated measures data (e.g., Rovine & Molenaar 1998) have become viable alternatives to the classic repeated measures analysis of variance (ANOVA) approach traditionally used by behavioral scientists. When one compares the LGM and repeated measures

ANOVA approaches, a variety of advantages associated with the former become apparent, including:

(*a*) the capacity to model and comparatively evaluate a broader array of growth functions (e.g., du Toit & Cudeck 2001, Ghisletta & McArdle 2001);

(*b*) a more flexible array of possible covariance structures for modeling random effects and residuals (e.g., Rovine & Molenaar 1998, Willett & Sayer 1994);

(*c*) the capacity to specify time-varying covariates (e.g., Curran & Hussong 2003, Muthén & Curran 1997);

(*d*) a greater ability to embed assessments of change in more complex causal models that assess predictors, mediators, and consequences of change (e.g., Curran & Hussong 2003);

(*e*) the ability to test models that include multiple levels of hierarchically structured data (Curran & Hussong 2003, Duncan et al. 2002);

(*f*) a better ability to assess the multivariate patterning of change across multiple measures (e.g., Willett & Sayer 1996);

(*g*) the capacity to assess whether higher-order constructs adequately account for relations among lower-order developmental functions (e.g., Duncan et al. 1999, McArdle 1988);

(*h*) access to better methods for the treatment of missing data (e.g., Duncan et al. 1999, McArdle & Hamagami 1991, see also the discussion of missing data below);

(*i*) greater statistical power according to the studies that have appeared to date (Duncan et al. 1999, Fan 2003, Muthén & Curran 1997).

Several of these benefits are quite relevant to the interests of clinical scientists. For example, via associative latent growth curve models, researchers can assess the degree to which two or more symptom dimensions demonstrate concordant patterns of change over time. While most examinations of comorbidity rely on correlations among measures assessed at either one point in time or aggregated across time, such examinations of developmental synchrony offer an additional window on the phenomenon. Higher-order growth models such as factor-of-curve and curve-of-factor models (e.g., Hancock et al. 2001) extend these capabilities and address the critical issue of the optimal level of the diagnostic or dimensional hierarchy for addressing questions pertaining to etiology, treatment, or prevention. For example, using a factor-of-curve LGM, Duncan et al. (1999) found that a general substance use factor strongly accounted for the relations among the growth trajectories of alcohol use, tobacco use, and marijuana use. As another example, multiple-group LGMs allow researchers to analyze cohort sequential designs that are otherwise quite difficult to handle statistically (Duncan et al. 1996, Ghisletta & McArdle 2001). By this means, segments of a limited amount of temporal data from different cohorts (e.g., those assessed between ages 9 and 11, between 10 and 12, etc.) can be linked together and used to approximate a longer-term developmental function.

BROADER MULTILEVEL MODELING CAPABILITIES Latent growth models are a specific type of multilevel model that is appropriate for clustered data structures in which repeated observations are nested within individuals. One recent development within the SEM domain is the capacity to model more general nested data structures (e.g., students nested in classrooms, individuals nested within families) beyond those that fall under the rubric of LGMs (e.g., Bentler & Liang 2002, Curran 2003, du Toit & du Toit 2004, Hox 2002, Liang & Bentler 2004, McDonald & Goldstein 1989, Muthén 1994). Such multilevel models have some similarity to multiple-group SEM models that have been available for a number of years. Conceptually, the distinction between the two parallels that between fixed-effects and random-effects designs in the context of the general linear model. Multilevel SEM models are appropriate for designs involving a large number of groups (e.g., 100–200), the effects of which are considered random. Multilevel SEM analyses provide aggregate estimates of within-group and between-group parameters but not separate estimates of the parameters for each group.

Such models help researchers to model behavioral phenomena when the experimental units are nonindependent. As in other statistical contexts, applications of SEM that do not account for dependencies among the experimental units are associated with problems (overestimation of model parameters, underestimation of standard errors, and inflated chi-square statistics; Julian 2001). In addition, SEM multilevel models can prevent significant distortions in results that occur when analyses fail to account for between-group heterogeneity (Muthén 1989). From a more substantive perspective, SEM multilevel models allow researchers to address interesting questions by providing separate estimates of within- and between-group relations. For example, using the multilevel approach, Duncan et al. (1996) assessed the relation between latent substance abuse and family conflict variables on both a between- and within-family basis. By this means, they were able to address two distinct questions: (*a*) Are those families characterized by the greatest overall levels of substance abuse also characterized by the highest levels of family conflict? (*b*) Within a family, do those siblings who demonstrate the highest levels of substance abuse also report the highest levels of family conflict? In light of the increasing interest in assessing the effects of shared versus nonshared environment (Plomin et al. 2001), we expect that multilevel covariance structure modeling will become an increasingly important tool for comparing between- and within-family influences on behavior.

SEM multilevel models are, however, associated with several limitations at their current stage of development. Because Muthén's (1989, 1994) MUML approach has been the most frequently used method to date, we will focus on its properties. This approach has several constraints: (*a*) Model setup and analysis are more complex than is typically the case with single-level SEM models. (*b*) It provides full-information maximum likelihood (FIML) estimates of within-group and between-group parameters when group sizes are equal, but only approximate FIML estimates when the design is unbalanced (i.e., group n's are unequal) (Muthén 1994). (*c*) When the design is unbalanced, chi-square tests of overall fit and of individual parameters are liable to have inflated Type 1 error rates (Hox &

Maas 2001). (*d*) It provides for random intercepts but not slopes (e.g., between-group variability in factor loadings and path coefficients cannot be modeled) (Hox 2002). (*e*) Even when the total sample size is rather large relative to the typical SEM study conducted by clinical scientists, various problems concerning the estimation of between-group parameters can arise when the number of groups is small (e.g., 50). These include inadmissible or biased parameter estimates and inaccurate standard errors (Hox & Maas 2001).

In recent years, several new approaches to estimation of SEM multilevel models have been developed that are designed to overcome the steep computational hurdles involved in generating true FIML estimates that are applicable even in the unbalanced case (du Toit & du Toit 2004, currently available in LISREL V8.5; Liang & Bentler 2004, soon to be available in EQS 6.00). Initial findings concerning the performance of these procedures are encouraging (e.g., Liang & Bentler 2004) and we anticipate continued innovations in this area.

COMPARISON TO HIERARCHICAL LINEAR MODEL APPROACHES Using software such as HLM (Raudenbush et al. 2000) or SAS PROC MIXED, many clinical scientists use a hierarchical linear model (HLM) approach (sometimes known as random regression modeling, linear mixed-effects modeling, or multilevel modeling) to perform growth curve or other types of multilevel analyses (for reviews, see, e.g., Raudenbush & Bryk 2002, Snijders & Bosker 1999). In the case of growth curve models, the parallels between the HLM and SEM approaches are particularly striking. Indeed, under a variety of conditions, the classic two-level HLM and SEM growth curve approaches yield essentially identical solutions (e.g., Curran 2003, Willett & Sayer 1994). More generally, as Curran (2003) has shown, when multilevel models are parameterized in a manner consistent with the LGM approach (i.e., effects of level-1 predictors are specified by fixed-factor loadings), a variety of SEM multilevel models can be specified that are analytically equivalent to those provided by HLM software.

These observations raise the question of the relative merits of the two general approaches (SEM or HLM) to multilevel models. In some respects, the SEM approach has several advantages, including a broader and more interpretable array of measures of overall model fit, more flexible modeling of residual structures and of growth functions (e.g., typically, some slope loadings can be freely estimated parameters), and a better overall capacity to model latent variables and their multivariate associations (e.g., Chou et al. 1998, Curran 2003, Willett & Sayer 1994). Conversely, HLM models are generally easier to specify, are less likely to be associated with estimation problems, and are able to perform certain types of analyses that the SEM approach cannot easily handle. There are additional points of comparison (e.g., robustness to assumption violations, relative power) between the two approaches that require further study by methodologists. Although such differences exist, the overriding point to emphasize here is that "the boundaries between these two modeling strategies are becoming increasingly porous" (Curran 2003, p. 565).

Modeling of Categorical Observed and Latent Variables

Another significant development that has been attributable to the work of Muthén (2001, 2002; Muthén & Muthén 2004) and other methodologists (e.g., Skrondal & Rabe-Hesketh 2004) is enhanced facility for modeling categorical observed and latent variables. These developments are particularly important, given the long-standing relevance of taxometric issues to clinical science (Waller & Meehl 1998). For example, using Mplus (Muthén & Muthén 2004), researchers can estimate structural equation mixture models to test hypotheses concerning the presence of unobserved latent classes characterized by different distributions on the variables of interest (for an application to clinical science, see van Lier et al. 2003). Latent growth mixture modeling is a particularly interesting variant that can be used to determine if subgroups can be identified that demonstrate distinct developmental trajectories (e.g., Li et al. 2001, Muthén 2001). Using a latent growth mixture modeling approach, Colder et al. (2002) delineated five longitudinal drinking patterns among adolescents and found that these different patterns had unique correlates.

Although these developments are exciting, these newer methods raise several potentially problematic issues. For example, SEM growth mixture modeling can lead to the discovery of spurious latent classes when the structural model is misspecified, when there are nonlinear relations among observed and latent variables, and when the distributions of latent variables are nonnormal (Bauer & Curran 2003, 2004). Even given these and other cautions that could be cited, we believe that the development of mixture modeling capabilities underscores that SEM is a broad framework that has grown progressively more inclusive with time.

From a historical perspective, these newer SEM capabilities represent the logical culmination of a long-term trend toward the development of increasingly general statistical models. For example, in the 1970s and 1980s, statisticians developed generalized linear models (GLMs), which represent a liberalization of the classical ordinary least-squares linear model to allow for nonlinear functional forms and nonnormal response distributions (e.g., Hardin & Hilbe 2001, Nelder & Wedderburn 1972). The GLM framework unifies several superficially disparate statistical techniques (e.g., linear regression, logistic regression, Poisson regression) under a single estimation framework (for a review, see Hardin & Hilbe 2001). More recent developments have allowed an even broader class of models to be incorporated into the GLM framework (e.g., repeated measures models and other models that include clustered data structures) (for reviews, see, e.g., Agresti 2002, Hardin & Hilbe 2003). From this perspective, the development of multilevel modeling and mixture modeling capabilities within the SEM domain represents a further extension of the GLM framework to incorporate latent continuous and categorical variables (Muthén & Muthén 2004, Skrondal & Rabe-Hesketh 2004). Thus, SEM represents arguably the most general data-analytic framework at the present point in time.

