Psychological Methods
2009, Vol. 14, No. 1, 6–23

1082-989X/09/$12.00 DOI: 10.1037/a0014694

Reporting Practices in Confirmatory Factor Analysis: An Overview and Some Recommendations

Dennis L. Jackson
University of Windsor

J. Arthur Gillaspy, Jr.
University of Central Arkansas

Rebecca Purc-Stephenson
University of Windsor

Reporting practices in 194 confirmatory factor analysis studies (1,409 factor models) published in American Psychological Association journals from 1998 to 2006 were reviewed and compared with established reporting guidelines. Three research questions were addressed: (a) how do actual reporting practices compare with published guidelines? (b) how do researchers report model fit in light of divergent perspectives on the use of ancillary fit indices (e.g., L.-T. Hu & P. M. Bentler, 1999; H. W. Marsh, K.-T., Hau, & Z. Wen, 2004)? and (c) are fit measures that support hypothesized models reported more often than fit measures that are less favorable? Results indicate some positive findings with respect to reporting practices including proposing multiple models a priori and near universal reporting of the chi-square significance test. However, many deficiencies were found such as lack of information regarding missing data and assessment of normality. Additionally, the authors found increases in reported values of some incremental fit statistics and no statistically significant evidence that researchers selectively report measures of fit that support their preferred model. Recommendations for reporting are summarized and a checklist is provided to help editors, reviewers, and authors improve reporting practices.

Keywords: confirmatory factor analysis, statistical reporting, structural equation models, research methods, construct validation

Confirmatory factor analysis (CFA) is a powerful statistical tool for examining the nature of and relations among latent constructs (e.g., attitudes, traits, intelligence, clinical disorders). In contrast to its analytic cousin, exploratory factor analysis, CFA explicitly tests a priori hypotheses about relations between observed variables (e.g., test scores or ratings) and latent variables or factors. CFA is often the analytic tool of choice for developing and refining measurement instruments, assessing construct validity, identifying method effects, and evaluating factor invariance across time and groups (Brown, 2006). Thus, CFA is a useful application for investigating issues of interest to most psychological researchers. Since the late-1990s, there has been a positive trend in the use of CFA, with most applications being in the area of scale development and construct validation (Brown, 2006; Russell, 2002).

CFA is part of the larger family of methods known as structural equation modeling (SEM) and plays an essential role in measurement model validation in path or structural analyses (Brown, 2006; MacCallum & Austin, 2000). When conducting SEM, researchers often first evaluate the measurement model (whether the measured variables accurately reflect the desired constructs or factors) before assessing the structural model. As noted by Thompson (2004), "It makes little sense to relate constructs within an SEM model if the factors specified as part of the model are not worthy of further attention" (p. 110). In many cases, problems with SEM models are due to measurement model issues that can be identified with CFA (Brown, 2006).

Dennis L. Jackson and Rebecca Purc-Stephenson, Department of Psychology, University of Windsor, Windsor, Ontario, Canada; J. Arthur Gillaspy, Jr., Department of Psychology and Counseling, University of Central Arkansas.

Rebecca Purc-Stephenson is now at the Department of Social Sciences, University of Alberta—Augustana Campus, Canada.

We would like to acknowledge the assistance of Shanesya Kean and Lori K. Gray for their work in identifying and collecting the articles reviewed for this study and Pamela G. Ing for help with coding some of the final articles. Interested readers may contact Dennis L. Jackson to obtain a list of the 194 studies reviewed in this article.

Correspondence concerning this article should be addressed to Dennis L. Jackson, Department of Psychology, 401 Sunset Ave., University of Windsor, Windsor ON N9B 3P4, Canada. E-mail djackson@uwindsor.ca

The growing application of CFA by psychological researchers follows the increased popularity of SEM in general. Hershberger (2003) noted the number of substantive and technical SEM articles referenced in the PsychINFO database increased by a factor of approximately 2.5 from 1994 to 2001. Similar trends have been noted within subdisciplines of psychology (e.g., Martens, 2005; Raykov, Tomer, & Nesselroade, 1991) as well as in other fields such as marketing and consumer research (Baumgartner & Homburg, 1996). As a special application of SEM, CFA represents a substantial subset of all SEM articles (Breckler, 1990; Martens, 2005; Tremblay, & Gardner, 1996). Due to the important role of CFA in measure development and the fact that proper measurement of constructs is foundational to all quantitative research in the social sciences, the proper conduct and reporting of CFA studies is essential to the scientific enterprise of psychology.

Guidelines for conducting SEM also apply to CFA studies. There are many excellent practice guidelines available for SEM (Boomsma, 2000; Breckler, 1990; Hoyle & Panter, 1995) and for CFA (Brown, 2006; Byrne, 2001; Kline, 2005; Thompson, 2004). An issue of particular importance across these guidelines is what should be reported in SEM/CFA studies. Good reporting practices form the basis of progress in science. Those wishing to contribute original research to a field or apply findings from previous research naturally want to comprehend the decisions made by researchers in order to understand the robustness of their findings. Likewise, when two studies arrive at contradictory conclusions, the natural course is to attempt to explain any differences in findings in terms of the methodology employed. A lack of clarity around methods prevents such an analysis; thus, one may have to launch another study merely to comprehend the original contradiction.

Unfortunately, the state of reporting results of SEM analyses has been relatively poor (Boomsma, 2000; MacCallum & Austin, 2000; McDonald & Ho, 2002; Steiger, 2001). In an examination of approximately 500 SEM studies published between 1993 and 1997, MacCallum and Austin (2000) were unable to determine the precise model tested in approximately 10% of studies and to which matrix models were fit (variance-covariance or correlation matrix) in approximately 25% of the studies. About half of the studies were incomplete in terms of parameter estimates. Additional problems included overgeneralization of findings, failure to consider equivalent models, use of directional influences in cross-sectional studies, and failure to take into account autoregressive effects in longitudinal designs.

More recently, McDonald and Ho (2002) reviewed SEM reporting practices for 41 articles in 13 journals (1995 to 1997). Their findings were similar to MacCallum and Austin's (2000). For example, checks for multivariate normality were reported in only 12% of the studies, parameter estimates were provided in 30%, and standard errors reported in only 12%. In addition, the correlation or covariance matrix was provided in less than half of the studies; the remaining studies made no mention of the availability of these data. A more positive finding was that most studies reported more than one fit index.

Existing reviews of SEM studies do not, however, inform researchers about the state of reporting practices in the CFA literature. McDonald and Ho (2002) excluded CFA studies from their sample, and MacCallum and Austin (2000) did not distinguish CFA applications from structural SEM. Given the prominent role of CFA in scale development and construct validity, CFA is an important analytic tool in its own right (Brown, 2006). Kenny (2006) even asserted that "the social and behavioral sciences have learned much more from CFA than from SEM" (p. ix). Thus, we argue that how CFA studies are reported warrants examination separate from SEM studies in general. Of particular importance is the degree to which CFA studies provide sufficient detail about model specification and evaluation to instill confidence in the results. Psychological measurement can have important implications for consumers of psychological services in a variety of settings. For instance, whether a measure of posttraumatic stress disorder (PTSD) for returning veterans is invariant across ethnic groups may be of great importance, not only to psychologists, but also to returning veterans. CFA is the main statistical method for evaluating such a question. Therefore, it is important to understand the state of reporting practices in CFA applications within psychological research.

To date, few reviews have investigated reporting practices specific to CFA. Existing reviews vary widely in focus, depth and breadth of variables coded, and number of studies reviewed. In the most detailed review of CFA studies, DiStefano and Hess (2005) examined 101 articles in four assessment journals (*Psychological Assessment, Journal of Psychoeducational Assessment, Assessment*, and *Journal of Psychopathology and Behavioral Assessment*) over a 12-year period (1990—2002). Variables were coded across four dimensions: study background and methods (models tested, estimation method, level of data), sample size and data preparation, results (fit indices, parameter estimates, standard errors), and conclusions (choice of final model). Positive reporting practices were found for specifying and testing theoretically relevant competing models, using adequate sample sizes, using multiple and different types of fit statistics to determine model fit, and identifying estimation procedure and software. Reporting practices were lacking, however, in terms of data screening (85% did not report checking for univariate or multivariate normality), estimation method (50% did not report), matrix analyzed (42% did not report), and a priori specification of cutoff criteria for fit measures (64% did not report).

