Psychological Methods
2002, Vol. 7, No. 1, 64–82

1082-989X/02/$5.00 DOI: 10.1037//1082-989X.7.1.64

Principles and Practice in Reporting Structural Equation Analyses

Roderick P. McDonald and Moon-Ho Ringo Ho
University of Illinois at Urbana–Champaign

Principles for reporting analyses using structural equation modeling are reviewed, with the goal of supplying readers with complete and accurate information. It is recommended that every report give a detailed justification of the model used, along with plausible alternatives and an account of identifiability. Nonnormality and missing data problems should also be addressed. A complete set of parameters and their standard errors is desirable, and it will often be convenient to supply the correlation matrix and discrepancies, as well as goodness-of-fit indices, so that readers can exercise independent critical judgment. A survey of fairly representative studies compares recent practice with the principles of reporting recommended here.

Structural equation modeling (SEM), also known as path analysis with latent variables, is now a regularly used method for representing dependency (arguably "causal") relations in multivariate data in the behavioral and social sciences. Following the seminal work of Jöreskog (1973), a number of models for linear structural relations have been developed (Bentler & Weeks, 1980; Lohmoller, 1981; McDonald, 1978), and work continues on these. Commercial statistical packages include LISREL (Jöreskog & Sörbom, 1989, 1996), EQS (Bentler, 1985, 1995), CALIS (Hartmann, 1992), MPLUS (Muthén & Muthén, 1998), RAMONA (Browne, Mels, & Cowan, 1994), SEPATH (Steiger, 1995), and AMOS (Arbuckle, 1997). Available freeware includes COSAN (Fraser & McDonald, 1988) and Mx (Neale, 1997).

McArdle and McDonald (1984) proved that different matrix formulations of a path model with latent variables are essentially equivalent. Programs such as those listed supply essentially the same basic information, with minor variations in the details supplied. Thus, the eight parameter LISREL model, which arose out of the work of Keesling and Wiley (see Wiley, 1973) and was subsequently developed to its current state by Jöreskog (see Jöreskog and Sörbom, 1996), the four-matrix model of Lohmoller (1981), the three-matrix EQS model of Bentler and Weeks (1980), and the two-matrix RAM model (see McArdle & McDonald, 1984) rearrange the same set of parameters. Not surprisingly—and perhaps not regrettably—user guides and texts on this topic are not in agreement in their recommendations about the style of presentation of results (e.g., see Bollen, 1989; Loehlin, 1992; Long, 1983a, 1983b). There is even less agreement in the form of the results actually reported in articles on applications.

It would be immodest for any journal article to offer a code of practice for the presentation of SEM results. It could also be counterproductive. (We note that for a long time there was a uniformly accepted convention for the publication of analysis of variance, or ANOVA results: the standard ANOVA table and the table of cell means. The near-disappearance of this from journals is regrettable.) Sound guidelines for the reporting of SEM results have been offered previously by Steiger (1988), Breckler (1990), Raykov, Tomer, and Nesselroade (1991), Hoyle and Panter (1995), and Boomsma (2000). MacCallum and Austin (2000) provided an excellent general survey of problems in applications of SEM.

Roderick P. McDonald and Moon-Ho Ringo Ho, Department of Psychology, University of Illinois at Urbana–Champaign.

We thank Jack McArdle, Malgorzata Szewczyk, and Bob Henson for their comments on this article. Any errors are our responsibility.

Correspondence concerning this article should be addressed to Roderick P. McDonald, Department of Psychology, University of Illinois at Urbana–Champaign, 603 East Daniel Street, Champaign, Illinois 61820. E-mail: rmcdonal@s.psych.uiuc.edu

Our objective is to review some principles for reporting SEM results. We also summarize some observations on the variety of treatments of results given in a selected set of reports on applications. The comparison of principles with practice is intended to underscore the importance of the recommendations made here. These are aimed at increasing the information supplied to the reader. Increased information should allow a more critical assessment of the study reported and serve to advance scientific knowledge. Naturally, some of our recommendations are essentially endorsements of those already given. Furthermore, this topic may need occasional revisiting until practice conforms with principles. We concentrate (appropriately) on matters in which there should be little disagreement. However, we do address a number of fresh elements here. We try to give fairly specific guidance on implementing previously published recommendations as well as our own. We also offer a few mild departures from previous suggestions. It is assumed that the reader is familiar with basic SEM method and with common terminology. However, we do offer occasional reminders about method, as well as remarks about recent developments that the reader may have missed.

From well-known SEM principles, we can formulate a list of results that we might hope to find in a comprehensive report, and we can check current practice against this list. Of course, there can be a conflict between an ideal of "completeness" and the understandable desire for conciseness in journal publications. This conflict is severe in the case of very large models.

We surveyed articles using SEM from 1995 to 1997 in the following journals: *British Journal of Psychology, Child Development, Developmental Psychology, Journal of Abnormal Psychology, Journal of Applied Psychology, Journal of Applied Social Psychology, Journal of Consulting and Clinical Psychology, Journal of Counseling Psychology, Journal of Educational Psychology, Journal of Family Psychology, Journal of Personality and Social Psychology, Journal of Research in Personality,* and *Psychological Assessment.* The method of search was simply to look for all path diagrams in the journals and period named. The intention was to find a reasonably representative sample, limited to path models with latent variables and to single-population studies.[1] This survey yielded 100 possible articles for review, of which 41 met the criteria: (a) path diagrams must be provided; (b) models being fitted must include both measurement and structural parts (pure factor analytic models or path analytic models without latent variables were excluded); and (c) multiple group comparisons were excluded (5 on this criterion).

The following discussion treats principles and practice together under familiar headings: model specification, identifiability, data and estimation, goodness of fit, parameters and their standard errors and, finally, alternative models. Each section concludes with specific recommendations.

Model Specification

Generally, a structural equation model is a complex composite statistical hypothesis. It consists of two main parts: The *measurement model* represents a set of p observable variables as multiple indicators of a smaller set of m latent variables, which are usually common factors. The *path model* describes relations of dependency—usually accepted to be in some sense causal—between the latent variables. We reserve the term *structural model* here for the composite SEM, the combined measurement and path models. This avoids the ambiguity that arises when the path model component is also labeled "the" structural model.

In most applications the measurement model is a conventional confirmatory factor model; the latent variables are just common factors and the error or specific terms are uncorrelated. The most common exception concerns longitudinal models with measurements replicated at two or more time points. Longitudinal models usually need specific components that are correlated across replicated measurements. Very commonly, the measurement model is an independent clusters model, that is, a factor model in which no indicator loads on more than one common factor.

In turn, the path model structures the correlation or

[1] Path models containing only observed variables—possibly weighted composite variables as in Wold's Partial Least Squares (PLS)—raise a separate set of issues from those conventionally recognized in SEM (see McDonald, 1996, for an account of some of these). Multiple population studies, likewise, raise questions of scale and of the structure of mean vectors, hence of goodness of fit, that would make them difficult to integrate into this account. It should be of interest to carry out reviews of articles written over later periods in the hope of finding improvement.

covariance matrix of the common factors. This structuring usually corresponds to a set of conjectured causal relations. The path model itself is also a composite hypothesis. It requires the specification both of a set of present versus absent directed arcs (paths) between latent variables, and a set of present versus absent nondirected arcs. Terminology in SEM is currently in a state of flux, because of the introduction of ideas from directed acyclic graph theory (for a general account, see Pearl, 1998, 2000). Here we follow Pearl in referring to a single direct connection between two variables in a path diagram as an *arc,* not a *path.* We reserve the term *path* for a sequence of arcs connecting two variables.

Pearl (2000) made a strong case for the position that a path model represents the operations of causality. The directed arcs indicate direct effects of conjectural (counterfactual) actions, interventions, or states of the world, whereas nondirected arcs represent correlated *disturbances,* random terms corresponding to variations not explained by the model. A directed arc in the graph of a path model is a single-headed arrow from one variable to another. A nondirected arc is usually drawn as a double-headed arrow. (The reader might wish to refer to Figures 1–3, where directed arcs drawn from the latent variable F_1 to F_2 and to F_3 and from F_1 to the indicator variables Y_1 and Y_2 are examples of directed arcs, whereas the double-headed arrow between F_2 and F_3 in Figure 1 or between d_2 and d_3 is an example of a nondirected arc in Figure 2.) On the causal interpretation, the absence of a directed arc from one variable to another implies the absence of a direct effect, whereas the absence of a nondirected arc implies that there are no omitted variables explaining their relationship. (These issues will be explored in greater detail later.)