SEM with Nonnormal Variables

In recent years, there have also been important developments in understanding the consequences of the violation of assumptions that underlie SEM analyses and in developing robust procedures that can be used when violations occur. Research on the normality assumption that underlies most SEM analyses conducted in practice is the prime example of this point. Despite the fact that psychological data are often poorly characterized by the normal distribution (Curran et al. 1996, Micceri 1989), most SEM applications rely on normal theory methods—such as maximum likelihood (ML) and generalized least squares (GLS)—when estimating model parameters and testing model goodness of fit. According to Yuan & Bentler (2001), researchers rarely consider the distributional properties of their data prior to fitting an SEM even though they have been encouraged to do so for at least 15 years (Breckler 1990).

Plotting distributions prior to fitting a parametric statistical model is wise statistical practice (Tukey 1980). Deviations from normality can also be detected by examining higher-order moments such as Mardia's multivariate skewness and kurtosis (Bollen 1989, p. 420; Mardia 1970, 1974). Prior to fitting a model, researchers should also screen for outliers (Bollen 1987, Lee & Xu 2003, Yuan & Bentler 2001) because "even if a proposed structure is correct for the majority of the data in a sample, a small proportion of outliers leads to biased estimators and significant test statistics" (Yuan & Bentler 2001, p. 161). For covariance structures, there are several methods for detecting outliers and influential data vectors (Lee & Xu 2003, Reise & Widaman 1999). EQS (Bentler 2004) is particularly strong in this area.

Normal theory methods, such as ML and GLS, are derived under the assumption that the data are multivariate normal (MVN), an assumption that is considerably more restrictive than univariate normality. Statisticians tend to favor normal theory methods, when they are available, because they yield parameter estimates that are (*a*) asymptotically unbiased (in large samples they are neither too large nor too small), (*b*) asymptotically efficient (in large samples they have the smallest sampling variability of any unbiased estimator), and (*c*) consistent (sample estimates converge to their population values as sample size increases) when the data are MVN *and* the model is correct. Normal theory methods for covariance structures also yield parameter standard errors and a test statistic (T_{ML}, T_{GLS}) that is distributed as a chi-square variate. Although T_{ML} in particular is often reported in empirical publications, researchers generally rely more on descriptive fit indices that are less sensitive to the effects of sample size (for discussion of other limitations associated with these test statistics, see Tomarken & Waller 2003).

Importantly, when data are not MVN, the desirable properties of normal theory estimators may not be realized. For instance, with nonnormal data, the ML test statistic (T_{ML}) tends to reject true models more frequently than the nominal (0.05) rejection rate (Curran et al. 1996, Fouladi 2000). ML standard errors also become attenuated when MVN is not satisfied. Standard errors that are biased downward

result in inflated Type I error rates when z and Wald tests are used to assess parameter significance. The GLS estimator yields similar findings. Nonnormal data can also have other undesirable effects. For example, as reviewed above, nonnormal latent variables can lead to the extraction of spurious latent classes in latent growth mixture modeling analyses (Bauer & Curran 2003, 2004). We should also note that nonnormality of the observed scores may or may not be due to nonnormal latent variables. For instance, nonnormal observed scores may represent coarsely categorized indicators (e.g., binary or Likert items) of underlying continua with normal distributions. Alternative estimators and test statistics can be used with such data (Muthén 1993; Muthén & Kaplan 1985, 1992).

The discussion above of the vitiating effects of nonnormal data in covariance structures has omitted some theoretically important details. Technically, in SEM, multivariate normality is a sufficient but not a necessary condition for realizing the desiderata of normal theory estimators (see Bollen 1989, pp. 126–128). A branch of statistics known as asymptotic robustness theory (Browne 1987, Browne & Shapiro 1988) has identified several conditions under which many (but not all) of the properties of ML estimators continue to hold with nonnormal data. Unfortunately, as noted by Bentler & Dudgeon (1996), "asymptotic robustness theory cannot be relied upon in practice, because it is practically impossible to evaluate whether its conditions are met" (p. 572). Consequently, if lack of MVN is a concern—as it should be in most clinical studies—researchers should consider the methods that are described in the following paragraphs.

One approach to dealing with nonnormal data is to use an estimator with less restrictive distributional assumptions. Browne's (1982, 1984) asymptotic distribution free method (called WLS in LISREL and AGLS in EQS) is perhaps the best-known method in this class. Unfortunately, the asymptotic distribution free method performs poorly in realistically sized samples (e.g., Chou et al. 1991, Curran et al. 1996, Fouladi 2000). This finding has led statisticians back to the ML estimator with an aim toward improving its performance with nonnormal data.

In SEM, two approaches for improving ML performance have shown promise. One approach applies scaling corrections to the ML test statistic, T_{ML}, and uses robust standard errors to mitigate bias. The Satorra-Bentler test statistic (T_{SB}) and robust standard errors (Satorra & Bentler 1986, 1988, 1994) are the best-studied corrections to normal theory estimators in covariance structures and both options are widely available in SEM packages. These statistics can also be calculated using the instructions in Bentler & Dudgeon (1996, pp. 587–588). Numerous Monte Carlo studies support the usefulness of these corrections (Chou et al. 1991, Curran et al. 1996, Fouladi 2000, Hu et al. 1992) and they are highly recommended when working with nonnormal data. In smaller samples ($N < 400$ with severely nonnormal data; see Boomsma & Hoogland 2001 for more details) the T_{SB} functions poorly; thus, in smaller samples with nonnormal data, the Yuan-Bentler Residual Based Test Statistic (Bentler & Yuan 1999, Yuan & Bentler 1998b) or the Yuan-Bentler Residual Based F Statistic may be preferable (Yuan & Bentler 1998a). These newer tests, as well as a variant of T_{SB} for missing data, are included in the

latest version of EQS and will undoubtedly undergo further scrutiny in upcoming years. Before leaving this topic, we should note that the T_{SB} should not be used for traditional chi-square differences tests because the difference between the T_{SB} of nested models is not distributed as a chi-square. Satorra & Bentler (2001, p. 511) describe an appropriate method for testing nested models with the T_{SB}.

Bootstrap methods (Efron 1979) represent a second choice when fitting covariance structures to nonnormal data (Bollen & Stine 1993, Yung & Bentler 1996, Yuan & Hayashi 2003). Such methods empirically generate sampling distributions via resampling without replacement from the original data. Using bootstrap samples, researchers can estimate accurate significance levels for T_{ML} and appropriate standard errors (i.e., with correct coverage probabilities) for various model parameters including direct and indirect effects (Bollen & Stine 1990). Simulation studies (Enders 2002, Nevitt & Hancock 2001) suggest that the bootstrap performs well in this context and that it outperforms the T_{SB} and robust standard errors in small samples (Enders 2002, Fouladi 2000, Nevitt & Hancock 2001). Unfortunately, in very small samples ($N < 100$), the bootstrap also yields inaccurate results for covariance structures.

Bootstrap functionality is now included in most SEM packages. Readers should be forewarned, however, that not all programs compute the correct bootstrap sampling distribution for test statistics such as T_{ML}. As described by Bollen & Stine (1993), bootstrap estimates of likelihood ratio statistics, such as T_{ML}, should be computed on model-consistent data matrices. By design, model-consistent data fit a model *exactly*. Bollen & Stine (1993; see also Enders 2002) show how to transform raw data into model-consistent data.

The Analysis of Missing Data

Just as nonnormal data are the rule rather than the exception in clinical research, missing data are quite common. This is particularly true of longitudinal studies, which are routinely subject to attrition and other factors that render data incomplete. Unfortunately, the ad hoc approaches to missing data that traditionally have been used by clinical scientists (e.g., listwise deletion, pairwise deletion, mean imputation) have several problems. Depending on the underlying reasons for missing data and the particular method used, such approaches can produce biased and inefficient parameter estimates, highly inaccurate standard errors, confidence intervals with poor coverage probabilities, and invalid hypothesis tests (e.g., Allison 2003, Schafer & Graham 2002, Wothke 2000).

Over the past 20 years, an important focus of statistical research has been the development of better methods for the principled treatment of missing data (e.g., Little & Rubin 1987, Rubin 1987, Schafer 1997). Such methods are applicable to a variety of contexts and thus are not intrinsically tied to SEM. Nevertheless, it is arguably the case that SEM has become the statistical framework most frequently used to demonstrate and compare these alternative approaches to the treatment of missing data. Correspondingly, an important innovation within the SEM domain

has been the development of software that allows researchers to use these newer methods. Four such methods are currently available in at least one SEM software package: (*a*) multisample analysis (Allison 1987, Muthén et al. 1987), which compares distinct groups that differ in missing data patterns; (*b*) full-information maximum likelihood (FIML) (sometimes referred to as casewise, direct, or raw ML) (Finkbeiner 1979), which generates ML estimates of the parameters of a specified model based on the available data per participant; (*c*) the expectation-maximization (EM) algorithm (Dempster et al. 1977), which is typically used to compute ML estimates of the covariance matrix and mean vector that can then be used as input to a SEM analysis; and (*d*) multiple imputation (MI) (Rubin 1987, Schaefer 1997), which first creates multiple samples in which all missing data values are estimated (i.e., imputed), then estimates the model of interest separately for each sample, and finally generates aggregate estimates of parameters, standard errors, and model fit by taking into account variability both between and within samples. (For descriptions of these four approaches, comparisons among them, and comparisons to traditional approaches, see Allison 2002, 2003; Collins et al. 2001; Duncan et al. 1998; Enders 2001a,b; Schaefer & Graham 2002; Schaefer & Olsen 1998; Sinharay et al. 2001; Wothke 2000; and Yuan & Bentler 2000.)

Several points concerning these newer approaches to missing data are salient. First, if the unobserved values are missing completely at random (MCAR) or missing at random (MAR), these approaches will generally yield unbiased estimates of population parameters, more accurate coverage probabilities for confidence intervals, and more efficient estimates (i.e., smaller standard errors) than the traditional methods (Allison 2003; Collins et al. 2001; Enders 2001a, 2002b; Sinharay et al. 2001; Wothke 2000). Indeed, the differences between the newer and more traditional methods can be quite striking. The available evidence also indicates that these newer approaches can often produce better results than the conventional methods when the data are not missing at random (MNAR) (Sinharry et al. 2001). Allison (2002, 2003), Collins et al. (2001), Schaefer & Graham (2002), and Sinharry et al. (2001) provide accessible definitions and examples of the MCAR, MAR, and MNAR categories.

On balance, we would recommend the FIML and MI approaches to SEM users because they can be more flexibly applied than the multisample alternative and generally produce more accurate estimates of standard errors than EM (but see Yuan & Bentler 2000 for a modified EM approach). A direct comparison of the FIML and MI approaches indicates contrasting strengths and weaknesses. For example, FIML is available, or soon will be available, in several commonly used software packages and is relatively easy for users to implement. On the other hand, MI can be used in an even broader array of situations than FIML (e.g., when estimation methods other than ML are used) and imputed data sets can subsequently be used in a variety of different types of analyses. However, this approach can be more difficult for users to implement than FIML (Allison 2003). In addition, because at present only one of the commonly used SEM packages (LISREL) has a built-in option for creating multiply imputed data sets, users

may have to use external software [Schafer's (1999) NORM program or SAS PROC MI] in conjunction with an SEM program. SAS users can combine the multiple-imputation programs PROC MI (which generates multiply-imputed data sets) and PROC MIANALYZE (which generates aggregated parameter estimates and standard errors across replicated data sets) with the SEM program PROC CALIS (for an example, see Allison 2003).

