STRUCTURAL EQUATION MODELING, *7*(3), 461–483

TEACHER'S CORNER

Reporting Analyses of Covariance Structures

Anne Boomsma
Department of Statistics, Measurement Theory & Information Technology
University of Groningen

This contribution is focused on how to write a research paper when structural equation models are being used in empirical work. The main question to be answered is what information should be reported and what results can be deleted without much loss of judgment about the quality of research and the validity of conclusions being made. The major conjecture is that all information should be reported, or referred to, that enables each member of the scientific community, at least in principle, to replicate the analysis as it is published. The recommendations are ordered in the framework of the empirical research cycle. They are meant for authors, in particular students employing structural equation models for their dissertation, as well as for editors and reviewers.

PRELIMINARY

When reading applied research concerning structural equation modeling (SEM), it is often difficult to judge its merits. This often occurs for several reasons: for instance, a lack of theoretical foundations for the postulated structural relations, inaccurate description of the model or the applied estimation methods, a lack of reporting the psychometric properties of scales, not providing the sample data or not even mentioning the sample size, giving an obscure delineation of the model modification process, or not describing the population under study. These deficiencies and lack of information make it very difficult and even annoying to evaluate the quality

Requests for reprints should be sent to Anne Boomsma, Department of Statistics, Measurement Theory & Information Technology, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands. E-mail: a.boomsma@ppsw.rug.nl

of such research because a proper scientific assessment cannot be done at all. Ideally the reader should be in a position to replicate the reported study. Frequently, however, this cannot be done.

How can such circumstances, which impair proper evaluation and replication of SEM, be avoided? A first step is to summarize the basic pieces of information that should be included in a publication about applied SEM methodology. A second step is to pay attention to some fundamental issues in doing that type of research, which should be reflected in its reporting. The purpose of this article is to offer some guidelines, or recommendations, regarding the content and organization of publications on applications of SEM. First some related publications are mentioned.

General directions for research publications can, for instance, be found in the *Publication Manual of the American Psychological Association* (1994). Since the release of this latest edition, however, progress has been made in terms of selecting proper fit indexes, hence hopefully the fledgling researcher will not only report what is enumerated in the manual (i.e., NFI, PFI; o.c., p. 134). Maxwell and Cole (1995) offered some general tips for writing methodological articles. They also highlighted several ways in which authors can make their methodological work more accessible, and less painful, as they put it, to readers who are not methodological experts. Raykov, Tomer, and Nesselroade (1991) proposed some guidelines for reporting SEM results in articles submitted to *Psychology and Aging*. Hoyle and Panter (1995) published a chapter on how to write about structural equation models, emphasizing the description of results of the analysis and model fit criteria in particular. Some of the material covered by the latter two publications coincides with the approach taken here.

The ordering of the contents of this article mirrors the empirical research cycle (observation, induction, deduction, testing, and evaluation) as discussed by De Groot (1966, 1969). In describing the different phases of this cycle, usually more general terms are employed: *Introduction* (the problem, theory, model, or hypotheses), *Method* (how to attack the problem, how to test the model), *Results* (estimates of model parameters and model fit), and *Discussion* (evaluating the results of the analysis). Figure 1 highlights main features of the research cycle in SEM.

With regard to Method and Results, the general stand is taken that all information should be reported that enables a researcher to replicate the published empirical research. If that is not possible in all detail, it should be indicated how the missing information can be obtained from the author. For example, references can be made to technical reports and unpublished manuscripts containing the sufficient statistics and more detailed information, or to the availability of data files on the World Wide Web or elsewhere (with reliable postal, e-mail, and Internet addresses being essential). Such publication standards serve to ensure the possibility of replication and thus of well-founded criticism and discussion. One of the criteria for manuscript consideration therefore can be that the model be replicated by an objective source, so as to prevent any flaws that might otherwise occur. If the crite-

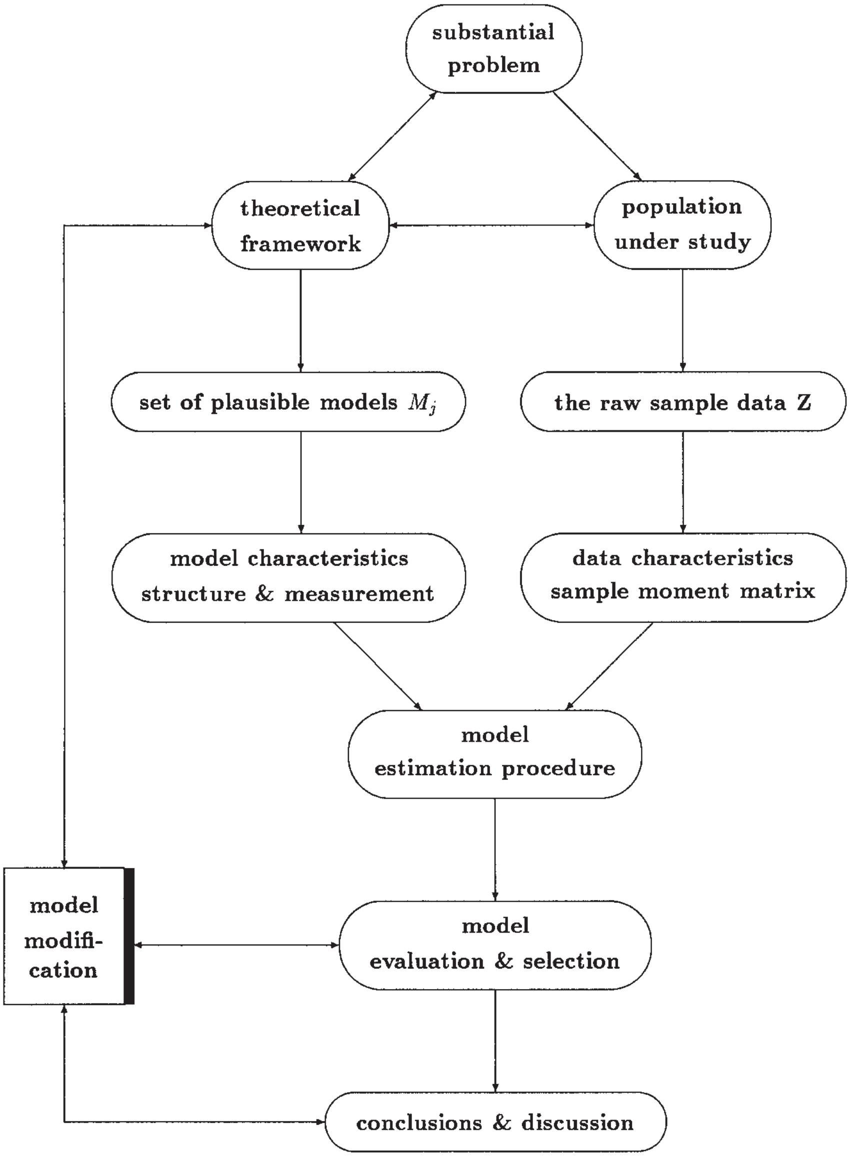

FIGURE 1 Flow diagram of the SEM process.

rion of replication cannot be met, the manuscript might as well be sent to the *Journal of Irreproducible Results* or to its successor, *The Annals of Improbable Research*.