Three additional smaller studies have reviewed CFA reporting practices. Russell (2002) reported the use of

CFA in factor analytic studies in *Personality and Social Psychology Bulletin* in 1996, 1998, and 2000. Nineteen CFA studies were reviewed (10 for scale validation; nine for measurement model as part of structural analysis). Of concern in terms of reporting practices were the following: 16% of studies did not report the software used; 58% did not report the estimation procedure; and 95% failed to report treatment of missing data. In a more qualitative study of 16 CFA applications in the *Journal of Educational Research* (1989—2004), Schreiber, Nora, Stage, Barlow, and King (2006) also found inconsistent reporting of basic information. Specific problem areas included no screening for univariate or multivariate normality, little mention of how missing data were handled, lack of clarity about sample size, and little discussion of cutoff values for fit indices. Only 50% of the articles reported the software used or specified the estimation method. Finally, Worthington and Whittaker (2006) reviewed 14 CFA studies in the context of new scale development published in the *Journal of Counseling Psychology* (1995—2005). The results indicated inconsistent use of fit indices and fit criteria to evaluate model fit, failure to include confidence intervals for fit statistics such as RMSEA, and a lack of information regarding cross-validation of results. Some positive reporting practices were also found, such as adequate participant-to-parameter ratios, specification of cutoff criteria for model fit (97% of studies), use of two or more fit indices (100%), and the lack of model modification after fit (86%).

Drawing clear conclusions about the state of CFA reporting practices in psychological research on the basis of these four reviews seems inadvisable for a number of reasons. First, each review focused on a fairly narrow field within psychology (assessment, personality and social psychology, counseling psychology, educational psychology). Second, the reviews included studies from only six journals. This may not be a large or representative enough sample of CFA studies to generalize to CFA reporting practices across a broad range of psychological research. Third, the methodology varied across reviews. Three studies (Russell, 2002; Schreiber et al. 2006; Worthington & Whittaker, 2006) were more qualitative or descriptive, providing nominal (yes/no) level data on the presence or absence of information. DiStefano and Hess (2005) coded more specific study variables such as the means and standard deviations of parameter estimates and standard errors. Most reviews did not report the actual values of specific statistics, such as fit indices. This type of data, for example, is essential for understanding the level of rigor researchers use to evaluate model fit.

In the present study, we expanded upon previous investigations of the CFA/SEM literature by comparing CFA reporting practices with existing guidelines using a large sample of studies ($N = 194$) drawn from all journals published by the American Psychological Association (APA). Noting McDonald and Ho's (2002) recommendation that SEM reporting practices "may need occasional revisiting until practice conforms with principles" (p. 65), we reviewed studies conducted from 1998 to 2006, the years since 1995–1997 studies covered in their review. We also extended upon the work of DiStefano and Hess (2005) by coding specific model characteristics (e.g., number of observed and latent variables) and values of fit statistics and by sampling a broader set of journals. In addition, whereas previous reviews provided data about only one model per study, many CFA articles test and report multiple models. To acknowledge this reality, we recorded data on all models tested within each study. Thus, the goal of this study was to develop a comprehensive understanding of CFA reporting practices in psychological research. From this perspective, our review may serve as a baseline for actual CFA practices. Such a baseline serves a number of purposes: (a) to identify areas of good CFA reporting practices, (b) to highlight areas that need improvement, (c) to increase knowledge of basic information necessary to evaluate CFA findings, and (d) to promote the use of existing reporting guidelines by both authors and journal editors. Finally, another way this study departs from previous reviews is that we provide a brief, generic checklist that may be used to guide report writing or the editorial review process.

Although the scope of this study is quite broad, we would like to clarify our intentions on several points. First, we are not offering our own or a new set of reporting guidelines. Excellent recommendations are readily available (e.g., Boomsma, 2000; Hoyle & Panter, 1995; McDonald & Ho, 2002), and we see little utility in duplicating these efforts. Thus, we focus our attention on increasing awareness of established reporting standards by relating them to actual practice. Second, we also recognize there may be a variety of reasons that reporting guidelines are not followed, such as lack of knowledge of standards, space limitations, journal focus, and editorial demand. Accordingly, it is not our intention to judge the quality of individual studies or journals but rather to better understand the state of CFA reporting in general. Finally, our offering of a checklist should not be construed as rigid criteria. There may be many cases in which items on the checklist would not be included in published articles. The checklist may, however, help authors, reviewers, and editors become more deliberate and conscious about their reporting decisions and lead to more complete reporting of CFA results.

The next section of this article provides a brief synthesis of existing reporting guidelines, after which we discuss the specific research questions in this study. We then report our review of CFA studies, make recommendations for reporting guidelines, and provide the aforementioned checklist.

Reporting Guidelines in SEM/CFA

What should be reported in SEM/CFA is not universally agreed upon; however, there is considerable consistency among authors who have addressed this question (e.g., Barrett, 2007; Bentler, 2007; Boomsma, 2000; Chin, 1998; Hoyle & Panter, 1995; MacCallum & Austin, 2000; McDonald & Ho, 2002; Medsker, Williams, & Holahan, 1994; Raykov et al., 1991; Thompson, 2004). Although readers are encouraged to consult these original articles to get a more comprehensive explanation for the importance of proper reporting with respect to some of the individual recommendations, a synthesis of these writings is provided later, along with some brief justification for the suggestions. We have opted to organize the recommendations by putting them into categories that reflect the activity surrounding a research project in which CFA is to be used. Naturally, many of the recommendations also apply to research in which any statistical technique is to be used.

Theoretical Formulation and Data Collection

An important aspect of CFA is that it allows researchers to specify precise and even highly complex hypotheses regarding the phenomenon under study. In order to test these hypotheses, they must be thoughtfully converted into a model. Boomsma (2000) suggested that researchers should clearly define and justify the models to be tested, including equivalent models, and ensure that all models can be estimated (i.e., are overidentified). This includes describing any hierarchically nested models and where the researcher's work fits on the continuum ranging from exploratory to confirmatory. Researchers are encouraged to not only identify equivalent models but also to identify competing theoretical models against which the fit of the model of interest can be compared (Bentler & Bonett, 1980; Hoyle & Panter, 1995). An evaluation of the plausibility of the results should include a decision about the credibility of the observed measures used to identify the latent variables under study. Therefore, the choice of observed measures to identify latent constructs should be justified. Another important point pertains more specifically to CFA studies. As mentioned above, a common application of CFA is for scale development and validation. Toward this end, one would naturally assume a greater amount of detail be reported about the results of CFA analyses when it is used to support the psychometric properties of existing or newly developed measures as opposed to, say, determining the adequacy of the measurement model portion of a structural equation model.

The types of activities that fit under data collection include ensuring adequate and appropriate sampling procedures and, when appropriate, utilizing some sort of power analysis to obtain estimates of an adequate sample size (e.g., MacCallum, Browne, & Sugawara, 1996; Muthén & Muthén, 2002). Additionally, researchers need to justify the choice of the population from which the sample was drawn. For instance, Hulland, Chow, and Lam (1996) reviewed SEM studies published from 1980 to 1994 in marketing research and found a heavier reliance on student samples in the latter years covered by the review. Depending upon the research question, this could represent a serious threat to external validity.

Data Preparation

Many activities fit under this heading, from assessing data integrity to evaluating the distributional assumptions of the estimation method to be used. Concerning the latter, the most common estimation procedure in SEM is maximum likelihood (ML), which carries with it the assumption of multivariate normality (MVN). Past research has found that the failure to meet the assumption of MVN can lead to an overestimation of the chi-square statistic and, hence, to an inflated Type 1 error (e.g., Curran, West, & Finch, 1996; Powell & Schafer, 2001) and downward biased standard errors (Bandalos, 2002; Kaplan, 2000; Nevitt & Hancock, 2001), and may undermine the assumptions inherent in several ancillary fit measures (Yuan, 2005). It should be noted, however, that ML estimation may perform well with mild departures from MVN (Chou, Bentler, & Satorra, 1991; Fan & Wang, 1998; Hu, Bentler, & Kano, 1992).

Another activity related to data preparation concerns the analysis and treatment of missing data. The effects of missing data depend on the method used to address it, which may include listwise deletion, pairwise deletion, mean substitution, multiple imputation, and expectation maximization. The most common approach is listwise deletion or available case analysis (McKnight, McKnight, Sidani, & Figueredo, 2007; Schaefer & Graham, 2002), in which cases with any missing data points involved in the analysis are removed. This is generally only an acceptable approach when data are missing completely at random (Schaefer & Graham, 2002). It is interesting that there is evidence that shows parameter estimates can be biased (Brown, 1994) or convergence failures can become more likely (Enders & Bandalos, 2001), depending upon the manner in which missing data is dealt with, even when it is missing at random. Hence, researchers should report on univariate and multivariate normality, criteria for deleting multivariate outliers and missing data issues. As well, other data manipulations should be described such as transformations like the square-root transformation designed to improve the distribution of measured variables and parceling measured variables together by summing or averaging individual items.