In contrast to path equations, regression equations are essentially predictive and correspond to conditioning on observations of explanatory variables without manipulation—actual or theoretical. This is a rather technical distinction. For our purposes we note that residuals in a linear regression equation are uncorrelated with the independent variables by definition. The disturbances (unexplained variations) in a path equation can be correlated with the causal variables in that equation. Pearl (2000) gave graph-theory conditions under which a path equation is a regression equation; McDonald (in press) gives algebraic conditions. Most applications use *recursive* models, with no closed cycles formed by directed paths. With commonly made assumptions, the path equations of recursive models satisfy the conditions for regression equations. In *nonrecursive* models there are directed paths consisting of a sequence of unidirectional arcs forming closed loops. The most common form of loop is the simple case in which there is a directed arc from variable *X* to variable *Y* and also from *Y* to *X.* More rarely, we may find cases in which there is a directed arc from *X* to *Y, Y* to *Z,* and *Z* to *X,* closing the loop. In applications, loops formed by more than three variables seem rare to the point of nonexistence. Generally, their path equations are not regression equations. For a nontechnical account of this point, see McDonald (1997).

On this causal view of path models, the specification of directed arcs rests on a substantive theory expressed as a rich set of causal (counterfactual) conjectures. Investigators of an empiricist persuasion may hold alternative theories—explicit or implicit—justifying the direction of a directed arc, without acknowledging a notion of causal effect. We cannot legislate this question in the current state of philosophical debate.

Ideally, we might hope that a published report gives theoretical (causal?) grounds for the presence or absence of every directed arc in the path model. This was strongly suggested by Hoyle and Panter (1995) and Boomsma (2000), and we endorse it. A careful reading of the selected studies suggests that the researchers chose to omit or include directed arcs with some substantive justification. We cannot question these choices, but we note that none of the 41 reports examined attempted to approach the ideal of accounting for every directed arc that was chosen or omitted. In the absence of an agreed methodology, some compromise seems inevitable in practice. And in the absence of the author's explicit account of the theoretical justification of the choice of each directed arc, the reader is usually responsible for checking the plausibility of the set of choices made. We are not in a position to apply the more tolerant criterion of plausibility across the wide range of applications examined. In the absence of a detailed account of the choices made, it is tempting to surmise that these could appear somewhat arbitrary, even to experts in the relevant theory.

Quite generally in applications, the choice of nondirected arcs lacks any explicit justification in the published reports. On reflection, it is easy to see that the decision to omit a nondirected arc requires as much theoretical justification as the decision to omit a directed arc. The omission of a directed arc corre-

sponds to a belief that there is no direct cause–effect relation between two variables. The omission of a nondirected arc corresponds to the belief that the variables have no common cause that has been omitted from the model (e.g., see McDonald, 1997, 1999, chap. 17; Pearl, 1998, 2000). Indeed, the possibility of unspecified omitted common causes is the Achilles heel of SEM. The substantive motivation for including a nondirected arc seems to rest on the belief that some unmeasured common cause creates an unexplained relationship between them.

We note that two opposite research strategies suggest themselves: (a) We can omit a nondirected arc, unless we are confident that an omitted common cause exists; and (b) we can include the arc unless we are confident that no such cause exists. The first strategy seems almost universally adopted in published applications, but the choice is not made explicit. Omission of nondirected arcs combines with the omission of directed arcs to make a testable model. With appropriate specification, the combination also makes an identified model. The first strategy, in the extreme, is a likely source of poorly fitting models. The second strategy, in the extreme, yields untestable, underidentified models. Even so, testability and identifiability should not be the primary motives for omitting parameters. To underline the dilemma facing the investigator we note that adding nondirected arcs does not change the causal hypothesis, yet it radically alters fit.

Short of the ideal of a complete accounting, we simply suggest what should be obvious, namely, that the theoretical justification of directed arcs should be given as much detail as possible, hence open to the plausibility judgment of a reader with relevant substantive expertise. The justification for omission or inclusion of nondirected arcs (correlated disturbances) seems a more difficult research task. Omission might be covered by a blanket admission of ignorance. This carries with it the recognition that all conclusions could be radically altered by adding relevant variables in future studies. Inclusion requires specific theoretical grounds. We will return to this issue in other contexts.

Identifiability

Desirably, the SEM report will contain an explicit account of the conditions on the model that will secure identifiability, which is logically prior to estimation. Pearl (1998, 2000) gave graphical conditions for the identifiability of directed arc or path coefficients in the path model, and McDonald (in press) gives corresponding algebraic conditions. (We say that the model is identified if every parameter is identified.) Bekker, Merckens, and Wansbeek (1994) gave very general, very technical methods for determining model identification. However, in most applications these technicalities can be avoided. Commonly, we just require a fairly simple treatment of three distinct problems, namely, identifiability of the measurement model, identifiability of the path model, and scaling of the latent variables. We now examine these in turn.

Identifiability of the Measurement Model

A known, strong, sufficient condition for the parameters of the measurement model to be identified (except for scaling) is the condition that the factor loadings form independent clusters. In an independent clusters model each observed variable loads on only one common factor, and we can call it a *pure indicator* of the factor. A weaker sufficient condition has been called an *independent clusters basis* by McDonald (1999). This requires each latent variable or common factor to have at least two pure indicators if the factors are correlated, and at least three if they are not. Fortunately, in applications, this condition (and the stronger independent clusters condition) commonly results from substantive considerations. The substantive motivation for choosing the pattern of zero versus nonzero factor loadings dates from the origins of the common factor model. By design of the measurements we can usually expect that a subset of indicator variables measures just one attribute of the examinees in common, although each also measures a unique component (and is possibly subject to an error of replication). Such a design will tend to give independent clusters. It should be noted that the classical Thurstonian conditions of simple structure are often confused with independent clusters. Simple structure is more general. It allows the possibility that every variable loads on more than one factor. Simple structure does not generally secure an identified measurement model, although many simple structures will do so.

Of the 41 studies in our survey, 25 have independent cluster structure, 2 have an independent clusters basis, and 8 contain mixed latent and observable variables in the path model, but are readily recognized to be identified by independent clusters criteria. Only 3 studies do not have a measurement model that is seen to be identified by inspection. Nevertheless, in nearly every case, the reader is responsible for verifying

identifiability because no account is provided in the report.

The recommendation that follows from these considerations is this: Authors have a duty to show evidence that the measurement model is identified. Commonly, this will be the presence of independent clusters, or at least an independent clusters basis. If authors fail to do this, the reader can usually check the point without inconvenience.

Identifiability of the Path Model

The identifiability of the path model rests crucially on the choice of nondirected arcs. To examine this problem, we first recall the distinction between exogenous and endogenous variables, and introduce a further useful distinction. In terms of the path diagram, an exogenous variable has no directed arc ending on it. Its entire variance is unexplained by variables in the set studied. An endogenous variable has at least one directed arc ending on it, originating from one or more exogenous or endogenous variables. (Again, see Figures 1–3 for illustrations; F_1 is an exogenous latent variable, whereas F_2 and F_3 are endogenous.) There is a further useful distinction concerning the order of the variables in the path model. A variable X precedes another variable Y if there is a directed path (a sequence of unidirectional arcs) from X to Y. Two such variables are causally ordered. The variables are fully ordered if we can list them so that there is a directed path to each variable from every variable listed before it. (In Figures 1–3, F_2 and F_3 are unordered, whereas F_1 precedes both F_2 and F_3, so F_1 and F_2, also F_1 and F_3 are causally ordered.)

A weak condition for the identifiability of a recursive path model is that every equation is a regression. This condition is difficult to test directly (see McDonald, 1997, in press). Fortunately, it is implied by the stronger condition that all covariances of disturbances of causally ordered variables are zero (McDonald, 1997, termed this the *precedence rule*). This condition is easily checked from the path diagram. The precedence rule is implied by the stronger condition that all disturbances of endogenous variables are uncorrelated (McDonald, 1997, termed this the *orthogonality rule*). If a model is fully ordered except for the exogenous variables, the precedence and orthogonality rules coincide. These rules are not well and widely known.