As is the case with complete-case SEM analyses, the ML-based missing data approaches (multiple-sample, EM, and FIML) assume multivariate normality and the most commonly used MI approaches generate imputed values under a multivariate normal model (Schaefer & Graham 2002). In the case of the ML-based methods, there is evidence that violations of the multivariate normality assumption can have effects on estimates, standard errors, and measures of model fit that parallel those reviewed above in the context of complete-case analyses (e.g., Enders 2001b, Yuan & Bentler 2000). Fortunately, recent evidence indicates that the bootstrapping procedures (e.g., Bollen & Stine 1993) and rescaled statistics (e.g., Chou & Bentler 1995) available for complete-case analyses of nonnormal data can be profitably extended to the missing data context (e.g., Enders 2001b, 2002; Yuan & Bentler 2000). We particularly recommend Enders's (2002) article and the accompanying SAS macro because no SEM programs currently provide bootstrap resampling when data are missing.

Over the past several years, an increasing number of SEM papers published by clinical scientists have made use of these newer missing data capabilities. Although this development is a positive one, we should caution against the blind application of such methods under the assumption that they will automatically provide unbiased and optimally efficient estimates. For example, when data are not missing by design but due to factors that are beyond the control of the researcher, it is likely that they are MNAR, rather than MCAR or MAR (Collins et al. 2001). These procedures produce unbiased estimates under only the latter two conditions. Unfortunately, our review of empirical applications indicated that users typically failed to address the issue of missing data mechanisms and thus interpreted their results without any qualifications. We also found that no empirical studies took explicit steps to improve the quality of estimation by including auxiliary variables in a model that are predictors of either missingness or the variables that have missing values. Such auxiliary variables can be important in minimizing biases because they render the situation more "MAR-like" even if the data are formally MNAR (e.g., Collins et al. 2001, Graham 2003).

LIMITATIONS IN APPLICATIONS

Although we have emphasized that SEM is a broad-analytic framework, it has been rarely used in some important design and analysis contexts. Below, we note two of the most important omissions and address the degree to which they reflect genuine limitations of the SEM approach versus other factors.

Interaction and Other Nonlinear Models

If there is one paradigmatic framework that has influenced thinking about the etiology of psychopathology for a number of years, it is the vulnerability-stress model (sometimes called the "diathesis-stress" model) (e.g., Alloy et al. 1999, Monroe & Simons 1991, Zubin & Spring 1977). According to this perspective, it is the specific combination of a predisposing vulnerability to disorder and subsequent exposure to stress that triggers the onset and/or maintenance of psychopathology. Although both additive and interactive models can be accommodated under the vulnerability-stress framework (e.g., Alloy et al. 1999, Monroe & Simons 1991), it most commonly predicts a statistical interaction between vulnerability factors and stressors. As such, a vulnerability-stress model is one type of moderator effect. Moderator effects are of broad interest to behavioral scientists in a variety of contexts (e.g., Baron & Kenny 1986) and are most commonly tested by interaction terms in general linear models.

Although interactions constitute a major class of hypotheses formulated by researchers, our review of empirical applications indicates that users have only rarely used SEM to test interaction hypotheses. This omission is certainly not due to the absence of statistical tests of interaction hypotheses from empirical publications. Such tests routinely appear in the literature but typically are performed by other analytic approaches.

When tests of interactions appear in SEM publications, they are almost always interactions between a categorical and continuous variable tested via a multiple-sample modeling approach. For example, to assess whether gender moderated the effects of coping styles on substance abuse, Wills et al. (2001) conducted a multiple-group analysis comparing coping effects in males and females. Although the multiple-group approach is a valuable one, it has its limitations when both predictors of interest are continuously distributed variables (e.g., Maxwell & Delaney 1993). In principle, the preferred alternative would be procedures that allow explicit specification of interactions between continuously distributed latent variables in a manner analogous to that traditionally used in multiple regression analyses. Indeed, SEM would appear to have some clear advantages as an approach to testing interactions. For example, interaction terms often have low reliability, which can bias estimates and compromise power (Moosbrugger et al. 1997). In theory, the ability to model interactions using latent variables should correct such effects.

In fact, a number of SEM procedures have been suggested for modeling interactions (e.g., Algina & Moulder 2001; Bollen & Paxton 1998; Jaccard & Wan 1995; Jöreskog 2000; Jöreskog & Yang 1996; Kenny & Judd 1984; Klein & Moosbrugger 2000; Lee et al. 2004; Ping 1996; Wall & Amemiya 2000, 2001). Unfortunately, the specification and estimation of SEM models with latent variable interactions are associated with potential problems and complexities that likely account for why researchers have tended to avoid their use (for reviews, see Moosbrugger et al. 1997, Schumaker & Marcoulides 1998). Some of the issues that have arisen are described in the list below.

(*a*) Most of the procedures that have been suggested involve nonstandard and complex model specifications that are challenging for the average user and thus susceptible to error. Indeed, errors have even been noted in the specifications developed by SEM specialists [see, e.g., Jöreskog & Yang's (1996) comments on the Kenny & Judd (1984) model].

(*b*) Convergence problems have been observed with some procedures (e.g., Algina & Moulder 2001, Lee et al. 2004).

(*c*) Because products of normally distributed observed and latent variables are themselves not normally distributed, standard errors and estimates of fit might not be accurate (for a review, see Moosbrugger et al. 1997). This problem will be more severe to the extent that the latent exogenous variables used to form the product term are highly correlated. Coping with this problem might ultimately require the computation of an estimated asymptotic covariance matrix from sample data (Jöreskog & Yang 1996) or restrictions on the number of observed product variables. Given the sample sizes commonly used in practice, however, the former approach might yield biased and inefficient estimates (Moosbrugger et al. 1997) and the latter approach can produce inconsistent results across the specific indicators chosen (Lee et al. 2004).

(*d*) If the latent variables that denote main effects are not normally distributed, the parameter estimates yielded by several procedures are not consistent (Wall & Amemiya 2001).

(*e*) Most of the methods proposed in the literature are applicable to a restricted class of measurement models (Wall & Amemiya 2000).

(*f*) Although a number of alternative procedures and options have been proposed, the selection of an optimal approach is made difficult by the absence of any one study or set of studies that compares all the viable alternatives that have been proposed to date under a variety of conditions.

(*g*) Some of the more promising approaches are not easily available in conventional SEM software.

Owing to these factors, users are likely to be legitimately confused about how best to proceed. On the positive side, we should add that latent variable interaction modeling is a very active area of research that has yielded several promising new developments in recent years (e.g., Jöreskog 2000; Klein & Moosbrugger 2000; Lee & Zhu 2002; Lee et al. 2004; Wall & Amemiya 2000, 2001). We look forward to a more comprehensive comparative evaluation of these new approaches and an increase in their accessibility to users via software developments and other means (e.g., tutorial papers).

We should note that the limitations in modeling interaction terms reflect more general difficulties in nonlinear modeling using SEM. Users can rather easily specify some types of nonlinear models that will yield valid estimates (e.g., LGM models that specify quadratic effects). For many other types of nonlinear models,

however, the same procedures that have been developed for interactions have to be used, and the same problems can occur. In addition, when nonlinear models are tested using SEM software, they have to be parameterized to meet the constraint that all measurement and structural equations must be linear in their parameters (compare Burchinal & Appelbaum 1991 and McDonald 1982). As a result, the most common nonlinear models other than interactions have been quadratic polynomials. In recent years, methodological specialists have developed several creative ways to specify intrinsically nonlinear (e.g., exponential) SEM models via nonlinear equality constraints, transformations of model parameters, and other means (e.g., Boker & Graham 1998, du Toit & Cudeck 2001, Muthén & Muthén 2004). We consider these promising developments particularly important because such models often yield more meaningful and interpretable patterns than their polynomial approximations (e.g., Cudeck & du Toit 2002). In addition, nonlinear dynamic models are increasingly applied to psychological phenomena that linear models do not handle well (e.g., Heath 2000). Indeed, the great majority of mathematical models that have proven influential in physics, engineering, medicine, and other domains are nonlinear. We expect that SEM will be perceived as a useful data-analytic technique by researchers outside the behavioral sciences to the extent that it can accommodate such models.

Is SEM Underutilized in Experimental Studies?

Our review of published papers revealed a second notable omission. SEM has been only rarely used in the context of true experiments; that is, studies in which participants are randomly assigned to treatments and independent variables are directly manipulated. For example, despite evidence that SEM can provide quite novel models for assessing treatment process and outcome (e.g., Khoo 2001, Muthén & Curran 1997), it has been rarely used in randomized treatment studies.

We speculate that there are several reasons why clinical scientists have tended not to use SEM in experimental studies. First, historically, it has been viewed as a technique for testing causal hypotheses in the context of nonexperimental studies that lack random assignment. Second, researchers are concerned that the inclusion of categorical variables denoting group status might violate the assumption of multivariate normality. Finally, researchers are concerned that their sample sizes are not sufficient for SEM, which is based on asymptotic theory.

Several of these issues can be rather easily addressed. First, SEM certainly is not limited to nonrandomized contexts. For example, both the LGM approaches discussed above and SEM alternatives to the MANOVA (Cole et al. 1993, Hancock et al. 2000, Kano 2001) can be used to compare groups in randomized experiments. These SEM approaches have several advantages relative to the procedures traditionally used by clinical scientists (e.g., increased power, relative freedom from certain assumptions) and do not necessarily violate the multivariate normality assumption, which is more circumscribed than many researchers commonly believe.

The issue of sample size is a more serious concern and may well be the most critical impediment to the use of SEM in experimental studies. Over the years, a number of simulation studies have assessed the effects of variations in sample size on SEM analyses (for reviews, see Boomsma & Hoogland 2001, Hoogland & Boomsma 1998). Unfortunately, one conclusion emanating from such studies is that there is no one recommended minimal sample size that is broadly applicable in all contexts. The prime reasons are that variations in sample size influence a number of factors (e.g., bias of parameter estimates, power, likelihood of inadmissible estimates) and interact with several other factors (e.g., degree of assumption violation, overall model complexity). If we were forced to quote a specific numeric range, based on the simulation studies reviewed by Boomsma & Hoogland (2001), we would recommend that sample sizes be at least in the 200 range even when relatively simple models (e.g., a confirmatory factor-analytic model with two factors and three to four indicators per factor) are tested. For more complex models and/or models for which the assumption of multivariate normality is likely violated, we would recommend larger N's—in some cases, much larger.

