Annu. Rev. Psychol. 2000. 51:201–226

APPLICATIONS OF STRUCTURAL EQUATION MODELING IN PSYCHOLOGICAL RESEARCH

Robert C. MacCallum and James T. Austin
The Ohio State University, Department of Psychology, 142 Townshend Hall, Columbus, Ohio 43210–1222; e-mail: maccallum.1@osu.edu, austin.38@osu.edu

Key Words structural equation models, covariance structures, LISREL

■ **Abstract** This chapter presents a review of applications of structural equation modeling (SEM) published in psychological research journals in recent years. We focus first on the variety of research designs and substantive issues to which SEM can be applied productively. We then discuss a number of methodological problems and issues of concern that characterize some of this literature. Although it is clear that SEM is a powerful tool that is being used to great benefit in psychological research, it is also clear that the applied SEM literature is characterized by some chronic problems and that this literature can be considerably improved by greater attention to these issues.

CONTENTS

0084–6570/00/0201–0201$12.00

INTRODUCTION

The development of structural equation modeling (SEM) methods and software has proceeded rapidly since the 1970s. A historical review of the development of SEM was provided by Bentler (1986), and annotated bibliographies of the technical SEM literature have been provided by Austin & Wolfle (1991) and Austin & Calderón (1996). Methodological developments also have been reviewed (Bentler 1980, Bentler & Dudgeon 1996). Software development, with respect to both implementation of methodological advances as well as improvement of user interfaces, has made this sophisticated and powerful method readily accessible to substantive researchers, and these researchers have found SEM to be well-suited to addressing a variety of questions arising in psychological research. This combination of methodological advances, improved software, and applicability to problems of interest has resulted in wide usage of SEM in psychology and related disciplines (Tremblay & Gardner 1996), which has in turn yielded advances in substantive knowledge. At the same time, this ease of access and application of such a complex and sophisticated technique has given rise to a variety of problems and chronic misuses and oversights in practice.

In this review we assess positive and negative aspects of the applied SEM literature in psychology. We describe a range of issues and designs to which SEM has been applied productively. We also discuss a number of problematic issues in this literature, ranging from concerns regarding perspective, design, and strategy to chronic errors in conducting analyses and presenting and interpreting results. Our goal is to promote improved usage of SEM by increasing awareness of the range of research questions that can be addressed using this method and by helping users avoid some common pitfalls.

OVERVIEW OF SEM

To provide a basis for subsequent discussion, we present a brief overview of SEM along with some important special cases and extensions of the usual model. SEM is a technique used for specifying and estimating models of linear relationships among variables. Variables in a model may include both measured variables (MVs) and latent variables (LVs). LVs are hypothetical constructs that cannot be directly measured. In SEM each such construct is typically represented by multiple MVs that serve as indicators of the construct. A structural equation model, then, is a hypothesized pattern of directional and nondirectional linear relationships among a set of MVs and LVs. Directional relationships imply some sort of directional influence of one variable on another. Nondirectional relationships are correlational and imply no directed influence. In the most common form of SEM, the purpose of the model is to account for variation and covariation of the MVs.

Some special cases of SEM are of particular interest and are commonly used in practice. For example, path analysis provides for the testing of models of relationships among MVs; no LVs are included in the model. Path analysis models are used when MVs are of primary interest or when multiple indicators of LVs are not available. Factor analysis, another special case, provides for testing models of relationships between LVs, which are common factors, and MVs, which are indicators of common factors. The factor analysis model allows for correlational (nondirectional) relationships among LVs but does not include directional influences as in general SEM.

The usual form of SEM can be extended to great advantage in a variety of ways. We focus here on two such extensions that are well developed, easily accessible, and probably underutilized in practice. These extensions have been described in various SEM texts (e.g. Bollen 1989, Kline 1998, Loehlin 1998, Maruyama 1998) as well as in the LISREL software manual (Jöreskog & Sörbom 1996). The first such extension is multisample modeling, wherein a model is fit simultaneously to sample data from different populations. A key aspect of this approach involves the testing of invariance of critical parameters across groups. Similarities in the covariance structure of the MVs across groups may be attributable to parameters that are invariant across groups, and differences attributable to parameters that vary across groups. A second extension involves the modeling of MV means as well as variances and covariances. This is especially useful in multisample models where, for example, group differences in means on MVs could be accounted for as a function of group differences in means on LVs. Modeling of means is also used extensively in analysis of repeated measures data. In this context, a model may account for variances and covariances among repeated measures of a MV as well as for change in mean level of the MV over time. By using one or both of these extensions in various ways and contexts, investigators have developed an array of powerful and elegant models applicable to particular research designs. Many of these models are described below.

There exist other extensions and generalizations of SEM that undoubtedly will eventually become widely used but that are not as yet fully developed or accessible. One involves the incorporation of interactions and other nonlinear effects among LVs into models. Early developments in this area (e.g. Kenny & Judd 1984) have proved difficult to implement in practice, and there is a clear need for better methods. Schumacker & Marcoulides (1998) reviewed and illustrated recent efforts to extend SEM in this way. Another developing extension involves the application of SEM to multilevel data structures, where units at one level of observation are nested within units at another (e.g. students within schools), and variables may be measured at each level. Some structural equation models for such data structures have been discussed by Muthén (1994, 1997) and McDonald (1994).

LITERATURE REVIEW

Previous Reviews of Applications

Several within-domain reviews of published applications of SEM have appeared. All limit themselves to four to six journals over variable time periods (6–18 years). Breckler (1990) reviewed 72 applications published between 1977 and 1987 in major journals in personality and social psychology. He presented problematic issues, guidelines for reporting, and an analysis of a typical application. Two evaluative reviews of applications have appeared in marketing journals. Hulland et al (1996) identified 186 applications published between 1980 and 1994, whereas Baumgartner & Homburg (1996) located 149 applications published between 1977 and 1994. Hulland et al coded models on dimensions of theory, measurement, evaluation, and descriptive adequacy, then attempted to replicate published results for a subset of 112 models; successful replication was achieved in only 75% of these cases. Baumgartner & Homburg (1996) coded articles in terms of specification, data screening, and model evaluation. Two reviews of applications of SEM in organizational research have appeared. James & James (1989) coded 55 applications (1978–1987) separately for manifest variable (72%) and latent variable (28%) designs. They found that most a priori models were disconfirmed and that specification searches were routinely undertaken without checking for assumptions. Similarly, Medsker et al (1994) coded 28 applications in five major journals (1988–1993). They described a variety of usage trends by comparing their sample to that of James & James.

Current Review

We reviewed approximately 500 published applications of SEM appearing in 16 psychological research journals from 1993 to 1997. The journals, grouped by domain, were *Journal of Applied Psychology, Journal of Vocational Behavior, Organizational Behavior & Human Decision Processes,* and *Personnel Psychology* (Industrial-Organizational); *Journal of Abnormal Psychology, Journal of Consulting and Clinical Psychology, Journal of Counseling Psychology,* and *Psychological Assessment* (Counseling-Clinical); *Psychology and Aging, Child Development,* and *Developmental Psychology* (Developmental); *Journal of Personality, Journal of Personality and Social Psychology, Journal of Research in Personality,* and *Personality and Social Psychology Bulletin* (Social-Personality). We also included *Multivariate Behavioral Research* and *Structural Equation Modeling.*Our collection of studies included those using the full SEM framework as well as those using special cases of SEM, such as path analysis and confirmatory (not exploratory) factor analysis.

USES OF SEM IN PSYCHOLOGICAL RESEARCH

In this section we organize, describe, and cite examples of usage of SEM in substantive research in psychology. By necessity our categories are fuzzy and overlapping and by no means exhaustive. Furthermore, although we have endeavored to identify and cite sound exemplars of various uses of SEM, virtually any such application might be critiqued on specific aspects of design, specification, or analysis. From this perspective, we begin with some comments about research design. In general, psychological research investigating relationships among variables can be categorized as either observational (correlational) or experimental. SEM is by far more heavily used in observational studies. In fact, there seems to be a common misconception that its use is restricted to that category. To the contrary, SEM can be used in experimental studies to great advantage, as we discuss below.