Frequently, editors of scientific journals will urge authors to be concise, which forces the latter to shorten their manuscript more or less drastically. In deleting material from a manuscript, a basic criterion to consider again is that fellow researchers must be able to replicate the analysis. The issue of being forced by editors to reduce the size of a paper may well serve to emphasize that it is not only the responsibility of authors, but even more strongly that of editors and other reviewers, to ascertain that the criterion of potential replication is being met. The recommendations or guidelines that follow would lose their prescriptive impact if only authors, but not the editors and reviewers involved, would keep up some basic standards of publication.

In writing about the analysis of structural equation models and trying to have its results published, authors could also benefit from the recommendations of Abelson (1995). He introduced the MAGIC criteria that govern the persuasive force while presenting the theory, the data, and the statistical and substantive analysis of the problem. He labels his five criteria as magnitude of effects, articulation, generality, interestingness, and credibility. Regarding the first four of these MAGIC criteria, he claims, “A good rule of thumb—the *rule of two criticisms*—is that two deficiencies among these four criteria will result in rejection by journal editors” (p. 170). Authors are warned not to neglect the implicit advice too easily.

INTRODUCING THE PROBLEM

In the introduction of a research paper about the application of structural equation models, usually some summarized background information is given of crucial results related to the problem under study. This means that a substantive background regarding the state of the art of the research subject is presented. In this way, the research questions under consideration are placed in a theoretical framework that fits closely the empirical knowledge gained so far. The goal of the introduction is to make it crystal clear what research questions are to be answered in the sequel, and what their potential importance is for the nomological network of theoretical knowledge (cf. De Groot, 1966, 1969).

THE THEORY

Next, but closely following the introduction of the research problem, a brief account is given of the substantive, theoretical foundations of the model, or a set of models, to be analyzed. In principle, this means that a theoretical justification is given for the imposed directed, structural relations between the constructs or variables of interest. This in turn implies that the direction, the sign, as well as the expected strength of such relations are discussed from a substantive point of view, as far and concise as possible. In the spirit of meta-analysis and Abelson's (1995) chapter on the magnitude of effects, speculating on the effect sizes with a confidence interval would further provide concise evidence of the researcher's confidence and expertise regarding the postulated model. Some justification for the nature of the proposed functional relations between variables could also be given. Feasible questions to be asked in this context are related to the validity of hypothesized linear or nonlinear relations, as well as the introduction of nonrecursive paths (cycles); see Baumgartner and Bagozzi (1995) or Yang Jonsson (1997, chapter 7) for an example.

Given the research questions posed in the introduction, it can be clarified—and that need not always be obvious—why it is useful or necessary to apply SEM. Why are competitive methods of analysis less appropriate to answer the research questions? For example, it can be elucidated why a simple correlation or classical regression type of approach is inadequate in the current research setting. This might give the reader an understanding why relatively complex approaches are preferred, over simpler, more familiar ones, for properly tackling the problem. See also the comments of the Task Force on Statistical Inference regarding the use of minimally sufficient designs and analytic strategies and issues related to computerized data analysis (the initial report of the Task Force of the American Psychological Association can be found at URL http://www.apa.org/science/tfsi.html).

Although some design details can be postponed, the introduction must clarify what kind of population the theory—the model, that is—applies to. The particular population of objects the research is concerned with has to be specified (e.g., men and women between 20 and 45 years old, growing up in Western societies). This clarification is important for the sampling design, as well as for the possible generalizibility of conclusions based on the current analysis.

THE SET OF MODELS UNDER STUDY

Often some basic model M_j for the structural relations is postulated (i.e., only one model is considered), which implies a strict confirmatory statistical analysis is being made. Frequently, however, it makes more sense not to analyze just a single model, but rather a set of plausible models $M_1, M_2, \ldots, M_m$ (possibly in some preferred sequence). From a modeling point of view, the latter implies a less confirmatory type of analysis, because no single model is preferred to any other beforehand: Each of them is defensible from a theoretical perspective and potentially adequate to fit the sample data.

At any rate, the complete set of models to be analyzed should be presented and justified in advance. For each model in this set, its relation to the underlying theory should be accounted for. It is also important to know whether the set is a nested sequence. If the set is nested, it is recommended to order the models from simple to more complex ones, that is, from parsimonious to complicated models.

At early stages of theory development, the researcher should prefer simple models, which reflect basic theoretical notions only. At a later stage, however, it is conceivable to study more elaborate models in the spirit of R. A. Fisher, as strikingly quoted by Cochran (1965): "when constructing a causal hypothesis one should envisage as many *different* consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold" (p. 252)—such an approach is generally less emphasized in SEM. Fisher's notion to elaborate theories is related to the concept of coherence and to the construction of focused hypotheses; see Rosenbaum (1995, chapter 8) for a discussion in the context of observational studies. Seen as risky predictions, the resulting rather specialized models are in general advocated by Popper (1974, 1983), as Rosenbaum demonstrated.

If no well-documented theory exists, or if hardly any explicit ideas about the underlying covariance structure can be formulated, undoubtedly the analysis is exploratory. Because of its statistical consequences, the position of the researcher on a confirmatory–exploratory type of dimension must be apparent from the start. This position has implications for the reliability and validity of the final conclusions as well: Exploratory analyses require replication or cross-validation, either by analysis of an alternative data set or by bootstrapping and other statistical techniques. As difficult as it may be, in the end some information about the stability of the results (conclusions) should be available.

Whether the researcher should specify or consider a set of equivalent models before collecting the sample data, as suggested by Hershberger (1995) following Stelzl (1986), is disputable (the term *equivalence* is used here in its narrow sense; Hayduk, 1996). Others might prefer to identify equivalent models during the estimation and model modification phase of the analysis. How to attack this problem is a matter of theoretical knowledge and methodological efficiency. It should thus be discouraged to specify equivalent models by relying purely on mathematics. If specific equivalent models are a priori invalid from a theoretical perspective, it makes no sense to incorporate them in the set of potential models or to consider them in a process of model modification. If the set of equivalent models is unknown, action might be taken to explore that set, for example by using the TETRAD program (Scheines, Spirtes, Glymour, & Meek, 1994); see Hayduk (1996), Freedman (1997), and Scheines, Spirtes, Glymour, Meek, and Richardson (1998) for discussions about the TETRAD approach.