Analysis Decisions

Once the data have been adequately prepared for analysis, the researcher still has some decisions to make. Two of the main decisions involve the choice of input matrix and the estimation method. The default choices tend to be the variance–covariance matrix with ML estimation. Even if this is the case, these choices should be stated explicitly. It is also recommended that the input matrix or equivalent information be provided or made available. MacCallum and Austin (2000) indicated that in 50% of the studies they reviewed, authors analyzed a correlation matrix, which requires the use of constrained estimation or an approach that ensures the model is scale invariant (Kline, 2005). Alternatives to ML are also available and can prove to be advantageous when, for instance, data do not follow a multivariate normal distribution (e.g., Satorra & Bentler, 1994). Additionally, there are a number of alternatives based on a weighted least squares approach (Browne, 1984). The behavior of these alternative methods are somewhat nuanced and under some conditions require very large sample sizes. The point remains that the reader should be able to determine the decisions that researchers have made with respect to these two choices. Finally, other aspects of the modeling process should be revealed, such as how latent variable scales were fixed and the type of software used.

Model Evaluation and Modification

Having derived model estimates, the researcher now needs to evaluate model fit. Aside from the chi-square goodness-of-fit test, there are numerous ancillary indices of global fit such as the goodness-of-fit index and adjusted goodness-of-fit index, (GFI, AGFI; Jöreskog & Sörbom, 1986), the comparative fit index (CFI; Bentler, 1990), and the root-mean-square error of approximation (RMSEA; Steiger & Lind, 1980). Many of the indices have different properties, and some have been recommended against, such as the GFI, AGFI, normed fit index (NFI; Bentler & Bonett, 1980), and Bollen's (1986) rho 1 (Hu & Bentler, 1998). Hu and Bentler recommended relying on fit indices that have different measurement properties, such as an incremental fit index (IFI; e.g., CFI) and a residuals-based fit index, such as the standardized root-mean-square residual (SRMR; Bentler, 1995; Jöreskog & Sörbom, 1986). Findings from Monte Carlo studies suggest that on the basis of effect size, direct measures of fit are more sensitive to model misspecifications than incremental fit measures (Fan, Thompson, & Wang, 1999; Jackson, 2007). Drawing from previous studies (Fan et al., 1999; Hu & Bentler, 1998, 1999; Jackson, 2007; Marsh, Balla, & Hau, 1996; Marsh, Hau, Balla, & Grayson, 1998), the following fit measures tend to perform well with respect to detecting model misspecification and lack of dependence on sample size: gamma hat (Steiger, 1989); RMSEA; centrality index (CI, McDonald, 1989); SRMR; Tucker–Lewis index (TLI)/nonnormed fit index (NNFI) (Tucker & Lewis, 1973; Bentler & Bonett, 1980); relative noncentrality index (RNI; McDonald & Marsh, 1990); CFI; and Bollen's delta 2, also referred to as the incremental fit index (Bollen, 1989).

Adding to the complexity of model evaluation, recommendations for cutoff values for some measures have changed over time. For instance, Bentler and Bonett (1980) recommended a cutoff of .90 for some incremental fit indices. More recently, Hu and Bentler (1999) recommended a cutoff of .95, and other authors have recommended cutoffs of .97 (Schermelleh-Engel, Moosbrugger, & Müller, 2003). Barrett (2007) suggested that ancillary fit indices should be abandoned altogether—citing recent articles that have highlighted the shortcomings of adopting strict cutoffs (e.g., Marsh et al., 2004; Yuan, 2005). Several authors responded to Barrett's article and indicated that such ancillary fit measures should not be abandoned (e.g., Bentler, 2007; Miles & Shevlin, 2007; Millsap, 2007; Steiger, 2007), whereas other authors expressed concerns about the ways that the indices were being used and the seeming instant dismissal of the chi-square test (e.g., Goffin, 2007; Markland, 2007). Cutoff values should be explicitly stated when ancillary fit measures are used. It is often recommended that in addition to examining global fit measures, researchers pay attention to other aspects of model fit such as examining the standardized residuals to determine whether specific variables or relations are being accounted for (Bollen, 1989; Hayduk, 1988) and parameter estimates to ensure they have the anticipated signs and magnitudes (Boomsma, 2000).

Another aspect of model fit concerns whether model modification is practiced. Ideally, researchers test several competing models so they are not in a position of having to modify a model to find acceptable fit. It is often noted that post hoc modifications to models, such as those based on modification indices, should be done sparingly and only when the modifications are theoretically and practically plausible (e.g., MacCallum, 1995). As researchers undertaking modifications may capitalize on chance variations in the obtained sample, any such modifications should be viewed as tentative until cross-validated on an independent sample (see, e.g., MacCallum, 1986). Additionally, Bentler (2007) noted that the test statistic, T, will not be distributed as χ^2 when it is based on post hoc model modifications. Whereas it seems clear to us that any model modifications should be clearly articulated in reporting the results of a study, Bentler went further and suggested that a letter should also be submitted with the manuscript verifying that each parameter in the model represents an a priori hypothesis, and, if not, all modifications should be adequately described.

Reporting Findings

In addition to clearly reporting on the activities outlined above, researchers must decide what, in their voluminous output, to report. It is recommended that parameter estimates (e.g., measurement model and structural paths), including variances of exogenous variables (which includes standard errors) be reported (Boomsma, 2000; Hoyle & Panter, 1995). Furthermore, it is useful to provide some indication of the variance accounted for in endogenous variables. Researchers may also choose to report structure coefficients in correlated models as suggested by Thompson (1997). This recommendation rests on two arguments: first, that it is common to report structure as well as pattern coefficients in other multivariate analyses, and second, that examining both can lead to an enhanced interpretation of the findings. Finally, authors should specify their preferred model and support their choice not only by achieving acceptable fit, but also by discussing the merits of their preferred model relative to equivalent and competing theoretical models.

The Current Study

First Research Question

The current study was undertaken to address three questions. To begin with, we were interested in how well some of the recommended guidelines for reporting on CFA studies are being observed in the psychology literature. More specifically, we assessed the extent to which published CFA articles reported on the broad categories outlined earlier, such as fully reporting parameter estimates, sample sizes, types and values of fit indices, and the quality of reporting regarding general analysis decisions (e.g., type of matrix, estimation method, and assessment of the assumptions of the technique).

Second Research Question

Another question had to do with understanding the methods used to assess model fit. We were originally interested in whether studies published after Hu and Bentler (1999) adopted higher fit index cutoffs for models deemed as acceptable and whether authors used Hu and Bentler's two-index presentation strategy. However, during the time when we were coding articles for this study, the recommendations made by Hu and Bentler came under criticism, most notably by Marsh et al. (2004) and Fan and Sivo (2005). These critiques make a compelling argument that adopting Hu and Bentler's recommendations as a "golden rule" is counterproductive. Despite the fact that Hu and Bentler cautioned against adopting their recommendations uncritically, concern was expressed that this has indeed occurred (Marsh et al., 2004), particularly given that their article has been referenced more than 2,000 times, according to PsychINFO. To date, however, this issue has not been examined empirically. Thus, given the dynamic nature of the area of fit assessment, and following recommendations that arose from the review process, we opted to broaden our original question to examine how researchers report fit indices in light of somewhat divergent perspectives, such as those prior to the timeframe covered by our study (e.g., Bentler & Bonett, 1980), as well as those introduced during the time period covered in this study (e.g., Barrett, 2007; Fan & Sivo, 2005; Hu and Bentler, 1999; Marsh et al., 2004).

We approached this second question in a variety of ways. To begin with, we examined whether authors changed the cutoffs for their fit measures in response to Hu and Bentler's article and also in response to Marsh et al.'s article. Specifically we asked, was there evidence that researchers began adopting higher cutoff values in the years after Hu and Bentler's (1999) publication; to what extent did researchers use their proposed two-index presentation strategy; and did it appear that researchers were willing to report on models that had lower fit indices following Marsh et al.'s (2004) article? Next, we examined the number and types of fit measures reported. Namely, whether in response to divergent perspectives on fit criteria, had researchers reported more fit indices in recent years? Finally, we were interested in whether empirically based recommendations for which fit measures have desirable properties, and which ones have undesirable properties, have made it into practice. Specifically, we examined the ratio of recommended fit measures to nonrecommended measures used from 1998 to 2006.

Third Research Question

For our final question, we were interested in whether authors selected fit measures that best supported their choice of preferred model. Since there are no firm rules about which fit indices to report, it is conceivable that when various fit indices present a contradictory picture of model fit, researchers might ignore fit indices that do not fall within generally accepted guidelines and report others that do. This represents but one possible way in which researchers could "game" their reporting in order to maximize the chances of having their research published. As a reviewer of an earlier version of this article pointed out, it is also possible that researchers whose results are disappointing with respect to model fit modify their model until they achieve acceptable fit. Although we were unable to assess this particular question in the current study, it must be kept in mind as another possible threat to the validity of findings published in peer-reviewed journals. Bentler (2007) seemed to share this concern when he suggested that authors essentially "certify" that each parameter represents an a priori hypothesis or detail any modifications that were made to the original model.