Within the realm of observational studies, designs can be broken down roughly into two categories: cross-sectional and longitudinal. SEM is heavily used for both kinds of designs, with applications to cross-sectional designs being more common in social/personality and industrial/organizational psychology, and applications to longitudinal designs being more common in studies of development and aging. We consider applications to these two designs in turn.

Cross-Sectional Designs

As the name suggests, a cross-sectional design is a single-occasion snapshot of a system of variables and constructs. Its key feature is the concurrent measurement of variables. The use of SEM in cross-sectional designs is common, with applications to manifest variable, latent variable, or measurement studies. Multisample models and models with structured means are often used. For example, Rice et al (1997) used a cross-sectional design with multiple groups to compare African-American and Caucasian, as well as male and female, adolescents to investigate predictors of emotional well-being. Judge & Locke (1993) used reports from individuals and significant others to estimate the effects of dysfunctional thought processes on subjective well-being and job satisfaction.

A notable feature of many models used in cross-sectional studies is the specification of directional influences among variables. The perspective that directional influences require some finite amount of time to operate suggests that interpretation of such effects in cross-sectional designs may be problematic because concurrent measurement of variables precludes such effects from occurring (Gollob & Reichardt 1987, 1991). Nevertheless, numerous cross-sectional studies that we reviewed framed models as involving noninstantaneous processes or change over time. This potential conflict between design and model is discussed in more detail below.

Longitudinal Designs

There are two sorts of longitudinal designs to which SEM is often applied. Both involve measurements obtained from the same individuals on repeated occasions. In one type of longitudinal design, which we call a sequential design, different variables are measured at successive occasions and the model specifies effects of variables at a given occasion on other variables at later occasions. The researcher is interested in the pattern of influences operating over time among different variables. The sequence and timing of measurements are designed to allow for these hypothesized effects to operate. In another type of longitudinal design, a repeated measures design, the same variable or variables are measured at each occasion. Here the researcher is interested in relationships among the repeated measures of the same variables as well as the pattern of change over time. Occasions of measurement are chosen to represent the range of time and intervals during which change is of interest. Of course these two kinds of designs are not mutually exclusive; many studies incorporate elements of both.

In applications of SEM to sequential designs, directional influences in a model are hypothesized as operating over some time interval, and fitting the model to observed data yields estimates of such effects. The interpretation of such effects is bolstered by the use of a design that allows appropriate time for the effects to occur. Examples include a study by Holohan et al (1997) of effects of social support and social stressors on subsequent depressive symptoms, a study by Gest (1997) of relationships between behavioral inhibition and peer relations over time, and a study by Gold et al (1995) of stability and predictors of intellectual abilities measured 40 years apart. An important aspect of such designs and models is the desirability of including autoregressive influences. That is, if one hypothesizes that variable A at time 1 (A_1) influences B at time 2 (B_2), one should also measure B_1 and include in the model the influence of B_1 on B_2 as well as the correlation of A_1 with B_1. Failure to do so can result in a highly biased estimate of the effect of A_1 on B_2 (Gollob & Reichardt 1991). That is, one might conclude that there is a strong influence of A_1 on B_2 when in fact that apparent influence is in part spurious and due to the autoregressive influence of B_1 on B_2 and the correlation of B_1 with A_1.

Turning to repeated measures designs, the application of SEM to such designs represents an area of highly creative development of models and novel applications in recent years. There are two general frameworks for model specification in such designs. In both approaches it is possible and often desirable to include means in the model, so that the model accounts for changes in the mean of the MV over time, as well as for variances and covariances of repeated measures of the MV. One general approach involves the use of autoregressive models (Jöreskog 1979, McArdle & Aber 1990). Given measures of X at occasions $1, 2, \ldots, t, \ldots, T$, such a model specifies a series of autoregressive effects wherein each X_t influences X_{t+1}. Such models are often called simplex models because they account for a simplex correlational structure, wherein the correlation between repeated measures decreases as the lag between the measures increases. The basic autoregressive model allows for many extensions and variations, many of which are described by Jöreskog (1979) and McArdle & Aber (1990). These include modeling of effects over various time lags, modeling of the structure of correlations among error terms, and evaluating whether parameters of interest are stable across occasions or intervals. One can also model autoregressive effects for LVs, where a given LV is represented by multiple indicators at each occasion. Examples of the use of autoregressive models in repeated measures designs include a study of drug abuse by Aiken et al (1994) and a study of anxiety in children by Lopez

& Little (1996). Both of these two-wave studies used multisample models with structured means. Examples of modeling data with more than two waves can be found in McArdle & Aber (1990) and Curran et al (1997b).

A second general class of models for repeated measures designs is latent curve models, often called growth curve models. Originating in early work by Tucker (1958) and Rao (1958), latent curve models represent a rich domain of methodological development in recent years (Collins & Horn 1991, Sayer & Collins 1999) and have been used in many elegant applications. A technical overview is provided by Meredith & Tisak (1990); less technical discussions are offered by Willett & Sayer (1994), Duncan et al (1994), and Duncan et al (1999). The objective of such models is to capture information about interindividual differences in intraindividual change (Nesselroade 1991). In their simplest forms, latent curve models can be viewed as a type of confirmatory factor analysis model, where the variances, covariances, and means of repeated measures of a MV are modeled. Factors, or latent curves, represent aspects of change over time, such as level, linear change, or acceleration. According to the model, each individual's vector of repeated measures on the MV is represented as a linear combination of these latent curves. Jones & Meredith (1996) presented an application of this model in a study of personality change across 30–40 years.

There is an array of important extensions of this model, many of which have been discussed and illustrated by McArdle & Aber (1990) and Duncan et al (1999). One can introduce additional variables representing predictors, correlates, or consequences of aspects of change. Such a model would involve an extension of the latent curve model into a full SEM model, specifying hypothesized associations between the latent curves and other LVs or MVs. Willett & Sayer (1994) described and illustrated this approach in a study using gender and exposure to deviant behavior to predict change in deviant behavior during adolescence. Raykov (1994) also used this method to study correlates of change in factors of fluid intelligence. Another extension involves investigating change on more than one MV simultaneously. In such multiple-outcome models, latent curves are specified for each MV, and one can then investigate relationships between latent curves representing different outcome variables. Such models have been discussed by MacCallum et al (1997), Tisak & Meredith (1990), and Willett & Sayer (1995). Curran et al (1997a,b) and Stoolmiller (1994) used this approach in empirical studies. Latent curve models can also be used in multisample designs, where repeated measures of a variable of interest are collected from samples of individuals from distinct populations. Such a design and model allow for the study of group differences in aspects of change. Muthén & Curran (1997) have proposed the use of such multisample latent curve models to analyze data from intervention studies. Given control and experimental groups, repeated measures of a dependent variable are obtained. The groups differ in that some treatment or intervention is applied in the experimental group after the baseline measure is observed. A latent curve model is specified to model change over time in the control group; the same latent curves are specified for the experimental group along with an additional latent curve to represent differential change after treatment. This approach had been used in an earlier application by Raykov (1995) to investigate the effect of a cognitive intervention on change in fluid intelligence during aging. Multisample latent curve models are also useful in sequential cohort designs in aging and developmental research. Such designs avoid the problems associated with obtaining repeated observations on a single sample for an extended period of time. Rather, multiple cohorts are followed for shorter periods of time, where cohorts differ with respect to age at the first occasion of measurement. Duncan et al (1996) described and illustrated the modeling of such data.

Measurement Studies

Confirmatory factor analysis (CFA) models, a special case of SEM, are widely used in measurement applications for a variety of purposes. Designs for construct validation and scale refinement, multitrait-multimethod validation, and measurement invariance can be evaluated through specification and testing of CFA models. We discuss each category in turn.