At some point, either after presenting the set of models or just before discussing the estimation procedure, attention should be given to the question of identification. It is the task of the researcher to try and examine whether a model is theoretically identified, which may or may not be a hard job to do. Efforts should be made to check whether necessary and sufficient conditions exists for identifiability. See, for example, Bollen (1989), Bekker (1994), and the book by Bekker, Merckens, and Wansbeek (1994) along with its software to control for conditions of identification. The advice is to try and reinstall a declining tradition to solve the identification problem *before* estimating the model. In addition, during the estimation phase of the analysis, as much information as possible should be gathered about the (empirical) identification of the model, for example by checking the occurrence of improper estimates (Heywood cases) and symptoms of multicollinearity (cf. Rindskopf, 1984).

STRUCTURAL AND MEASUREMENT PARTS OF THE MODEL

By definition, latent variables are labels for the hypothetical constructs or theoretical concepts under study. As far as these latent variables do not coincide with observed, measured variables, a difference can be made between the structural and the measurement part of the model. If that is the case, as it most often will be, it should be indicated unequivocally how latent variables are being measured, that is, what the indicators for each of the latent variables are. Whenever possible, theoretical justifications for the decisions being made here are reported briefly. The position taken is that measurements of latent variables are part of the theory, albeit its measurement part.

If available, known aspects of the validity and reliability of the measurements are mentioned or referred to. Composite reliability and discriminant and construct validity are important issues in the establishment of durable constructs in a nomological network of knowledge (e.g., Bollen [1989] and Raykov [1997]). Also, if a specific instrument is being used, it should be clear whether the indicators are single items, total scores, or some other scale measure of that instrument. Familiarity with the effects of the number of indicators per factor and the degree of item parceling on model estimates is incumbent (Marsh, Hau, Balla, & Grayson, 1998).

The way in which latent variables are scaled should also be reported: either direct scaling of latent variables to a variance of one, or indirect scaling by fixation of factor loadings of specific observed variables (the researcher should indicate which ones). Thus, based on a description of the research, information is now available concerning what the vector of k observed variables $\boldsymbol{z} = (\boldsymbol{y}', \boldsymbol{x}')'$ and what the postulated relations between observed and latent variables in model M_j ($j = 1, 2, \ldots, m$) look like from a substantive point of view.

At least for the basic model in the set of models under consideration, a path diagram should be presented rather than the mathematical model equations, which are far more difficult to grasp. The path diagram should be complete; that is, it should include both the structural and measurement errors in the model and possible covariances among them. It is conceivable, however, that the model is so complex that it is impossible to display it completely, or that editorial policies impose certain restrictions here. In such cases, it remains necessary that the reader can deduce the complete model from its verbal or mathematical description.

Clearly, a path diagram tends to show the presence of hypothesized effects. It should be emphasized, however, that in developing a model, and in reporting that process, paths that are absent are as important as those that are present. In practice, authors should make every effort to justify the absence of effects: Presence of relations is, in principal, not more obvious than lack of associations; both need theoretical justification and empirical backup.

THE POPULATION AND THE SAMPLE

Given the model under study, the population covariance matrix $\Sigma = [\sigma_{ij}]$ $(k \times k)$ of the observed, random variables $\boldsymbol{z} = (\boldsymbol{y}', \boldsymbol{x}')'$ can be written as a function of the parameters of the model. That is, each element σ_{ij} can be written as a function of the unknown population parameters $\theta(t \times 1)$; in short $\Sigma = \Sigma(\theta)$. Given the preceding model description, the reader should be able to do this; otherwise crucial information is lacking.

In principle, a random sample of size N is taken from the population under study, which has covariance structure $\Sigma = \Sigma(\theta)$ by assumption. First, the population from which a sample is taken is described; an approximation of its size and nature should be given in terms of relevant background variables. As indicated earlier, this is all very important for a proper understanding of the type of generalizations that can be made after, or in the process of, formulating substantive conclusions. The sample size N should always be mentioned. Aspects of the sampling procedure could be summarized next, for example, the sampling procedure not being random or being stratified and the response rate. In case of multisample or multigroup analyses (cf. Jöreskog & Sörbom, 1996b), the information referred to previously is reported for each group. This also applies when multilevel structural equation models are analyzed (cf. Ernste, 1996; Hox, 1994; Jöreskog, Sörbom, Du Toit, & Du Toit, 1999; McArdle & Hamagami, 1996; Muthén, 1994).

FEATURES OF THE OBSERVATIONS

Features of the distribution of observations in the sample $\boldsymbol{Z}$ $(N \times k)$ should be described briefly. The ultimate goal here is to justify the plausibility of distributional assumptions of the estimation method that has to be selected in relation to such features. The measurement level of the observations (interval, ordinal, or categorical) is one such feature; the number of categories of ordinal variables is another one. Some further examples and possibilities are enumerated now.

If the assumption of multivariate normality of the observed variables is made, information about normality characteristics based on the sample of observations can be presented (for example, descriptive statistics like skewness and kurtosis, or test statistics for univariate or multivariate normality). If transformations of observed variables are made, they should also be described (purpose and formulas). The LISREL 8.30 program has an option to normalize variables before analysis, thus providing a way to deal with nonnormality in samples of small and moderate size (Jöreskog et al., 1999).

If the asymptotically distribution-free or weighted least squares estimation method is chosen, apart from the number of categories of ordinal and nominal variables, descriptive measures of skewness and kurtosis could be of interest here too, because their size has consequences also for the robustness of that estimation method (see Hoogland & Boomsma, 1998). And if polychoric correlations are used (or if variables are clearly censored), the assumptions regarding the distribution of underlying variables might be checked.

The treatment of missing values, if any, should be reported. In the structural modeling programs Amos (Arbuckle, 1997) and Mx (Neale, 1994) a direct, model-based, maximum likelihood (ML) estimation procedure is implemented; see Arbuckle (1996) for details. The M*plus* program (Muthén & Muthén, 1998) also uses an ML approach for missing values. The research of Verleye (1996) and Duncan, Duncan, and Li (1998) showed that this ML procedure yields comparatively good estimates. So far, programs for covariance structure analysis have no options to employ the multiple imputation approach to missing data straightforwardly (Little & Rubin, 1987, 1990; Rubin, 1987, 1996; Schafer, 1996): Additional software is needed there. Such indirect methods are therefore more cumbersome; they also seem to be less efficient than the direct ML approach (Duncan et al., 1998; Verleye, 1996). If imputation techniques are employed, details should be given or referred to. See also Brown (1994) for comparisons of different procedures for handling missing data and Marsh (1998) on possible problems with pairwise deletion procedures in SEM.

Criteria or arguments for deleting outlying observations should be provided along with the number of deleted cases, if any. Often, deletion of outlying observations is not an easy task; it should be handled carefully (see Barnett & Lewis, 1994). Some users might find it helpful here when programs calculate histograms, or univariate and multivariate measures for skewness and kurtosis; the PRELIS program (Jöreskog & Sörbom, 1996c), for example, has many options. The EQS program exposes the five case numbers with the largest contributions to a Mardia-based normalized multivariate kurtosis coefficient (Bentler, 1995). In the same context, it might be considered to take advantage of likelihood-based procedures that replace ordinary sample covariances by robust estimates of covariances (Yuan & Bentler, 1998b, 1998c).

