Psychological Methods
1998, Vol. 3, No. 4, 424-453

1082-989X/98/$3.00

Fit Indices in Covariance Structure Modeling: Sensitivity to Underparameterized Model Misspecification

Li-tze Hu
University of California, Santa Cruz

Peter M. Bentler
University of California, Los Angeles

This study evaluated the sensitivity of maximum likelihood (ML)-, generalized least squares (GLS)-, and asymptotic distribution-free (ADF)-based fit indices to model misspecification, under conditions that varied sample size and distribution. The effect of violating assumptions of asymptotic robustness theory also was examined. Standardized root-mean-square residual (SRMR) was the most sensitive index to models with misspecified factor covariance(s), and Tucker–Lewis Index (1973; TLI), Bollen's fit index (1989; BL89), relative noncentrality index (RNI), comparative fit index (CFI), and the ML- and GLS-based gamma hat, McDonald's centrality index (1989; Mc), and root-mean-square error of approximation (RMSEA) were the most sensitive indices to models with misspecified factor loadings. With ML and GLS methods, we recommend the use of SRMR, supplemented by TLI, BL89, RNI, CFI, gamma hat, Mc, or RMSEA (TLI, Mc, and RMSEA are less preferable at small sample sizes). With the ADF method, we recommend the use of SRMR, supplemented by TLI, BL89, RNI, or CFI. Finally, most of the ML-based fit indices outperformed those obtained from GLS and ADF and are preferable for evaluating model fit.

This study addresses the sensitivity of various fit indices to underparameterized model misspecification. The issue of model misspecification has been almost completely neglected in evaluating the adequacy of fit indices used to evaluate covariance structure models. Previous recommendations on the adequacy of fit indices have been primarily based on the evaluation of the effect of sample size, or the effect of estimation method, without taking into account the sensitivity of an index to model misspecification. In other words, virtually all studies of fit indices have concentrated their efforts on the adequacy of fit indices under the modeling null hypothesis, that is, when the model is correct. Although such an approach is useful, as noted by Maiti and Mukherjee (1991), it misses the main practical point for the use of fit indices, namely, the ability to discriminate well-fitting from badly fitting models. Of course, it is certainly legitimate to ask that fit indices reliably reach their maxima when the model is correct, for example, under variations of sample size, but it seems much more vital to assure that a fit index is sensitive to misspecification of the model, so that it can be used to determine whether a model is incorrect. Maiti and Mukherjee term this characteristic *sensitivity*. Thus, a good index should approach its maximum under correct specification but also degrade substantially under misspecification. As far as we can tell, essentially no studies have inquired to what extent this basic requirement is met by the many indices that have been proposed across the years. Maiti and Mukherjee have provided an analysis of only a few indices under very restricted modeling conditions.

In this study, the sensitivity of four types of fit indices, derived from maximum-likelihood (ML), generalized least squares (GLS), and asymptotic distribution-free (ADF) estimators, to various types of underparameterized model misspecification is examined. Note that in an *underparameterized* model, one

Li-tze Hu, Department of Psychology, University of California, Santa Cruz; Peter M. Bentler, Department of Psychology, University of California, Los Angeles.

This research was supported by a grant from the Division of Social Sciences, by a Faculty Research Grant from the University of California, Santa Cruz, and by U.S. Public Health Service Grants DA00017 and DA01070. The computer assistance of Shinn-Tzong Wu is gratefully acknowledged.

Correspondence concerning this article should be addressed to Li-tze Hu, Department of Psychology, University of California, Santa Cruz, California 95064. Electronic mail may be sent to lth@cats.ucsc.edu.

or more parameters whose population values are nonzero are fixed to zero. In addition, we evaluate the adequacy of these four types of fit indices under conditions such as violation of underlying assumptions of multivariate normality and asymptotic robustness theory, providing evidence regarding the efficacy of the often stated idea that a model with a fit index greater than (or, in some cases, less than) a conventional cutoff value should be acceptable (e.g., Bentler & Bonett, 1980). Also, for the first time, we evaluated several new and supposedly superior indices (i.e., gamma hat, McDonald's [1989] centrality index [Mc], and root-mean-square error of approximation [RMSEA]) that have been recommended with little or no empirical support. We present here a nontechnical summary of the methods and the results of our study. Readers wishing a more detailed report of this study should consult our complete technical report (Hu & Bentler, 1997).