Construct Validation and Scale Refinement Development of a measure of a construct requires validation of the hypothesized relationship between the construct and its indicators. The indicators are typically single items or composites consisting of multiple items. One or more models of this relationship are often evaluated using CFA. Methods for developing and validating multi-item assessment scales using factor analysis have been described by Floyd & Widaman (1995). Jackson et al (1993) used such methods to develop five scales for assessing job control, cognitive demand, and production responsibility. They tested a series of nested measurement models using two samples of shop employees, confirming a hypothesized five-factor model as well as invariance for four of the five factors across their samples. Novy et al (1994) used CFA to evaluate a theoretical model of ego development. Neuberg et al (1997) illustrated how CFA can be used to critique an existing scale, specifying and supporting a two-dimensional interpretation for a scale of need for closure that had been thought to be unidimensional. Similarly, some CFA analyses of personality inventories (e.g. Church & Burke 1994, Parker et al 1993) have yielded poor support for the popular five-factor theory of personality. McCrae et al (1996) criticized usage of CFA to validate such dimensional theories and recommended instead the use of exploratory factor analysis with target rotation. In our view, the phenomenon underlying such controversies involves the lack of correspondence between the highly structured nature of a confirmatory factor model and the somewhat looser criteria used to develop assessment instruments. Floyd & Widaman (1995) have discussed some aspects of this issue.

If construct validity is supported by confirmation of a hypothesized dimensional structure, other types of scale refinement or assessment may be considered. These include the development of short forms, as done by Donders (1997) with the WISC-III, or the assessment of construct validity in a different context, as done by Meyer et al (1993) in extending their measure of the construct of commitment from the context of organizations to the context of occupations. On the other hand, if CFA results do not support construct validity, it is possible to use results to revise the scale for future reevaluation. For example, items with a complex loading structure or high unique variance might be omitted, or additional items might be constructed to better represent a LV. Walker et al (1997) used this approach in refining a pain-response scale for children.

Multitrait-Multimethod Studies Multitrait-multimethod (MTMM) studies represent a specific design for construct validation, in which multiple traits are measured by multiple methods (Campbell & Fiske 1959). Models for such matrices represent trait and method factors (LVs) as combining either additively (Widaman 1985) or multiplicatively (Cudeck 1988) to yield observed correlations among MVs. Reichardt & Coleman (1995), Kenny & Kashy (1992), and Widaman (1985) showed how evidence for convergent and discriminant validity can be obtained through the comparative evaluation of alternative additive models. Brannick & Spector (1990) demonstrated problems with fitting the additive formulation. They described common identification and specification problems and reanalyzed data from 18 published MTMM studies, finding estimation problems in 17 of 18. Applications of the additive model for MTMM designs typically encounter problems of empirical indeterminacy unless restrictions such as equality constraints are imposed on model parameters. An application by Coovert et al (1997) compared the additive and multiplicative models for performance rating data across seven military jobs and concluded that the multiplicative formulation performed better.

Measurement Invariance When an investigator wishes to use a given measure or set of measures to make comparisons across populations or across time, the validity of those comparisons depends on the assumption that the same construct is being measured in different groups or at different occasions. This assumption of measurement invariance can be tested using CFA (Meredith 1993), usually including mean structures so as to allow for comparison of latent variable means across groups or occasions. The assessment of measurement invariance across groups involves the use of multisample CFA. Methods have been described and illustrated by Widaman & Reise (1997) and Reise et al (1993). Horn & McArdle (1992) and Tisak & Meredith (1990) described and illustrated methods for studying measurement invariance over time. In either context, one is interested in testing the invariance of parameters of the factor model, including factor means, across groups or time. Studies of measurement invariance are common in the literature. Little (1997) investigated invariance of factors of control expectancy across gender and four cultural groups. Kim et al (1996) studied invariance of world views and religious beliefs of older adults over time. Pentz & Chou (1994) described and illustrated the study of measurement invariance across both groups and time simultaneously in a study of adolescent drug abuse. Millsap (1997) extended research on invariance by examining the relationship between measurement invariance and predictive bias. Given a predictor and criterion set measured in multiple groups, he found that measurement invariance ordinarily implies predictive bias (group differences in regression coefficients), and that an absence of predictive bias ordinarily implies an absence of measurement invariance. Millsap showed how these conditions can be investigated using simple structural equation models.

Experimental Studies

We return now to the issue of observational versus experimental designs. Most of the applications described to this point involve observational studies. As mentioned earlier, there seems to be a common misconception that SEM is applicable only to observational studies and not to experimental studies. This misconception apparently stems from the fact that models for data gathered in experimental studies would incorporate categorical independent variables, and the inclusion of such variables appears to violate the assumption of multivariate normality underlying the commonly used maximum likelihood method of estimation. However, as clarified by Bollen (1989:126–28), the assumption of multivariate normality need not apply to exogenous measured variables. Rather, this assumption can be stated as referring to all other variables conditional on the levels of exogenous measured variables. Thus, given experimentally manipulated independent variables, one can use SEM to model the relationships of those variables, represented by coded variables such as dummy variables, to other variables, including covariates, mediators, and outcomes. Russell et al (1998) provided a discussion of procedures, issues, and advantages in the use of SEM in experimental studies, along with illustrations. The social cognition literature provides numerous examples of the use of SEM to study mediational effects in experimental designs. For example, Prussia & Kinicki (1996) used SEM to investigate mediators of effects of performance feedback and vicarious experience on group effectiveness, and Hoyle (1993) used SEM to study mediational effects of perceived attitudinal similarity on attraction toward a stranger. Another approach to using SEM in experimental studies is to treat different levels of experimental variables as representing different populations and to conduct multisample SEM, where the model of interest is fit simultaneously to samples representing different experimental conditions. Muthén & Curran (1997) used this approach to fit latent curve models to control and treatment groups. They illustrated this method using data from a study of the effectiveness of an intervention intended to reduce aggressive behavior in school children. In general, the application of SEM in experimental studies represents a significant but relatively untapped potential area of application. The conceptual boundary that is usually defined between observational and experimental research is probably far too rigid. SEM clearly cuts across that boundary.

Twin Studies

In recent years SEM has become a valuable tool in twin and family designs to model genetic and environmental influences on variables of interest (Loehlin 1989, Plomin & Rende 1991). In the most basic model, relevant data consist of measures on an observed variable collected from samples of siblings of different levels of relatedness (monozygotic twins, dizygotic twins, nontwin siblings, etc). Genetic and environmental influences are then investigated using a form of multisample confirmatory factor analysis, in which groups are defined according to relatedness, and factors represent shared or nonshared genetic and environmental influences. Results provide estimates of the impact of each source of variation and covariation. Methodological details have been provided by Loehlin (1998), Neale & Cardon (1992), McArdle & Prescott (1997), and McArdle et al (1998). There are many applications in the literature, often including other capabilities of SEM described above. Examples include a study of temperament (Saudino et al 1995), a study of memory performance including a multisample analysis of Swedish and American twin data (Finkel et al 1995), and a study of genetic and environmental components of latent curves representing aspects of change in alcohol consumption (Prescott & Kendler 1996).

Other Areas of Application

We reemphasize that our discussion of types of applications of SEM is by no means exhaustive, and that many applications involve aspects of several different categories described above. In addition, many applications and special cases address problems that fall outside the domain we have described. Such methods include, for instance, the use of SEM in meta-analysis (Rounds & Tracey 1993) and test-retest designs (McArdle & Woodcock 1997), the use of multisample models to study moderator effects (e.g. Eisenberg et al 1997, Harnish et al 1995), and the longitudinal study of reliability and validity (Tisak & Tisak 1996).

PROBLEMATIC ISSUES IN APPLICATIONS OF SEM

We now turn to methodological problems and issues of concern that emerged in our review of published applications. We consider several global concerns before turning to problems involving details of analysis, interpretation, and presentation of results.