In case of applications of structural equation models to clustered sampling designs, attention should be given to not only the sample sizes of the different groups, but also to the intraclass correlations so as to judge whether a multilevel analysis is appropriate (see Muthén [1994] and Kaplan & Elliott [1997] for strategies and examples). The features of the observations, which include the sample size, can (or even should) have a strong impact on the choice of the estimation method. Neglecting such features can have serious consequences, because biased estimates may very well lead to wrong conclusions.

THE MOMENT MATRIX TO BE ANALYZED

There are several types of moment matrices, for example, a covariance matrix **S**, a correlation matrix **R**, or an augmented moment matrix **A**. The type of moment matrix to be analyzed always has to be specified, and if population means are analyzed along with covariances, this should be brought up as well. If the number of observed variables *k* is not too large, the moment matrix (and the vector of means) must be presented in a table, and if it is too large, then the data should be made available via the Internet or by contacting the author. At any rate, the reader should be able to inspect these sample data or have access to them in principle. It has to be known also whether the sample size on which the moment matrix is based equals the size of the original sample, or whether it is smaller due to missing observations or to deletion of outlying observations.

If a correlation matrix **R** is analyzed, the reasons for doing so should be mentioned (cf. Cudeck, 1989; Jöreskog & Sörbom, 1989). In this context, it would also be valuable to know whether the model is scale invariant and whether parameters are scale-free (cf. Bollen, 1989; Jöreskog & Sörbom, 1989). If **R** is being analyzed, one might still be interested in the analysis of the corresponding covariance matrix **S**, at least as far as that makes sense given the measurement scale of variables. If appropriate, it is therefore recommended also to report the vector of standard deviations of the observed variables along with **R**. But even if the covariance matrix is the unit of analysis, many readers and researchers still appreciate the publication of the correlation matrix as it is much easier to get a feeling for intercorrelations of variables via **R** as opposed to **S**. Therefore, in general it is recommended to publish the correlation matrix along with standard deviations, if appropriate.

THE ESTIMATION PROCEDURE

Given the model under study and the plausibility of the statistical assumptions to be made in light of variable characteristics (level of measurement), the distributional features of the sample from the population (including sample size *N*), and the moment matrix to be analyzed, an appropriate choice of the estimation method is made. In making such a decision, robustness questions play a crucial role; see Hoogland and Boomsma (1998) for an overview. The latter amounts to trying to answer the question of how robust the substantive conclusions of the analysis are against violations of the statistical assumptions of potential estimation methods.

The method of ML assumes multivariate normality of the observed variables and not too small a sample size (several hundred). Using the LISREL or EQS program, the same holds for the generalized least squares estimation method (cf. Bentler, 1995; Jöreskog & Sörbom, 1996b). On the other hand, the distribution-free or weighted least squares method needs a very large sample: in general several thousand if the number of variables $k \geq 15$; however, for a small model with observed variables having no strong kurtosis $N = 800$ seems large enough. See, for example, Hu, Bentler, and Kano (1992), West, Finch, and Curran (1995), Curran, West, and Finch (1996), Hoogland and Boomsma (1998), and Hoogland (1999). In view of these previous general requirements, it is important to have knowledge of the recent work of Marsh et al. (1998) and Marsh and Hau (1999), confirming the findings of Boomsma (1985). Their results indicate that a higher number of indicators per factor ratio in confirmatory factor analysis may compensate for small *N*, and larger *N* may compensate for a small number of indicators per factor ratio.

Nowadays, programs like EQS and LISREL (back to version 8.20) have the nice feature of calculating scaled test statistics and so-called asymptotically robust estimates of standard errors. These statistics appear to be (asymptotically) robust against deviations from normality (see Browne, 1984; Chou & Bentler, 1995; Chou, Bentler, & Satorra, 1991; Satorra, 1990, 1992, 1993; Satorra & Bentler, 1994). Whenever appropriate, such estimates should be preferred over less robust estimates; among other factors, the preference for robust estimates depends on the sample size and the complexity of the model (cf. Hu et al., 1992). For more recent developments in this area, see Yuan and Bentler (1995, 1997, 1998a; and Bentler & Yuan, 1999).

It should always be mentioned explicitly which computer program was used to estimate the model. This should include not only the name of the program, but also its version number (e.g., LISREL 8.30), because estimation results may differ from version to version. If the weighted least squares estimation method is used in LISREL, the PRELIS version (Jöreskog & Sörbom, 1996c) also has to be reported, because this preprocessor program produces the required estimates of the asymptotic covariances of elements of the moment matrix to be analyzed.

If the researcher deviates from program default values in estimating the model, the reader ought to know; for example, if nonautomatic starting values or a nonstandard convergence criterion or iteration method (cf. Jöreskog & Sörbom, 1996b) are being used.

The strategy for analyzing a postulated model is also important. One strategy would be to estimate the full model at once; another would be first to estimate the measurement part of the model and, after possible modifications, to analyze measurement and structural parts simultaneously. In the literature, ample discussions about such multistep procedures can be found. See, for example, the exchanges between Anderson and Gerbing (1988, 1992) and Fornell and Yi (1992a, 1992b), as well as Hayduk's comments (1996). Hayduk and Glaser (2000) presented a leading article on a much-debated four-step procedure, followed by reflections of Mulaik and Millsap (2000) and other contributors, which make up Volume 7, Number 1, of *Structural Equation Modeling*. Verschuren (1991) also gave an extensive treatment of analysis strategies.

ESTIMATES OF THE MODEL

It is not feasible or necessary to present all estimation results as obtained with a structural equation estimation program, not even when only default options are being used. A journal cannot have them all—and neither can its readers, for that matter. Given the model under study and the specific research questions posed, often different aspects of estimation results need to be emphasized. Some general guidelines can be given though.

First of all, if any kind of irregularities occur in estimating the model, the reader needs to know. This would include, for example, convergence problems, inadmissible solutions, and indications of empirical underidentification (see Rindskopf, 1984). This is what Jöreskog and Sörbom (1989) call "examination of the solution" (p. 41): by inspecting estimates of parameters, standard errors, correlations between parameter estimates, and multiple correlation coefficients, possible irregular, unexpected results are located and explicated.

After that, most often the following estimation results are of primary interest: estimates of the fit of the model, estimates of model parameters, and estimates of the (asymptotic) standard errors of parameter estimates. The presentation of these results can be organized around four main questions.

- How well does the model fit the sample data? The researcher wants to know whether there is not too large a discrepancy between the theoretical and the observed relations. Some relevant statistics and indexes to evaluate the fit of the model are therefore needed.