Clearly, these recommended sample sizes are larger than those used in most laboratory experiments and in many randomized treatment or prevention studies. One important direction for future inquiry is to assess how the SEM alternatives to the ANOVA and MANOVA perform under sample size conditions that are more like those used in practice. In such contexts, bootstrapping approaches to SEM (e.g., Bollen & Stine 1993, Nevitt & Hancock 2001) may prove to be a valuable option. Even given these qualifications and uncertainties, we believe that SEM should be more frequently used in treatment and prevention studies, many of which have sample sizes that could be considered within an acceptable range. Muthén & Curran (1997) have provided an excellent overview of some of the possible applications of SEM in such contexts.

MORE GENERAL LIMITATIONS, CONSTRAINTS, AND MISCONCEPTIONS

SEM has more general limitations in addition to the fact that it may not be optimally applied in specific data-analytic contexts. Although users are typically aware of the strengths of SEM that we have detailed at various points, they are often unaware of such limitations and they are subject to several additional misconceptions. As a result, they tend to overstate both the strength and certainty of the conclusions yielded by SEM analyses. Although a number of potential issues could be addressed, we focus on those factors that we deem most relevant to clinical scientists.

Omitted Variables

As several methodologists have emphasized over the years, structural models—like all statistical models—are typically only approximations of reality (e.g., Browne

& Cudeck 1993, Cudeck & Henley 1991, MacCallum 2003, MacCallum & Austin 2000, Meehl & Waller 2002). One way that SEM models are approximations is by omitting variables that are implicated in the causal processes or other features of a model. Such omissions present a misleading picture of the measurement and/or causal structure and, in addition, commonly result in biased parameter estimates and inaccurate estimates of standard errors (e.g., Kaplan 1989, Mauro 1990, Reichardt 2002). Although the problem of omitted variables is certainly not unique to SEM, it is a major feature of well-known criticisms of SEM as an approach to model testing (e.g., Cliff 1983, Freedman 1987).

It is likely that the great majority of the SEM models specified and tested by clinical scientists omit important variables. This conclusion is particularly true of causal models, which are commonly specified by clinical scientists. Surprisingly, however, our review of published applications indicated that omitted variables are rarely acknowledged by clinical scientists. One reason may be an assumption that, if a model fits well, it must include all the necessary and important variables implicated in the hypothesized structure. As Tomarken & Waller (2003) have pointed out, however, while the conventionally used fit indices are sensitive to omitted variables in many contexts, in other cases they may be insensitive. Thus, good fit by no means guarantees the inclusion of all relevant variables in a model.

A second reason why users might downplay omitted variables is that structural models routinely include residual terms that denote the composite effects of the unmeasured influences on a given variable. The variances of such residual terms are typically freely estimated parameters in structural models. Covariances among residuals can also be specified that denote omitted common causes that contribute to the covariance between two constructs. Because researchers can use such residual terms to account for omitted variables, they may believe that the latter are not a problem.

Unfortunately, the provision of residual variance and covariance terms does not necessarily solve the problem of biased parameter estimates and inaccurate standard errors introduced by omitted variables. As reviewed by Tomarken & Waller (2003), there are commonly constraints on both the number and types of covariances involving residual terms that can be specified. As a result, biased estimates can still be obtained—and most likely are, in the majority of cases.

There is also a more general point to consider. In our experience, users often underestimate the importance of residual variance and covariance terms for generating a model with acceptable fit. Were such terms excluded, a high proportion of the models tested by researchers would fit poorly or demonstrate other problems (e.g., inadmissible estimates). Note, however, the paradox here: Terms denoting unknown and unmeasured variables that have been omitted from the researcher's theoretical model can contribute significantly to the fit of that model. In essence, the residual parameterizations afforded by SEM software can mask the limitations of a rather incomplete model.

For all the reasons that we have noted, it is important for authors of SEM articles to acknowledge the high likelihood of omitted variables and their deleterious effects

on parameter estimates, standard errors, and broader inferences about structure (Tomarken & Waller 2003). Authors might also consider using various types of sensitivity analyses to assess the possible biases induced by omitted variables on coefficients (e.g., Mauro 1990, Scheines et al. 1994).

Does SEM Encourage Neglect of Lower-Order Model Components?

Although measures of global fit test the validity of model-imposed restrictions on the covariance matrix, they do not directly test what might be considered lower-order components of a model (Tomarken & Waller 2003). Such components include specific model parameters (e.g., path coefficients) and relevant quantities that can be derived from such parameters, such as the proportion of variance in an endogenous variable that is accounted for by the specified predictors in the model (excluding, of course, residual terms). In our experience, users are sometimes unaware that a model can fit perfectly yet be associated with problematic lower-order components (e.g., parameter estimates that are biased, small in magnitude, or opposite to theoretical expectations) (e.g., Bollen 1989).

For example, it is possible for a model to fit perfectly, yet account for well below 1% of the variance of the primary endogenous variables (Tomarken & Waller 2003). This situation can arise due to a factor that we have already highlighted in the discussion of omitted variables: the ability to specify residual terms. In this case, the critical factor is residual variance parameters. Even if the explicitly specified causes of an endogenous variable account for only a small proportion of its variance, the implied and observed variances can be equal if the residual variance is sufficiently large. Because residual variances are usually just-identified parameters with few restrictions, they can often rather easily fill in the difference and contribute to an implied variance that equals the observed variance (Tomarken & Waller 2003).

A second factor that enables the combination of good fit and small magnitudes of association is the nature of the structural equation models specified and tested by researchers. Such models are less precise than the detailed graphic depictions appearing in journal articles might imply. They generally include many more free parameters than fixed parameters, and the former often are the researcher's primary interest. By the very nature of the term, free parameters can conceivably take on a range of possible values, with the optimal (e.g., maximum likelihood) estimate in any given context being determined by numerical algorithms designed to maximize fit to the observed data. Thus, the SEM analyst may predict that construct X causes construct Y, but rarely predicts a precise point value for the causal coefficient or a range of plausible or acceptable values.

This characteristic of most SEM models tested by researchers is by no means unique to SEM. It reflects the more general absence of point or range predictions in clinical psychology and related disciplines (e.g., Meehl 1978, Roberts & Pashler 2000). In addition, we should acknowledge that many SEM models do incorporate

theoretically meaningful fixed or constrained coefficients beyond those needed simply to identify a model (e.g., behavioral genetics models constitute a prime example of this point). Even given these considerations, most of the critical specifications of most SEM models tested in practice are highly flexible—perhaps overly flexible—and accommodate a broad range of possible outcomes (e.g., Roberts & Pashler 2000). For this reason, coefficients that are small in magnitude or even opposite in sign to theoretical expectations could conceivably occur in the context of a well-fitting model.

As these points suggest, researchers should focus more on measures of effect size and of association, confidence intervals, and other lower-order components (e.g., Bollen 1989, Tomarken & Waller 2003) when evaluating a model. Unfortunately, our review of SEM papers indicated that clinical scientists often failed to report confidence intervals and measures of association (e.g., proportion of variance accounted for) or effect size.

There is an additional respect in which SEM analyses conducted by behavioral researchers have tended to ignore lower-order components of the data. SEM analyses generally focus on summary statistics; that is, covariances and, in some cases, means. For example, raw data values are not needed as input for the great majority of analyses and the core component of most fit indices is the discrepancy between the sample covariance matrix and covariance matrix implied by the model. Unfortunately, the reliance on summary statistics and measures of global fit may lead researchers to ignore the issue of how well models fit at the level of the individual participant. One likely reason for this omission is that SEM analyses rarely have been used for the selection or prediction of individuals in real-world contexts (Reise & Widaman 1999).

We believe that an assessment of person-fit can potentially be quite valuable in the SEM context. It might ultimately sharpen models and predictions by identifying subgroups of experimental participants who do not fit the model currently under consideration but who fit modified models or demonstrate aberrant response patterns. Recently, methodologists have developed measures of person-fit that can be used in the SEM context (Neale 2000, Reise & Widaman 1999). For example Reise & Widaman (1999) have proposed a measure (IND_{CHI}) that is computed as –2 times the difference between the likelihood of an individual's response to the estimated model of interest and the likelihood of his or her response to a saturated model (that fits perfectly). Mathematically, this measure indicates the contribution of each individual to the chi-square test of exact fit that is computed across the entire sample. Reise & Widaman's (1999) preliminary assessment of the properties of this measure was encouraging, although they did find that it was only weakly correlated with a measure of fit derived from item-response theory. To date, few studies in the SEM area have been conducted using this or other possible measures of person-fit. The recently developed SEM mixture modeling capabilities will also likely prove important in the future in this regard (Neale 2000).

In sum, we are concerned that SEM analyses encourage researchers to focus on global fit at the expense of an assessment of various lower-order features of

the data that also have an important bearing on the evaluation of a model. In this respect, SEM may actually compare unfavorably to more conventional statistical procedures used by clinical scientists. For example, when conducting an ordinary least squares regression analysis, researchers routinely report R^2 values (indicating magnitude of association) and focus on the values of individual regression coefficients. In addition, multiple regression is associated with a rich array of procedures for the analysis of residuals and other individual-level features of data (e.g., Draper & Smith 1998). In these respects, a typical multiple regression analysis might actually allow for a more comprehensive and rigorous evaluation of a model and its assumptions than would an SEM analysis (Friedman 1987).

Problems with Estimates and Tests of Parameters

Consistent with the need to attend to lower-order model components, researchers need to be aware of several problematic issues concerning the estimation and testing of individual parameters that are commonly ignored in SEM applications. First, given the observation that SEM models are approximations, it is relevant to note that the parameter estimates and associated standard errors yielded by analyses are unbiased only under the assumption that the specified model is correct. Magnifying the problem here is the phenomenon of propagation of specification errors. The estimators most commonly used in empirical applications of SEM (e.g., maximum likelihood) use all available information in the covariance matrix of the observed variables to generate parameter estimates. While this feature is associated with several advantages (e.g., smaller standard errors when models are correct), it also allows the effects of a misspecified parameter to be propagated beyond the specific equation in which it occurs (e.g., Kaplan 1988, 1989). Thus, for example, an omitted path from a given latent variable to another latent variable could potentially bias estimates of other structural or measurement parameters that would appear to be far downstream from the misspecified parameter. In other words, the costs of misspecification are not nearly as localized as many users might hope.

Several other issues are associated with estimation and testing of parameters and the formation of confidence intervals that have been highlighted in recent years by methodologists. For example, even when alternative ways to identify a model produce identical fit, the Wald *z* tests for free parameters that are commonly reported in SEM software are often not invariant (Gonzalez & Griffin 2001, Neale & Miller 1997). Thus, different ways of identifying a model may produce different ratios of a given parameter to its standard error. In addition, although the Wald statistic assumes that the sampling distribution of a parameter is normal, it is likely that such distributions often are not symmetrical given the sample sizes used in practice. The assumption of symmetry is particularly inappropriate when a parameter is close to its lower or upper bound (e.g., close to 0 in the case of variance parameter). When computed correctly (e.g., Self & Liang 1987), likelihood-ratio tests and confidence intervals for individual parameters are typically superior to

Wald-based tests and intervals (Gonzalez & Griffin 2001, Neale & Miller 1997). Unfortunately, with the exception of MX (Neale et al. 2003) and LISREL (Jöreskog & Sörbom 1996), LR tests and intervals for a large number of individual parameters can be cumbersome to compute using SEM software. In addition, even LR tests are subject to problems under certain conditions (e.g., when sample sizes are small).