Generalizability of Findings

In our view, much of the applied SEM literature is characterized by inadequate understanding or acknowledgment of the limitations of single studies. Even in a well-designed study where analyses are conducted properly, conclusions may be limited to the particular sample, variables, and time frame represented by the design. Such limitations are seldom acknowledged, let alone addressed as a topic of research. Rather, results are usually interpreted as if substantial generalizability exists.

More specifically, researchers using SEM must recognize that results are subject to sampling or selection effects with respect to at least three aspects of a study: individuals, measures, and occasions (Nesselroade 1991). Most researchers are familiar with the notion of sampling effects with respect to individuals. Such effects are taken into account via sampling procedures and the use of inferential statistics. Another mechanism for taking such effects into account in SEM is cross-validation (Browne & Cudeck 1989, Cudeck & Browne 1983). Researchers can use the expected cross-validation index (ECVI) (Browne & Cudeck 1989, 1993), which is computed from a single sample, as an index of how well a solution obtained in one sample is likely to fit an independent sample. This index is useful for comparison of alternative models, especially when sample size is not large, providing an indication of which model yields a solution with greatest generalizability.

Another aspect of selection effects at the level of observations involves the population of interest. A structural equation model is a hypothesis about the structure of relationships among MVs in a specific population. Researchers should explicitly define the population of interest, although this is often not done in practice, and should acknowledge that the generalizability of a model beyond that population may be uncertain. Of course, many studies do address this issue explicitly, especially those that involve multisample SEM, where the focus of the study is the evaluation of model fit and parameter estimates across samples from distinct populations (e.g. Aiken et al 1994, Lindenberger & Baltes 1997).

Selection effects are also inherent in the choice of measured variables used in a given study. In SEM this issue is especially prominent with regard to the choice of indicators to represent latent variables. In any given study, a particular latent variable is effectively defined as that which its indicators have in common. The nature of the construct can shift with the choice of indicators, which in turn can influence results and interpretation. Clearly, valid results and interpretation depend on having appropriate operationalizations of the latent variables under study. Little et al (1999) discussed and demonstrated selection effects involving indicators and offered guidelines for selection of good indicators.

An additional aspect of selection effects involves occasions of measurement. In any study where one investigates effects that operate over time, those effects may vary with the length of the time interval (Gollob & Reichardt 1987, 1991). From this perspective, there is no single true effect of one variable on another, unless the variables themselves do not change over any time period of interest. In addition to acknowledging this point explicitly, researchers could conduct studies of effects of interest over different time lags to more fully understand the nature of such effects.

In general, we urge researchers to be more aware of the limitations of single studies with respect to these issues. Such an awareness could manifest itself via explicit acknowledgments, qualified interpretations, and studies that investigate generalizability of findings with respect to these issues.

Confirmation Bias

Results of our review suggested that researchers using SEM are quite susceptible to a confirmation bias (Greenwald et al 1986), i.e. a prejudice in favor of the model being evaluated. We see two symptoms of this bias: (*a*) a moderately frequent, overly positive evaluation of model fit, to be discussed below; and (*b*) a routine reluctance to consider alternative explanations of data. Reichardt (1992) discussed the fallibility of judgments about models, noting that it is easy to accept an explanation that fits our data well, and that researchers are not then motivated to consider alternatives. This is especially problematic in SEM. Given a study where support is obtained for a specific model, other models that would fit the data equally well, or approximately so, routinely exist. It is important for researchers to consider this possibility and to employ strategies and methods that provide for examination of alternatives. One such approach is to specify and evaluate multiple a priori models. Although many studies use this strategy (about 50% in our review), most such studies investigate a series of nested models, where parameters of one model in the series are a subset of those in the next model. Although this approach helps to indicate which parameters are needed to account for observed data, it does not address the possible existence of qualitatively distinct models of approximately equal complexity that provide alternative meaningful explanations of data. Millsap & Meredith (1994) provide an illustration of the investigation of qualitatively distinct nonnested models in their study of models of salary discrimination.

An especially interesting phenomenon in the context of alternative models is the existence of equivalent models, which are alternative models that fit any data to the same degree. Such models can be distinguished only in terms of substantive meaning. Lee & Hershberger (1990) provided some simple rules for generating equivalent models from a given model. In a study of published applications of SEM, MacCallum et al (1993) showed that equivalent models occur routinely in practice, often in very large numbers. Investigators are apparently almost universally unaware of this phenomenon, or else they choose to ignore it. We urge researchers to generate and evaluate the substantive meaningfulness of equivalent models in empirical studies. Ruling out their existence or meaningfulness would strengthen the support of a favored model. More generally, any effort to examine alternative models can provide some protection against a confirmation bias and bolster support of a favored model.

The Issue of Time

Gollob & Reichardt (1987, 1991) suggested that directional effects in structural equation models can be considered as causal effects in a loose sense wherein a change in one variable somehow results in a change in another variable, and they described three properties of such effects: (*a*) these effects take some finite amount of time to operate; (*b*) a variable may be influenced by the same variable at an earlier point in time, i.e. autoregressive effects may exist; and (*c*) the magnitude of an effect may vary as a function of the time lag. Our review showed that directional effects are routinely studied using cross-sectional designs, which as pointed out by Gollob & Reichardt (1987, 1991) cannot incorporate any of the three properties just defined. Cross-sectional designs allow only for the evaluation of relationships among variables at one point in time and do not allow for autoregressive effects or time lags. Thus, it may be problematic to infer causality or directional influence in cross-sectional studies. For such an inference to be valid, a researcher must take one of two positions. One option is to argue that the time lag during which the causal influence operates is essentially instantaneous, thereby justifying concurrent measurement of variables in a cross-sectional design. Such situations may not be unusual. For example, Pajares & Miller (1995), in a study of the effect of mathematics self-efficacy on performance on mathematics tests, argued explicitly that the effect of interest was essentially instantaneous, meaning that the variables should be measured as closely together as possible in time. If the time lag of interest is not of extremely short duration, then to justify the study of directional influences in cross-sectional designs the investigator must assume that the causal variables under study do not change over the time period of interest, i.e. between the time the causal effect occurs and the time the causal variable is measured. If this assumption is not valid or if effects of interest are not essentially instantaneous, estimates of directional effects obtained in a cross-sectional design may be highly biased (Gollob & Reichardt 1987, 1991). Given the prevalence in the literature of models incorporating directional effects estimated and tested using data from cross-sectional studies, such biased estimates of effects are probably a chronic problem in the literature.

We suggest that one or the other of these justifications should be stated explicitly for cross-sectional models that include directional influences. In the absence of such justification, a cross-sectional design may be inappropriate and a longitudinal design preferred to allow for representation of the properties described by Gollob & Reichardt (1987, 1991). Furthermore, as noted earlier, longitudinal designs and models should provide for assessment of autoregressive effects. Our review showed that such effects are routinely overlooked in practice, resulting in substantial potential bias in estimates of directional influences (for further discussion of this point and for an application, see Wagner et al 1994). Also, in designing longitudinal studies to evaluate causal effects, one must not forget Cliff's (1983) admonition that a temporal sequence in itself is not sufficient to imply causality. Even given a temporal sequence, an apparent causal influence may be due to an intervening variable or to some correlate of the putative cause, so it is important to make efforts to avoid omitting important variables of this kind. The critical general point is that many studies show inadequate consideration of the issue of time in design, and that greater consideration of this issue would result in improved information about directional influences among variables.

Model Specification, Design, and Analysis Issues

Latent Variable Versus Measured Variable Models Approximately 25% of the studies we reviewed used path analysis models, with no LVs. Many researchers may be unaware of advantages of LV models over MV models. Theories in psychology typically postulate patterns of relationships among LVs. Single MVs generally do not provide adequate representation of such constructs because of imperfect (often modest) reliability and validity. Nevertheless, path analysis models treat single MVs as exact, error-free representations of the constructs of interest. This approach can result in estimates of effects that are highly biased due to the influence of error (Bollen 1989:151–76, Maruyama 1998:79–87). An alternative is to obtain multiple indicators of each LV. A given LV is then defined in effect as whatever its indicators have in common. Little et al (1999) have discussed issues and important phenomena relevant to the selection of indicators. A full LV model then specifies relationships of the indicators to the LVs as well as relationships of the LVs to each other. Such a model allows for estimation of the unique variance in each indicator, and estimates of relationships among LVs are not biased by the presence of error in the indicators.