It is rather well known that it is not easy to evaluate the fit of a structural equation model to the sample data. For a number of guidelines and considerations, see, for instance, Bagozzi and Li (1988), the second issue of Volume 25 of *Multivariate Behavioral Research*, a number of contributions in Bollen and Long (1993), Marsh and Balla (1994), Ding, Velicer, and Harlow (1995), Hu and Bentler (1995), Marsh, Balla, and Hau (1996), Marsh and Hau (1996), and Fan, Thompson, and Wang (1999).

In making a reasonable choice about which indexes to report from the redundant set of available measures and statistics, aspects of sample size, power, and the complexity of the model must be kept in mind. It also makes a difference whether alternative, nested models are being assessed. Clearly, each type of index refers to different aspects of model fit; see Boomsma (1996) for an overview.

The adequacy of conventional cutoff criteria for fit indexes—rules of thumb to facilitate decision making—was studied by Hu and Bentler (1999). They made a plea for a two-index presentation strategy, which includes the ML-based standardized root mean squared residuals supplemented by some other index, for example the old Tucker–Lewis index, which is similar to the non-normed fit index, or the root mean squared error of approximation (RMSEA). Based on extensive Monte-Carlo studies, they propose new cutoff values for various (combinations of) indexes. The researcher should take notice of other findings from this study as well: Some indexes, like the ML-based Tucker–Lewis index and the RMSEA, are less preferable when the sample size is small, as they tend to overreject the population model.

Conditional on such notions, the following statistics or indexes should always be considered: the χ^2 goodness-of-fit test statistic (including the number of degrees of freedom, and its *p* value), the estimated value of the RMSEA, or preferably a 90% confidence interval for the RMSEA, and features of the residuals, in particular the standardized root mean squared residuals. It may happen, of course, that the χ^2 goodness-of-fit test statistic is relatively large, while at the same time residuals are very small. Such seemingly contradicting results need not be discouraging; a search for possible explanations might be revealing.

If required, it is feasible to summarize the size of squared multiple correlation coefficients (R^2), which are indicative of "the fit of separate equations," or for the percentage of variance accounted for. Notice that Saris and Stronkhorst (1984) demanded percentages no smaller than 90%, which is a burden in social science applications, for sure. However, it has been noticed that Goldberger (1991), for example, was far less demanding on the size of R^2s; see also O'Grady (1982). Further reference is made to Hoyle and Panter (1995), who discussed these matters more extensively from a reporting point of view.

If a number of nested models are compared, the results of chi-square difference tests should be reported in a surveyable way (cf. Bollen, 1989), along with some comparative indexes, like Akaike's Information Criterion, Schwarz' Bayesian Information Criterion, or single sample approximations of cross-validation indexes (Browne & Cudeck, 1989; De Gooijer, 1995; De Gooijer & Koopman, 1988). Usually from the set of nested models a "final" model is selected. In most cases, it is only for this model that more detailed information is given.

- If a model gives not too bad an approximation of the observed structures, the next question refers to the size of the model parameters. How strong are the postulated relations between variables of primary interest? Therefore, estimates of the unknown parameters, $\hat{\theta}_j$ ($j = 1, 2, \ldots, t$), unstandardized, standardized, or both, are reported. Which parameters to select here depends on the size of the model and the main research questions. For a small model, all estimates might be presented. If indirect effects are of importance, their estimates should be included too.

• Given the parameter estimates, it is uncertain to what extent they might fluctuate from sample to sample. An answer is needed to the following question: How reliable are the parameter estimates? Therefore, estimates of the standard errors of the (primary) parameter estimates, denoted as $\hat{s}(\hat{\theta}_j)$, are presented.

• Given estimates of parameters and their corresponding standard errors, under appropriate circumstances hypothesis tests regarding specific population parameters might be performed. For example, the question might be whether a parameter deviates significantly from zero. Estimates of the test statistics for parameter estimates, the so-called *t* values, defined as $t(\hat{\theta}_j) = \hat{\theta}_j \,/\, \hat{s}(\hat{\theta}_j)$, are reported to examine such hypotheses. Some authors, like Bollen (1989), Kaplan (1989), and Bentler (1995), referred to these values as *z* scores, because the statistic is assumed to have approximately a standard normal distribution.

Whether *p* values should be given too is a matter of statistical taste. Recall that parameter estimates are mutually dependent; see Kaplan (1980), Cudeck and O'Dell (1994), and Hancock (1999) on the control of the probability of a Type I error. The size of the *t* values is often sufficient to give a rough indication of the statistical significance of model parameters , which should be evaluated along with the substantive relevance of their estimates. The occurrence of huge *t* values deserves special consideration. Significance indicators like asterisks should be avoided throughout.

In the SEM literature, the issue of prediction is hardly ever raised, but if prediction indeed would be the goal of the modeling process, the focus should be on the actual size and sign of the estimates and on their interpretation and reliability. See Kaplan and Elliott (1997) for an example in the context of examining the predictive validity of indicator variables. In such a situation, but also in general, the fit of the model should never be predominant in describing results.

EVALUATION AND MODEL MODIFICATION

Once the model has been estimated, the researcher should evaluate the results of the analysis. This phase in the empirical cycle is characterized by feedback to the theoretical postulates that constituted the model in the first place (see Figure 1), which means that the results have to be evaluated within the theoretical nomological framework that gave rise to the model under study. Craftsmanship in theory evaluation and theory building (i.e., model modification) is required here. It is almost an art to find a fragile balance between dutiful theoretical considerations and statistical interpretations of single sample estimates and to be concise and to the point in the description of such treacherous rope-walking.

If the model fits reasonably well, the question is raised as to how close parameter estimates match the theoretical expectations in terms of both sign and size. If appropriate, it could be decided to simplify the postulated model by deleting some structural relations (fixing parameters to zero). On the other hand, if the model has a bad fit, model expansions may come into focus (adding some hypothesized relations). Beforehand it should be realized, however, that apart from real specification errors there are potentially many other reasons why a model may not fit: small sample size, nonnormality, missing data, multilevel data, and so on. Kaplan (1990) argued cogently that before considering some type of model modification, these other reasons need to be ruled out first.

Both in the case of model simplification and model expansion, the postulated modifications should be defensible primarily from a theoretical point of view. Of course, statistical information from the estimated model could be of help too: Estimated *t* values might suggest certain simplifications, and estimated modification indexes along with expected parameter change statistics could give clues to specific model expansions. Kaplan (1990, 1991), for example, strongly recommended a combined use of both the modification index and the expected parameter change statistic for model evaluation and modification. It cannot be emphasized enough, however, that purely data-driven decisions to model modifications are indefensible: The addition or deletion of a relation or association should have to make substantive sense, and parameters should ideally have clear interpretations that can directly be linked to subject-matter questions (cf. Cox & Wermuth, 1996). One major pitfall is when researchers allow error variances to correlate, solely as a tempting means to improve fit, without providing an empirical or theoretical rationale.