Historical Background

Structural equation modeling has become a standard tool in psychology for investigating the plausibility of theoretical models that might explain the interrelationships among a set of variables. In these applications, the assessment of goodness-of-fit and the estimation of parameters of the hypothesized model(s) are the primary goals. Issues related to the estimation of parameters have been discussed elsewhere (e.g., Bollen, 1989; Browne & Arminger, 1995; Chou & Bentler, 1995); our discussion here focuses on those issues that are critical to the assessment of goodness-of-fit of the hypothesized model(s).

The most popular ways of evaluating model fit are those that involve the chi-square goodness-of-fit statistic and the so-called fit indices that have been offered to supplement the chi-square test. The asymptotic chi-square test statistic was originally developed to serve as a criterion for model evaluation or selection. In its basic form, a large value of the chi-square statistic, relative to its degrees of freedom, is evidence that the model is not a very good description of the data, whereas a small chi-square is evidence that the model is a good one for the data. Unfortunately, as noted by many researchers, this simple version of the chi-square test may not be a reliable guide to model adequacy. The actual size of a test statistic depends not only on model adequacy but also on which one among several chi-square tests actually is used, as well as other conceptually unrelated technical conditions, such as sample size being too small or violation of an assumption underlying the test, for example, multivariate normality of variables, in the case of the standard chi-square test (e.g., Bentler & Dudgeon, 1996; Chou, Bentler, & Satorra, 1991; Curran, West, & Finch, 1996; Hu, Bentler, & Kano, 1992; Muthen & Kaplan, 1992; West, Finch, & Curran, 1995; Yuan & Bentler, 1997). Thus, a significant goodness-of-fit chi-square value may be a reflection of model misspecification, power of the test, or violation of some technical assumptions underlying the estimation method. More important, it has been commonly recognized that models are best regarded as approximations of reality, and hence, using chi-square to test the hypothesis that the population covariance matrix matches the model-implied covariance matrix, $\Sigma = \Sigma(\theta)$, is too strong to be realistic (e.g., de Leeuw, 1983; Jöreskog, 1978). Thus the standard chi-square test may not be a good enough guide to model adequacy.

As a consequence, alternative measures of fit, namely, so-called fit indices, were developed and recommended as plausible additional measures of model fit (e.g., Akaike, 1987; Bentler, 1990; Bentler & Bonett, 1980; Bollen, 1986, 1989; James, Mulaik, & Brett, 1982; Jöreskog & Sörbom, 1981; Marsh, Balla, & McDonald, 1988; McDonald, 1989; McDonald & Marsh, 1990; Steiger & Lind, 1980; Tanaka, 1987; Tanaka & Huba, 1985; Tucker & Lewis, 1973). However, despite the increasing popularity of using fit indices as alternative measures of model fit, applied researchers inevitably face a constant challenge in selecting appropriate fit indices among a large number of fit indices that have recently become available in many popular structural equation modeling programs. For instance, both LISREL 8 (Jöreskog & Sörbom, 1993) and the PROC CALIS procedure for structural equation modeling (SAS Institute, 1993) report the values of about 20 fit indices, and EQS (Bentler & Wu, 1995a, 1995b) prints the values of almost 10 fit indices. Frequently, the values of various fit indices reported in a given program yield conflicting conclusions about the extent to which the model matches the observed data. Applied researchers thus often have difficulties in determining the adequacy of their covariance structure models. Furthermore, as noted by Bentler and Bonett (1980), who introduced several of these indices and popularized the ideas, fit indices were designed to avoid some of the problems of sample size and distributional misspecification on evaluation of a model. Initially, it was hoped that

these fit indices would more unambiguously point to model adequacy as compared with the chi-square test. This optimistic state of affairs is unfortunately also not true.

The Chi-Square Test

The conventional overall test of fit in covariance structure analysis assesses the magnitude of discrepancy between the sample and fitted covariance matrices. Let S represent the unbiased estimator of a population covariance matrix, Σ, of the observed variables. The population covariance matrix can be expressed as a function of a vector containing the fixed and free model parameters, that is, θ: $\Sigma = \Sigma(\theta)$. The parameters are estimated so that the discrepancy between the sample covariance matrix S and the implied covariance matrix $\Sigma(\hat{\theta})$ is minimal. A discrepancy function $F = F[S, \Sigma(\theta)]$ can be considered to be a measure of the discrepancy between S and $\Sigma(\theta)$ evaluated at an estimator $\hat{\theta}$ and is minimized to yield $F_{\min}$. Under an assumed distribution and the hypothesized model $\Sigma(\theta)$ for the population covariance matrix Σ, the test statistic $T = (N - 1)F_{\min}$ has an asymptotic (large sample) chi-square distribution. The test statistic T is usually called the chi-square statistic by other researchers. In general, the null hypothesis $\Sigma = \Sigma(\theta)$ is rejected if T exceeds a value in the chi-square distribution associated with an α level of significance. The T statistics can be derived from various estimation methods that vary in the degrees of sensitivity to the distributional assumptions. The T statistic derived from ML under the assumption of multivariate normality of variables is the most widely used summary statistic for assessing the adequacy of a structural equation model (Gierl & Mulvenon, 1995).