Users should also be aware that the statistical theory underlying SEM pertains to covariance, rather than correlation, matrices (Cudeck 1989). Unfortunately, our review of SEM papers indicated that correlation matrices were analyzed in some studies whereas in others it was unclear precisely whether covariance or correlation matrices were used. When correlation matrices are analyzed as if they were covariance matrices, standard errors of parameter estimates are usually inaccurate. Under specific conditions (e.g., when equality constraints on parameters are tested), biased parameter estimates and inaccurate estimates of fit can also occur (Cudeck 1989, MacCallum & Austin 2000). Researchers who want to analyze correlation matrices or obtain valid tests of models in which both exogenous and endogenous latent variables are measured in a standardized metric should use constrained estimation methods (Steiger 2002). Two software packages currently available, RAMONA (Browne & Mels 1999) and SEPATH (Steiger 1995), include constrained estimation routines that allow users to easily impose the appropriate restrictions necessary to analyze standardized data. As an alternative, it is possible for users to specify explicitly the complex nonlinear constraints required for correct standardization by using some other software packages. If these different options are not available or practical, we echo MacCallum & Austin (2000) by recommending that psychopathologists fit models to covariance matrices and sacrifice some ease of interpretation.

In sum, users need to be aware of the variety of problematic issues that exists concerning the estimation and testing of parameters. Caution is also necessary when interpreting parameter estimates and associated statistics that appear in many published applications of SEM.

Other Models Will Also Fit Well

It is impossible to prove that a model is correct using statistical analyses or other means. Alternative models may be available that could fit the data equally well or better. Unfortunately, this conclusion is too often ignored by researchers, who tend to overstate the certainty and strength of the conclusions yielded by a SEM analysis. As discussed by Tomarken & Waller (2003), the two primary problems in the SEM context are: *(a)* equivalent models that impose the same restrictions on the implied covariance matrix as the target model and thus will always yield identical measures of fit (e.g., Breckler 1990, Hershberger 1994, Lee & Hershberger 1990, MacCallum et al. 1993, Stelzl 1986); and (*b*) alternative nonequivalent models that might fit the data as well as or better than the target model under consideration (e.g., MacCallum & Austin 2000, Meehl & Waller 2002, Waller & Meehl 2002). Concerning the latter, although researchers commonly conduct nested tests to

comparatively evaluate at least some of the alternatives, such tests comprise only a small subset of the possible comparisons. Because these issues have been reviewed extensively in the literature, we direct readers to the sources cited above, as well as the discussion by Tomarken & Waller (2003).

Rules of Thumb Can Be Inaccurate

In many statistical contexts, researchers use rules of thumb to guide decision making and justify the decisions made. Unfortunately, in many cases, such rules of thumb are oversimplified or simply erroneous (e.g., MacCallum et al. 2001, Marsh et al. 2004). SEM is no exception. For example, in our experience, users commonly believe that alternative ways to identify a model (e.g., fixing a factor loading at one versus fixing a factor variance at one) always produce identical results. This principle is not universally true. As noted above, tests of individual parameters may not be invariant across alternative identification schemes even when estimates of overall fit are invariant (e.g., Gonzalez & Griffin 2001). Further, as Steiger (2002) has pointed out, in certain contexts that are important in practice (e.g., when equality constraints are imposed on parameters), different ways of identifying latent variables might subtly introduce corresponding differences in the specific restrictions implied by a model. As a result, chi-square tests and other measures of fit will also differ.

The assessment of fit is arguably the area in which researchers have most consistently used rules of thumb. For a number of years, the most common criteria for fit indices that have been used by behavioral researchers are rules of thumb that lack a detailed mathematical or empirical justification. For example, one rule of thumb that has been frequently used over the years is that values of incremental fit indices [e.g., the Tucker-Lewis Index (Tucker & Lewis 1973), the Comparative Fit Index (Bentler 1990)] greater than 0.90 indicate acceptable fit. Unfortunately, several simulation studies have indicated that (*a*) such rules of thumb are often inaccurate (typically they are too lenient); (*b*) the optimal cutoff criteria for most fit indices are conditional upon a variety of factors including the estimation method used, sample size, model complexity, and the degree to which the assumption of multivariate normality is violated; and (*c*) some commonly used fit indices (e.g., the Goodness of Fit Index; Jöreskog & Sörbom 1996) are insufficiently sensitive to misspecifications whereas the rarely reported standardized root mean squared residual (Bentler 1995) is often appropriately sensitive (e.g., Browne et al. 2002; Fan et al. 1999; Hu. & Bentler 1998, 1999; Marsh et al. 1996, 2004). In short, these results indicate that the conventional rules of thumb and guidelines used by researchers for the selection and interpretation of fit indices are often erroneous or oversimplified. Given the complexity of the issues here, we agree with Marsh et al. (2004) that even under the best of circumstances a healthy dose of subjectivity is involved in determining whether a model fits well.

Researchers need to be cautious even when applying guidelines that appear more firmly grounded than rules of thumb. Monte Carlo simulation studies are a

major source of knowledge about various aspects of structural equation modeling. In the majority of cases, such studies have assessed performance when the model being tested is correctly specified in all respects. As such, they do not accurately reflect the reality noted above that the SEM models posited by researchers do not hold exactly in the population but are approximations. Clearly, then, it is important to study various SEM features when the model in question is *not* correct in the population (MacCallum 2003).

Consistent with this reasoning, those simulation studies that have been conducted assessing performance under conditions of model misspecification have contributed important new insights—for example, the evidence reviewed above that conventional rules of thumb for fit indices are often inaccurate (e.g., Hu & Bentler 1998, 1999). Although simulation studies that incorporate model error as well as sampling error can be more complex to conduct, methods have been developed to generate population covariance matrices that are approximated to precisely specified degrees by factor-analytic models (for a review, see MacCallum 2003) and broader structural equation models (Cudeck & Browne 1992). Studies using these methods have shown that the results and conclusions yielded by simulations performed when a model is misspecified may be quite different from those yielded by assessments of performance when the specified model fits perfectly (Cudeck & Brown 1992). In our view, it is important that future simulation studies attempt to mirror the reality of models as approximations.

Structural Equation Modeling Cannot Compensate for Limitations in Design and Method

As statistical consultants are well aware, sophisticated statistical procedures cannot rescue a poorly designed study. In the SEM context, even a completely correct theoretical model (e.g., one that includes all necessary variables and paths) can fit poorly and yield highly biased estimates if the study is poorly designed. Unfortunately, authors of empirical SEM papers often fail to discuss the specific rationale for decisions about design and method, and the potential impact of the choices made on results and conclusions.

Mediational models are very popular among clinical scientists and serve to illustrate the importance of design features. Some mediational studies use cross-sectional designs while others use longitudinal designs. As Cole & Maxwell (2003, see also Gollob & Reichardt 1987) have shown, cross-sectional mediational designs will yield unbiased estimates of direct and indirect causal effects only under highly restricted conditions unlikely to be met in practice and, even when longitudinal designs are used, the specific estimates obtained are highly dependent on the relation between the temporal lag of the actual causal effects and the time lags between measurement occasions. Thus, conceivably, one could have a model that specifies the correct constructs and relations between constructs, yet does not yield good estimates of direct and indirect effects because the measurement occasions are not appropriately lagged. Cole & Maxwell (2003) have also highlighted

several other design features that have a large impact on the results and conclusions reached in mediational studies. We consider their paper required reading for clinical researchers who test mediational hypotheses.

Design factors are also important because of their impact on statistical power and the sensitivity of measures of fit to misspecifications. As Tomarken & Waller (2003) noted, when power is discussed in applications of SEM, researchers tend to focus solely on the effects of sample size on the chi-square test of exact fit. By emphasizing sample size to the exclusion of other factors, SEM users demonstrate a bias that is evident in a variety of statistical contexts (McClelland 1997). However, there is ample evidence that design factors such as the reliability and number of observed indicators and the number of time points assessed in repeated measures studies are significant influences on power and sensitivity in the SEM context (e.g., Mandys et al. 1994, Matsueda & Bielby 1986, Raykov 2000, Tomarken & Waller 2003).

It is important for researchers to ensure that the experimental design and overall analytic approach confer sufficient sensitivity to detect misspecifications that would be deemed nontrivial (Tomarken & Waller 2003). Ideally, this question should be addressed while an SEM study is being planned. During this phase, two very helpful skills to exercise are the ability to conduct power analyses and computer simulations. Fortunately, such analyses can be conducted relatively conveniently using most conventional software packages. For interested readers, we recommend several excellent pedagogical papers that have appeared in recent years on these topics (MacCallum et al. 1996, Muthén & Curran 1997, Muthén & Muthén 2002, Paxton et al. 2001).

SUMMARY AND CONCLUSION

As a data-analytic approach, SEM has a number of appealing features. For example, it is arguably the most broadly applicable statistical procedure currently available and it has a number of unique and flexible capabilities. SEM has become a particularly attractive data-analytic option in recent years because of the development of several new types of models and software capabilities that are particularly well suited to the research interests of clinical scientists.

SEM is not, however, a statistical magic bullet. It cannot be used to prove that a model is correct and it cannot compensate for a poorly designed study. In addition, even a well-fitting SEM model can have problematic lower-order components and omit important variables. Ironically, several of the limitations and misconceptions that we have identified have been well known to methodologists for many years (e.g., Breckler 1990, Cliff 1983, Freedman 1987, MacCallum et al. 1993) and arise in other statistical contexts (e.g., Judd et al. 1995). In a sense, then, one could describe SEM as a cutting-edge statistical technique that is subject to some very old and familiar problems, constraints, and misconceptions. It is important for users to become aware of both the strengths and limitations of SEM. Indeed, we

consider such a balanced perspective a necessary condition for the appropriate use of this powerful statistical technique.

The *Annual Review of Clinical Psychology* is online at http://clinpsy.annualreviews.org

LITERATURE CITED

Agresti A. 2002. *Categorical Data Analysis*. Hobokin, NJ: Wiley. 2nd ed.