There is a commonly accepted view that in a recursive path model identifiability may or should be secured by the orthogonality rule, but in a nonrecursive model the orthogonality rule cannot be employed. Instead, a rather technical inquiry is suggested. With all disturbances allowed to be nonzero, the rank and order rules are applied to see if the exogenous variables can yield identifiability by serving as instruments. These rules are too technical to be described here; for an introduction to this method, see Bollen (1989). (We note that if we follow these rules, the covariances between disturbances cannot be accounted for by omitted variables.) The contrary view is that we may assume orthogonality in nonrecursive models. This belief is at least implicit in the work of a number of authors. For example, Spirtes, Richardson, Meek, Scheines, and Glymour (1998) used it in a graph-theory account, and MacCallum, Wegener, Uchino, and Fabrigar (1993) used it in an account of equivalent models.[2] There appear to be unresolved problems in the foundations of nonrecursive models. These have been interpreted in a number of ways, including "feedback," systems measured in equilibrium, or averaged time series (see McDonald, 1997, for a nontechnical account of some of these issues, and Fisher, 1970, for a technical account of the averaged time series; see, also, Gollob & Reichardt, 1987, for the problem of time lags in such models). In the current state of knowledge we cannot adjudicate between these views, but it is fair to state that it is much easier to draw a nonrecursive loop in a path model than to motivate it rigorously.

Of the 41 studies surveyed, 10 have fully ordered models (except exogenous variables). Of these, all except 1 followed the orthogonality rule (and, equivalently, the precedence rule). The exception gave no substantive reason for including a nondirected arc between two variables connected by a directed arc (and no discussion of a likely identifiability problem). Of the remaining 31 studies that have one or more unordered pairs of latent variables, 7 chose nondirected arcs for the unordered pairs, as if they were following the precedence rule, and 24 had no correlated distur-

[2] These opposing views seem to have arisen because the assumption that every path equation is a regression is implied by the assumption that the disturbances are orthogonal if the model is recursive, and contradicted by the latter assumption if the model is nonrecursive. The contradiction can be resolved by denying either assumption. The instrumental variables treatment abandons both assumptions in favor of the assumption that the exogenous variables are orthogonal to the disturbances. It does not seem that any account of foundations has been given to adjudicate between these choices (see McDonald, 1997).

bances, as if they were following the orthogonality rule. Six longitudinal studies appropriately included correlated disturbances corresponding to replicated observations. As already noted, explicit reasons for the choices to include or to omit nondirected arcs are lacking in all of the studies.

Our recommendation on this matter is as follows: In the common case of a recursive model, the choice to include or omit a nondirected arc should, in principle, rest on substantive grounds for supposing the existence or nonexistence of unmeasured common causes, and not on securing identifiability. If the resulting model has no nondirected arcs between causally ordered variables, it is certainly identified. (This principle is not widely recognized and may not be known to the general user.) In a case with a nondirected arc between ordered variables, investigation is desirable using the rather technical methods of Pearl (1998, 2000), McDonald (in press), or Bekker et al. (1994). In the case of a nonrecursive model there is currently no clear identifiability rule. Perhaps there is also no single clear motive for the use of such a model.

Identifiability and Scaling

It is easy to choose the scale of exogenous latent variables: simply set their variances to unity in the computer program. The variance of an endogenous variable is not a parameter in the model, so fixing the scale on which it is measured becomes problematic. Currently, there are three main procedures for dealing with this problem. Browne and Du Toit's (1992) general method can be used to set all variances of endogenous latent variables to unity by a constrained minimization procedure. McDonald, Parker, and Ishizuka's (1993) reparameterization method also sets these variances to unity, but is restricted to recursive models. It is also possible to scale by setting chosen factor loadings or error variances to a constant (usually unity) during minimization and then to rescale the latent variables after fitting it. Browne and Du Toit's method is implemented in RAMONA and SEPATH; McDonald et al.'s method can be implemented in PROC CALIS, COSAN, or any program allowing suitable FORTRAN-like supplementary programming. The rescaling method is implemented in LISREL and EQS. The first two methods share the advantage that they yield correct standard errors in a fully standardized model. The rescaling method obtains correct parameters in a completely standardized model, but without corresponding standard errors.

Of the 41 reports in the survey, 40 gave standardized solutions, the exception being also the only nonrecursive model. All these are presumably computed by the rescaling method, as they all used either LISREL or EQS for the analyses.

It is certainly an old psychometric custom to interpret the numerical values of standardized factor loadings, standardized regression coefficients, and standardized path coefficients. It seems as though these are often thought to be metrically comparable, although their unstandardized counterparts are clearly not. This is an unsuitable place for a lengthy discussion of the dilemmas underlying the use of standardized coefficients, unstandardized coefficients, or variance explained as measures of the importance of an explanatory variable in a regression or in a path model.

We accept that standardization either before or after estimation is virtually unavoidable for applications of a path model with latent variables. Standardization avoids underidentifiability due to arbitrariness of scale. Experience seems to show that a completely standardized solution also aids interpretation of the results. As noted, some computer software is available that uses methods for obtaining standardized solutions with correct standard errors. We hope that the use of these methods will increase in the near future.

Data and Estimation

In the commonly employed estimation procedures, a sample of observations of size N on p variables gives a $p \times p$ sample covariance or correlation matrix S, with elements s_{jk}. A computer program then minimizes some function of the discrepancies $s_{jk} - \sigma_{jk}$ between these and the fitted covariances σ_{jk} implied by the composite structural model. Possibilities include maximum likelihood (ML), ordinary (unweighted) least squares (OLS), generalized least squares (GLS), and a variety of weighted least squares (WLS) discrepancy functions intended to give good estimates without requiring multivariate normality; these last stem from the seminal work of Browne (1984).

Multivariate Normality

Both ML and GLS estimation in SEM require the assumption of multivariate normality. However, as Micceri (1989) suggested, much social and behavioral science data may fail to satisfy this assumption. Several studies of the robustness of the multivariate nor-

mality assumption (Amemiya & T. W. Anderson, 1990; T. W. Anderson, 1989; Browne & Shapiro, 1988; Satorra & Bentler, 1994) have found that parameter estimates remain valid under reasonable assumptions even when the data are nonnormal, whereas standard errors do not. (These assumptions are difficult to test in applications.) A number of simulation studies (Chou, Bentler, & Satorra, 1991; Hu & Bentler, 1995; West, Finch, & Curran, 1995) suggest that ML and GLS estimation can give biased standard errors and incorrect test statistics in the presence of excessive skewness and/or kurtosis in the data.

The multivariate normality assumption may be evaluated univariately by checking the marginal distribution of each variable or by Mardia's (1970) multivariate skewness and kurtosis coefficients. Outliers, a likely source of skewed data, can be detected from the univariate distribution of the variables. Bollen and Arminger (1991) suggested the use of factor scores for outlier detection. Of the 41 studies, just 5 reported a test of multivariate normality in justification of the ML procedure, applied in all cases.

To deal with nonnormal data, Browne (1984) developed an asymptotically distribution-free (ADF) estimator, but a very large sample size is needed to obtain reliable weight matrices. For elliptical distributions, Browne (1984), Kano (1992), Shapiro and Browne (1987), and Satorra and Bentler (1994) have developed corrections to the normal likelihood ratio statistic. (To avoid technicalities, we simply remark that if a set of variables follows an elliptical distribution, all its marginal distributions are symmetric with the same kurtosis. Estimation based on an elliptical distribution can be thought of as intermediate between estimation using normality and ADF estimation.) Simulation studies by Curran, West, and Finch (1996), and by Hu, Bentler, and Kano (1992) suggest that the Satorra and Bentler rescaled statistic works well over a variety of distributions (see, also, Bentler & Dudgeon, 1996). In the context of analysis of variance and multivariate regression problems, transformation of variables is a common method for dealing with nonnormality of data. Mooijaart (1993) proposed the use of univariate Box-Cox transformations. Yuan, Chan, and Bentler (2000) proposed a robust transformation method.

The presence of categorical variables or indicators may cause nonnormality. Muthén (1984) developed a continuous/categorical variable methodology (CVM) estimator, which allows the analysis of any combination of dichotomous, ordered polytomous, and measured variables. Like Browne's (1984) ADF estimator, it requires a very large sample size to obtain reliable weight matrices. Simulation studies (Muthén, 1989; Muthén & Kaplan, 1992) suggest that the CVM estimator outperforms the Satorra and Bentler (1994) and ADF estimators when the number of categories of the variables are few (< 5).

A mild dilemma stems from the fact that ML estimation and its associated statistics seem fairly robust against violations of normality, whereas the use of ADF estimators requires extremely large samples for reliable weight matrices, far larger than are commonly available in current SEM applications.

Accordingly, we hesitate to make firm recommendations for the resolution of this dilemma, beyond noting that Mardia's (1970) test of multivariate skewness and kurtosis is well known and implemented in available computer software. It should, therefore, be easy for the investigator to see if a problem appears to exist and to report it to the reader. But in many cases the sample size will require the investigator to rely on the robustness of ML/GLS methods.