Types of Fit Indices

Unlike a chi-square test that offers a dichotomous decision strategy implied by a statistical decision rule, a fit index can be used to quantify the degree of fit along a continuum. It is an overall summary statistic that evaluates how well a particular covariance structure model explains sample data. Like R^2 in multiple regression, fit indices are meant to quantify something akin to variance accounted for, rather than to test a null hypothesis $\Sigma = \Sigma(\theta)$. In particular, these indices generally quantify the extent to which the variation and covariation in the data are accounted for by a model. One of the most widely adopted dimensions for classifying fit indices is the *absolute* versus *incremental* distinction (Bollen, 1989; Gerbing & Anderson, 1993; Marsh et al., 1988; Tanaka, 1993). An absolute-fit index directly assesses how well an a priori model reproduces the sample data. Although no reference model is used to assess the amount of increment in model fit, an implicit or explicit comparison may be made to a saturated model that exactly reproduces the observed covariance matrix. As a result, this type of fit index is analogous to R^2 by comparing the goodness of fit with a component that is similar to a total sum of squares. In contrast, an incremental fit index measures the proportionate improvement in fit by comparing a target model with a more restricted, nested baseline model. Incremental fit indices are also called *comparative* fit indices. A null model in which all the observed variables are allowed to have variances but are uncorrelated with each other is the most typically used baseline model (Bentler & Bonett, 1980), although other baseline models have been suggested (e.g., Sobel & Bohrnstedt, 1985).

Incremental fit indices can be further distinguished among themselves. We define three groups of indices, Types 1–3 (Hu & Bentler, 1995).[1] A Type 1 index uses information only from the optimized statistic T, used in fitting baseline (T_B) and target (T_T) models. T is not necessarily assumed to follow any particular distributional form, though it is assumed that the fit function F is the same for both models. A general form of such indices can be written as Type 1 incremental indices $= |T_B - T_T|/T_B$. The ones we study in this article are the normed fit index (NFI; Bentler & Bonett, 1980) and a fit index by Bollen (1986; BL86).

[1] The terminology of Type 1 and Type 2 indices follows Marsh et al. (1988), although our specific definitions of these terms are not identical to theirs. Their Type 2 index has some definitional problems, and its proclaimed major example is not consistent with their own definition. They define Type 2 indices as $|T_T - T_B|/|E - T_B|$, where T_T is the value of the statistic for the target model, T_B is the value for a baseline model, and E is the expected value of T_T if the target model is true. Note first that E may not be a single quantity: Different values may be obtained depending on additional assumptions, such as on the distribution of the variables. As a result, the formula can give more than one Type 2 index for any given absolute index. In addition, the absolute values in the formula have the effect that their Type 2 indices must be nonnegative; however, they state that an index called the Tucker–Lewis Index (TLI; discussed later in text) is a Type 2 index. This is obviously not true because TLI can be negative.

Table 1 contains algebraic definitions, properties, and citations for all fit indices considered in this article.

Type 2 and Type 3 indices are based on an assumed distribution of variables and other standard regularity conditions. A Type 2 index additionally uses information from the expected values of T_T under the central chi-square distribution. It assumes that the chi-square estimator of a valid target model follows an asymptotic chi-square distribution with a mean of df_T, where df_T is the degrees of freedom for a target model. Hence, the baseline fit T_B is compared with df_T, and the denominator in the Type 1 index is replaced by $(T_B - df_T)$. Thus, a general form of such indices can be written as Type 2 incremental fit index $= |T_B - T_T|/(T_B - df_T)$. On the basis of the work of Tucker and Lewis (1973), Bentler and Bonett (1980) called such indices *nonnormed fit* indices, because they need not have a 0–1 range even if $T_B \geq T_T$. We study their index (NNFI or TLI) and a related index developed by Bollen (1989; BL89).