Algina J, Moulder BC. 2001. A note on estimating the Jöreskog-Yang model for latent variable interaction using LISREL 8.3. *Struct. Equ. Model.* 8:40–52

Allison PD. 1987. Estimation of linear models with incomplete data. In *Sociological Methodology*, ed. CC Clogg, 17:71–103. San Francisco: Jossey-Bass

Allison PD. 2002. *Missing Data*. Thousand Oaks, CA: Sage

Allison PD. 2003. Missing data techniques for structural equation modeling. *J. Abnorm. Psychol.* 112:545–57

Alloy LB, Abramson LY, Raniere D, Dyller I. 1999. Research methods in adult psychopathology. In *Handbook of Research Methods in Clinical Psychology*, ed. PC Kendall, JN Butcher, GN Holmbeck, pp. 466–98. New York: Wiley. 2nd ed.

Anthony JL, Lonigan CJ, Hecht SA. 1999. Dimensionality of posttraumatic stress disorder symptoms in children exposed to disaster: results from confirmatory factor analyses. *J. Abnorm. Psychol.* 108:326–36

Arbuckle J, Wothke W. 1999. *AMOS 4 User's Reference Guide*. Chicago: Smallwaters Corp.

Baron RM, Kenny DA. 1986. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Personal. Soc. Psychol.* 51:1173–82

Bauer DJ, Curran PJ. 2003. Distributional assumptions of growth mixture models: implications for over-extraction of latent trajectory classes. *Psychol. Methods* 8:338–63

Bauer DJ, Curran PJ. 2004. The integration of continuous and discrete latent latent variable models: potential problems and promising opportunities. *Psychol. Methods* 9:3–29

Bentler PM. 1980. Multivariate analysis with latent variables: causal modeling. *Annu. Rev. Psychol.* 31:419–56

Bentler PM. 1990. Comparative fit indexes in structural models. *Psychol. Bull.* 107:238–46

Bentler PM. 1995. *EQS Structural Equation Program Manual.* Encino, CA: Multivariate Software

Bentler PM. 2005. *EQS 6 Structural Equations Program Manual.* Encino, CA: Multivariate Software. In press

Bentler PM, Dudgeon P. 1996. Covariance structure analysis: statistical practice, theory, and directions. *Annu. Rev. Psychol.* 47:563–92

Bentler PM, Liang J. 2002. Two-level mean and covariance structures: maximum likelihood via an EM algorithm. In *Multilevel Modeling: Methodological Advances, Issues, and Applications*, ed. SP Reise, N Duan, pp. 53–70. Mahwah, NJ: Erlbaum

Bentler PM, Yuan K-H. 1999. Structural equation modeling with small samples: test statistics. *Multivar. Behav. Res.* 34:181–97

Boker SM, Graham JW. 1998. A dynamical systems analysis of adolescent substance use. *Multivar. Behav. Res.* 33:479–507

Bollen KA. 1987. Outliers and improper solutions: a confirmatory factor analysis example. *Sociol. Methods Res.* 15:375–84

Bollen KA. 1989. *Structural Equations with Latent Variables*. New York: Wiley

Bollen KA, Long JS, eds. 1993. *Testing Structural Equation Models.* Newbury Park, CA: Sage

Bollen KA, Paxton P. 1998. Interactions of

latent variables in structural equation models. *Struct. Equ. Model.* 5:266–93

Bollen KA, Stine RA. 1990. Direct and indirect effects: classical and bootstrap estimates of variability. In *Sociological Methodology*, ed. CC Clogg, 20:115–40. Oxford: Blackwell Sci.

Bollen KA, Stine RA. 1993. Bootstrapping goodness-of-fit measures in structural equation models. See Bollen & Long 1993, pp. 111–35

Boomsma A, Hoogland JJ. 2001. The robustness of LISREL modeling revisited. See Cudeck et al. 2001, pp. 139–68

Breckler SJ. 1990. Applications of covariance structure modeling in psychology: cause for concern? *Psychol. Bull.* 107:260–73

Brown TA, Chorpita BF, Barlow DH. 1998. Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. *J. Abnorm. Psychol.* 107:179–92

Browne MW. 1982. Covariance structures. In *Topics in Multivariate Analyses*, ed. DM Hawkins, pp. 72–141. New York: Cambridge Univ. Press

Browne MW. 1984. Asymptotic distribution free methods in analysis of covariance structures. *Br. J. Math. Stat. Psychol.* 37:62–83

Browne MW. 1987. Robustness of statistical inference in factor analysis and related models. *Biometrika* 74:375–84

Browne MW, Cudeck R. 1993. Alternative ways of assessing model fit. See Bollen & Long 1993, pp. 136–92

Browne MW, MacCallum RC, Kim CT, Anderson BL, Glaser R. 2002. When fit indices and residuals are incompatible. *Psychol. Methods* 7:403–21

Browne MW, Mels G. 1999. Path analysis (RAMONA). In *SYSTAT 10*: *Statistics II*, pp. II-233–91. Chicago: SPSS

Browne MW, Shapiro A. 1988. Robustness of normal theory methods in the analysis of linear latent variable models. *Br. J. Math. Stat. Psychol.* 41:193–208

Büchel C, Friston KJ. 1997. Modulation in connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modeling and fMRI. *Cereb. Cortex* 7:768–78

Bullmore E, Horwitz B, Honey G, Brammer M, Williams S, Sharma T. 2000. How good is good enough in path analysis of fMRI data? *NeuroImage* 11:289–301

Burchinal M, Appelbaum M. 1991. Estimating individual developmental functions: methods and their assumptions. *Child Dev.* 6:23–43

Burgess PW, Veitch E, de Lacy Costello A, Shallice T. 2000. The cognitive and neuroanatomical correlates of multitasking. *Neuropsychologia* 38:848–63

Chassin L, Pitts S, DeLucia C, Todd M. 1999. A longitudinal study of children of alcoholics: predicting young adult substance use disorders, anxiety, and depression. *J. Abnorm. Psychol.* 108:106–19

Chou C-P, Bentler PM, Pentz MA. 1998. Comparison of two statistical approaches to study growth curves: the multilevel model and the latent curve analysis. *Struct. Equ. Model.* 5:247–66

Chou C-P, Bentler PM, Satorra A. 1991. Scaled test statistics and robust standard errors for nonnormal data in covariance structure analysis: a Monte Carlo study. *Br. J. Math. Stat. Psychol.* 44:347–57

Cliff N. 1983. Some cautions concerning the application of causal modeling methods. *Multiv. Behav. Res.* 18:81–105

Colder CR, Campbell RT, Ruel E, Richardson JL, Flay BR. 2002. A finite mixture model of growth trajectories of adolescent alcohol use: predictors and consequences. *J. Consult. Clin. Psychol.* 70:976–85

Cole DA, Maxwell SE. 2003. Testing mediational models with longitudinal data: myths and tips in the use of structural equation modeling. *J. Abnorm. Psychol.* 112:558–77

Cole DA, Maxwell SE, Arvey R, Salas E. 1993. Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychol. Bull.* 114:174–84

Collins LM, Schafer JL, Kam CM. 2001. A

comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychol. Methods* 6:330–51

Cudeck R. 1989. Analysis of correlation matrices using covariance structure models. *Psychol. Bull.* 105:317–27

Cudeck R, Browne MW. 1992. Constructing a covariance matrix that yields a specified minimizer and a specified minimum discrepancy function value. *Psychometrika* 57:357–69

Cudeck R, du Toit SHC. 2002. A version of quadratic regression with interpretable parameters. *Multiv. Behav. Res.* 37:501–19

Cudeck R, du Toit SHC, Sorbom D. 2001. *Structural Equation Modeling: Present and Future*. Chicago: Sci. Software Int.

Cudeck R, Henley SJ. 1991. Model selection in covariance structure analysis and the "problem" of sample size: a clarification. *Psychol. Bull.* 109:512–19

Curran PJ. 2003. Have multilevel models been structural equation models all along? *Multiv. Behav. Res.* 38:529–69

Curran PJ, Hussong AM. 2003. The use of latent trajectory models in psychopathology research. *J. Abnorm. Psychol.* 112:526–44

Curran PJ, West SG, Finch JF. 1996. The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychol. Methods* 1:16–29

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. Ser. B* 39:1–38

DeShon RP. 1998. A cautionary note on measurement error corrections in structural equation models. *Psychol. Methods* 3:412–23

Draper NR, Smith H. 1998. *Applied Regression Analysis*. New York: Wiley. 3rd ed.

Dumenici L, Windle M. 1998. A multitrait multioccasion generalization of the latent trait-state model: description and application. *Struct. Equ. Model.* 5:391–410

Duncan SC, Duncan TE, Hops H. 1996. Analysis of longitudinal data within accelerated longitudinal designs. *Psychol. Methods* 1:236–48

Duncan TE, Alpert A, Duncan SC, Hops H. 1996. Multilevel covariance structure analysis of sibling substance use and intrafamily conflict. *J. Psychopathol. Behav. Assess.* 18:347–69

Duncan TE, Duncan SC, Li F. 1998. A comparison of model- and multiple-imputation-based approaches to longitudinal analyses with partial missingness. *Struct. Equ. Model.* 5:1–21

Duncan TE, Duncan SC, Okut H, Strycker LA, Li F. 2002. An extension of the general latent variable growth modeling framework to four levels of the hierarchy. *Struct. Equ. Model.* 9:303–26

Duncan TE, Duncan SC, Strycker LA, Li F, Alpert A. 1999. *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Applications.* Mahwah, NJ: Erlbaum

du Toit SHC, Cudeck R. 2001. The analysis of nonlinear random coefficient regression models with LISREL using constraints. See Cudeck et al. 2001, pp. 259–78

du Toit SHC, du Toit M. 2005. Multilevel structural equation modeling. In *Handbook of Quantitative Multilevel Analysis*, ed. J de Jeeuw, IGG Kreft. Boston: Kluwer Acad. In press

Efron B. 1979. Bootstrap methods: another look at the jackknife. *Ann. Stat.* 7:1–26

Enders CK. 2001a. A primer on maximum likelihood algorithms available for use with missing data. *Struct. Equ. Model.* 8:128–41

Enders CK. 2001b. The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychol. Methods* 6:352–70

Enders CK. 2002. Applying the Bollen-Stine Bootstrap for goodness-of-fit measures to structural equation models with missing data. *Multivar. Behav. Res.* 37:359–77

Fan X. 2003. Power of latent growth modeling for detecting group differences in linear growth trajectory parameters. *Struct. Equ. Model.* 10:380–400

Fan X, Thompson B, Wang L. 1999. Effects of sample size, estimation methods, and model

specification on structural equation modeling fit indices. *Struct. Equ. Model.* 6:56–83

Finkbeiner C. 1979. Estimation for the multiple factor model when data are missing. *Psychometrika* 44:409–20

Finn PR, Sharkansky EJ, Brandt K, Turcotte N. 2000. The effects of familial-risk, personality, and alcohol expectancies on alcohol use and abuse. *J. Abnorm. Psychol.* 109:122–33

Fouladi RT. 2000. Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Struct. Equ. Model.* 7(3):356–410

Freedman DA. 1987. As others see us: a case study in path analysis. *J. Educ. Stat.* 12:101–28