However, let us consider those situations in which only a single indicator is available for each LV of interest. The relatively common use of path analysis in such settings is of concern because results are susceptible to the biasing effects of error just described. Fortunately, there is a simple method for resolving this problem in many cases. Given a LV represented by a single multiitem scale, one approach to obtaining multiple indicators might be to use each item as an indicator. However, this strategy is often ineffective because of the potentially large number of items, the lack of unidimensionality of the scale, or the relatively low reliability of single items. An alternative approach that is often viable is to construct parcels. A parcel is simply a sum of a subset of items from the scale. Multiple parcels can be defined by aggregating distinct subsets of items, and the parcels then serve as multiple indicators of the given LV. An investigator can thus gain the advantages inherent in full LV models and avoid some of the difficulties associated with MV path analysis models. Our review indicated that such an approach could have been taken in many published studies using path analysis. Methods for constructing parcels have been described by Kishton & Widaman (1994). Examples of the use of parcels can be found in Lent et al (1997) and Lopez & Little (1996).

Sample Size Our review indicated a wide range of sample sizes used in SEM applications, with small-sample studies not being uncommon. About 18% of the studies we reviewed used samples of fewer than 100 individuals. Published applications rarely include explicit consideration of whether the available sample is sufficiently large, and few well-founded guidelines are available. Developments in recent years make it feasible to address this question from a power analysis perspective. MacCallum et al (1996) provided a method for determining the minimum sample size necessary to achieve a given level of power for tests of model fit. As noted by MacCallum et al, however, a minimum sample size determined by power analysis for tests of overall fit is not necessarily adequate for other purposes, such as obtaining sufficiently precise parameter estimates. In any case, lacking a power analysis or other justification, SEM analyses of small samples are almost certainly problematic. We are reluctant to recommend rules of thumb regarding sample size in SEM. Recent work (MacCallum et al 1999) on the question of sample size in factor analysis has shown rules of thumb to be generally invalid in that context, which is a special case of SEM, and that minimum sample size necessary to accurately recover population factor loadings is highly dependent on characteristics such as communality level of the MVs. We expect that similar phenomena operate in the more general context of SEM, and we anticipate further research and improved guidelines regarding the sample size question.

The issue of sample size is also relevant in the context of model evaluation. Hu & Bentler (1998) reviewed literature regarding effects of sample size on fit indexes. Although important effects exist, investigators should not necessarily favor measures of fit that are independent of sample size, especially when comparing alternative models. This issue has been examined carefully by Cudek & Henly (1991). Consider the case where sample size is relatively small and one wishes to compare alternative models of varying complexity, where more complex models have more parameters. As one estimates larger numbers of parameters, one loses some precision in those estimates, more so with smaller samples. A small sample may not be large enough to support the estimation of more complex models. Thus, the model that can be best supported may depend on sample size, with simpler models favored when sample size is smaller. In comparing models when sample size is small, it is advisable to use an index such as ECVI (Browne & Cudeck 1989, 1993), which is sensitive to the phenomenon just described.

Strategy Jöreskog & Sörbom (1996) described three strategies in model specification and evaluation: (*a*) strictly confirmatory, wherein a single a priori model is studied; (*b*) model generation, wherein an initial model is fit to data and then modified as necessary until it fits adequately well; and (*c*) alternative models, wherein multiple a priori models are specified and evaluated. Our review showed approximately the following percentages of studies using each of these strategies: strictly confirmatory, 20%; model generation, 25%; alternative models, 55%. The relatively common usage of the first two strategies is unfortunate. The strictly confirmatory strategy is highly restrictive, requiring the investigator to evaluate a single model in isolation and leaving little recourse if that model does not work well. The model generation strategy is potentially misleading and easily abused. Studies have shown that such data-driven model modifications may lack validity (MacCallum 1986) and are highly susceptible to capitalization on chance (MacCallum et al 1992). Any use of a model generation strategy must be subject to conditions. First, it must be acknowledged that the resulting model is in part data driven; second, modifications must be substantively meaningful; and third, the modified model must be evaluated by fitting it to an independent sample. Although some applications using this strategy meet the first and second conditions, very few meet the third. Fortunately, there are some pleasant exceptions providing cross-validation of modified models (e.g. Belsky et al 1996, Marsiske et al 1997). The alternative models strategy provides an attractive alternative to the other strategies, avoiding the difficulties just described and providing comparative information about alternative explanations of the data, thus lending some protection against a confirmation bias.

Correlation Versus Covariance Matrices In typical applications of SEM, users must decide whether to fit a model to a covariance matrix, S, or a correlation matrix, R. In some designs, such as multisample or repeated measures designs, it is necessary to use covariance matrices so as to retain information about variances of variables. Differences in variances of variables among groups or across time represent important information to be accounted for by a model. This recommendation is followed closely in practice. In most other designs the user may choose to analyze either S or R, but there are interpretational advantages to using R. If latent variables are standardized and the model is fit to R, then parameter estimates can be interpreted in terms of standardized variables. In practice, about 50% of the published applications we reviewed fit models to correlation matrices. Unfortunately, however, most users seem unaware that fitting a model to R versus S introduces a subtle but potentially serious problem (Cudeck 1989, Lawley & Maxwell 1971:100–3). Conventional estimation methods in SEM are based on statistical distribution theory that is appropriate for S but not for R. It is not correct to fit a model to R while treating R as if it were a covariance matrix. As discussed by Cudeck (1989), the consequences of treating R as a covariance matrix depend on properties of the model being fit. In all cases, standard errors of parameter estimates as well as confidence intervals and test statistics for parameter estimates will be incorrect. Correct standard errors will generally be smaller than the incorrect values, resulting in narrower confidence intervals and larger test statistics. In some cases, parameter estimates themselves may be incorrect, and in some cases values of fit indexes will also be incorrect. As noted by Cudeck, these facts have significant implications for published applications of SEM where popular programs are used to fit models to correlation matrices; such studies undoubtedly include some incorrect results.

At the time of this writing (August 1999) various computer programs for SEM deal with this issue in different ways. Currently, two programs automatically provide for correct estimation when analyzing a correlation matrix (RAMONA) Browne & Mels 1998; SEPATH, Steiger 1999). This same facility is expected to be implemented in the next release of EQS (PM Bentler, personal communication). Correct estimation of a correlation structure can be done in LISREL (Jöreskog & Sörbom 1996) and Mx (Neale 1997) but requires the user to introduce specific constraints on parameters. The AMOS program (Arbuckle & Wothke

1999) does not accept correlation matrices for analysis. Given this variation in treatment of this issue and the rapid pace of software development, along with the potential impact of incorrect analyses of correlation structures, we offer a test whereby users can determine whether a specific SEM program provides correct estimation of a model fit to a correlation matrix. Jöreskog & Sörbom (1996) present the widely used stability-of-alienation example (Wheaton et al 1977). Their model B (Figure 6.3, p. 218, excluding error covariances) can be fit to the correlation matrix obtained by rescaling the covariance matrix in their Table 6.5 (p. 216). A check of correctness of results can be made with reference to a single parameter. The correct maximum likelihood estimate of the factor loading of Alienation67 on Powerlessness67 is 0.999. The correct estimate of the standard error for this loading is 0.036. If the correlation matrix is treated incorrectly as a covariance matrix, the resulting incorrect estimate of the standard error is 0.047. If this latter value is obtained, the implication is that the software in question provides incorrect results for analysis of correlation matrices, as described above and as discussed in detail by Cudeck (1989). We urge users fitting models to correlation matrices to be certain that their SEM software treats such matrices correctly. Otherwise, it would be preferable to fit models to covariance matrices, ensuring correct results but sacrificing some ease of interpretation.