A Type 3 index uses Type 1 information but additionally uses information from the expected values of T_T or T_B, or both, under the relevant noncentral chi-square distribution. A *noncentrality* fit index usually involves first defining a population-fit-index parameter and then using estimators of this parameter to define the sample-fit index (Bentler, 1990; McDonald, 1989; McDonald & Marsh, 1990; Steiger, 1989). When the assumed distributions are correct, Type 2 and Type 3 indices should perform better than Type 1 indices because more information is being used. We study Bentler's (1989, 1990) and McDonald and Marsh's (1990) relative noncentrality index (RNI) and Bentler's comparative fit index (CFI). Note also that Type 2 and Type 3 indices may use inappropriate information, because any particular T may not have the distributional form assumed. For example, Type 3 indices make use of the noncentral chi-square distribution for T_B, but one could seriously question whether this is generally its appropriate reference distribution. We also study several absolute-fit indices. These include the goodness-of-fit (GFI) and adjusted-GFI (AGFI) indices (Bentler, 1983; Jöreskog & Sörbom, 1984; Tanaka & Huba, 1985); Steiger's (1989) gamma hat; a rescaled version of Akaike's information criterion (CAK; Cudeck & Browne, 1983); a cross-validation index (CK; Browne & Cudeck, 1989); McDonald's (1989) centrality index (Mc); Hoelter's (1983) critical N (CN); a standardized version of Jöreskog and Sörbom's (1981) root-mean-square residual (SRMR; Bentler, 1995); and the RMSEA (Steiger & Lind, 1980).

Issues in Assessing Fit by Fit Indices

There are four major problems involved in using fit indices for evaluating goodness of fit: sensitivity of a fit index to model misspecification, small-sample bias, estimation-method effect, and effects of violation of normality and independence. The issue on sensitivity of fit index to model misspecification has long been overlooked and thus deserves careful examination. The other three issues are a natural consequence of the fact that these indices typically are based on chi-square tests: A fit index will perform better when its corresponding chi-square test performs well. Because, as noted above, these chi-square tests may not perform adequately at all sample sizes and also because the adequacy of a chi-square statistic may depend on the particular assumptions it requires about the distributions of variables, these same factors can be expected to influence evaluation of model fit.

Sensitivity of Fit Index to Model Misspecification

Among various sources of effects on fit indices, the sensitivity of fit indices to model misspecification (Gerbing & Anderson, 1993; i.e., the effect of model misspecification) has not been adequately studied because of the intensive computational requirements. A correct specification implies that a population exactly matches the hypothesized model and also that the parameters estimated in a sample reflect this structure. On the other hand, a model is said to be misspecified when (a) one or more parameters are estimated whose population values are zeros (i.e., an overparameterized misspecified model), (b) one or more parameters are fixed to zeros whose population values are nonzeros (i.e., an underparameterized misspecified model), or both. In the very few studies that have touched on such an issue, the results are often inconclusive due either to the use of an extremely small number of data sets (e.g., Marsh et al., 1988; Mulaik et al., 1989) or to the study of a very small number of fit indices under certain limited conditions (e.g., Bentler, 1990; La Du & Tanaka, 1989; Maiti & Mukherjee, 1991). For example, using a small number of simulated data sets. Marsh et al. (1988) reported that sample size was substantially associated with several fit indices under both true and false models. They showed also that the values of most of the absolute-

Table 1
Algebraic Definitions, Properties, and Citations for Incremental and Absolute-Fit Indices

Algebraic definition	Property	Citation
Incremental fit indices		
Type 1		
NFI = $(T_B - T_T)/T_B$	Normed (has a 0–1 range)	Bentler & Bonett (1980)
BL86 = $[(T_B/df_B) - (T_T/df_T)]/(T_B/df_B)$	Normed (has a 0–1 range)	Bollen (1986)
Type 2		
TLI (or NNFI) = $[(T_B/df_B) - (T_T/df_T)]/[(T_B/df_B) - 1]$	Nonnormed (can fall outside the 0–1 range) Compensates for the effect of model complexity	Tucker & Lewis (1973) Bentler & Bonett (1980)
BL89 = $(T_B - T_T)/(T_B - df_T)$	Nonnormed Compensates for the effect of model complexity	Bollen (1989)
Type 3		
RNI = $[(T_B - df_B) - (T_T - df_T)]/(T_B - df_B)$	Nonnormed Noncentrality based	McDonald & Marsh (1990) Bentler (1989, 1990)
CFI = $1 - \max[(T_T - df_T), 0]/\max[(T_T - df_T), (T_B - df_B), 0]$	Normed (has a 0–1 range) Noncentrality based	Bentler (1989, 1990)
Absolute fit indices		
$GFI_{ML} = 1 - [\mathrm{tr}(\Sigma^{-1}S - I)^2/\mathrm{tr}(\Sigma^{-1}S)^2]$	Has a maximum value of 1.0 Can be less than 0	Jöreskog & Sörbom (1984)
$AGFI_{ML} = 1 - [p(p+1)/2df_T](1 - GFI_{ML})$	Has a maximum value of 1.0 Can be less than 0	Jöreskog & Sörbom (1984)
Gamma hat = $p/\{p + 2[(T_T - df_T)/(N-1)]\}$	Has a known distribution Noncentrality based	Steiger (1989)
CAK = $[T_T/(N-1)] + [2q/(N-1)]$	Compensates for the effect of model complexity	Cudeck & Browne (1983)
CK = $[T_T/(N-1)] + [2q/(N-p-2)]$	Compensates for the effect of model complexity	Browne & Cudeck (1989)
Mc = $\exp\{-1/2[(T_T - df_T)/(N-1)]\}$	Noncentrality based Typically has the 0–1 range (but it may exceed 1)	McDonald (1989)
CN = $\{(z_{crit} + \sqrt{2df-1})^2/[2T_T/(N-1)]\} + 1$	A CN value exceeding 200 indicates a good fit of a given model	Hoelter (1983)
$SRMR = \sqrt{\{2\sum_{i=1}^{p}\sum_{j=1}^{i}[(s_{ij} - \hat{\sigma}_{ij})/(s_{ii}s_{jj})]^2\}/p(p+1)}$	Standardized root-mean-square residual	Jöreskog & Sörbom (1981) Bentler (1995)
RMSEA = $\sqrt{\hat{F}_0/df_T}$, where $\hat{F}_0 = \max[(T_T - df_T)/(N-1), 0]$	Has a known distribution Compensates for the effect of model complexity Noncentrality based	Steiger & Lind (1980) Steiger (1989)