Garber J, Keiley MK, Martin NC. 2002. Developmental trajectories of adolescents' depressive symptoms: predictors of change. *J. Consult. Clin. Psychol.* 70:79–95

Ghisletta P, McArdle JJ. 2001. Latent growth curve analyses of the development of height. *Struct. Equ. Model.* 8:531–55

Gollob HF, Reichardt CS. 1987. Taking account of time lags in causal models. *Child Dev.* 58:80–92

Gonzalez R, Griffin D. 2001. Testing parameters in structural equation modeling: Every "one" matters. *Psychol. Methods* 6:258–69

Graham JW. 2003. Adding missing-data-relevant variables to FIML-based structural equation models. *Struct. Equ. Model.* 10:80–100

Hancock GR, Kuo W-L, Lawrence FR. 2001. An illustration of second-order latent growth models. *Struct. Equ. Model.* 8:470–89

Hancock GR, Lawrence FR, Nevitt J. 2000. Type 1 error and power of latent mean methods and MANOVA in factorially invariant and noninvariant latent variable systems. *Struct. Equ. Model.* 7:534–56

Hardin J, Hilbe J. 2001. *Generalized Linear Models and Extensions.* College Station, TX: Stata

Hardin J, Hilbe J. 2003. *Generalized Estimating Equations*. Boca Raton, FL: Chapman & Hall

Hayduk LA. 1996. *LISREL: Issues, Debates, and Strategies*. Baltimore, MD: Johns Hopkins Univ. Press

Heath RA. 2000. *Nonlinear Dynamics: Techniques and Applications in Psychology*. Mahwah, NJ: Erlbaum

Hershberger SL. 1994. The specification of equivalent models before the collection of data. See von Eye & Clogg 1994, pp. 68–108

Hershberger SL. 2003. The growth of structural equation modeling: 1994–2001. *Struct. Equ. Model.* 10:35–46

Hoogland JJ, Boomsma A. 1998. Robustness studies in covariance structure modeling: an overview and a meta-analysis. *Sociol. Methods Res.* 26:329–67

Hox JJ. 2002. *Multilevel Analysis: Techniques and Applications.* Mahwah, NJ: Erlbaum

Hox JJ, Maas CJM. 2001. The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Struct. Equ. Model.* 8:157–74

Hoyle RH. 1994. Introduction to the special section: structural equation modeling in clinical research. *J. Consult. Clin. Psychol.* 62:427–28

Hu LT, Bentler PM. 1998. Fit indices in covariance structural modeling: sensitivity to underparameterized model misspecification. *Psychol. Methods* 3:424–53

Hu LT, Bentler PM. 1999. Cutoff criteria to fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model.* 6:1–55

Hu LT, Bentler PM, Kano Y. 1992. Can test statistics in covariance structure models be trusted? *Psychol. Bull.* 112:351–62

Jaccard J, Wan CK. 1995. Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: multiple indicator and structural equation approaches. *Psychol. Bull.* 117:348–57

Jackson KM, Sher KJ, Wood PK. 2000. Prospective analyses of comorbidity: tobacco and alcohol use disorders. *J. Abnorm. Psychol.* 109:679–94

Jöreskog KG. 2000. *Latent Variable Scores and*

Their Uses. Lincolnwood, IL: Sci. Software Int.

Jöreskog KG, Sörbom D. 1996. *LISREL 8 User's Reference Guide*. Chicago: Sci. Software Int.

Jöreskog KG, Yang F. 1996. Nonlinear structural equation models: the Kenny and Judd model with interaction effects. See Marcoulides & Schumacker 1996, pp. 57–88

Judd CM, McClelland GH, Culhane SE. 1995. Data analysis: continuing issues in the everyday analysis of psychological data. *Annu. Rev. Psychol.* 46:433–65

Julian MW. 2001. The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Struct. Equ. Model.* 8:325–52

Kano Y. 2001. Structural equation modeling for experimental data. See Cudeck et al. 2001, pp. 381–402

Kaplan D. 1988. The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivar. Behav. Res.* 23:69–86

Kaplan D. 1989. A study of the sampling variability and *z*-values of parameter estimates from misspecified structural equation models. *Multivar. Behav. Res.* 24:41–57

Kenny DA, Judd CM. 1984. Estimating the nonlinear and interactive effects of latent variables. *Psychol. Bull.* 96:201–10

Kenny DA, Zautra A. 1995. The trait-state-error model for multiwave data. *J. Consult. Clin. Psychol.* 63:52–59

Khoo ST. 2001. Assessing program effects in the presence of treatment-baseline interactions: a latent curve approach. *Psychol. Methods* 6:234–57

King DW, King LA, Gudanowski DG, Vreven D. 1995. Alternative representations of war zone trauma: outcomes in male and female Vietnam veterans. *J. Abnorm. Psychol.* 104:184–96

Klein A, Moosbrugger H. 2000. Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika* 65:457–74

Kline RB. 1998. *Principles and Practice of Structural Equation Modeling*. New York: Guilford

Krueger RF, Hicks BM, Patrick CJ, Carlson SR, Iacono WG, McGue M. 2002. Etiologic connections among substance dependence, antisocial behavior, and personality: modeling the externalizing spectrum. *J. Abnorm. Psychol.* 111:411–24

Lee S, Hershberger S. 1990. A simple rule for generating equivalent models in covariance structural modeling. *Multivar. Behav. Res.* 25:313–34

Lee S-Y, Song X-Y, Poon W-Y. 2004. Comparison of approaches in estimating interaction and quadratic effects of latent variables. *Multivar. Behav. Res.* 39:37–67

Lee S-Y, Xu L. 2003. Case-deletion diagnostics for factor analysis models with continuous and ordinal categorical data. *Sociol. Methods Res.* 31:389–419

Lee S-Y, Zhu HT. 2002. Maximum likelihood estimation of nonlinear structural equation models. *Psychometrika* 67:189–210

Li F, Duncan TE, Duncan SC. 2001. Latent growth modeling of longitudinal data: a finite growth mixture modeling approach. *Struct. Equ. Model.* 8:493–530

Liang J, Bentler PM. 2004. An EM algorithm for fitting two-level structural equation models. *Psychometrika* 69:101–22

Little RJA, Rubin DB. 1987. *Statistical Analysis with Missing Data.* New York: Wiley

Little TD, Lindenberger U, Nesselroade JR. 1999. On selecting indicators for multivariate measurement and modeling with latent variables: when "good" indicators are bad and "bad" indicators are good. *Psychol. Methods* 4:192–211

Little TD, Schnabel KU, Baumert J, eds. 2000. *Modeling Longitudinal and Multilevel Data.* Mahwah, NJ: Erlbaum

Llabre MM, Spitzer S, Siegel S, Saab PG, Schneiderman N. 2004. Applying latent growth curve modeling to the investigation of individual differences in cardiovascular recovery from stress. *Psychosom. Med.* 66:29–41

Loehlin JC. 2004. *Latent Variable Models: An*

Introduction to Factor, Path, and Structural Analysis. Mahwah, NJ: Erlbaum. 3rd ed.

MacCallum RC. 2003. Working with imperfect models. *Multivar. Behav. Res.* 38:113–39

MacCallum RC, Austin JT. 2000. Applications of structural equation modeling in psychological research. *Annu. Rev. Psychol.* 51: 201–26

MacCallum RC, Browne MW, Sugawara HM. 1996. Power analysis and determination of sample size for covariance structure modeling. *Psychol. Methods* 1:130–49

MacCallum RC, Wegener DT, Uchino BN, Fabrigar LR. 1993. The problem of equivalent models in applications of covariance structure analysis. *Psychol. Bull.* 114:185–99

MacCallum RC, Widaman KF, Zhang S, Hong S. 2001. Sample size in factor analysis. *Multivar. Behav. Res.* 36:611–37

Mandys F, Dolan CV, Molenaar PCM. 1994. Two aspects of the simplex model: goodness of fit to linear growth curve structures and the analysis of mean trends. *J. Educ. Behav. Stat.* 19:201–15

Marcoulides GA, Schumacker RE, eds. 1996. *Advanced Structural Equation Modeling: Issues and Techniques*. Mahwah, NJ: Erlbaum

Mardia KV. 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57:519–30

Mardia KV. 1974. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhya B* 36:115–28

Marsh HW, Balla JR, Hau K-T. 1996. An evaluation of incremental fit indices: a clarification of mathematical and empirical properties. See Marcoulides & Schumacker 1996, pp. 315–53

Marsh HW, Hau K-T, Wen ZL. 2004. In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's 1999 findings. *Struct. Equ. Model.* 11:320–41

Matsueda RL, Bielby WT. 1986. Statistical power in covariance structure models. In *Sociological Methodology 1986*, ed. NB Tuma, 16:120–58. San Francisco: Jossey-Bass

Mauro R. 1990. Understanding L.O.V.E. (left out variables error): a method for estimating effects of omitted variables. *Psychol. Bull.* 108:314–29

Maxwell SE, Delaney HD. 1993. Bivariate median splits and spurious statistical significance. *Psychol. Bull.* 113:181–90

McArdle JJ. 1988. Dynamic but structural equation modeling of repeated measures data. In *Handbook of Multivariate Experimental Psychology*, ed. JR Nesselroade, RB Cattell, pp. 561–614. New York: Plenum. 2nd ed.

McArdle JJ. 2001. A latent difference score approach to longitudinal dynamic structural analyses. See Cudeck et al. 2001, pp. 342–80

McArdle JJ, Hamagami F. 1991. Modeling incomplete longitudinal data using latent growth structural equation models. In *Best Methods for the Analysis of Change*, ed. L Collins, JJ Horn, pp. 276–304. Washington, DC: Am. Psychol. Assoc.

McClelland GH. 1997. Optimal design in psychological research. *Psychol. Methods* 2:3–19

McDonald RP. 1982. Linear versus nonlinear models in item response theory. *Appl. Psychol. Meas.* 6:379–96

McDonald RP, Goldstein H. 1989. Balanced versus unbalanced designs for linear structural relations in two-level data. *Br. J. Math. Stat. Psychol.* 42:215–32

Meehl PE. 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *J. Consult. Clin. Psychol.* 46:806–34

Meehl PE, Waller NG. 2002. The path analysis controversy: a new statistical approach to strong appraisal of verisimilitude. *Psychol. Methods* 7:283–300

Micceri T. 1989. The unicorn, the normal curve, and other improbable creatures. *Psychol. Bull.* 105:156–66

Monroe SM, Simons AD. 1991. Diathesis-stress theories in the context of life stress

research: implications for depressive disorder. *Psychol. Bull.* 110:406–25

Moosbrugger H, Schermelleh-Engel K, Klein A. 1997. Methodological problems of estimating latent interaction effects. *Methods Psychol. Res. Online* 2:95–111

Muthén BO. 1989. Latent variable modeling in heterogeneous populations. *Psychometrika* 54:557–85

Muthén BO. 1993. Goodness of fit with categorical and other nonnormal variables. See Bollen & Long 1993, pp. 205–34

Muthén BO. 1994. Multilevel covariance structure analysis. *Sociol. Methods Res.* 22:376–98

Muthén BO. 2001. Second-generation structural equation modeling with a combination of categorical and continuous latent variables: new opportunities for latent class/latent growth modeling. In *New Methods for the Analysis of Change*, ed. LM Collins, A Sayer, pp. 291–322. Washington, DC: Am. Psychol. Assoc.