Method

Sample Selection

To obtain a broad sample of CFA studies across psychological research, we searched the PsychINFO database (1998–2006) using the keywords *confirmatory factor analysis* and *CFA*. This approach to identifying studies was consistent with previous reviews (Hershberger, 2003). We also limited our search to APA journals, which tend to be highly selective when it comes to accepting research manuscripts (23.6% acceptance rate; APA, 2007). Only articles using CFA techniques were selected, excluding literature reviews and articles that dealt with theoretical or methodological issues. Further, we excluded any articles that did not use an SEM approach to CFA (e.g., studies using Procrustes rotation).

Coding

A coding sheet was developed that closely mirrored the reporting categories outlined earlier (theoretical formulation and data collection, data preparation, analysis decisions, model evaluation and modification, and reporting findings). Because we did not possess the breadth of content expertise to evaluate the theoretical justification of the CFA models, we focused on coding the methodological aspects of each study. These fields are reported in Table 1 by category. We also recorded as much information as was provided for each model tested and noted which model the researchers preferred as the final model.

Table 1
Coded Features of Confirmatory Factor Analysis Studies

Theoretical formulation and data collection:
Number of models posited a priori
Sample size
Type of model tested
Whether structural equation modeling was used for subsequent validation
Data preparation:
Assessment of multivariate normality
Assessment of univariate normality
Identification of multivariate outliers
Discussion of missing data and how it was addressed
Data transformations
Parceling
Analysis decisions:
Type of matrix analyzed
Estimation method
Software brand and version
Method of fixing scale of latent variables
Model evaluation and modification:
Number, type, and values of fit indices reported
Used Hu & Bentler's (1999) strategy for assessing fit
Cutoffs for fit indices were stated clearly and a priori
Rationale for the fit indices reported
Researchers engaged in model modification
Reporting findings:
Latent variable correlations
Factor loadings (pattern loadings)
Standard errors
Graphic representation of model
Structure coefficients

Procedure

The CFA articles identified were randomly assigned to the three authors for coding. We calibrated definitions for the coding sheet by independently coding a common set of articles and then comparing and discussing our results. When there were discrepancies, we discussed the reasons until we reached agreement. At the beginning of the coding process, these discussions served to help us refine and standardize our coding definitions. For the duration of the project, authors met regularly via conference calls to discuss coding. These meetings were used to discuss issues or problems related to the review process and as a periodic check on agreement. As an example, we originally attempted to code the orientation of the research, such as whether it was strictly confirmatory, exploratory, or something in between (see Jöreskog, 1993). However, we ultimately abandoned this because we found it too difficult to determine the orientation due to a lack of information provided by authors.

Out of concern for the integrity of the coding process, we examined the level of coder agreement. A research assistant independently assigned a subset of articles to both of the first two authors. Three of these articles (18 models) were examined to check for coder agreement. Thirty-six fields from the coding sheet were compared (excluding obvious fields such as year of publication and journal), and the percent agreement scores were calculated and then averaged across the 18 models. Overall agreement rate was .88 (range = .78–.97). The third author joined the study later and was trained by the first author, who also spot-checked the third author's coding sheets as part of the training.

Data Analysis

There were three levels of analysis relevant to the current study. First, there was the *study level.* Data analyzed at the study level, where each study (i.e., published article) counted equally, were used to answer questions such as which journals published the most CFA studies, or which software applications were most frequently used. Second, because several studies included substudies, in which authors might have more than one "final" or "preferred" model (e.g., confirming a measure on several independent samples or under different conditions), data could also be analyzed at

the *preferred-model level*. We analyzed data at this preferred-model level to answer questions such as whether authors reported parameter estimates for their models. Finally, authors may have tested several alternative models that were not selected as preferred models. For instance, it was fairly common for researchers to compare an oblique factor model to an orthogonal model and to a single-factor model, where the oblique model might have been the model hypothesized to be the correct one, with the other two serving as conceivable alternatives. Thus, analyses could also be performed at the *total-model level*. We will use these terms—*study level*, *preferred-model level*, and *total-model level*—when reporting results to ensure clarity about what the various statistics reported represent.

Results

A total of 212 studies were initially identified from the literature search. Eighteen studies were eliminated as they were not CFA studies, for instance, technical articles or articles in which the methodology, while confirmatory in nature, did not incorporate an SEM approach. Therefore, a total of 194 studies, published in 24 journals, composed the final sample. These 194 studies included a total of 1,409 models. The journals that published the highest number of CFA studies were *Psychological Assessment* ($n = 69$), *Journal of Applied Psychology* ($n = 19$), *Journal of Counseling Psychology* ($n = 14$), and the *Journal of Personality and Social Psychology* ($n = 13$). We begin with a summary of reporting practices represented in the 194 articles reviewed for this study. The number published by year ranged from a low of 13 in 2002 to a high of 33 in 2000. There were only 2 years in which fewer than 20 were published (2001 and 2002). The mean number of articles per year, for the 9 years covered by this study, was 21.6.

Theoretical Formulation and Data Collection

The majority of studies reviewed (*75.5%*) focused on validating or testing the factor structure of an instrument. Of these, most (77.7%) used an existing instrument, with the remainder (22.3%) reporting on validation of a new measure. The remaining studies, which did not involve instrument validation, used CFA to either examine constructs or theories (15.8%) or to assess a measurement model prior to conducting a SEM (8.8%). Authors of the studies reviewed often validated their findings, such as by creating scale scores and correlating those measures with other criteria. However, it was far less common to conduct this subsequent validation within the SEM framework, such as by specifying a model that included the latent variables and allowing them to correlate with other latent variables of interest. At the study level, less than one in five studies (15.5%) incorporated this approach. The majority of the articles (63.8%) posited more than one a priori model. However, it was often difficult to assess specifically what models or how many models were to be tested due to lack of sufficient information. For instance, in some studies, new models were introduced in the results section but labeled as a priori. Examining all models tested in articles published from 1998 through 2006 ($n = 1,409$), the most commonly tested models were correlated factor models (50.5%), followed by orthogonal (12.0%), hierarchical (10.6%), multisample (9.8%), single factor (9.5%), and multitrait multimethod (2.3%). We were unable to determine the nature of 5.3% of the models. For example, in some studies, latent variable correlations were not reported, and it was not clearly stated whether an orthogonal model was being tested or whether latent variable covariances were free to be estimated.

Based on the preferred-model level of analysis, the median sample size was 389, with a minimum of 58 and a maximum of 46,133. One-fifth (20.3%) of the preferred models were tested on samples smaller than 200. At the other extreme, 14.7% of the models were tested on samples greater than 1,000. To put this in perspective, in terms of the size of models being tested, the median number of measured variables was 17 (with 12 and 24 representing the 25th and 75th percentiles) and the median number of first-order latent variables was three (with 89.6% having six or fewer). It should be noted that when very small samples were used, authors typically addressed the reasons for this and justified their use of CFA.

Data Preparation

To answer questions with regard to data preparation, we utilized the study level of analysis ($n = 194$). For the vast majority of studies reviewed here, little information was provided regarding the extent to which the assumptions of CFA were met. For instance, approximately one-fifth (21.6%) of the studies clearly indicated that measured variables had been examined for univariate normality. In only 13.4% of the articles was it clear that MVN had been examined, and in even fewer instances (3.6%) was it mentioned that data had been screened for multivariate outliers.

A similar pattern was observed for missing data. The majority of studies (64.9%) did not report whether they had missing data or, obviously, what measures they may have taken to address missing data. Of those researchers who specifically indicated a method for dealing with missing data ($n = 68$), the most popular approach was listwise deletion (66.2%), followed by mean substitution (10.3%). In total, few researchers reported using a more sophisticated method such as expectation maximization (5.9%), or multiple imputation (4.4%).

Finally, in terms of data preparation, the majority of studies (86.1%) did not indicate the use of any data transformations, whereas a few reported doing either a square-

root or logarithmic transformation (5.1%). A number of studies (16.0%) used item parcels. A larger number of studies (38.0%) used either parcels or scales for measured variables, rather than item-level data.

Analysis Decisions

Using study-level data we found that, not surprisingly, ML was the most commonly used estimation method (41.8%), followed by the Satorra–Bentler ML correction (14.4%). It should be noted, however, that the method of estimation used was not reported in 33.0% of studies. Additionally, in most cases, either a covariance matrix was analyzed (34.5%) or the type of matrix analyzed was not reported (59.8%). Only a small proportion (1.5%) reported analyzing a Pearson correlation matrix. A few studies reported analyzing polychoric or tetrachoric correlation matrices (4.1%). In about one-fifth of the studies (18.6%), authors supplied the covariance matrix (or equivalent information), indicated that it would be provided upon request or used a matrix that was readily available elsewhere (e.g., from a test manual).