Sampling and Missing Data

A common problem in SEM applications is missing data. There are many possible reasons why the data matrix may be incomplete. Rubin (1976) proposed a classification scheme for missing data mechanisms and argued that missing data can be ignored (i.e., unbiased estimates can be obtained) under the condition that data are missing completely at random (MCAR) and missing at random (MAR). MCAR refers to missing data on a variable where presence or absence of the observation is independent of other observed variables and the variable itself. This is a stringent assumption that may not be justifiable in practice (Muthén, Kaplan, & Hollis, 1987). Less restrictively, MAR allows the presence or absence of an observation to depend on other observable variables, but not the variable from which the value is missing.

Listwise and pairwise deletion for missing data can lead to biased parameter estimates under MAR, but unbiased parameter estimates can be obtained under MCAR. Under MCAR and MAR, unbiased and more efficient parameter estimates can be obtained by full information maximum likelihood estimation (FIML), also known as the individual raw-score likelihood method (for simulation studies, see, e.g., Enders & Bandalos, 2001). Alternatively, if there are not too many distinct patterns of missing data, multiple-group

SEM may be used. This requires a different group for each pattern of missing data, with equality constraints imposed across groups. The implementation of this method can be found in Allison (1987), McArdle (1994), and Muthén et al. (1987). The expectation–maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) may also be employed. Recently, multiple imputation has become a popular method to deal with missing data. Multiple data sets are created with a set of plausible values replacing the missing values, then standard analysis is applied to the complete data set (for details of this procedure, see Graham & Hofer, 2000). Currently, available software for dealing with missing data includes Mplus, AMOS, and Mx, using full information, whereas LISREL provides full information and multiple imputation; all of these assume MAR.

If the missing data mechanism is nonignorable (not MAR), there is no correction available. We may guess that this will be the common case in applications (for further discussion, see Heckman, 1979; Heckman & Robb, 1986; McArdle, 1994; Rubin, 1987, 1991).

The problem of missing data is closely related to the problem of sampling as such in SEM studies. In principle, we scale a measurement model by choosing a zero mean and unit standard deviation for the latent variables in the population sampled. The likely effect of nonrandom sampling (which is surely common in applications) is to change (in an unknown way) the variances and covariances of the latent variables. The metric of the model will then fail to correspond to any well-defined norming population. However, in many applications there is no defined norming population in the first place. For applications such as testing a causal model, this may not be critical. Random sampling will be of primary importance mainly in certain test theory applications, where tests are scaled on the basis of a reference or calibration population. In such cases there will likely be no accompanying causal model, the object being primary test development (see McDonald, 1999).

The effect of nonrandom sampling includes the effect of listwise deletion. Technically, the likely effect of either is to produce a change (of unknown extent) in the variances and covariances of the latent variables. Because we scale the latent variables to have unit variance, the result will be proportional changes (again of unknown extent) in factor loadings. This expectation follows from the classical account by Meredith (1964) of the effect of selection on a factor model. We hesitate to make firm recommendations on the problem of missing data, beyond the obvious point that a report should indicate the extent to which there is a missing data problem and describe the method used to deal with it. The report should also include some account of the extent to which nonrandom sampling (whether caused by missing data or not) could significantly work against the purposes of the study.

Reporting Data

Subject to editors' concerns about space, the sample covariance or correlation matrix gives the reader a great deal of freedom to formulate and evaluate plausible alternative models. Mere inspection of this information can often be useful to the reader.

Either the sample correlation matrix with or without standard deviations or the sample covariance matrix was supplied in 19 of the 41 reports ($M = 14.74$ variables, $SD = 6.22$, range = 9–28). The 22 that did not supply covariances or correlations have a mean of 21.74 ($SD = 8.55$, range = 10–40), with only 4 in excess of 28; none of the latter indicated availability of their data.

It is desirable that this information be readily available. Few authors justify their model in detail, and it is generally easy to postulate equally plausible, and possibly better-fitting alternatives on the basis of such theory as is available. (Any instructor in quantitative methods who has given this task as an exercise to graduate students will confirm this statement.) We suggest that there is a strong case for publishing the correlations, possibly accompanied by means and standard deviations, for up to 30 variables. In the case of a large set of variables, the author can alternatively indicate a means to access the covariance or correlation matrix if the reader wishes to do so; possibilities include application to the author or placing the information in the author's, or possibly the journal's, website. In the case of truly large studies, a more comprehensive account than is possible in a journal might be made available in this way. Information technology is changing too rapidly for this advice to be more precise.

Goodness of Fit

Except for OLS estimates as usually treated, estimation methods also yield an asymptotic chi-square and asymptotic standard errors for the parameters in an identified model. It has long been recognized that all SEMs are simplified approximations to reality, not hypotheses that might possibly be true. Accordingly,

an abundance of indices has been developed as measures of the goodness or badness of the approximation to the distribution from which the sample was drawn, and it is very easy to invent many more (see, e.g., Bollen & Long, 1993; McDonald & Marsh, 1990). Available computer programs supply varying subsets of these. We may distinguish absolute and relative fit indices: Absolute indices are functions of the discrepancies (and sample size and degrees of freedom); relative indices compare a function of the discrepancies from the fitted model to a function of the discrepancies from a null model. The latter is almost always the hypothesis that all variables are uncorrelated. Some psychometric theorists have prescribed criterion levels of fit indices for a decision to regard the approximation of the model to reality as in some sense "close." This is considered to make the decision objective (see, e.g., Hu & Bentler, 1999). There are four known problems with fit indices. First, there is no established empirical or mathematical basis for their use. Second, no compelling foundation has been offered for choosing a relative fit index over an absolute index, or for regarding uncorrelated variables as a null model. Third, there is not a sufficiently strong correspondence between alternative fit indices for a decision based on one to be consistent with a decision based on another; the availability of so many could license a choice of the best-looking in an application, although we may hope this does not happen. Fourth, and perhaps most important, a given degree of global misfit can originate from a correctable misspecification giving a few large discrepancies, or it can be due to a general scatter of discrepancies not associated with any particular misspecification. Clear misspecifications can be masked by the indexed fit of the composite structural model. It is impossible to determine which aspects of the composite hypothesis can be considered acceptable from the fit indices alone. Along with checking these, we recommend examining the (standardized) discrepancies in the measurement model and the individual discrepancies between the latent variable correlations in the measurement model and the fitted correlations constrained by the path model.

Global Fit Indices

As we would expect, the question of goodness of fit is resolved by different investigators in quite different ways. Unsurprisingly, all 41 studies report the global chi-square and degrees of freedom for the composite structural model (the measurement model and path model combined). Of these, 5 were in the enviable position of having a conventionally nonsignificant chi-square, at sample sizes 70, 165, 193, 330, and 461. However, these, as well as the others, also reported some indices of goodness or badness of global approximation. In terms of simple popularity, the index independently given by McDonald and Marsh (1990) as the unbiased relative fit index (URFI) and Bentler (1990) as the comparative fit index (CFI)[3] was most used (21 of 41 studies) and next, used in 20 of 41 studies, was the root mean square error of approximation (RMSEA), which originated with Steiger and Lind, and is accessible in published work by Browne and Cudeck (1993). Other indices commonly used are the goodness-of-fit index (GFI; Jöreskog & Sörbom, 1989; 15 of 21 studies), the Tucker-Lewis index (TLI; Tucker & Lewis, 1973; sometimes referred to as the nonnormed fit index, or NNFI, attributed to Bentler & Bonett, 1980; 13 of 41 studies), and the normed fit index (NFI; Bentler & Bonett, 1980; 9 of 41 studies). Most investigators reported at least two such measures. Those relying on the RMSEA generally accepted the authoritative claim that an RMSEA less than .05 corresponds to a "good" fit and an RMSEA less than .08 corresponds to an "acceptable" fit. Most authors who used an index scaled up to unity for "perfect" fit regarded these (URFI/CFI, GFI, etc.) as acceptable if they were greater than .9. It is sometimes suggested that we should report a large number of these indices, apparently because we do not know how to use any of them. It appears that the only metacriterion we can use to evaluate these conventions is the principle that a model is acceptable if the discrepancies are too small to support a more complex model (see McDonald, 1999, chap. 17). None of these studies give discrepancies (or, it seems, use their distributions to aid judgment), so we cannot check them against such a criterion. Simulation studies based on restrictive models do not help because the fit indices were invented to deal with the fact that no restrictive model fits real data.