Interpretation of Results Our review produced a variety of concerns about assessment of model fit and interpretation of parameter estimates. With respect to model fit, researchers do not seem adequately sensitive to the fundamental reality that there is no true model (Browne & Cudeck 1993, Cudeck & Henly 1991), that all models are wrong to some degree, even in the population, and that the best one can hope for is to identify a parsimonious, substantively meaningful model that fits observed data adequately well. At the same time, one must recognize that there may well be other models that fit the data to approximately the same degree. Given this perspective, it is clear that a finding of good fit does not imply that a model is correct or true, but only plausible. These facts must temper conclusions drawn about good-fitting models.

In addition, a finding of good fit does not imply that effects hypothesized in the model are strong. In fact, it is entirely possible for relationships among variables to be weak, or even zero, and for a hypothesized model to fit extremely well. The resulting parameter estimates would reflect weak relationships among variables and large residual variances for endogenous variables. Thus, it is critically important to pay attention to parameter estimates, even when fit is very good. We found this to be a common oversight in applications. For instance, in nearly 50% of the studies reviewed, researchers failed to report residual variances for endogenous variables. For example, a model might hypothesize an influence of LV *A* on LV *B*. The model would be found to fit well; the researcher would report the estimate of the effect of *A* on *B* but would not report the residual variance in *B*, which may well be substantial. Good fit does not imply at all that such residual variances are small. Such information is critical to a full understanding of the magnitude of effects and should always be reported and discussed.

We observed a wide variety of measures of fit being used, as well as a range of criteria for determining what constitutes good fit. There was little consistency in choice of fit indexes or criteria for their evaluation. We encountered numerous cases of clearly overly positive assessment of fit, with authors stating that fit was acceptable even though values of fit indexes fell well short of any proposed criteria for adequate fit. Researchers using SEM need better guidelines regarding selection and interpretation of fit measures. Some useful information has been provided in recent papers by Hu & Bentler (1998, 1999). In studies of the performance of various fit measures with respect to sensitivity to model misspecification, Hu & Bentler (1998) recommended the use of the standardized root mean square residual (SRMSR) in tandem with one of several other indexes that they see as more or less interchangeable. These include the non-normed fit index (NNFI; Bentler & Bonett 1980) and the root mean square error of approximation (RMSEA; Browne & Cudeck 1993, Steiger & Lind 1980), among others. Significantly, Hu & Bentler (1998) recommended against usage of some common indexes such as the goodness of fit index (GFI) and adjusted goodness of fit index (AGFI), measures that are widely used in the SEM literature. In addition to being insufficiently and inconsistently sensitive to model misspecification, these indexes have also been shown to be strongly influenced by sample size (Marsh et al 1988). Hu & Bentler (1998, 1999) also emphasized that commonly used criteria for interpretation of popular fit indexes have a clear tendency to result in acceptance of poorly specified models. They propose rules of thumb that are more conservative than popular criteria and are also more conservative than criteria recommended in the heavily cited early paper by Bentler & Bonett (1980). We would add one strong recommendation to those of Hu and Bentler. We especially encourage use of RMSEA for several reasons: (*a*) It appears to be adequately sensitive to model misspecification (Hu & Bentler 1998); (*b*) commonly used guidelines for interpretation seem to yield appropriate conclusions about model quality (Hu & Bentler 1998, 1999); and (*c*) most importantly, a confidence interval is available. This confidence interval provides important information about precision of the estimate of fit, which is not available for almost all other fit indexes. We urge the routine use of RMSEA and the reporting and discussion of the associated confidence interval. Finally, we also encourage use of methods for the assessment of model fit at the individual level (Neale 1997, Reise & Widaman 1999). These procedures provide a measure of the contribution of each individual to the overall lack of model fit, allowing for identification of individuals for whom a model fits well or poorly, as well as subsequent analysis of such information to assess possible predictors or correlates of person-fit.

Reporting of Results

In our review we encountered many difficulties associated with presentation of information about models, methods, analyses, and results. For example, in about

10% of the articles reviewed we were unable to determine precisely either the model or the indicators of LVs. In about 25% we could not determine if the model was fit to R or S. In about 50% reporting of parameter estimates was incomplete (omission of nonsignificant estimates, unique variances, and/or residual variances). Many of these issues can be resolved by attention to published guidelines for presenting results of SEM (Hoyle & Panter 1995, Raykov et al 1991). We urge investigators as well as journal editors to adhere to these guidelines and suggest that every application of SEM should provide at least the following information: a clear and complete specification of models and variables, including a clear listing of the indicators of each LV; a clear statement of the type of data analyzed, with presentation of the sample correlation or covariance matrix (or making such data available upon request); specification of the software and method of estimation; and complete results. We define complete results to mean multiple measures of fit as recommended above, with confidence intervals when available, along with all parameter estimates and associated confidence intervals or standard errors. Criteria for evaluating values of fit indexes should be clearly stated.

CONCLUSION

Our review confirmed that SEM is a highly versatile tool heavily used in the psychology research literature to investigate a variety of problems. Although there are high-quality applications that provide important insights or advances in particular substantive areas, there are also problematic aspects of this literature. These range from problems of perspective, design, and strategy to mechanical aspects of model specification, data analysis, interpretation, and presentation. The problems we have described can have a substantial impact on the quality of information produced in these applications as well as on the validity of interpretations and conclusions. Paying attention to the concerns raised in this review should enhance the quality of applications of SEM and in turn increase the quality of knowledge gained from its use.

ACKNOWLEDGMENTS

We wish to thank the following individuals for valuable comments and suggestions provided at various stages of this project: Michael Browne, Roger Millsap, Chip Reichardt, Keith Widaman, Joe Rodgers, Jack McArdle, John Nesselroade, Jim Steiger, and Ken Bollen. We also thank Scott Davies for his enormous assistance in conducting the literature review and Tom Sidick for assistance in checking the manuscript and references.

LITERATURE CITED

Aiken L, Stein JA, Bentler PM. 1994. Structural equation analyses of clinical subpopulation differences and comparative treatment outcomes: characterizing the daily lives of drug addicts. *J. Consult. Clin. Psychol.* 62:488–99

Arbuckle JL, Wothke W. 1999. *Amos 4.0 User's Guide.* Chicago: SmallWaters

Austin JT, Calderón RF. 1996. Theoretical and technical contributions to structural equation modeling: an updated annotated bibliography. *Struct. Equ. Model.* 3: 105–75

Austin JT, Wolfle LM. 1991. Annotated bibliography of structural equation modeling: technical work. *Br. J. Math. Stat. Psychol.* 44:93–152

Baumgartner H, Homburg C. 1996. Applications of structural equation modeling in marketing and consumer research: a review. *Int. J. Res. Market.* 13:139–61

Belsky J, Hsieh K, Crnic K. 1996. Infant positive and negative emotionality: one dimension or two? *Dev. Psychol.* 32:289–98

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

Bentler PM. 1986. Structural modelling and psychometrika: an historical perspective on growth and achievements. *Psychometrika* 51:35–51

Bentler PM, Bonett DG. 1980. Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychol. Bull.* 88:588–606

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

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

Brannick MT, Spector PE. 1990. Estimation problems of the block diagonal model of the multitrait-multimethod matrix. *Appl. Psychol. Meas.* 14:325–40

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

Browne MW, Cudeck R. 1989. Single sample cross-validation indices for covariance structures. *Multivar. Behav. Res.* 24: 445–55

Browne MW, Cudeck R. 1993. Alternative ways of assessing model fit. In *Testing Structural Equation Models,* ed. KA Bollen, JS Long, pp. 136–61. Newbury Park, CA: Sage

Browne MW, Mels G. 1998. Path analysis: RAMONA. In *SYSTAT for Windows: Advanced Applications* (Ver 8), pp. 607–65. Evanston, IL: SYSTAT

Campbell DT, Fiske DW. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56:81–105

Church AT, Burke PJ. 1994. Exploratory and confirmatory tests of the Big Five and Tellegen's three- and four-dimensional models. *J. Pers. Soc. Psychol.* 66:93–114

Cliff N. 1983. Some cautions concerning the application of causal modelling methods. *Multivar. Behav. Res.* 18:115–26

Collins LM, Horn JL, eds. 1991. *Best Methods for the Analysis of Change.* Washington, DC: Am. Psychol. Assoc.