The breakdown of software used was as follows: LISREL (28.9%), EQS (23.2%), Amos (17.0%), MPlus (7.2%), SAS PROC CALIS (3.6%), and RAMONA (0.5%). Nearly one fifth (19.6%) did not report the type of software used. Additionally, in the vast majority of studies (77.8%), it was not indicated how the scales of latent variables had been fixed. Of those who did report this information, the favored method seemed to be fixing the latent variable variances to 1.0 (69.8%).

Model Evaluation and Modification

We used study-level data ($n = 194$) to determine the types of fit measures reported and preferred-model-level data ($n = 389$) to determine the average fit measures for these preferred models. Table 2 summarizes the fit measures reported as well as the means for these measures. As seen in this table, nearly all authors reported chi-square values (89.2%). After chi-square values, the most commonly reported measures of fit were CFI (78.4%), RMSEA (64.9%), and TLI (46.4%). The information-based fit measures such as the Akaike information criterion (AIC; Akaike, 1987) tended to be least frequently reported. Null-model statistics were reported in 7.2% of studies. The modal number of fit measures reported was three to four (25.3% each), with 92.8% of the authors reporting more than one type of fit measure, such as an absolute measure and an incremental measure (chi-square values counted as an absolute measure of fit).

We examined model modification practices using total-model-level data. Authors indicated that they modified

Table 2
Descriptive Statistics for Values and Frequency of Reporting Various Fit Measures

Index	Studies		No. of preferred models	*M*	*SD*
	n	%			
Chi-square (χ^2)	173	89.2	357	678.759	1277.162
Degrees of freedom (*df*)	173	89.2	357	228.661	439.411
χ^2/df ratio	42	21.6	62	3.034	2.492
Root-mean-square error of approximation (RMSEA)	126	64.9	231	0.062	0.026
Root-mean-square residual (RMR)	29	14.9	39	0.060	0.042
Standardized RMR	45	23.2	84	0.054	0.022
Goodness of fit index (GFI)	66	34.0	117	0.906	0.062
Adjusted GFI	39	20.1	66	0.862	0.080
McDonald's centrality index	2	1.0	2	0.936	0.020
Normed fit index	46	23.7	109	0.912	0.067
Tucker–Lewis index (TLI)	90	46.4	194	0.925	0.053
Comparative fit index (CFI)	152	78.4	320	0.933	0.059
Relative noncentrality index (RNI)	4	2.1	8	0.951	0.028
Bollen's rho 1 (relative fit index)	3	1.5	4	0.951	0.012
Bollen's delta 2 (incremental fit index)	17	8.8	33	0.938	0.044
Akaike information criteria (AIC)	20	10.3	36	881.582	3495.553
Consistent AIC	3	1.5	6	164.790	235.018
Schwarz's Bayesian criteria (SBC)	1	0.5	2	−74.940	44.746
Expected cross-validation index (ECVI)	13	6.7	29	1.983	2.380
Absolute standardized residuals	3	1.5	20	0.278	0.345
Null model χ^2	14	7.2	25	3738.213	4011.448
Null model *df*	14	7.2	25	308.880	477.191

Note. The frequency and percentage reporting each measure are based on the study level ($N = 194$), whereas the mean and standard deviations of each measure are based on the number reporting the measure at the preferred model level (*N* of preferred models).

13.2% of the 1,409 models tested. Further, cross-validation of models on separate samples was conducted for 18.7% of the models. There was a weak correlation between whether a model was modified and whether it was cross-validated ($\Phi = .088$, $p < .01$). Only 27.6% of the modified models were cross-validated, whereas 17.4% of models that were not modified were cross-validated.

To further comment on the methods-of-fit assessment, we examined study-level data to determine whether cutoff criteria had been clearly established for accepting models and whether some rationale for the choice of fit measures had been provided. About half of the studies (57.2%) stated an explicit cutoff criteria and approximately one-third (36.1%) provided a rationale for their choice of fit measures.

Reporting Findings

Table 3 provides a summary of reporting for the preferred-model-level data. It should be noted that researchers did not always report all parameter estimates and in many cases did not clearly state whether estimates were standardized or unstandardized (e.g., 20.8% did not specify the type of loading reported, and 42.7% did not report the loadings). Further, it was less common for researchers to report standard errors (12.6%). Finally, it was not common practice to provide a graphic representation of the models tested (i.e., only 30.1% depicted models), though this might be excused in the case of CFA, where models may be inferred from the written description or tables of parameter estimates (when provided).

Trends in Use of Fit Indices

The second research question involved how model fit was assessed in light of divergent perspectives offered in the literature. First, we approached this question by looking at the effects of Hu and Bentler's (1999) article recommending higher values for incremental fit measures and the two-index presentation strategy. Average fit index values for selected indices are presented in Table 4. With respect to the two-index presentation strategy, we looked at the number of studies published after 1999 in which this strategy was used (on the basis of study-level data). The first instances of this approach appeared in 2000 and only 11.8% of studies published from 2000 through 2006 reported using this strategy. Further, we examined use of the strategy by year and found no clear indication of an increasing trend in the use of this approach; it remained infrequently used through 2006, though there is a peak in the last year (see Table 4).

Table 3
Percentage of Preferred Models Reporting Descriptive Attributes

Report	Frequency	Percent
Latent variable correlations	106	48.6
Standardized loadings	122	31.4
Unstandardized loadings	20	5.1
Either standardized or unstandardized loadings (not specified by authors)	81	20.8
No loadings	166	42.7
Measurement errors	49	12.6
Graphic depiction	117	30.1
Structure coefficients	2	1.0

Note. Percentages are based on all preferred models ($N = 389$) except for those pertaining to latent variable correlations and structure coefficients ($N = 218$), where percentages are based on oblique models only.

Second, we examined the percentage of preferred models that would have been acceptable based on Hu and Bentler's (1999) recommended cutoffs for RMSEA (.06), TLI (.95), and CFI (.95). These acceptance rates are reported in Table 5. Although the percentage of models meeting or bettering these cutoffs did increase on average for the incremental measures, beginning with articles published in 2000, there was not a clear increasing trend. Collapsing across years, in order to compare 1998 and 1999 with 2000 through 2006, the increase seems clearer for TLI and CFI in that a higher percentage of models exceeded Hu and Bentler's cutoffs. The only effect that was significant was for the TLI: CFI: $\chi^2(1, N = 320) = 2.263$, $p > .05$; TLI: $\chi^2(1, N = 194) = 6.314$, $p < .05$; RMSEA: $\chi^2(1, N = 231) = 3.116$, $p > .05$. Hence, there was a significant increase in the proportion of models meeting or bettering the cutoffs for the TLI after the publication of Hu and Bentler's article but not for the CFI and RMSEA. This makes sense in that their recommendations with regard to RMSEA were similar to previously recommended cutoffs (e.g., Browne & Cudeck, 1993); however, their recommendations regarding incremental fit measures were substantively higher than previous recommendations (e.g., Bentler & Bonett, 1980). Furthermore, RMSEA was less frequently reported prior to 2002.

Third, we sought to examine whether there was any evidence that the percentage of models meeting or exceeding the higher cutoffs for IFIs decreased at all after the publication of the Marsh et al.'s (2004) article, which cautioned against overinterpreting Hu and Bentler's (1999) recommendations. Although we did not have a great amount of data for this analysis (only 2005 and 2006 articles), we did attempt it. We found that the percentage of preferred models exceeding the .95 cutoffs for TLI and CFI actually increased in 2005 and 2006. This suggests that either recommendations against overinterpreting Hu and Bentler have not yet had an effect or models tested by researchers and selected for publication tend to be better fitting. The percentage of models that surpass Hu and Bentler's recommended cutoffs by year are presented in Table 5.