If an article does supply the sample correlations, it

[3] There is a minor difference between the CFI and the URFI. If the URFI exceeds unity, corresponding to an overfitting model, the CFI is reset to unity, leaving the extent of overfit unknowable. Hoyle and Panter (1995) gave this difference as a reason for preferring the CFI. We mildly demur, suggesting it is a reason for preferring the URFI as being more informative.

is very easy for the authors to complete the table by including the standardized discrepancies in the complementary triangle. This enables the reader to form an independent judgment of the relationships not well explained by the model. It becomes possible to judge whether a marginal or low index of fit is due to a correctable misspecification of the model, or to a scatter of discrepancies, which suggests that the model is possibly the best available approximation to reality. If a covariance matrix has been analyzed, the computer program may supply variance-standardized discrepancies for this purpose.

Table 1 illustrates the format of a table presenting correlations, variances, and standardized discrepancies, and is adapted from McDonald (1999, pp. 388–389), where details can be found. Such a table presents the information in a very compact form. We note that in this (empirical) example the two nonzero discrepancies correspond directly to missing arcs between the pairs of variables. The remaining discrepancies are necessarily exact zeros in this model, therefore, a global fit index would not seem very helpful.

For large numbers of variables the report can give summary information about the discrepancies, from which the reader can endorse or disagree with the author's judgment. Listings of the largest discrepancies are given by some computer programs, and in any case are easy for authors to create. Of the 19 of 41 studies giving the correlation matrix, none gave the discrepancies or summary information about them. (Given all the parameters, the reader can construct the discrepancies, but not without considerable effort.)

Our comments on this issue might seem intended to discourage the use of fit indices. Our intention is, instead, to warn that the issue is unsettled. More constructively, but tentatively, we offer some recommendations as follows: It is our belief that no global index of fit (together with a criterion for its acceptability) can substitute for a detailed examination of the discrepancies. However, if inspection shows that these are well scattered, they are adequately summarized in the root mean square residual (RMR), which is an immediately interpretable measure of the discrepancies. In turn, the GFI is a function of the RMR and the corresponding root mean square of the sample correlations. The GFI is therefore acceptable to those who believe that for a given RMR, fit is better if the correlations explained are larger. As shown in some detail by McDonald and Marsh (1990), most other fit indices can be expressed as functions of the noncentrality parameter connected to ML/GLS estimation. An unbiased estimate of the noncentrality parameter is given by $d = (\chi^2 - df)/N$ (McDonald, 1989). This parameter is also a norm on the sizes of the discrepancies. Accordingly, if the discrepancies are well scattered, such indices capture their general spread well enough. Those indices that have been shown (as in McDonald & Marsh, 1990) to be free of sampling bias, for example, the RMSEA and the URFI (and CFI if not reset to unity because of overfitting), can be recommended as supplements to the investigator's primary judgment based on the discrepancy matrix.

Table 1
An Example of Reporting Correlations and Discrepancies

Correlations and discrepancies				
5.86	.0	.0	.0	.097
.370	6.08	.0	.0	−.084
−.400	−.290	3.00	.0	.0
−.320	−.420	.550	7.90	.0
−.210	−.410	.630	.730	7.54

Note. Sample correlations are in lower triangle; variances are in diagonal and discrepancies are in upper triangle. An exact zero is represented by .0. (Example adapted from McDonald, 1999, pp. 388–389.)

Path Model Fit

As already noted, a structural model is a composite of a measurement model and a path (causal) model. Accordingly, it might be useful to separate measures of fit into parts corresponding at least to these two major components. Surely the primary objective of an SEM study is to give supporting evidence for the specified path model. The (hopefully well designed) set of measurements is essentially subservient to this aim.

J. C. Anderson and Gerbing (1988) suggested a sequential testing procedure based on the recognition that the structural model is nested within the measurement model. (See also Fornell & Yi, 1992, and the reply by J. C. Anderson & Gerbing, 1992, as well as Bentler, 2000; Bollen, 2000; Hayduk & Glaser, 2000a, 2000b; Herting & Costner, 2000; Mulaik & Millsap, 2000, for further discussion of this question.) The asymptotic distribution of the ML or GLS discrepancy function for a composite structural model can be decomposed into independent additive noncentral chi-squares, one for the measurement model, and one for the path model (see Steiger, Shapiro, & Browne, 1985). The path model component is estimated as the difference between the discrepancy functions for the structural model and the measurement

model. The degrees of freedom are correspondingly additive, of course.

Separate chi-squares and degrees of freedom for the measurement model, with unconstrained latent variable correlations, were given in 14 of the 41 studies. These studies all provided some comparison of the global and measurement chi-squares, but their final conclusions were based only on the goodness of approximation of the composite structural model.

Table 2 gives the chi-squares and degrees of freedom for the structural and measurement model and also the chi-squares and degrees of freedom for the

Table 2
Reexamination of the Goodness of Fit of the Structural (s), Measurement (m), and Path (p) Models for 14 Selected Studies

Study	*N*	Model[a]	χ^2	*df*	*p*	*d*	RMSEA
1	461	s	124.8	106	.001	.041	.020
		m	27.7	97	1.000	−.150	
		p	97.1	9	.000	.191	.146
2	357	s	980.7	391	.000	1.651	.065
		m	638.1	361	.000	.678	.046
		p	342.6	30	.000	.876	.171
3	465	s	1368.2	472	.000	1.929	.064
		m	1284.1	467	.000	1.757	.061
		p	84.1	5	.000	.170	.184
4	326	s	212.0	89	.000	.377	.065
		m	158.0	81	.000	.236	.054
		p	54.0	8	.000	.141	.133
5	507	s	400.4	161	.000	1.180	.054
		m	306.8	155	.000	.299	.044
		p	93.6	6	.000	.173	.170
6	81	s	112.0	56	.000	.691	.111
		m	51.3	42	.153	.115	.052
		p	60.8	14	.000	.576	.203
7	237	s	519.2	340	.000	.756	.047
		m	514.8	333	.000	.767	.048
		p	4.6	7	.727	−.011	
8	289	s	521.0	180	.000	1.180	.081
		m	434.9	175	.000	.899	.072
		p	86.1	5	.000	.280	.237
9	330	s	288.5	214	.000	.226	.032
		m	284.4	201	.000	.252	.035
		p	4.0	13	.991	−.026	
10	197	s	685.8	300	.000	1.958	.081
		m	562.3	270	.000	1.940	.085
		p	33.5	30	.299	.018	.025
11	377	s	161.1	80	.000	.215	.052
		m	141.2	79	.000	.165	.046
		p	19.9	1	.000	.050	.224
12	1556	s	725.0	269	.000	.293	.033
		m	577.0	247	.000	.212	.029
		p	148.0	22	.000	.081	.061
13	84	s	50.2	43	.209	.086	.045
		m	44.7	38	.211	.080	.046
		p	5.5	5	.356	.006	.035
14	70	s	41.9	21	.000	.298	.119
		m	12.4	17	.000	−.066	
		p	29.5	4	.000	.365	.302

Note. *d* = noncentrality parameter; RMSEA = root mean square error of approximation.

path model, obtained by difference, for the 14 studies in our survey that provided the necessary information. Not surprisingly, the degrees of freedom associated with the path model are generally much smaller than the degrees of freedom for the measurement model. Accordingly, the fit of the composite structural model can appear satisfactory when the few constraints implied by the path model are not, in fact, correctly specified. Conversely, if the fit of the composite model appears unacceptable, it is important to know if the misfit is due primarily to a misspecified measurement model, a misspecified path model, or both.

As previously noted, the URFI (McDonald & Marsh, 1990), the equivalent CFI (Bentler, 1990), and most current fit indices are functions of the noncentrality parameter. This parameter also has an additive property. The noncentrality of the structural model is the sum of the noncentralities of the measurement and the path models. Table 2 includes the unbiased estimates of the (additive) noncentralities as well as the resulting RMSEAs (which, of course, are not additive) for these studies. If we are willing to rest on the established properties of the RMSEA, the results suggest that in all but a few cases the goodness of fit of the composite structural model, with its large number of degrees of freedom, conceals the badness of fit of the path model, with its much smaller number of degrees of freedom. By a natural extension of theory, we may regard the additive noncentrality parameters from the measurement and path model discrepancy functions as representing the errors of approximation of these two independent components of the model. We tentatively conclude that in the majority of studies for which the measurement model information is available, the goodness of approximation of the path model may be unacceptable, contrary to the published conclusions. By extrapolation, this may be true of those studies for which this information is not available.

It certainly seems desirable that existing SEM studies be reevaluated with attention to direct assessment of the fit of the path model. For this purpose, we might use a supplementary analysis by a two-stage procedure. We first fit the measurement model, and then fit the path model to the latent variable correlation matrix in order to study in detail the pattern of the discrepancies. We recommend such a two-stage procedure in future studies.