Coovert MD, Craiger JP, Teachout MS. 1997. Effectiveness of the direct product versus confirmatory factor model for reflecting the structure of multimethod-multirater job performance data. *J. Appl. Psychol.* 82:271–80

Cudeck R. 1988. Multiplicative models and MTMM matrices. *Multivar. Behav. Res.* 13:131–47

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

Cudeck R, Browne MW. 1983. Cross-validation of covariance structures. *Multivar. Behav. Res.* 18:507–34

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

Curran P, Harford T, Muthén BO. 1997a. The relation between heavy alcohol use and bar patronage: a latent growth model. *J. Stud. Alcohol* 57:410–18

Curran P, Stice E, Chassin L. 1997b. The relation between adolescent alcohol use and peer alcohol use: a longitudinal random coefficients model. *J. Consult. Clin. Psychol.* 65:130–40

Donders J. 1997. A short form of the WISC-III for clinical use. *Psychol. Assess.* 9:15–20

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

Duncan TE, Duncan SC, Stoolmiller M. 1994. Modeling developmental processes using latent growth structural equation methodology. *Appl. Psychol. Meas.* 18:343–54

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

Eisenberg N, Guthrie IK, Fabes RA, Reiser M, Murphy BC, et al. 1997. The relations of regulation and emotionality to resilience and competent social function in elementary school children. *Child Dev.* 68:295–311

Finkel D, Pederson N, McGue M. 1995. Genetic influences on memory performance in adulthood: comparison of Minnesota and Swedish twin data. *Psychol. Aging* 10:437–46

Floyd FJ, Widaman KF. 1995. Factor analysis in the development and refinement of clinical assessment instruments. *Psychol. Assess.* 7:286–99

Gest SD. 1997. Behavioral inhibition: stability and associations with adaptation from childhood to early adulthood. *J. Pers. Soc. Psychol.* 72:467–75

Gold DP, Andres D, Etezadi J, Arbuckle T. 1995. Structural equation model of intellectual change and continuity and predictors of intelligence in older men. *Psychol. Aging* 10:294–303

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

Gollob HF, Reichardt CS. 1991. Interpreting and estimating indirect effects assuming time lags really matter. In *Best Methods for the Analysis of Change,* ed. LM Collins, JL Horn, pp. 243–59. Washington, DC: Am. Psychol. Assoc.

Greenwald AG, Pratkanis AR, Leippe MR, Baumgardner MH. 1986. Under what conditions does theory obstruct research progress? *Psychol. Rev.* 93:216–29

Harnish JD, Dodge KA, Valente E. 1995. Mother-child interaction quality as a partial mediator of the roles of maternal depressive symptomatology and socioeconomic status in the development of child behavior problems. *Child Dev.* 66:739–53

Holohan CJ, Moos RH, Holohan CK, Brennan P. 1997. Social context, coping strategies, and depressive symptoms: an expanded model with cardiac patients. *J. Pers. Soc. Psychol.* 72:918–28

Horn JL, McArdle JJ. 1992. A practical guide to measurement invariance in research on aging. *Exp. Aging Res.* 18:117–44

Hoyle RH. 1993. Interpersonal attraction in the absence of explicit attitudinal information. *Soc. Cogn.* 11:309–20

Hoyle RH, Panter AT. 1995. Writing about structural equation modeling. In *Structural Equation Modeling: Concepts, Issues, and Applications,* ed. RH Hoyle, pp. 158–76. Thousand Oaks, CA: Sage

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

Hu L, Bentler PM. 1999. Cutoff criteria for fit indexes in covariance structure analysis:

conventional criteria versus new alternatives. *Struct. Equ. Mod.* 6:1–55

Hulland J, Chow YH, Lam S. 1996. Use of causal models in marketing research: a review. *Int. J. Res. Mark.* 13:181–97

Jackson PR, Wall TD, Martin R, Davids K. 1993. New measures of job control, cognitive demand, and production responsibility. *J. Appl. Psychol.* 78:753–62

James LR, James LA. 1989. Causal modelling in organizational research. In *International Review of Industrial and Organizational Psychology 1989*, ed. C Cooper, I Robertson, pp. 371–404. Chichester, UK: Wiley

Jones C, Meredith W. 1996. Patterns of personality change across the life span. *Psychol. Aging* 11:57–65

Jöreskog KG. 1979. Statistical models and methods for analysis of longitudinal data. In *Advances in Factor Analysis and Structural Equation Models*, ed. KG Jöreskog, D Sörbom, pp. 129–69. Cambridge, MA: Abt

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

Judge TA, Locke EA. 1993. Effect of dysfunctional thought processes on subjective well-being and job satisfaction. *J. Appl. Psychol.* 78:475–90

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

Kenny DA, Kashy D. 1992. Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychol. Bull.* 112:165–72

Kim J, Nesselroade JR, Featherman DL. 1996. The state component in self-reported worldviews and religious beliefs of older adults: the MacArthur successful aging studies. *Psychol. Aging* 11:396–407

Kishton JM, Widaman KF. 1994. Unidimensional versus domain representative parceling of questionnaire items: an empirical example. *Educ. Psychol. Meas.* 54:757–65

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

Lawley DN, Maxwell AE. 1971. *Factor Analysis as a Statistical Method.* New York: Elsevier

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

Lent RW, Brown SD, Gore PA. 1997. Discriminant and predictive validity of academic self-concept, academic self-efficacy, and mathematics-specific self-efficacy. *J. Counsel. Psychol.* 44:307–15

Lindenberger U, Baltes PB. 1997. Intellectual functioning in old and very old age: cross-sectional results for the Berlin Aging Study. *Psychol Aging* 12:410–32

Little TD. 1997. Mean and covariance structures (MACS) analyses of cross-cultural data: practical and theoretical issues. *Multivar. Behav. Res.* 32:53–76

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

Loehlin JC. 1989. Partitioning environmental and genetic contributions to behavioral development. *Am. Psychol.* 44:1285–92

Loehlin JC. 1998. *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis.* Hillsdale, NJ: Erlbaum. 3rd ed.

Lopez DF, Little TD. 1996. Children's action-control beliefs and emotional regulation in the social domain. *Dev. Psychol.* 32:299–312

MacCallum RC. 1986. Specification searches in covariance structure modelling. *Psychol. Bull.* 100:107–20

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, Kim C, Malarkey WB, Kiecolt-Glaser JK. 1997. Studying multivariate change using multilevel models and latent curve models. *Multivar. Behav. Res.* 32:215–53

MacCallum RC, Roznowski M, Necowitz LB. 1992. Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychol. Bull.* 111:490–504

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. 1999. Sample size in factor analysis. *Psychol. Methods* 4:84–99

Marsh HW, Balla JR, McDonald RP. 1988. Goodness-of-fit indexes in confirmatory factor analysis: the effect of sample size. *Psychol. Bull.* 103:391–410

Marsiske M, Klumb P, Baltes MM. 1997. Everyday activity patterns and sensory functioning in old age. *Psychol. Aging* 12:444–57

Maruyama G. 1998. *Basics of Structural Equation Modeling.* Thousand Oaks, CA: Sage

McArdle JJ, Aber MS. 1990. Patterns of change within latent variable structural equation modeling. In *New Statistical Methods in Developmental Research,* ed. A von Eye, pp. 151–224. New York: Academic

McArdle JJ, Prescott CA. 1997. Contemporary models of the biometric genetic analysis of intellectual abilities. In *Contemporary Intellectual Assessment: Theories, Tests, and Issues,* ed. DP Flanagan, JL Genshaft, PL Harrison, pp. 403–36. New York: Guilford