Note. NFI = normed fit index; T_B = T statistic for the baseline model; T_T = T statistic for the target model; BL86 = fit index by Bollen (1986); df_B = degrees of freedom for the baseline model; df_T = degrees of freedom for the target model; TLI = Tucker–Lewis index (1973); NNFI = nonnormed fit index; BL89 = fit index by Bollen (1989); RNI = relative noncentrality index; CFI = comparative fit index; GFI = goodness-of-fit index; ML = maximum likelihood; tr = trace of a matrix; AGFI = adjusted-goodness-of-fit index; CAK = a rescaled version of Akaike's information criterion; q = no. parameters estimated; CK = cross-validation index; Mc = McDonald's centrality index; CN = critical N; z_{crit} = critical z value at a selected probability level; SRMR = standardized root-mean-square residual; s_{ij} = observed covariances; $\hat{\sigma}_{ij}$ = reproduced covariances; s_{ii} and s_{jj} = observed standard deviations; RMSEA = root-mean-square error of approximation. The formulas for generalized least squares and asymptotic distribution-free versions of GFI and AGFI are shown in Hu and Bentler (1997).

and Type 2 fit indices derived from true models were significantly greater than those derived from false models. La Du and Tanaka (1989, Study 2) studied the effects of both overparameterized and underparameterized model misspecification (both with misspecified path[s] between observed variables) on the ML- and GLS-based GFI and NFI. No significant effect of overparameterized model misspecification on these fit indices was found. A very small but significant effect of underparameterized model misspecification was observed for some of these fit indices (i.e., the ML-based NFI and ML-/GLS-based GFI). The ML-based NFI also was found to be more sensitive to this type of model misspecification than was the ML- and GLS-based GFI. Marsh, Balla, and Hau (1996) found that degrees of model misspecification accounted for a large proportion of variance in NFI, BL86, TLI, BL89, RNI, and CFI. Although their study included several substantially misspecified models, their analyses failed to reveal the degree of sensitivity of these fit indices for a less misspecified model. In our study, the sensitivity of various fit indices to model misspecification, after controlling for other sources of effects, are examined.

Small-Sample Bias

Estimation methods in structural equation modeling are developed under various assumptions. One is that the model $\Sigma = \Sigma(\theta)$ is true. Another is the assumption that estimates and tests are based on large samples, which will not actually obtain in practice. The adequacy of the test statistics is thus likely to be influenced by sample size, perhaps performing more poorly in smaller samples that cannot be considered asymptotic enough. In fact, the relation between sample size and the adequacy of a fit index when the model is true has long been recognized; for example, Bearden, Sharma, and Teel (1982) found that the mean of NFI is positively related to sample size and that NFI values tend to be less than 1.0 when sample size is small. Their early results pointed out the main problem: possible systematic fit-index bias.