Muthén BO. 2002. Beyond SEM: general latent variable modeling. *Behaviormetrika* 29:81–117

Muthen BO, Curran PJ. 1997. General longitudinal modeling of individual differences in experimental designs: a latent variable framework for analysis and power estimation. *Psychol. Methods* 2:371–402

Muthén BO, Kaplan D. 1985. A comparison of some methodologies for the factor analysis of non-normal Likert variables. *Br. J. Math. Stat. Psychol.* 38:171–89

Muthén BO, 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. *Br. J. Math. Stat. Psychol.* 45:19–30

Muthén BO, Kaplan D, Hollis M. 1987. On structural equation modeling with data that are not missing completely at random. *Psychometrika* 52:431–62

Muthén LK, Muthén BO. 2002. How to use a Monte Carlo study to decide on sample size and determine power. *Struct. Equ. Model.* 9:599–620

Muthén LK, Muthén BO. 1998–2004. *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén. 3rd ed.

Neale MC. 2000. Individual fit, heterogeneity, and missing data in multigroup structural equation modeling. See Little et al. 2000, pp. 249–67

Neale MC, Boker SM, Xie G, Maes HH. 2003. *Mx: Statistical Modeling.* Richmond, VA: VCU Dept. Psychiatry

Neale MC, Cardon LR. 1992. *Methodology for Genetic Studies of Twins and Families.* London: Kluwer

Neale MC, Miller MB. 1997. The use of likelihood-based confidence intervals in genetic models. *Behav. Genet.* 27:113–20

Nelder JA, Wedderburn RWM. 1972. Generalized linear models. *J. R. Stat. Soc. Ser. A* 135:370–84

Nevitt J, Hancock GR. 2001. Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Struct. Equ. Model.* 8:353–77

Paxton P, Curran PJ, Bollen KA, Kirby J, Chen F. 2001. Monte Carlo experiments: design and implementation. *Struct. Equ. Model.* 8:288–312

Ping RA. 1996. Latent variable interaction and quadratic effect estimation: a two-step technique using structural equation analysis. *Psychol. Bull.* 119:166–75

Plomin R, DeFries JC, McClearn GE, eds. 2001. *Behavioral Genetics.* New York: Worth. 3rd ed.

Raudenbush SW, Bryk AS. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Thousand Oaks, CA: Sage. 2nd ed.

Raudenbush SW, Bryk AS, Cheong YF, Congdon RT. 2000. *HLM5: Hierarchical Linear and Nonlinear Modeling*. Lincolnwood, IL: Sci. Software Int.

Raykov T. 2000. On sensitivity of structural equation modeling to latent variable misspecifications. *Struct. Equ. Model.* 7:596–607

Reichardt CS. 2002. The priority of

just-identified recursive models. *Psychol. Methods* 7:307–15

Reise SP, Widaman KF. 1999. Assessing the fit of measurement models at the individual level: a comparison of item response theory and covariance structure approaches. *Psychol. Methods* 4:3–21

Roberts S, Pashler H. 2000. How persuasive is good fit: a comment on theory testing. *Psychol. Rev.* 107:358–67

Rovine MJ, Molenaar PCM. 1998. A LISREL model for the analysis of repeated measures with a patterned covariance matrix. *Struct. Equ. Model.* 5(4):318–43

Rubin DB. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley

Satorra A, Bentler PM. 1986. Some robustness properties of goodness of fit statistics in covariance structure analysis. *ASA Proc. Bus. Econ. Stat. Sect.*, pp. 549–54. Alexandria, VA: Am. Stat. Assoc.

Satorra A, Bentler PM. 1988. Scaling corrections for chi-square statistics in covariance structure analysis. *ASA Proc. Bus. Econ. Stat. Sect.*, pp. 308–13. Alexandria, VA: Am. Stat. Assoc.

Satorra A, Bentler PM. 1994. Corrections to test statistics and standard errors in covariance structure analysis. See von Eye & Clogg 1994, pp. 399–419

Satorra A, Bentler PM. 2001. A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika* 66:507–14

Schafer JL. 1997. *Analysis of Incomplete Multivariate Data.* London: Chapman & Hall

Schafer JL. 1999. *NORM: multiple imputation of incomplete multivariate data under a normal model* [computer software]. University Park: Penn. State Univ. Dept. Stat.

Schafer JL, Graham JW. 2002. Missing data: our view of the state of the art. *Psychol. Methods* 7:147–77

Schafer JL, Olsen MK. 1998. Multiple imputation for multivariate missing-data problems: a data-analyst's perspective. *Multivar. Behav. Res.* 33:545–71

Scheines R, Spirtes P, Glymour C, Meek C. 1994. *TETRAD II: Tools for Discovery*. Hillsdale, NJ: Erlbaum

Schumaker RE, Marcoulides GA. 1998. *Interaction and Nonlinear Effects in Structural Equation Modeling.* Mahwah, NJ: Erlbaum

Self SG, Liang HY. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.* 82:606–10

Sher KJ, Wood MD, Wood PK, Raskin G. 1996. Alcohol outcome expectancies and alcohol use: a latent variable cross-lagged panel study. *J. Abnorm. Psychol.* 105:561–74

Shipley B. 2000. *Cause and Correlation in Biology*. Cambridge, UK: Cambridge Univ. Press

Sinharay S, Stern HS, Russell D. 2001. The use of multiple imputation for the analysis of missing data. *Psychol. Methods* 6:317–29

Skrondal A, Rabe-Hesketh S. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Boca Raton, FL: Chapman & Hall

Slutske WS, Heath AC, Dinwiddie SH, Madden PAF, Bucholz KK, et al. 1998. Common genetic risk factors for conduct disorder and alcohol dependence. *J. Abnorm. Psychol.* 107:363–74

Snijders T, Bosker R. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Thousand Oaks, CA: Sage

Steiger JH. 1995. Structural equation modeling (SEPATH). In *Statistica 5, Vol. III*, pp. 3539–688. Tulsa, OK: Statsoft

Steiger JH. 2001. Driving fast in reverse: the relationship between software development, theory, and education in structural equation modeling. *J. Am. Stat. Assoc.* 96:331–38

Steiger JH. 2002. When constraints interact: a caution about reference variables, identification constraints, and scale dependencies in structural equation modeling. *Psychol. Methods* 7:210–27

Steiger JH, Fouladi RT. 1997. Noncentrality interval estimation and the evaluation of statistical models. In *What If There Were No*

Significance Tests, ed. L Harlow, S Mulaik, JH Steiger, pp. 221–57. Mahwah, NJ: Erlbaum

Stelzl I. 1986. Changing a causal hypothesis without changing the fit: some rules for generating equivalent path models. *Multivar. Behav. Res.* 21:309–31

Tomarken AJ, Baker TB. 2003. Introduction to the special section on structural equation modeling. *J. Abnorm. Psychol.* 112:523–25

Tomarken AJ, Waller NG. 2003. Potential problems with well fitting models. *J. Abnorm. Psychol.* 112:578–98

Trull TJ. 2001. Structural relations between borderline personality disorder features and putative etiological correlates. *J. Abnorm. Psychol.* 110:471–81

Tucker LR, Lewis C. 1973. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika* 38:1–10

Tukey JW. 1980. We need both exploratory and confirmatory. *Am. Stat.* 34:23–25

van Lier PAC, Verhulst FC, Crijnen AAM. 2003. Screening for disruptive behavior syndromes in children: the application of latent class analyses and implications for prevention programs. *J. Consult. Clin. Psychol.* 71:353–63

von Eye A, Clogg CC, eds. 1994. *Latent Variable Analysis: Applications to Developmental Research.* Thousand Oaks, CA: Sage

Wall MW, Amemiya Y. 2000. Estimation for polynomial structural equation models. *J. Am. Stat. Assoc.* 95:929–40

Wall MW, Amemiya Y. 2001. Generalized appended product indicator procedure for nonlinear structural analysis. *J. Educ. Behav. Stat.* 26:1–29

Waller NG, Meehl PE. 1998. *Multivariate Taxometric Procedure: Distinguishing Types from Continua.* Thousand Oaks, CA: Sage

Waller NG, Meehl PE. 2002. Risky tests, verisimilitude, and path analysis. *Psychol. Methods* 7:323–37

Willett JB, Sayer AG. 1994. Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychol. Bull.* 116:363–81

Willett JB, Sayer AG. 1996. Cross-domain analyses of change over time: combining growth modeling and covariance structure analysis. See Marcoulides & Schumacker 1996, pp. 125–57

Wills TA, Sandy JM, Yaeger AM. 2002. Moderators of the relation between substance use level and problems: test of a self-regulation model in middle adolescence. *J. Abnorm. Psychol.* 111:3–21

Wills TA, Sandy JM, Yaeger AM, Cleary SD, Shinar O. 2001. Coping dimensions, life stress, and adolescent substance use: a latent growth analysis. *J. Abnorm. Psychol.* 110:309–23

Wothke W. 2000. Longitudinal and multigroup modeling with missing data. See Little et al. 2000, pp. 219–40

Yuan K-H, Bentler PM. 1998a. Normal theory based test statistics in structural equation modeling. *Br. J. Math. Stat. Psychol.* 51:289–309

Yuan K-H, Bentler PM. 1998b. Robust mean and covariance structure analysis. *Br. J. Math. Stat. Psychol.* 51:63–88

Yuan K-H, Bentler PM. 1999. F-tests for mean and covariance structure analysis. *J. Educ. Behav. Stat.* 24:225–43

Yuan K-H, Bentler PM. 2000. Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. In *Sociological Methodology*, ed. ME Sobel, MP Becker, 30:165–200. Washington, DC: Am. Sociol. Assoc.

Yuan K-H, Bentler PM. 2001. Effect of outliers on estimators and tests in covariance structure analysis. *Br. J. Math. Stat. Psychol.* 54:161–75

Yuan K-H, Hayashi K. 2003. Bootstrap approach to inference and power analysis based on three statistics for covariance structure models. *Br. J. Math. Stat. Psychol.* 56:93–110

Yung Y-F, Bentler PM. 1996. Bootstrapping techniques in analysis of mean and covariance structures. See Marcoulides & Schumacker 1996, pp. 125–57

Zubin J, Spring B. 1977. Vulnerability: a new view of schizophrenia. *J. Abnorm. Psychol.* 86:103–26

Annual Review of Clinical Psychology
Volume 1, 2005

CONTENTS