McArdle JJ, Prescott CA, Hamagami F, Horn JL. 1998. A contemporary method for developmental-genetic analyses of age changes in intellectual abilities. *Dev. Neuropsychol.* 14:69–114

McArdle JJ, Woodcock RW. 1997. Expanding test-retest designs to include developmental time-lag components. *Psychol. Methods* 2:403–35

McCrae RR, Zonderman AB, Costa PT, Bond MH, Paunonen SV. 1996. Evaluating replicability of factors in the revised NEO Personality Inventory: confirmatory factor analysis versus Procrustes rotation. *J. Pers. Soc. Psychol.* 70:552–66

McDonald RP. 1994. The bilevel reticular action model for path analysis with latent variables. *Sociol. Methods Res.* 22:399–413

Medsker GM, Williams LJ, Holohan P. 1994. A review of current practices for evaluating causal models in organizational behavior and human resources management research. *J. Manage.* 20:439–64

Meredith W. 1993. Measurement invariance, factor analysis, and factorial invariance. *Psychometrika* 58:525–43

Meredith W, Tisak J. 1990. Latent curve analysis. *Psychometrika* 55:107–22

Meyer JP, Allen NJ, Smith CA. 1993. Commitment to organizations and occupations: extension and test of a three-component conceptualization. *J. Appl. Psychol.* 78:475–90

Millsap RE. 1997. Invariance in measurement and prediction: their relationship in the single-factor case. *Psychol. Methods* 3:248–60

Millsap RE, Meredith W. 1994. Statistical evidence in salary discrimination studies: nonparametric inferential conditions. *Multivar. Behav. Res.* 29:339–64

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

Muthén BO. 1997. Latent variable modeling of longitudinal and multilevel data. In *Sociological Methodology,* ed. A Raftery, 27:453–80. Oxford, UK: Basil Blackwell

Muthén 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

Neale MC. 1997. *Mx: Statistical Modeling.* Richmond, VA: Dept. Psychiatry, Med. Coll. Virginia

Neale MC, Cardon LR, eds. 1992. *Methodology for Genetic Studies of Twins and Families. NATO ASI Ser. D, Vol. 67.* Dordrecht, The Netherlands: Kluwer

Nesselroade JR. 1991. Interindividual differences in intraindividual change. In *Best Methods for the Analysis of Change,* ed. LM Collins, JL Horn, pp. 92–105. Washington, DC: Am. Psychol. Assoc.

Neuberg SL, Judice TN, West SG. 1997. What the need for closure scale measures and what it does not: toward differentiating among related epistemic motives. *J. Pers. Soc. Psychol.* 72:1396–1412

Novy DM, Frankiewicz RG, Francis DJ, Liberman D, Overall JE, Vincent KR. 1994. An investigation of the structural validity of Loevinger's model and measure of ego development. *J. Pers.* 62:86–118

Pajares F, Miller MD. 1995. Mathematics self-efficacy and mathematics performances: the need for specificity of assessment. *J. Couns. Psychol.* 42:190–98

Parker JDA, Bagby RM, Summerfeldt LJ. 1993. Confirmatory factor analysis of the Revised NEO Personality Inventory. *Pers. Individ. Differ.* 15:463–66

Pentz MA, Chou C. 1994. Measurement invariance in longitudinal clinical research assuming change from development and intervention. *J. Consult. Clin. Psychol.* 62:450–62

Plomin R, Rende R. 1991. Human behavioral genetics. *Annu. Rev. Psychol.* 42:161–90

Prescott CA, Kendler KS. 1996. Longitudinal stability and change in alcohol consumption among female twins: contributions of genetics. *Dev. Psychopathol.* 8:849–66

Prussia GE, Kinicki AJ. 1996. A motivational investigation of group effectiveness using social cognitive theory. *J. Appl. Psychol.* 81:187–98

Rao CR. 1958. Some statistical methods for comparison of growth curves. *Biometrics* 14:1–17

Raykov T. 1994. Studying correlates and predictors of longitudinal change using structural equation modeling. *Appl. Psychol. Meas.* 18:63–77

Raykov T. 1995. Multivariate structural modeling of plasticity in fluid intelligence of aged adults. *Multivar. Behav. Res.* 30:255–87

Raykov T, Tomer A, Nesselroade JR. 1991. Reporting structural equation modeling results in *Psychology and Aging:* some proposed guidelines. *Psychol. Aging* 6:499–503

Reichardt CS. 1992. The fallibility of our judgments. *Eval. Pract.* 13:157–63

Reichardt CS, Coleman SC. 1995. The criteria for convergent and discriminant validity in a multitrait-multimethod matrix. *Multivar. Behav. Res.* 30:513–38

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

Reise SP, Widaman KF, Pugh RH. 1993. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychol. Bull.* 114:552–66

Rice KG, Cunningham TJ, Young MB. 1997. Attachment to parents, social competence, and emotional well-being: a comparison of black and white late adolescents. *J. Couns. Psychol.* 44:89–101

Rounds J, Tracey TJ. 1993. Prediger's dimensional representation of Holland's RIASEC circumplex. *J. Appl. Psychol.* 78:875–90

Russell DW, Kahn JH, Spoth R, Altmaier EM. 1998. Analyzing data from experimental studies: a latent variable structural equation modeling approach. *J. Couns. Psychol.* 45:18–29

Saudino KF, McGuire S, Reiss D, Hetherington EM, Plomin R. 1995. Parent ratings of EAS temperaments in twins, full siblings, half siblings, and step siblings. *J. Pers. Soc. Psychol.* 68:723–33

Sayer A, Collins LM, eds. 1999. *New Methods for the Analysis of Change.* Washington, DC: Am. Psychol. Assoc.

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

Steiger JH. 1999. Structural equation modeling (SEPATH). In *Statistica for Windows,* Vol. 3. Tulsa, OK: StatSoft

Steiger JH, Lind A. 1980. *Statistically based tests for the number of common factors.* Presented at the Annu. Meet. Psychometric Soc., Iowa City, IA

Stoolmiller M. 1994. Antisocial behavior, delinquent peer association, and unsupervised wandering for boys: growth and change from childhood to early adolescence. *Multivar. Behav. Res.* 29:263–88

Tisak J, Meredith W. 1990. Descriptive and associative developmental models. In *Statistical Methods in Longitudinal Research,* ed. A von Eye, 2:387–406. San Diego, CA: Academic

Tisak J, Tisak M. 1996. Longitudinal models of reliability and validity: a latent curve approach. *Appl. Psychol. Meas.* 20:275–88

Tremblay PF, Gardner RC. 1996. On the growth of structural equation modeling in psychological journals. *Struct. Equ. Model.* 3:93–104

Tucker LR. 1958. Determination of parameters of a functional relationship by factor analysis. *Psychometrika* 23:19–23

Wagner RK, Torgeson JK, Rashotte CA. 1994. Development of reading-related phonological processing abilities: new evidence of bidirectional causality from a latent variable longitudinal study. *Dev. Psychol.* 30:73–87

Walker LS, Smith CA, Garber J, Van Slyke DA. 1997. Development and validation of the Pain Response Inventory for children. *Psychol. Assess.* 9:392–405

Wheaton B, Muthén B, Alwin D, Summers G. 1977. Assessing reliability and stability in panel models. In *Sociological Methodology,* ed. DR Heise, pp. 84–136. San Francisco: Jossey-Bass

Widaman KF. 1985. Hierarchically nested covariance structure models for multitrait-multimethod data. *Appl. Psychol. Meas.* 9:1–26

Widaman KF, Reise S. 1997. Exploring the measurement invariance of psychological instruments: applications in the substance use domain. In *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse,* ed. KJ Bryant, M Windle, SG West, pp. 281–324. Washington, DC: Am. Psychol. Assoc.

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. 1995. Cross-domain analyses of change over time: combining growth modeling and covariance structure analysis. In *Advanced Structural Equation Modeling: Issues and Techniques,* ed. GA Marcoulides, RE Schumacker, pp. 125–57. Mahwah, NJ: Erlbaum