With respect to the number and types of fit measures reported, and given the divergent perspectives on assessing fit and cutoffs for fit measures, it is possible that researchers might choose to report more fit measures in an effort to

Table 4
Values of Selected Fit Measures by Year of Publication for Preferred Models

	RMSEA			TLI			CFI			Studies using two-index strategy	
Year	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	%
1998	30	.065	.027	28	.873	.124	28	.940	.049	0	0.0
1999	39	.072	.045	31	.916	.055	33	.928	.038	0	0.0
2000	57	.062	.038	52	.906	.077	52	.922	.051	3	9.1
2001	67	.055	.027	58	.912	.067	59	.931	.064	1	5.9
2002	26	.068	.019	25	.904	.059	25	.925	.053	2	15.4
2003	33	.066	.031	32	.879	.103	32	.903	.088	1	4.3
2004	33	.066	.025	29	.922	.056	29	.940	.041	4	16.0
2005	32	.065	.024	28	.914	.060	29	.928	.051	1	5.0
2006	40	.062	.040	32	.907	.084	34	.944	.055	6	28.6
Total	357	.064	.033	315	.904	.079	322	.929	.057	18	11.8

Note. RMSEA = root-mean-square error of approximation; TLI = Tucker–Lewis index, sometimes labeled nonnormed fit index (NNFI); CFI = comparative fit index. When these values were not reported, an attempt was made to impute them on the basis of information provided by authors. The total percentage of studies using Hu & Bentler's (1999) two-index cutoff strategy was calculated using studies published after 1999 ($N = 153$), and all percentages were based on one representative model per study in order to weight each study equally.

satisfy editors and reviewers of the adequacy of their model. Similarly, we wanted to know whether there was a tendency for authors to report more fit measures that have tended to perform well in simulation studies and fewer fit measures that have not faired as well. Toward this end, we counted the total number of fit measures for each study (study level) as well as the number of ancillary fit measures that have tended to perform well (RMSEA, SRMR, CFI, TLI, Bollen's delta 2, CI) and those that have tended to not perform as well (GFI, AGFI, NFI, Bollen's rho 1). These values are reported in Table 6 as well as the ratio of recommended to not-recommended fit measures. There is a trend toward authors reporting more fit measures, which might be an artifact of computer software development. There is, however, also a trend toward authors reporting more of the recommended fit measures and fewer of the not-recommended measures. This ratio was lowest in 1998, less than one, and then rose to between two and three from 1999 through 2003, and exceeded four from 2004 through 2006. It is worth noting here that there has been an increase in the frequency of reporting RMSEA. From 1998 through 2000, only 37.3% of the studies reported RMSEA values, but for the most recent 3 years (2004–2006), that number increased to 83.3%.

Table 5
Number and Percentage of Models Meeting or Bettering Cutoff Criteria Proposed by Hu and Bentler (1999)

	RMSEA		TLI		CFI	
Year	*N*	%	*N*	%	*N*	%
1998	11	81.8	6	33.3	25	60.0
1999	18	61.1	20	10.0	35	25.7
2000	24	41.7	31	19.4	42	40.5
2001	26	65.4	46	32.6	66	54.6
2002	25	40.0	7	14.3	24	29.2
2003	29	44.8	18	38.9	29	51.7
2004	26	53.9	15	53.3	31	51.6
2005	31	54.8	13	38.5	32	46.9
2006	41	56.1	24	75.0	36	72.2
1998 & 1999	29	69.0	26	17.2	60	40.0
2000–2006	202	51.5	165	41.8	260	50.8
Total	231	53.7	194	38.1	320	48.8

Note. RMSEA = root-mean-square error of approximation; TLI = Tucker–Lewis index, sometimes labeled nonnormed fit index (NNFI); CFI = comparative fit index. Values represent the percentage of models that authors identified as preferred or accepted that would be acceptable according to cutoffs proposed by Hu and Bentler (1999). Cutoffs are .06 for RMSEA and .95 for both TLI and CFI. The frequencies (*N*) are the number of preferred models where authors reported each fit measure (e.g., six preferred models had TLI values reported in 1998).

Selection Bias in Reporting Fit Measures

The third research question was whether the researchers reported fit measures that supported the preferred or accepted model versus fit indices that did not support the model. As indicated, authors seldom reported all available fit indices and sometimes left out commonly used measures of fit. However, enough information was provided in many studies so that the values of missing fit statistics could be estimated. For instance, one can estimate the null model chi-square when incremental fit measures are reported, which, when coupled with the degrees of freedom for both the null model and the model under investigation, and the chi-square for the model under investigation, allows one to construct other incremental fit measures. Likewise, one can calculate the RMSEA by knowing the chi-square, degrees of

Table 6
Number of Fit Indices Total, Recommended, and Not Recommended in the Literature by Year

Year	Total fit measures reported	Fit measure		Ratio of recommended to not-recommended fit measures
		Recommended	Not recommended	
1998	2.700	1.000	1.100	0.909
1999	3.136	1.818	0.818	2.222
2000	3.303	1.939	0.879	2.206
2001	3.412	2.177	0.765	2.846
2002	4.000	2.462	1.000	2.462
2003	4.261	2.565	0.870	2.948
2004	3.560	2.560	0.640	4.000
2005	4.050	2.750	0.650	4.231
2006	4.000	2.905	0.476	6.103

Note. Values are based on the study level ($N = 194$). Total fit measures reported include all fit indices/statistics reported (including the chi-square test). The recommended fit measures include the root-mean-square error of approximation, standardized root-mean-square residual, confidence fit index, Tucker–Lewis index, Bollen's delta 2, and the confidence interval. The not-recommended fit measures include the goodness-of-fit index, adjusted goodness-of-fit index, normed fit index, Bollen's rho 1. The ratio of recommended to not-recommended fit measures is the recommended value divided by the not-recommended value, such that higher values indicate a higher number of recommended fit measures were reported relative to not-recommended measures.

freedom, and sample size. Using this approach, we calculated fit indices for all preferred models, and independent samples t tests[1] were used to compare RMSEA, $t(147) = 1.152$, $p > .05$; CFI, $t(14.426) = -1.482$, $p > .05$; and TLI, $t(131) = -1.091$, $p > .05$, for models for which authors reported these measures and models for which authors did not report these measures. There were no statistically significant differences between the fit measures reported and those not reported. However, the mean values of the fit measures not reported—RMSEA = .0691; CFI = .9054; TLI = .9034—tended to be less favorable than the fit measures that the authors reported—RMSEA = .0630; CFI = .9325; TLI = .9156.

Discussion

Since the advent of modern computer software, CFA has become an essential tool for psychological researchers interested in construct validity, be it in the context of scale development, construct validation, or measurement model validation in SEM applications. This article reviewed CFA reporting practices in a relatively large sample of studies published in APA journals from 1998 to 2006. The primary objective of the current study was to assess the degree to which researchers using CFA techniques follow guidelines proposed for reporting studies involving SEM/CFA procedures. This work also serves as a baseline measure of CFA reporting that could be used in future reviews of CFA practice. Overall, we found many instances of what could only be considered excellent reporting. Unfortunately, these instances were not representative of the CFA studies reviewed as a whole. Consistent with previous reviews of CFA studies (e.g., DiStefano & Hess, 2005; Russell, 2002; and Schreiber et al., 2006), our findings revealed a variety of reporting problems.

Initially, however, we highlight some more encouraging findings. For example, the majority of studies posited more than one model a priori and reported the chi-square values, degrees of freedom, and p values associated with the models tested. Additionally, nearly all studies reported multiple fit measures from different families, namely absolute and incremental. Furthermore, the most commonly reported fit measures were those that have been found to perform generally well in Monte Carlo studies (e.g., Fan et al., 1999; Hu & Bentler, 1999).

Another encouraging finding was that more than two thirds of studies (67.0%) reported the estimation procedure. This is somewhat consistent with DiStefano and Hess (2005) who found 62% and is higher than the 42% reported by Russell (2002) or the 50% by Schreiber et al. (2006). In addition, the vast majority of studies used fairly large sample sizes, with few (7.7%) using very small samples ($n < 100$). Even when small samples were used, reports tended to provide a credible rationale for the small sample, as well as the decision to proceed with

[1] Since several articles had multiple preferred models, this represented a threat to the assumption of independence of observations, as many of the multiple studies involved data sampled from the same population, use of identical instruments, and other study characteristics that could not be considered independent. Therefore, for these studies with multiple preferred models, means for each fit measure were computed across the multiple models so that each article only contributed one final model.

CFA (e.g., data that were rare or difficult to collect). Finally, most studies described the tested models in sufficient detail so as to allow the reader to comprehend what was being tested, although there were exceptions to this. For instance, there were cases in which we could not tell if latent variables were allowed to correlate or exactly which measured variables were specified to load on each latent variable.

In our view, however, these positive practices were far outweighed by the paucity of reported information in most studies. Our findings in this regard are consistent with previous reviews of CFA applications. For example, like DiStefano and Hess (2005), we found that most studies did not specify the type of matrix analyzed. This finding is of concern as analyzing a correlation matrix, depending upon the nature of the model and the software used, may lead to incorrect parameter estimates and even fit indices (Cudeck, 1989; MacCallum & Austin, 2000), and it is generally deemed inappropriate when studies involve multiple samples (Loehlin, 2004). Further, nearly half of the studies did not report either standardized or unstandardized factor loadings—a rather surprising finding given that these were CFA studies, with the majority focused on psychometric evaluation of measurement instruments. Similarly, approximately half of the studies failed to report latent variable correlations (for oblique models). One might argue that authors did not mention certain aspects of their study because they viewed their decisions as rather standard, such as using a covariance matrix with ML estimation. Nevertheless, when reading studies in which basic information is absent, it is impossible to tell whether this is the case or to evaluate whether reasonable analytic decisions were made.