If the mean of a fit index, computed across various samples under the same condition when the model is true, varies systematically with sample size, such a statistic will be a biased estimator of the corresponding population parameter. Thus, the decision for accepting or rejecting a particular model may vary as a function of sample size, which is certainly not desirable. The general finding seems to be a positive association between sample size and the goodness-of-fit fit index size for Type 1 incremental fit indices. Obviously, Type 1 incremental indices will be influenced by the badness of fit of the null model as well as the goodness of fit of the target model, and Marsh et al. (1988) have reported this type of effect. On the other hand, the Type 2 and Type 3 indices seem to be substantially less biased. The results on absolute indices are mixed.

A few key studies can be mentioned. Bollen (1986, 1989, 1990) found that the means of the sampling distributions of NFI, BL86, GFI, and AGFI tended to increase with sample size. Anderson and Gerbing (1984) and Marsh et al. (1988) showed that the means of the sampling distributions of GFI and AGFI were positively associated with sample size whereas the association between TLI and sample size was not substantial. Bentler (1990) also reported that TLI (and NNFI) outperformed NFI on average; however, the variability of TLI (and NNFI) at a small sample size (e.g., $N = 50$) was so large that in many samples, one would suspect model incorrectness and, in many other samples, overfitting. Cudeck and Browne (1983) and Browne and Cudeck (1989) found that CAK and CK improved as sample size increased. Bollen and Liang (1988) showed that Hoelter's (1983) CN increased as sample size increased. McDonald (1989) reported that the value of Mc was consistent across different sample sizes. Anderson and Gerbing (1984) found that the mean values of RMR (the unstandardized root-mean-square residual; Jöreskog & Sörbom, 1981) was related to the sample size. J. Anderson, Gerbing, and Narayanan (1985) further reported that the mean values of RMR were related to the sample size and model characteristics, such as the number of indicators per factor, the number of factors, and indicator loadings. In one of the major studies that investigated the effect of sample size on the older fit indices, Marsh et al. (1988) found that many indices were biased estimates of their corresponding population parameters when sample size was finite. GFI appeared to perform better than any other stand-alone index (e.g., AGFI, CAK, CN, or RMR) studied by them. GFI also underestimated its asymptotic value to a lesser extent than did NFI.

The Type 2 and Type 3 incremental fit indices, in general, perform better than either the absolute or Type 1 incremental indices. This is true for the older indices such as TLI, as noted above, but appears to be especially true for the newer indices based on noncentrality. For example, Bentler (1990) reported that FI (called RNI in this article), CFI, and IFI (called

BL89 in this article) performed essentially with no bias, though by definition CFI must be somewhat downward biased to avoid out-of-range values greater than 1, which can occur with FI. The bias, however, is trivial, and it gains lower sampling variability in the index. The relation of RNI to CFI has been spelled out in more detail by Goffin (1993), who prefers RNI to CFI for model-comparison purposes.

Estimation-Method Effects

As noted above, the three major problems involved in using fit indices are a natural consequence of the fact that these indices typically are based on chi-square tests. This rationale is elaborated through a brief review of the ML, GLS, and ADF estimation methods, as well as their relationships to the chi-square statistics. For a more technical review of each method, readers are encouraged to consult Hu et al. (1992), Bentler and Dudgeon (1996), or, especially, the original sources.

Estimation methods such as ML and GLS in covariance structure analysis are traditionally developed under multivariate normality assumptions (e.g., Bollen, 1989; Browne, 1974; Jöreskog, 1969). A violation of multivariate normality can seriously invalidate normal-theory test statistics. ADF methods therefore have been developed (e.g., Bentler, 1983; Browne, 1982, 1984) with the promising claim that the test statistics for model fit are insensitive to the distribution of the observations when the sample size is large. However, empirical studies using Monte Carlo procedures have shown that when sample size is relatively small or model degrees of freedom are large, the chi-square goodness-of-fit test statistic based on the ADF method may be inadequate (Chou et al., 1991; Curran et al., 1996; Hu et al., 1992; Muthen & Kaplan, 1992; Yuan & Bentler, 1997).

The recent development of a theory for the asymptotic robustness of normal-theory methods offers hope for the appropriate use of normal-theory methods even under violation of the normality assumption (e.g., Amemiya & Anderson, 1990; T. W. Anderson & Amemiya, 1988; Browne, 1987; Browne & Shapiro, 1988; Mooijaart & Bentler, 1991; Satorra & Bentler, 1990, 1991). The purpose of this line of research is to determine under what conditions normal-theory-based methods such as ML or GLS can still correctly describe and evaluate a model with nonnormally distributed variables. The conditions are technical but require the very strong condition that the latent variables (common factors or errors) that are typically considered as simply uncorrelated must actually be mutually independent, and common factors, when correlated, must have freely estimated variance–covariance parameters. Independence exists when normally distributed variables are uncorrelated. However, when nonnormal variables are uncorrelated, they are not necessarily independent. If the robustness conditions are met in large samples, normal-theory ML and GLS test statistics still hold, even when the data are not normal. Unfortunately, because the data-generating process is unknown for real data, one cannot generally know whether the independence of factors and errors, or of the errors themselves, holds, and thus, the practical application of asymptotic robustness theory is unclear.