Parameters and Standard Errors

The parameters of an SEM are the independently estimated loadings and error variances and covariances in the measurement model, and the independently estimated directed arc coefficients and disturbance variances and covariances in the path model. Special cases include: (a) pure measurement models, that is, traditional confirmatory common factor models; (b) path models for observable variables; and (c) mixed models in which some variables in the path model are observable and some are latent.

We note in passing that the possibility of mixed models underlines a dilemma in SEM, needing further research. As pointed out by McDonald (1996), the researcher usually has a choice between making a path model with latent variables and a path model with composite variables: simple or weighted sums of indicators. The common choice seems to be the former, because it estimates causal relations between attributes that are corrected for errors of measurement. For these causal relations to be taken seriously, we must be able to suppose we could, in principle, add enough further indicators to make error-free measurements. If the number of indicators of each attribute is small (as seems common in applications), such corrections may themselves be unreliable. In a model in which some attributes are measured by a single total test score, we can take scores from items or subtests of the test as multiple indicators, if we wish to allow for errors of measurement. In the rare case where a single item is the only measure of an attribute, there does not yet seem to be any treatment available for its error of measurement.

Just 12 of the 41 studies report all parameters, whereas 20 omit both error and disturbance parameters, 2 omit error variances, 3 omit disturbance variances, 2 give only the directed arc coefficients, and 2 omit parameter estimates in the measurement part of the model. Even in a fully standardized model, the reader would find it difficult to construct the error variances as unit complements of functions of the other parameters. Generally, we would wish to verify that the unique variances are not close to zero, corresponding to an improper (Heywood) solution, as would also be indicated by large standard errors. The disturbance variances also allow the reader to see what proportions of variance of the endogenous variables are accounted for by the model (only 2 of the 41 reports explicitly addressed this question).

Standard errors of some parameters were reported in only 5 of the 41 studies. Of these 5, none reported all standard errors. Usually, standard errors of unique and disturbance variances are not reported. Generally, there appears to be nothing preventing such a report.

Standard errors can be included in tables of the parameter values or, following one convention, put in parentheses attached to the parameters in a path diagram. However, unless the scaling problem is solved, either by the constrained minimization method of Browne and Du Toit (1992) or by the reparameterization method of McDonald et al. (1993), standard errors are not available for the standardized solution. If the standardized parameters are obtained by rescaling, it would still be possible to examine and report the standard errors of the unstandardized parameters, and to describe statistical inferences based on these before interpreting the corresponding standardized coefficients. This would follow the comparable procedure in regression programs.

The obvious and conveniently implemented recommendation is to include all the parameters and their standard errors in the research report. It is difficult to see what could prevent this.

The parameters (and standard errors if supplied) can, with equal convenience and approximately equal space on the page, be presented in tables or attached to directed and nondirected arcs in path diagrams. Path diagrams originated with Sewell Wright (1921), and have come into current SEM practice rather informally and haphazardly. There is at least some agreement that the network of relationships in a path model is most easily appreciated in the form of a path diagram, a picture of the graph. Formal directed acyclic graph (DAG) theory (see Pearl, 1988, 2000; Scheines, Spirtes, Glymour, Meek, & Richardson, 1998) is becoming more widely recognized by SEM researchers. Recognition of this work should soon lead to a greater formality in the pictorial representation of the hybrid graphs (graphs containing both directed and nondirected arcs) that structure SEMs, and to their direct use in determining the precise form of the constraints on the covariances.

We comment, and invite possible disagreement, that in the common case where the measurement model is a standard independent cluster model, there is little to be gained from including the relations between common factors or latent variables and their indicators in the path diagram. A good case can be made for presenting the measurement model in the tabular form of the older factor analysis tradition, and drawing an uncluttered path diagram containing only the latent variables (or any observable variables included in the path model), with their arcs and disturbance terms. This allows a much clearer appreciation of the relations postulated in the path model than if the diagram is complicated by the measurement relationships.

It cannot be claimed that there is a "right" way to draw a path diagram. There appear to be three main conventions currently available, in addition to the simple picture of a graph favored by DAG theorists (e.g., Pearl, 2000). (A number of minor variants in use can be characterized by incompleteness.)

Figures 1–3, which are largely self-explanatory, illustrate these. (In applications, a diagram of the theoretical model also contains parameter names close to the arcs, and a diagram of output has the parameter values—and possibly standard errors in parentheses—similarly associated with the arcs.) Convention 1 (Figure 1), due to McArdle (1980), most closely corresponds to the picture of a graph used in DAG theory. It does not seem to have been widely adopted, possibly because it uses a nondirected arc between two variables (latent or observable) to represent the covariance of their disturbance or error terms, and a closed loop for variance. This could seem confusing to some users, and logical to others. Convention 2 (Figure 2) graphs the disturbance and error terms, together with the observed and latent variables, with unit path coefficients from the former to the latter. This is a complete counterpart of the equations of the model. In this version, on the face of it, error terms are exogenous latent variables. Convention 3 (Figure 3) distinguishes disturbances and error terms from latent variables. However, its mode of representation makes it at least inconvenient to add values of error variances or covariances (and their standard errors) to the diagram. Possibly this convention actively serves to discourage users from presenting error and disturbance variances or from including disturbance covariances in the path model they draw.

Of the 41 cases, 1 used Convention 1, 4 used Convention 2, 5 used Convention 3, and the rest exhibited varying degrees of incompleteness, or in a few cases did not follow any established convention. Of the 10 with complete path diagrams, 8 gave complete sets of parameters. Of the 31 with incomplete path diagrams, 6 gave all parameters. A simple chi-square test gives a significant association, but does not establish that the use of an incomplete path diagram is the cause of an incomplete account of the parameters!

As a recommendation, we simply repeat the truism that the report should give all parameters and their standard errors. Although the method of presentation can be regarded as a matter of personal taste, it should be consciously chosen with recognition of its advan-

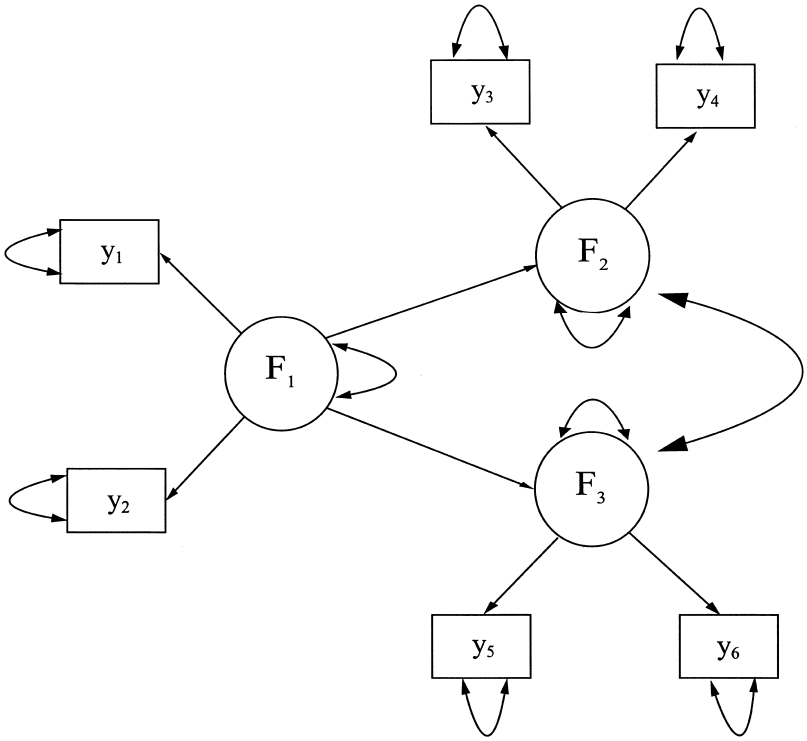

Figure 1. Path diagram, Convention 1: Disturbance variances modeled as closed, nondirected arcs.

tages and disadvantages, and there should be no ambiguity from the reader's viewpoint.

Alternative and Equivalent Models

For any set of multivariate data there will almost always be more than one plausible structural model. Both misfit and overfit of a single target model can be taken to imply model modification (adding or deleting parameters). Model modification indices, supplied by a number of the available computer programs, can be used to generate alternative models mechanically.

Seminal work by Stelzl (1986) and by Verma and Pearl (1990) has shown how to generate alternative, equivalent SEMs that cannot be distinguished empirically. MacCallum et al.'s (1993) review should have drawn the question of equivalent models to the attention of a wide range of SEM users, with a reasonably immediate influence on the practice of SEM.