Some of the more glaring omissions had to do with data preparation. Although ML may be robust to mild violations of normality (Chou et al., 1991; Fan & Wang, 1998; Hu et al., 1992), researchers seldom mentioned whether they had examined their data for normality. This finding supports previous negative reviews of screening for univariate and multivariate normality in CFA and SEM studies (DiStefano & Hess, 2005; McDonald & Ho, 2002; Schreiber et al., 2006). Furthermore, the overwhelming majority of studies used Likert-type survey items, and most based their analyses on item-level data. However, few studies indicated analyzing polychoric correlation matrices. Although basing analyses on polychoric correlation matrices may not always be advisable, there are applications under which it can lead to less biased test statistics and parameter estimates (Flora & Curran, 2004). A more thorough discussion of these issues can be found in DiStefano (2002).

Another consistent omission concerned the treatment of missing data. The vast majority of the studies utilized questionnaire data. Data collected by this method nearly always yields some level of missing data. It is unreasonable to assume that well over half the studies examined here (64.9%) had no missing data problems. As in previous reviews (McDonald & Ho, 2002; Schreiber et al., 2006), the reporting of missing data continues to be an issue in CFA and SEM studies. We recommend that researchers report their efforts with respect to determining the mechanism for missing data and any judgment with regard to whether the pattern of missing data and the solution for it is likely to affect aspects of the solution, such as parameter estimates. Furthermore, we suggest that researchers utilize methods of addressing missing data besides listwise deletion. A number of options for addressing missing data are reviewed by Schaefer and Graham (2002). They recommend that currently the best choices available to researchers are ML-based methods such as full-information maximum likelihood (Arbuckle, 1996) and multiple imputation methods (e.g., Rubin, 1978). Monte Carlo studies specifically aimed at examining the relative effectiveness of different methods of dealing with missing data in SEM appear to favor the full-information maximum likelihood approach (Brown, 1994; Enders, 2001; Enders & Bandalos, 2001).

It was mentioned above that the studies tended to report some of the more frequently recommended measures of fit (e.g., CFI, TLI, chi-square values, and RMSEA), which is an encouraging finding. However, only about half of studies reported explicit cutoff standards for fit indices. Although this is a higher estimate than those in previous reviews (DiStefano & Hess, 2005; Russell, 2002; Schreiber et al., 2006), it is still low, as specifying fit criteria a priori helps readers understand the context of decisions about model fit.

This study also explored how researchers report model fit in light of the divergent perspectives present in the literature. We found that model fit measures in the later years covered by this study were more consistent with Hu and Bentler's (1999) recommendations; however, the change was not statistically significant, with the exception of the TLI. We also did not find evidence that warnings about strict adherence to Hu and Bentler's suggestions were being heeded, though it may be too early to detect such an effect. It bears mentioning that Marsh et al. (2004) recommended using a norm-reference approach to evaluating model fit, such that fit indices around .90 might be acceptable in areas where that is the norm, whereas in other areas such an index might be considered deficient. One would hope that such contextual factors are considered by authors, reviewers and editors. This issue would be an excellent focus for future reviews of CFA/SEM studies.

Finally, this study also examined the relation between what fit measures were reported and their support (or lack thereof) for the preferred or final model. In other words, were some fit measures not reported (or suppressed) because they were inconsistent with other fit measures? This seems a reasonable question given the variety of fit indices computed by most software applications and the choice this

affords to researchers. Descriptively speaking, there was a trend for those indices that were not reported to be less favorable when compared with studies in which they were reported. However, if authors were consistently choosing fit measures that supported their preferred model, we would expect to see significant differences in at least two, if not all three, of the most commonly reported indices. Overall, our findings do not support the notion that "cherry picking" fit measures is a common practice.

Recommendations

It makes sense that such a review of reporting practices should conclude with some concrete recommendations for reporting. At the same time, it seems superfluous to make recommendations that have been so eloquently proposed by other authors (e.g., Boomsma, 2000; Hoyle & Panter, 1995; McDonald & Ho, 2002). In short, we feel that authors should generally follow recommendations set forth in the articles just cited. Additionally, we propose that journal editorial teams may benefit from adopting some minimum guidelines for reporting CFA and SEM research. Although the latter recommendation could be viewed as being overly prescriptive, guidelines need not be hard and fast criteria that must be met for every research study, nor do they preclude editors and reviewers from taking into account other factors, such as journal space, author intention, or journal mission. Some suggestions for consideration follow.

With respect to model description, we propose that authors should clearly define all models they propose to test and label any post hoc modifications as such. This includes clearly defining which measures identify each latent variable, revealing whether latent variables are correlated, and indicating any other pertinent information, such as whether some error terms are allowed to covary and whether any constraints are used. We also suggest that authors strongly consider including alternative models that are theoretically plausible and identify plausible equivalent models. In terms of data preparation, some description of data cleaning and assessment of univariate and multivariate normality should be provided. Additionally, we feel it is important for authors to clearly identify the extent of missing data, any analyses conducted to assess whether the missing data were deemed ignorable, and how missing data were handled. Granted, this requires substantially more work for authors who are currently dealing with missing data through listwise deletion or accepting some default method of dealing with missing data, but it is difficult to argue that research findings are relevant without making mention of how missing data problems were addressed.

For analysis decisions, we propose that authors minimally reveal the type of matrix they use as input, the estimation procedure used, and how latent variables are scaled. This requires very little effort or journal space and ensures the reader can properly evaluate and replicate research findings. Furthermore, either a covariance matrix or equivalent information (i.e., correlations and standard deviations) should be either included in the article or made available on a Web site or upon request.[2] Finally, there are some differences among software applications, thus the brand and version used should be communicated.

For model evaluation, authors should indicate the cutoff values for fit measures they intend to use. Although there is no universally agreed upon number of fit indices to report, a minimal set would include the chi-square value and the associated degrees of freedom and probability value, an index to describe incremental fit, such as the TLI, CFI (or RNI), or Bollen's delta 2, and a residuals-based measures (e.g., RMSEA and its associated confidence intervals or SRMR). Other approaches to assessing fit should also be described, such as examining the residual matrix and determining whether the magnitudes and signs of parameter estimates are appropriate. The examination of the residual matrix is performed to ensure that one does not overly emphasize global fit at the expense of fit of all the relations among measured variables (Kline, 2005).

Finally, researchers should report all parameter estimates necessary for the reader to make an interpretation of the results. This minimally would include either standardized or unstandardized loading estimates for manifest variables on latent variables, structural regression coefficients, and latent variable covariances or correlations. Whether authors choose to report standardized or unstandardized coefficients depends upon which better enables interpretation for their particular model. Other parameter estimates can be helpful to readers, such as standard errors, confidence intervals, and squared multiple correlation coefficients (R^2) values for endogenous variables. This information is especially useful when considering models that fit well, to understand whether they also predict well.

It may seem burdensome for reviewers to keep track of even a trimmed-down version of good reporting practices. Furthermore, it seems inadequate for us to submit yet another review lamenting the poor state of reporting in CFA/SEM, followed by a reiteration of good reporting practices. Toward this end, we have supplied a generic checklist that could be used in the review process to improve reporting practices (see Appendix). We hope that such a checklist could be easily incorporated into the writing or review process to improve reporting for studies using SEM/CFA.

[2] A reviewer of an earlier version of this article recommended that the covariance matrices should be made available on the journal Web sites in order to avoid problems with faculty/researcher mobility and Web address changes and to also save journal space.

For example, the checklist might be used by authors to ensure appropriate information is reported prior to submission for publication.

Limitations and Future Research

A potential limitation of the current study is that we located articles using keyword searches. Thus, it is likely that there were other CFA studies that we did not review. This might have included studies in which CFA was not viewed as a central aspect of the study, for example, when used as a preliminary analysis to a SEM. Or the use of CFA might be viewed as routine and thus not warrant mention as a keyword. We feel that in cases such as these, reporting would likely be even sparser, given that CFA was not viewed as a critical component of the study.

Another issue related to this study has to do with limiting our search to APA journals. Many other journals publish psychological research using CFA techniques, and we could have extended our search beyond APA journals. We decided, however, that we could better control for the possible confounds of journal quality and variable editorial standards by choosing only APA journals. APA regularly assesses and publishes statistics on journal operations (i.e., acceptance and rejection rates). Thus, we reasoned that APA journals would provide a consistent and high-quality sample of CFA applications. Our results should not, however, be generalized beyond APA journals. Future reviews may also consider broadening the scope of journals sampled.