In a process of model modification, subsequent changes are preferably made one at a time. Changing one parameter most often triggers changes in many other parameters as well. It's like touching a spider's web: a minor parameter change can have significant and often unpredictable effects in any part of the model. A careful, single-step modification approach is therefore recommended (see also Kaplan & Wenger, 1993).

In a process of purely data-driven model modification, it is an illusion to expect that the researcher ends up with an approximately correct model, even when the correct model is part of the set but not the baseline model. There is empirical evidence that this is most frequently not the case (see Luijben, 1989; Luijben & Boomsma, 1988; MacCallum, 1986; Silvia & MacCallum, 1988).

If after a number of model modifications a decision is made to stick to a "final" model, it is the researcher's responsibility to answer questions about the validity of that model. Serious attempts should be made to realize some form of cross-validation with respect to the set of models considered in a sequence of model evaluations. Camstra and Boomsma (1992) gave an overview of cross-validation techniques in SEM; Bentler (1980) and MacCallum, Roznowski, Mar, and Reith (1994) discussed partial cross-validation techniques. When cross-validation procedures are being used for the purpose of model selection, Camstra (1993, 1998) showed empirically that the single-sample approximations of cross-validation indexes, as proposed by De Gooijer and Koopman (1988), can be highly effective compared to other indexes, like the information criteria of Akaike or Schwarz, for example.

To test the validity of the proposed modified models, ideally an independent, fresh sample from the same population should be available (e.g., even a validation sample hold out from the complete set of observations, while doing model exploration on the other, independent calibration part first). By testing modified models on the same sample data, pitfalls like chance capitalization can be very serious (see MacCallum, Roznowski, & Necowitz, 1992). It is a small step then to move from a confirmatory statistical analysis to an adventurous data exploration.

In conclusion, it should be obvious that explorative modification procedures cry for some sort of cross-validation (see also Browne & Cudeck, 1989; Cudeck & Browne, 1983). The concerns about the validity of a model after exploratory model modifications are seemingly less serious if a fixed set of plausible models based on solid theory is postulated from the start, but they are still present. A statement of Steiger (1990) reflected the relative importance of replication and cross-validation most forcefully: “An ounce of replication is worth a ton of inferential statistics” (p. 176).

During the process of model modification, and in reflection of some final model, questions regarding equivalent models might be raised again: Are there equivalent models, not too much in contradiction with theoretical postulates, that fit the sample data equally as well as the final model? Breckler (1990) claimed that it is rarely noted that the fit of a favored model is identical for a potentially large number of equivalent models. Problems related to the occurrence of equivalent models are discussed by MacCallum, Wegener, Uchino, and Fabrigar (1993), Williams, Bozdogan, and Aiman-Smith (1996), and Hayduk (1996). The recurrent and primary criterion in model selection and evaluation of equivalent models is the congruity with a theoretical perspective as presented in advance.

A final issue of concern in the evaluation and modification phase of SEM is statistical power analysis (Kaplan, 1995; MacCallum, Browne, & Sugawara, 1996; Saris & Satorra, 1993; Saris, Satorra, & Sörbom, 1987; Satorra & Saris, 1985). There is ample evidence that the power of model tests depends on specific model characteristics. Saris and Satorra (1987) emphasized that a model test cannot be used without knowledge of the power of the test and that the varying sensitivity of the test for different types of specification errors leads to serious problems. In applied research this issue is seldomly addressed, let alone designated as a leading criterion in the guidance of model modification. One reason is that it is rather difficult here to formulate general directives on how to proceed and how to report. Saris and Stronkhorst (1984) distinguished four different test situations and discussed possible actions related to power and model fit. Although such considerations are likely to be made a posteriori, after estimating the model, it seems far more adequate to reflect on the power issue in relation to the sample size a priori, before collecting the data.

In any event, the process of model modification and its theoretical and statistical justification should be properly described. The steps leading from a basic model, or a sequence of models, to some final structure should not be hidden. It is difficult though, to give a general recommendation as to the extent to which this process should be covered, or what to report at each step. Clearly, both subject matter and statistical considerations have to be disclosed and exposed concisively.

DISCUSSION

In the preceding section, quite a bit of model evaluation and interpretation came into play in the discussion of the process of model selection and model modification. Therefore, there is some inevitable overlap with that section when moving to a general, concluding discussion.

The researcher can evaluate and interpret the implications of the results in an overview. It seems appropriate to start the discussion with a summary of the substantive conclusions of the analysis. In the case of a strictly confirmative analysis, a conclusive statement should be made about the extent to which the original theoretical model is supported or falsified. If the analysis was more exploratory in character, or if a number of model modifications were made, the conclusions should be far more tentative. The firmness of conclusions very much depends on the researcher's position on a confirmatory–exploratory type of dimension; this is similar to the distinction Jöreskog and Sörbom (1996a) made for three situations: (a) a strictly confirmative situation, (b) testing alternative or competing models, and (c) a model generating situation.

In a summarized feedback to the theory, the research questions posed at the start of the study have to be answered and reflected on (see Figure 1). Feedback to similar or related published work of others should be given by discussing similarities and differences between earlier and present results. Viable explanations for differences, contradictions, or paradoxes need to be presented with conclusive arguments, and as briefly as possible.

Further general guidelines for this descriptive phase can be found in the *Publication Manual of the American Psychological Association* (1994). According to that manual, in the discussion section researchers should be guided by three questions: "What have I contributed here? How has my study helped to resolve the original problem? What conclusions and theoretical implications can I draw from my study?" (p. 19).

It is encouraged to pay some attention to the important but difficult question regarding the discrepancy between the simplified, formalized theoretical model and the actual operating processes that govern the phenomena under study in the real empirical world. This is what Bollen (1989) called "model-reality consistency" (p. 67ff.). If necessary and appropriate, aspects of causal inference might be touched on; see Sobel (1994, 1995, 1996) for some first introductions and an overview and Cox (1992) or Cox and Wermuth (1996) for some cautions to be employed.

A concise summary of weak and strong spots in the analysis can be presented. Given the results of the analysis, directives or concrete plans for future research, which might include aspects of cross-validation and replication, are briefly mentioned. By now the investigations are about to be returned to the first phase of the empirical research cycle.

It is stressed that the discussion section should not be too long. It can be expected that a curious reader tends to read the abstract of an article first, and that an immediate next jump is made to the discussion section. There, the reader expects to find a more elaborate summary of the conclusions and a recapitulation of the positive and the negative aspects of the research, both from a theoretical and methodological and statistical point of view.

ACKNOWLEDGMENTS

I would like to thank Jeffrey Hoogland, David Kaplan, and Ivo Molenaar for their valuable comments on an earlier version of this article and two anonymous reviewers for a number of useful suggestions.