Although Hu et al. (1992) have examined the adequacy of six chi-square goodness-of-fit tests under various conditions, not much is known about estimation effects on fit indices. Even if the distributional assumptions are met, different estimators yield chi-square statistics that perform better or worse at various sample sizes. This may translate into differential performance of fit indices based on different estimators. However, the overall effect of mapping from chi-square to fit index, while varying estimation method, is unclear. In pioneering work, Tanaka (1987) and La Du and Tanaka (1989) have found that given the same model and data, NFI behaved erratically across ML and GLS estimation methods. On the other hand, they reported that GFI behaved consistently across the two estimation methods. Their results must be due to the differential quality of the null model chi-square used in the NFI but not the GFI computations.[2] On the basis of these results, Tanaka and Huba (1989) have suggested that GFI is more appropriate than NFI in finite samples and across different estimation methods. Using a large empirical data set, Sugawara and MacCallum (1993) have found that absolute-fit indices (i.e., GFI and RMSEA) tend to behave more consistently across estimation methods than do incremental fit indices (i.e., NFI, TLI, BL86, and BL89). This phenomenon is especially evident when there is a good fit between the hypothesized model and the observed data. As the degree of fit between hypothesized models and observed data decreases, GFI and RMSEA behave less consistently

[2] Earlier versions of EQS also incorrectly computed the null model chi-square under GLS, thus affecting all incremental indices.

across estimation methods. Sugawara and MacCallum have stated that the effect of estimation methods on fit is tied closely to the nature of the weight matrices used by the estimation methods. Ding, Velicer, and Harlow (1995) found that all fit indices they studied, except the TLI, were affected by estimation method.

Effects of Violation of Normality and Independence

An issue related to the adequacy of fit indices that has not been studied is the potential effect of violation of assumptions underlying estimation methods, specifically, violation of distributional assumptions and the effect of dependence of latent variates. The dependence condition is one in which two or more variables are functionally related, even though their linear correlations may be exactly zero. Of course, with normal data, a linear correlation of zero implies independence. Nothing is known about the adequacy of fit indices under conditions such as dependency among common and unique latent variates, along with violations of multivariate normality, at various sample sizes.

Study Questions and Performance Criteria

This study investigates several critical issues related to fit indices. First, the sensitivity of various incremental and absolute-fit indices derived from ML, GLS, and ADF estimation methods to underparameterized model misspecification is investigated. Two types of underparameterized model misspecification are studied: *simple* misspecified models (i.e., models with misspecified factor covariance[s]) and *complex* misspecified models (i.e., models with misspecified factor loading[s]). Second, the stability of various fit indices across ML, GLS, and ADF methods (i.e., the effect of estimation method on fit indices) is studied. Third, the performance of these fit indices, derived from the ML, GLS, and ADF estimators under the following three ways of violating theoretical conditions, is examined: (a) Distributional assumptions are violated, (b) assumed independence conditions are violated, and (c) asymptotic sample-size requirements are violated. Our primary goals are to recommend fit indices that perform the best overall and to identify those that perform poorly. Good fit indices should be (a) sensitive to model misspecification and (b) stable across different estimation methods, sample sizes, and distributions. Finally, attempts are also made to evaluate the "rule of thumb" conventional cutoff criterion for a given fit index (Bentler & Bonett, 1980), which has been used in practice to evaluate the adequacy of models.

Method

Two types of confirmatory factor models (called *simple* model and *complex* model), each of which can be expressed as $x = \Lambda\xi + \varepsilon$, were used to generate measured variables x under various conditions on the common factors ξ and unique variates (errors) ε. That is, the vector of observed variables (xs) was a weighted function of a common-factor vector (ξ) with weights given by the factor-loading matrix, Λ, plus a vector of error variates (ε). The measured variables for each model were generated by setting certain restrictions on the common factors and unique variates. Several properties are noted in the usual application of these types of factor analytic approaches. First, factors are allowed to be correlated and have a covariance matrix, Φ. Second, errors are uncorrelated with factors. Third, various error variates are uncorrelated and have a diagonal covariance matrix, Ψ. Consequently, the hypothesized model can be expressed as $\Sigma = \Sigma(\theta) = \Lambda\Phi\Lambda' + \Psi$, and the elements of θ are the unknown parameters in Λ, Φ, and Ψ.