Tests of significance on differences in chi-squares were used in 33 of the 41 studies to choose among nested models; 6 of these were aided by Wald tests. Only 4 of the 41 studies surveyed gave some recognition to the existence of equivalent models (2 from 1996 and 2 from 1997) from a set with 7 in 1995, 16 in 1996, and 17 in 1997. We thus add further evidence to that given by MacCallum et al. (1993) of persistent neglect of this question.

Endorsing remarks by Hoyle and Panter (1995) and by Boomsma (2000), we note that desirably, plausible competing models (nested or nonnested) should be specified a priori by the investigator, along with the preferred "target" model, and the relative goodness of fit of those reported. Readers are reminded of an important study by MacCallum, Roznowski, and Necowitz (1992), which advised against taking a quasi-random walk through a sequence of models generated mechanically by model-modification indices. In view of earlier remarks advocating an examination of discrepancies as a basis for judging fit, we note that this warning would also apply to the mechanical use of large standardized discrepancies for this purpose. At the least, theoretical justification of model modifications needs to be found and reported. We suspect that post facto this task is commonly all too easy. As MacCallum et al. pointed out, the prior specification of a set of alternative models would generally be safer than model modification after the facts

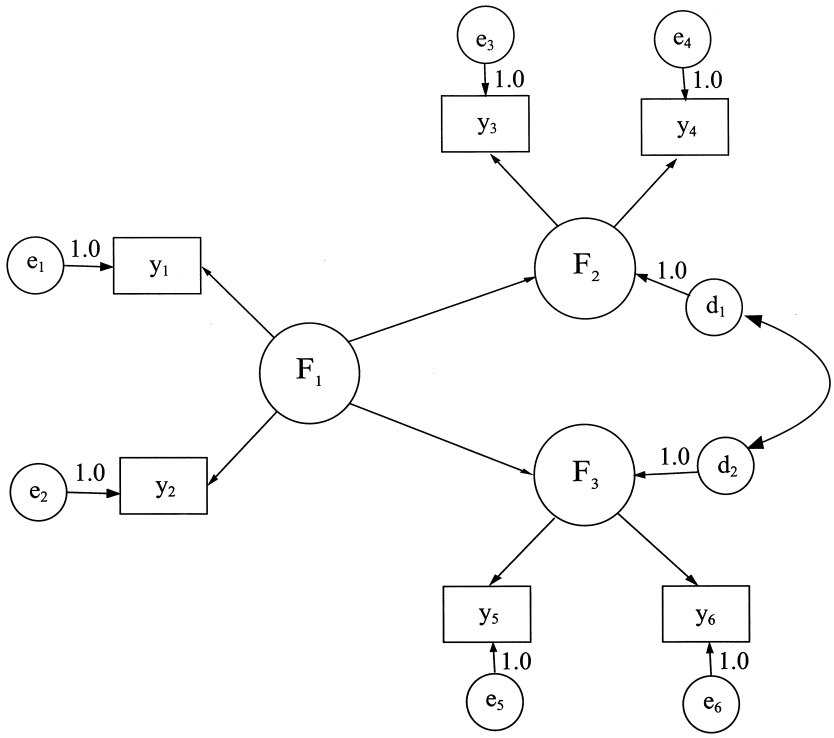

Figure 2. Path diagram, Convention 2: Disturbances modeled as latent variables.

(and, as noted, Hoyle & Panter, 1995, and Boomsma, 2000, are in agreement).

A more tolerant recommendation would allow the researcher a few modifications of an initial model, provided that a clear theoretical justification can be found, and provided that a clear history of the decision steps is given to the reader, who may not share the authors' enthusiasm for their conclusions. Both Hoyle and Panter (1995) and Boomsma (2000) specifically advised against adding disturbance covariance to improve fit. We endorse the general principle that model modification should not be purely data-driven. However, as noted previously, the addition of nondirected arcs does not change the causal model, and a nondirected arc corresponds to a specified or unspecified omitted variable. It is not unreasonable to add such an arc if a plausible theory can be suggested for it and further work is implied that explores the possibility of measuring it.

Conclusion

We claim no special authority to offer a code of practice for the reporting of SEM results. Recommendations, with varying degrees of confidence, have been offered at the end of each section of this article. Unquestionably, practice could be greatly improved simply by a conscious intention on the part of investigators to make reasoned choices on each aspect of their work and to report the basis of those choices. Such choices include decisions to report sample covariances (or correlations with standard deviations) and standardized discrepancies, or at the very least to indicate means for the reader to access this information. The report should certainly include all the parameters and their standard errors. There should also be a reasoned choice of a clear and complete form of the path model structure, reported possibly as a conventional path diagram, and of the measurement model, reported possibly in traditional tabular form. Completeness is essential. The form of representation remains a matter of taste, as long as it leaves no ambiguity. Again, a careful rational choice is needed for conclusions about fit of the structural model, with separate attention to its measurement and path components, and attention to individual discrepancies. Global indices of fit, with criteria for their acceptability, do not in the present state of knowledge substitute for a more detailed examination and careful judgment. Investigators should also provide reasoned justifica-

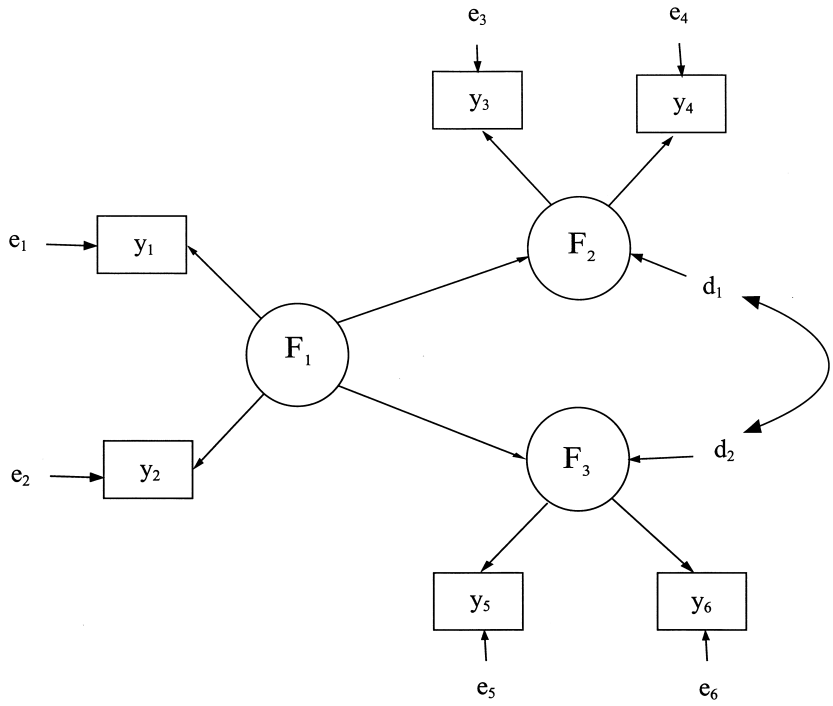

Figure 3. Path diagram, Convention 3: Disturbances distinguished from latent variables.

tions for omitted directed and nondirected arcs, which jointly create a testable model, paying attention to substantively plausible alternative and equivalent models. An explicit account of the identifiability of the model is very desirable, although in many cases the careful reader will be able to determine this by inspection.

SEM has been characterized as "a dangerously conjectural technique for asking essential research questions which otherwise are impossible to consider" (McDonald, 1999, p. 367). This review indicates some steps that can be taken to reduce the manifest dangers accompanying its use.

References

Allison, S. (1987). Estimation of linear models with incomplete data. In C. C. Clogg (Ed.), *Sociological methodology* (pp. 71–103). San Francisco: Jossey-Bass.

Amemiya, Y., & Anderson, T. W. (1990). Asymptotic chi-square tests for a large class of factor analysis models. *Annals of Statistics, 18,* 1453–1463.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411–423.

Anderson, J. C., & Gerbing, D. W. (1992). Assumptions and comparative strengths of the two-step approach. *Sociological Methods and Research, 20,* 321–333.

Anderson, T. W. (1989). Linear latent variable models and covariance structures. *Journal of Econometrics, 41,* 91–119.

Arbuckle, J. L. (1997). *Amos user's guide.* Chicago: Small-Waters.

Bekker, P. A., Merckens, A., & Wansbeek, T. J. (1994). *Identification, equivalent models, and computer algebra.* San Diego, CA: Academic Press.

Bentler, P. M. (1985). *Theory and implementation of EQS: A structural equations program.* Los Angeles: BMDP Statistical Software.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246.

Bentler, P. M. (1995). *EQS structural equations programs manual.* Encino, CA: Multivariate Software, Inc.