REFERENCES

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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: Comment on Fornell and Yi. *Sociological Methods & Research, 20,* 321–333.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.)*, Advanced structural equation modeling: Issues and techniques* (pp. 243–277). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Arbuckle, J. L. (1997). *Amos user's guide. Version 3.6.* Chicago: SPSS.

Bagozzi, R. P., & Li, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 16,* 74–94.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (2nd ed.). Chichester, England: Wiley.

Baumgartner, H., & Bagozzi, R. P. (1995). Specification, estimation, and testing of moment structure models based on latent variables involving interactions among exogenous constructs. *Sociological Methods, 24,* 187–213.

Bekker, P. A. (1994). Counting rules for identification in linear structural models. *Computational Statistics & Data Analysis, 18,* 485–498.

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

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology, 31,* 419–456.

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

Bentler, P. M., & Yuan, K. -H. (1999). Structural modeling with small samples: Test statistics. *Multivariate Behavioral Research, 34,* 181–197.

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

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

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika, 52,* 345–370.

Boomsma, A. (1996). De adequaatheid van covariantiestructuurmodellen: een overzicht van maten en indexen [The adequacy of models for covariance structure: An overview of measures and indexes]. *Kwantitatieve Methoden, 52,* 7–52.

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

Brown, R. L. (1994). Efficacy of the indirect approach for estimating structural equation models with missing data: A comparison of five methods. *Structural Equation Modeling, 1,* 287–316.

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. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24,* 445–455.

Camstra, A. (1993). The use of cross-validation techniques for model selection in covariance structure analysis. In R. Steyer, K. F. Wender, & K. F. Widaman (Eds.), *Psychometric methodology: Proceedings of the 7th European Meeting of the Psychometric Society in Trier* (pp. 85–89). Stuttgart: Fischer.

Camstra, A. (1998). *Cross-validation in covariance structure analysis.* Unpublished doctoral dissertation, University of Groningen, The Netherlands.

Camstra, A., & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis: An overview. *Sociological Methods & Research, 21,* 89–115.

Chou, C.-P., & Bentler, P. M. (1995). Estimation and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 37–55). Thousand Oaks, CA: Sage.

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.

Cochran, W. G. (1965). The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, Series A, 128,* 234–265.

Cox, D. R. (1992). Causality: Some statistical aspects. *Journal of the Royal Statistical Society, Series A, 155,* 291–301.

Cox, D. R., & Wermuth, N. (1996). *Multivariate dependencies—Models, analysis and interpretation.* London: Chapman & Hall.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105,* 317–327.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18,* 147–167.

Cudeck, R., & O'Dell, L. L. (1994). Applications of standard error estimates in unrestricted factor analysis: Significance tests for factor loadings and correlations. *Psychological Bulletin, 115,* 475–487.

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.

De Gooijer, J. G. (1995). Cross-validation criteria for covariance structures. *Communications in Statistics: Part B. Simulation and Computation, 24,* 1–16.

De Gooijer, J. G., & Koopman, S. J. (1988). Cross-validation criteria and the analysis of covariance structures. In M. G. H. Jansen & W. H. van Schuur (Eds.), *The many faces of multivariate analysis: Proceedings of the SMABS–88 Conference* (Vol. II, pp. 296–311). Groningen: RION.

De Groot, A. D. (1966). *Methodologie. Grondslagen van onderzoek en denken in de gedragswetenschappen (derde druk)* [Methodology: Foundations of inference and research in the behavioral sciences (3rd ed.)]. 'S-Gravenhage: Mouton.

De Groot, A. D. (1969). *Methodology: Foundations of inference and research in the behavioral sciences.* The Hague: Mouton.

Ding, L., Velicer, W. V., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling, 2,* 119–143.

Duncan, T. E., Duncan, S. C., & Li, F. (1998). A comparison of model- and multiple imputation-based approaches to longitudinal analyses with partial missingness. *Structural Equation Modeling, 5,* 1–21.

Ernste, H. (Ed.). (1996). *Multilevel analysis with structural equation models*. Zürich: Gruppe für Quantitative Geographie und Humanökologie, Geographisches Institut ETH.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling, 6,* 56–83.

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

Fornell, C., & Yi, Y. (1992b). Assumptions of the two-step approach: Reply to Anderson and Gerbing. *Sociological Methods & Research, 20,* 334–339.

Freedman, D. A. (1997). From association to causation via regression (with discussion). In V. R. McKim & Turner, S. P. (Eds.), *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences* (pp. 113–182). Notre Dame, IN: University of Notre Dame Press.

Goldberger, A. S. (1991). *A course in econometrics.* Cambridge, MA: Harvard University Press.

Hancock, G. R. (1999). A sequential Scheffe-type respecification procedure for controlling Type I error in exploratory structural equation modification. *Structural Equation Modeling, 6,* 158–168.

Hayduk, L. A. (1996). *LISREL issues, debates, and strategies.* Baltimore, MA: Johns Hopkins University Press.

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

Hershberger, S. L. (1995). The specification of equivalent models before the collection of data. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 68–105). Newbury Park, CA: Sage.

Hoogland, J. J. (1999). *The robustness of estimation methods for covariance structure analysis.* Unpublished doctoral dissertation, University of Groningen, The Netherlands.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26,* 329–367.

Hox, J. J. (1994). *Applied multilevel analysis.* Amsterdam: TT-Publikaties.

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

Jöreskog, K. G., & Sörbom, D. (1996a). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Chicago: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (1996b). *LISREL 8: User's reference guide* (2nd ed.). Chicago: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (1996c). *PRELIS 2: User's reference guide* (3rd ed.). Chicago: Scientific Software International.

Jöreskog, K. G., Sörbom, D., Du Toit, S., & Du Toit, M. (1999). *LISREL 8: New statistical features.* Chicago: Scientific Software International.

Kaplan, D. W. (1989). A study of the sampling variability and *z*-values of parameter estimates from misspecified structural equation models. *Multivariate Behavioral Research, 24,* 41–57.

Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation (with discussion). *Multivariate Behavioral Research, 25,* 137–155.

Kaplan, D. (1991). On the modification and predictive validity of covariance structure models. *Quality & Quantity, 25,* 307–314.

Kaplan, D. (1995). Statistical power in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 100–117). Thousand Oaks, CA: Sage.

Kaplan, D., & Elliott, P. R. (1997). A model-based approach to validating education indicators using multilevel structural equation modeling. *Journal of Educational and Behavioral Statistics, 22,* 323–347.

Kaplan, D., & Wenger, R. N. (1993). Asymptotic independence and separability in covariance structure models: Implications for specification error, power, and model modification. *Multivariate Behavioral Research, 28,* 467–482.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Little, R. J. A., & Rubin, D. B. (1990). The analysis of social science data with missing data. *Sociological Methods & Research, 18,* 292–326.

Luijben, T. C. W. (1989). *Statistical guidance for model modification in covariance structure analysis.* Amsterdam: Sociometric Research Foundation (Doctoral dissertation, University of Groningen, The Netherlands).