A related issue is that because of page limits and the focus on content in APA journals, authors may have been discouraged from reporting technical details (e.g., multivariate normality, missing data issues, matrix analyzed, actual data or matrices) as well as reporting model modifications. With regards to model modifications, Baumgartner and Homburg (1996) suggested in their review of SEM studies that it was unlikely that the model initially specified by authors was the one ultimately represented as the most parsimonious summary of the data. They observed that whereas some authors discuss modifications in great detail, others presented only the final model. Similarly, we cannot rule out this possibility for our sample. According to simulation work by MacCallum (1986), specification searches can fail to uncover the correct underlying model especially when the author's search was motivated by a need to improve overall fit of the model, as this approach to model modification may be capitalizing on chance. We agree with Baumgartner and Homburg's (1996) recommendation that model modifications must be guided by careful consideration that is theoretically meaningful. Further, our ability to evaluate the extent to which authors modified their initial model is limited. We believe that future research aimed at answering this question would be useful. It also deserves to be mentioned that even though page limits may be perceived to hinder more thorough reporting with regard to analysis decisions, it requires relatively little space to at least mention that data were screened to ensure assumptions of the statistical techniques were met or to indicate the estimation method used in the analysis for example.

This study highlights important discrepancies between established CFA reporting guidelines and actual reporting practices and contributes to our understanding of the conduct of CFA studies. Although there are some positive findings, many problems persist, often involving information that is relatively easy to report (e.g., method of estimation, screening data for multivariate normality, management of missing data, and what matrix was analyzed). This type of information is often vital to understanding researchers' decision making and evaluating the validity of the results. Studies such as this one should be conducted periodically to monitor CFA practices and to encourage the reporting of more complete information. As McDonald and Ho (2002) eloquently stated, "Completeness is essential" (p. 78).

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52,* 317–332.

American Psychological Association. (2007). Summary report of journal operations, 2006 [and] summary report of division journal operations, 2006. *American Psychologist, 62,* 543–544.

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.* Mahwah, NJ: Erlbaum.

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling, 9,* 78–102.

Barrett, P. (2007). Structural equation modeling: Adjudging model fit. *Personality and Individual Differences, 42,* 815–824.

Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing, 13,* 139–161.

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

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

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences, 42,* 825–829.

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

Bollen, K. A. (1986). Sample size and Bentler and Bonett's nonnormed fit index. *Psychometrika, 51,* 375–377.

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

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.

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.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York: Guilford.

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). Newbury Park, CA: Sage.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming.* Mahwah, NJ: Erlbaum.

Chin, W. W. (1998). Issues and opinion on structural equation modeling. *MIS Quarterly, 22,* vii–xvi.

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

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

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.

DiStefano, C. (2002). Effects of ordered categorical data with confirmatory factor analysis. *Structural Equation Modeling, 9,* 327–346.

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment, 23,* 225–241.

Enders, C. K. (2001). The impact of non-normality on Full Information Maximum-Likelihood Estimation for structural equation models with missing data. *Psychological Methods, 6,* 352–370.

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.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indices to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling, 12,* 343–367.

Fan, X., Thompson, B., & Wang, L. (1999). The effects of sample size, estimation methods, and model specification on SEM fit indices. *Structural Equation Modeling, 6,* 56–83.

Fan, X., & Wang, L. (1998). Effects of potential confounding factors on fit indices and parameter estimates for true and misspecified SEM models. *Educational and Psychological Measurement, 58,* 701–735.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466–491.

Goffin, R. D. (2007). Assessing the adequacy of structural equation models: Golden rules and the editorial policies. *Personality and Individual Differences, 42,* 831–839.

Hayduk, L. A. (1988). *Structural equation modeling with LISREL: Essentials and advances.* Baltimore, MD: The Johns Hopkins University Press.

Hershberger, S. L. (2003). The growth of structural equation modeling: 1994–2001. *Structural Equation Modeling, 10,* 35–46.

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.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424–453.

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

Hulland, J., Chow, Y. H., & Lam, S. (1996). Use of causal models in marketing research: A review. *International Journal of Research in Marketing, 13,* 181–197.

Jackson, D. L. (2007). The effect of the number of observations per parameter in misspecified confirmatory factor analytic models. *Structural Equation Modeling, 14,* 48–76.

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.

Jöreskog, K. G., & Sörbom, D. (1986). *LISREL VI: Analysis of linear structural relationships by maximum likelihood and least squares methods.* Mooresville, IN: Scientific Software.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions.* Thousand Oaks, CA: Sage.

Kenny, D. A. (2006). Series editor's note. In T. A. Brown (Ed.), *Confirmatory factor analysis for applied research* (pp. ix–x). New York: Guilford.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis* (4th ed.). Mahwah, NJ: Erlbaum.

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

MacCallum, R. C. (1995). Model specification: Procedures, strategies, and related issues. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications.* Thousand Oaks, CA: Sage.

MacCallum, R. C., & Austin, J. T. (2000). Applications of struc-

tural equation modeling in psychological research. *Annual Review of Psychology, 51,* 201–226.

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.

Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on Paul Barrett's recommendations for reporting model fit in structural equation modeling. *Personality and Individual Differences, 42,* 851–858.

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

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.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320–341.

Martens, M. P. (2005). The use of structural equation modeling in counseling psychology research. *The Counseling Psychologist, 33,* 269–298.

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

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7,* 64–82.

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

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction.* New York: Guilford Press.

Medsker, G. J., Williams, L. J., & Holahan, P. J. (1994). A review of current practices for evaluating causal models in organizational behavior and human resources management research. *Journal of Management, 20,* 439–464.

Miles, J., & Shevlin, M. (2007). A time and a place for incremental fit indices. *Personality and Individual Differences, 42,* 869–874.

Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences, 42,* 875–881.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9,* 599–620.

Nevitt, J., & Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling, 8,* 353–377.

Powell, D. A., & Schafer, W. D. (2001). The robustness of the likelihood ratio chi-square test for structural equation models: A meta-analysis. *Journal of Educational and Behavioral Statistics, 26,* 105–132.

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. B. (1978). Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonresponse. *Proceedings of the Research Methods Section of the American Statistical Association,* 20–34.

Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in *Personality and Social Psychology Bulletin*. *Personality and Social Psychology Bulletin, 28,* 1629–1646.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. *Proceedings of the Business and Economic Statistics Section of the American Statistical Association,* 308–313.

Schaefer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research, 8,* 23–74.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research, 99,* 323–337.

Steiger, J. H. (1989). *EzPath: A supplementary module for SYSTAT and SYGRAPH.* Evanston, IL: SYSTAT.

Steiger, J. H. (2001). Driving fast in reverse: The relationship between software development, theory, and education in structural equation modeling. *Journal of the American Statistical Association, 96,* 331–338.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences, 42,* 893–898.

Steiger, J. H., & Lind, J. M. (1980, June). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Thompson, B. (1997). The importance of structure coefficients in structural equation modeling confirmatory factor analysis. *Educational and Psychological Measurement, 57,* 5–19.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC: American Psychological Association.

Tremblay, P. F., & Gardner, R. C. (1996). On the growth of structural equation modeling in psychological journals. *Structural Equation Modeling, 3,* 93–104.

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

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34,* 806–838.

Yuan, K. H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40,* 115–148.

Appendix

Confirmatory Factor Analysis Reporting Guidelines Checklist

Theoretical formulation and data collection
- _ Theoretical/empirical justification of models tested
- _ Number and type of models tested (correlated, orthogonal, hierarchical)
- _ Specification of models tested (explicit relationships between observed and latent variables)
- _ Graphic representation of models tested
- _ Sample characteristics (justification, sampling method, sample size)
- _ Identification of equivalent and theoretically alternative models
- _ Specification of model identifiably (can models be tested)?

Data preparation
- _ Screening for univariate and multivariate normality and outliers
- _ Analysis of missing data and method for addressing
- _ Scale of observed variables (nominal, ordinal, interval, ratio; range of values)
- _ Description of data transformations (include parceling)

Analysis decisions
- _ Type of matrix analyzed (covariance, correlation)
- _ Matrix included or available upon request
- _ Estimation procedure and justification given normality assessment (ML, S-B ML, WLS)
- _ Scale of latent variables
- _ Software and version

Model evaluation
- _ Inclusion of multiple fit indices (e.g., chi-square, *df*, *p;* RMSEA, CFI, TLI)

Note. ML = maximum likelihood; S-B ML = Satorra–Bentler maximum likelihood; WLS = weighted least squares; *df* = degrees of freedom; RMSEA = root-mean-square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index.

Received July 19, 2007
Revision received October 1, 2008
Accepted November 7, 2008 ■