Bentler, P. M. (2000). Rites, wrong, and gold in model testing. *Structural Equation Modeling, 7,* 82–91.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Bentler, P. M., & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Reviews of Psychology, 47,* 563–592.

Bentler, P. M., & Weeks, D. G. (1980). Linear structural equations with latent variables. *Psychometrika, 45,* 289–308.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bollen, K. A. (2000). Modeling strategies: In search of the Holy Grail. *Structural Equation Modeling, 7,* 74–81.

Bollen, K. A., & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 235–262). Cambridge, MA: Blackwell.

Bollen, K. A., & Long, J. S. (Eds.). (1993). *Testing structural equation models.* Thousand Oaks, CA: Sage.

Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling, 7,* 461–483.

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin, 107,* 260–273.

Browne, M. W. (1984). Asymptotically distribution free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62–83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Thousand Oaks, CA: Sage.

Browne, M. W., & Du Toit, S. H. C. (1992). Automated fitting of nonstandard models. *Multivariate Behavioral Research, 27,* 269–300.

Browne, M. W., Mels, G., & Cowan, M. (1994). *Path analysis: RAMONA: SYSTAT for DOS advanced applications* (Version 6, pp. 167–224). Evanston, IL: SYSTAT.

Browne, M. W., & Shapiro, A. (1988). Robustness of normal theory methods in the analysis of linear latent variate models. *British Journal of Mathematical and Statistical Psychology, 41,* 193–208.

Chou, C. P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 44,* 347–357.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1,* 16–29.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39,* 1–38.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8,* 430–457.

Fisher, F. M. (1970). A correspondence principle for simultaneous equation models. *Econometrica, 38,* 73–92.

Fornell, C., & Yi, Y.-J. (1992). Assumptions of the two-step approach to latent variable modeling. *Sociological Methods and Research, 20,* 291–320.

Fraser, C., & McDonald, R. P. (1988). COSAN: Covariance structure analysis. *Multivariate Behavioral Research, 23,* 263–265.

Gollob, H. F., & Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Development, 58,* 80–92.

Graham, J. W., & Hofer, S. M. (2000). Multiple imputation in multivariate research. In T. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches and specific examples* (pp. 201–218). Mahwah, NJ: Erlbaum.

Hartmann, W. M. (1992). *The CALIS procedure: Extended user's guide.* Cary, NC: SAS Institute.

Hayduk, L. A., & Glaser, D. N. (2000a). Jiving the four-step waltzing around factor analysis, and other serious fun. *Structural Equation Modeling, 7,* 1–35.

Hayduk, L. A., & Glaser, D. N. (2000b). Doing the four-step, right 2-3, wrong 2-3: A brief reply to Mulaik and Millsap; Bollen; Bentler; and Herting and Costner. *Structural Equation Modeling, 7,* 111–123.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrika, 45,* 153–161.

Heckman, J. J., & Robb, R. (1986). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In H. Wainer (Ed.), *Drawing inference from self-selected samples* (pp. 63–107). New York: Springer.

Herting, J. R., & Costner, H. L. (2000). Another perspective on "the proper number of factors" and the appropriate number of steps. *Structural Equation Modeling, 7,* 92–110.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 158–176). Thousand Oaks, CA: Sage.

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 76–99). Thousand Oaks, CA: Sage.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112,* 351–362.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85–112). New York: Seminar Press.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd ed.). Chicago: SPSS Inc.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: Users' reference guide.* Chicago: Scientific Software International.

Kano, Y. (1992). Robust statistics for test-of-independence and related structural models. *Statistics and Probability Letters, 15,* 21–26.

Loehlin, J. C. (1992). *Latent variable models: An introduction to factor, path, and structural analysis.* Mahwah, NJ: Erlbaum.

Lohmoller, J. B. (1981). *LVPLS 1.6 program manual: Latent variables path analysis with partial least squares estimation.* Munich, Germany: Hochschule der Bundeswehr.

Long, J. S. (1983a). *Confirmatory factor analysis: A preface to LISREL.* Thousand Oaks, CA: Sage.

Long, J. S. (1983b). *Covariance structure models: An introduction to LISREL.* Thousand Oaks, CA: Sage.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51,* 201–226.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111,* 490–504.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure models. *Psychological Bulletin, 114,* 185–199.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57,* 519–530.

McArdle, J. J. (1980). Causal modeling applied to psychonomic systems simulation. *Behavior Research Methods and Instrumentation, 12,* 193–207.

McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research, 29,* 409–454.

McArdle, J. J., & McDonald, R. P. (1984). Some algebraic properties of the reticular action model for moment structures. *British Journal of Mathematical and Statistical Psychology, 37,* 234–251.

McDonald, R. P. (1978). A simple comprehensive model for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 31,* 59–72.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6,* 97–103.

McDonald, R. P. (1996). Path analysis with composite variables. *Multivariate Behavioral Research, 31,* 239–270.

McDonald, R. P. (1997). Haldane's lungs: A case study in path analysis. *Multivariate Behavioral Research, 32,* 1–38.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

McDonald, R. P. (in press). What can we learn from the path equations? *Psychometrika.*

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin, 107,* 247–255.

McDonald, R. P., Parker, P. M., & Ishizuka, T. (1993). A scale-invariant treatment for recursive models. *Psychometrika, 58,* 431–443.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 29,* 177–185.

Micceri, T. (1989). The unicorn, the normal curve and other improbable creatures. *Psychological Bulletin, 105,* 156–165.

Mooijaart, A. (1993). Structural equation models with transformed variables. In K. Haagen, D. J. Bartholomew, & M. Deistler (Eds.), *Statistical modeling and latent variables* (pp. 249–258). Amsterdam: North Holland.

Mulaik, S. A., & Millsap, R. E. (2000). Doing the four-step right. *Structural Equation Modeling, 7,* 36–74.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika, 49,* 115–132.

Muthén, B. (1989). Multiple group structural modeling with non-normal continuous variables. *British Journal of Mathematical and Statistical Psychology, 42,* 55–62.

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45,* 19–30.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52,* 431–462.

Muthén, B., & Muthén, L. (1998). *Mplus user's guide.* Los Angeles, CA: Muthén & Muthén.

Neale, M. C. (1997). *Mx: Statistical modeling* (4th ed.). Richmond, VA: Author.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* San Mateo, CA: Morgan Kaufman.

Pearl, J. (1998). Graphs, causality, and structural equation

models. *Sociological Methods and Research, 27,* 226–284.

Pearl, J. (2000). *Causality: Models, reasoning and inference.* Cambridge, England: Cambridge University Press.

Raykov, T., Tomer, A., & Nesselroade, J. R. (1991). Reporting structural equation modeling results in *Psychology and Aging:* Some proposed guidelines. *Psychology and Aging, 6,* 499–503.

Rubin, D. (1976). Inference and missing data. *Biometrika, 63,* 581–592.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rubin, D. (1991). EM and beyond. *Psychometrika, 56,* 241–254.

Satorra, A., & Bentler, P. M. (1994). Corrections to standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications to developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Scheines, R., Spirtes, P., Glymour, C., Meek, C., & Richardson, T. (1998). The TETRAD project: Constraint-based aids to causal model specification. *Multivariate Behavioral Research, 33,* 65–117.

Shapiro, A., & Browne, M. W. (1987). Analysis of covariance structures under elliptical distributions. *Journal of the American Statistical Association, 82,* 1092–1097.

Spirtes, P., Richardson, T., Meek, C., Scheines, R., & Glymour, C. (1998). Path diagrams as a structural equation modeling tool. *Sociological Methods and Research, 27,* 182–225.

Steiger, J. H. (1988). Aspects of person–machine communication in structural modeling of correlations and covariances. *Multivariate Behavioral Research, 23,* 281–290.

Steiger, J. H. (1995). *SEPATH: STATISTICA version 5.* Tulsa: StatSoft, Inc.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50,* 253–264.

Stelzl, I. (1986). Changing a causal hypothesis without changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research, 21,* 309–331.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.

Verma, T., & Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the fifth conference on uncertainty in artificial intelligence* (pp. 220–227). Amsterdam: Elsevier.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with non-normal variables. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 56–75). Thousand Oaks, CA: Sage.

Wiley, D. E. (1973). The identification problem for structural equation models with unmeasured variables. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 69–83). New York: Seminar Press.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20,* 557–585.

Yuan, K., Chan, W., & Bentler, P. F. (2000). Robust transformation with applications to structural equation modeling. *British Journal of Mathematical and Statistical Psychology, 53,* 31–50.

Received May 12, 2000
Revision received October 29, 2001
Accepted November 1, 2001 ■