Luijben, T. C. W., & Boomsma, A. (1988). Statistical guidance for model modification in covariance structure analysis. In D. Edwards & N. E. Raun (Eds.), *COMPSTAT 1988: Proceedings in Computational Statistics* (pp. 335–340). Heidelberg: Physica.

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100,* 107–120.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130–149.

MacCallum, R. C., Roznowski, M., Mar, C. M., & Reith, J. V. (1994). Alternative strategies for cross-validation of covariance structure models. *Multivariate Behavioral Research, 29,* 1–32.

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 analysis. *Psychological Bulletin, 114,* 185–199.

Marsh, H. W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling, 5,* 22–36.

Marsh, H. W., & Balla, J. R. (1994). Goodness of fit in confirmatory factor analysis: The effects of sample size and parsimony. *Quality & Quantity, 28,* 185–217.

Marsh, H. W., Balla, J. R., & Hau, K.-T. (1996). An evaluation of incremental fit indexes: A clarification of mathematical and empirical properties. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 315–353). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Marsh, H. W., & Hau, K.-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *The Journal of Experimental Education, 64,* 364–390.

Marsh, H. W., & Hau, K.-T. (1999). Confirmatory factor analysis: Strategies for small sample sizes. In R. H. Hoyle (Ed.), *Statistical strategies for small sample size* (pp. 251–306). Thousand Oaks, CA: Sage.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181–220.

Maxwell, S. E., & Cole, D. A. (1995). Tips for writing (and reading) methodological articles. *Psychological Bulletin, 118,* 193–198.

McArdle, J. J., & Hamagami, F. (1996). Multilevel models from a multiple group structural equation perspective. In G. A. Marcoulides & R. A. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 89–124). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Mulaik, S. A., & Millsap, R. E. (2000). Doing the four-step right. *Structural Equation Modeling, 7,* 36–73.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22,* 376–398.

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

Neale, M. C. (1994). *Mx: Statistical modeling* (2nd ed.). Richmond, VA: Medical College of Virginia, Department of Psychiatry.

O'Grady, K. E. (1982). Measures of explained variance: Cautions and limitations. *Psychological Bulletin, 92,* 766–777.

Popper, K. (1974). *Conjectures and refutations: The growth of scientific knowledge* (5th ed.). London: Routledge & Kegan Paul.

Popper, K. (1983). *Realism and the aim of science.* (W. W. Bartley III, Ed.). [Post-script to The logic of scientific discovery (Vol. 1)]. London: Hutchinson.

Publication manual of the American Psychological Association. (4th ed.). (1994). Washington, DC: American Psychological Association.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21,* 173–184.

Raykov, T., Tomer, A., & Nesselroade, J. R. (1991). Reporting structural equation modeling results in *Psychology and Aging*: Some proposed guidelines. *Psychology and Aging, 6,* 499–533.

Rindskopf, D. (1984). Structural equation models: Empirical identification, Heywood cases, and related problems. *Sociological Methods & Research, 13,* 109–119.

Rosenbaum, P. R. (1995). *Observational studies.* New York: Springer.

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

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91,* 473–489.

Saris, W. E., & Satorra, A. (1987). Characteristics of structural equation models which affect the power of the likelihood ratio test. In W. E. Saris & I. N. Gallhofer (Eds.), *Sociometric research* (Vol. 2., pp. 220–236). London: Macmillan.

Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 181–204). Newbury Park, CA: Sage.

Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. In C. C. Clogg (Ed.), *Sociological methodology 1987* (pp. 105–129). Washington, DC: The American Sociological Association.

Saris, W. E., & Stronkhorst, H. (1984). *Causal modelling in nonexperimental research: An introduction to the LISREL approach.* Amsterdam: Sociometric Research Foundation.

Satorra, A. (1990). Robustness issues in structural equation modeling: A review of recent developments. *Quality & Quantity, 24,* 367–386.

Satorra, A. (1992). Asymptotic robust inferences in the analysis of mean and covariance structures. In P. V. Marsden (Ed.), *Sociological Methodology 1992.* Oxford, England: Blackwell.

Satorra, A. (1993). Multi-sample analysis of moment structures: Asymptotic validity of inferences based on second order moments. In K. Haagen, D. J. Bartholomew, & M. Deistler (Eds.), *Statistical modelling and latent variables* (pp. 283–298). Amsterdam: Elsevier.

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

Satorra, A., & Saris, W. E. (1985). The power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50,* 83–90.

Schafer, J. L. (1996). *Analysis of incomplete multivariate data.* London: Chapman & Hall.

Scheines, R., Spirtes, P., Glymour, C., & Meek, C. (1994). *TETRAD II: Tools for causal modeling. User's manual.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Silvia, E. S. M., & MacCallum, R. C. (1988). Some factors affecting the success of specification searches in covariance structure modeling. *Multivariate Behavioral Research, 23,* 297–326.

Sobel, M. E. (1994). Causal inference in latent variable models. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 3–35). Newbury Park, CA: Sage.

Sobel, M. E. (1995). Causal inference in the social and behavioral sciences. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 1–38). New York: Plenum.

Sobel, M. E. (1996). An introduction to causal inference. *Sociological Methods & Research, 24,* 353–379.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173–180.

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

Verleye, G. (1996). *Missing at random data problems in attitude measurement using maximum likelihood structural equation modelling.* Unpublished doctoral dissertation, Vrije Universiteit Brussel, Faculteit Psychologie en Opvoedkunde.

Verschuren, P. J. M. (1991). *Structurele modellen tussen theorie en praktijk* [Structural models between theory and practice]. Utrecht: Het Spectrum.

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

Williams, L. J., Bozdogan, H., & Aiman-Smith, L. (1996). Inference problems with equivalent models. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 279–314). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Yang Jonsson, F. (1997). *Non-linear structural equation models: Simulation studies of the Kenny–Judd model* (Acta Universitatis Upsaliensis. Studia Statistica Upsaliensia 4). Doctoral dissertation, University of Uppsala, Department of Statistics.

Yuan, K.-H., & Bentler, P. M. (1995). *Robust methods for mean and covariance structure analysis* (UCLA Statistical Series, No. 195). Los Angeles: University of California.

Yuan, K.-H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association, 92,* 767–774.

Yuan, K.-H., & Bentler, P. M. (1998a). Normal theory based test statistics in structural equation modelling. *British Journal of Mathematical and Statistical Psychology, 51,* 289–309.

Yuan, K.-H., & Bentler, P. M. (1998b). Robust mean and covariance structure analysis. *British Journal of Mathematical and Statistical Psychology, 51,* 63–88.

Yuan, K.-H., & Bentler, P. M. (1998c). Structural equation modeling with robust covariances. In A. E. Raftery (Ed.), *Sociological methodology 1998* (pp. 363–396). Oxford, England: Blackwell.