Study Design

Simple and complex models are both confirmatory factor analytic models based on 15 observed variables with three common factors. Although many other model types are possible, most models used in practice involve latent variables, and the confirmatory factor model is most representative of such models. For example, variants of confirmatory factor models have been the typically studied models in the new journal *Structural Equation Modeling,* in the special section on "Structural Equation Modeling in Clinical Research" (Hoyle, 1994) published in the *Journal of Consulting and Clinical Psychology,* and in the larger models among the approximately two dozen modeling articles published in the *Journal of Personality and Social Psychology (JPSP)* during 1995. In practice, correlations among factors may be replaced by hypothesized paths, and correlated residuals may be added. Such models also form the basis of many recent simulation studies (e.g., Curran et al., 1996; Ding et al., 1995; Marsh et al., 1996). It is important to choose a number of variables that is not too small (e.g., Hu et al., 1992) yet remains practical in the context of a large simulation. We chose a number larger than the median number of variables (9–10)

$$\begin{bmatrix} .70 & .70 & .75 & .80 & .80 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 \\ .00 & .00 & .00 & .00 & .00 & .70 & .70 & .75 & .80 & .80 & .00 & .00 & .00 & .00 & .00 \\ .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .70 & .70 & .75 & .80 & .80 \end{bmatrix}$$

used in *JPSP*'s 1995 modeling studies but smaller than many ambitious studies (e.g., Hoyle's special section contains five studies with 24 or more variables). We also chose distributional conditions and samples sizes to cover a wider range of practical relevance. Figure 1 displays the structures of true-population and misspecified models used in this study.

The factor-loading matrix (transposed) Λ' for the simple model had the structure shown at the top of the page.

The structure of the factor-loading matrix (transposed) Λ' for the complex model was as shown at the bottom of the page.

For both the simple model and complex model, variances of the factors were 1.0, and the covariances among the three factors were 0.30, 0.40, and 0.50. The unique variances were taken as values that would yield unit-variance measured variables under normality for the simple model. For the complex model, the unique variances were taken as values that would yield unit variance for most measured variables (except for the 1st, 4th, and 9th observed variables in the model) under normality. The unique variances for the 1st, 4th, and 9th observed variables were 0.51, 0.36, and 0.36, respectively. In estimation, the factor loading of the last indicator of each factor was fixed for identification at 0.80, and the remaining nonzero parameters were free to be estimated.

Two hundred replications (samples) of a given sample size were drawn from a known population model in each of the seven distributional conditions as defined by Hu et al. (1992). The first was a baseline distributional condition involving normality, the next three involved nonnormal variables that were independently distributed when uncorrelated, and the final three distributional conditions involved nonnormal variables that, although uncorrelated, remained dependent.

Distributional Condition 1. The factors and errors and hence measured variables are multivariate normally distributed.

Distributional Conditions 2. Nonnormal factors and errors, when uncorrelated, are independent, but asymptotic robustness theory does not hold because the covariances of common factors are not free parameters. The true excess kurtoses for the nonnormal factor in the population are −1.0, 2.0, and 5.0. The true excess kurtoses for the unique variates are −1.0, 0.5, 2.5, 4.5, 6.5, −1.0, 1.0, 3.0, 5.0, 7.0, −0.5, 1.5, 3.5, 5.5, and 7.5.

Distributional Condition 3. Nonnormal factors and errors are independent but not multivariate normally distributed. The true kurtoses for the factors and unique variates are identical to those in Distributional Condition 2.

Distributional Condition 4. The errors and hence the measured variables are not multivariate normally distributed. The true kurtoses for the unique variates are identical to those in Distributional Conditions 2 and 3, but the true kurtoses for the factors are set to zero.

Distributional Condition 5. An elliptical distribution: Factors and errors are uncorrelated but dependent on each other.

Distributional Condition 6. The errors and hence the measured variables are not multivariate normally distributed, and both factors and errors are uncorrelated but dependent on each other.

Distributional Condition 7. Nonnormal factors and errors are uncorrelated but dependent on each other.

In Distributional Conditions 5–7, the factors and error variates were divided by a random variable, $z = [\chi^2(5)]^{1/2}/\sqrt{3}$, that was distributed independently of the original common and unique factors. The division was made so that the variances and covariances of the factors remained unchanged but the kurtoses of the factors and errors became modified. As a consequence of this division, the factors and errors were

$$\begin{bmatrix} .70 & .70 & .75 & .80 & .80 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 \\ .00 & .00 & .00 & .70 & .00 & .70 & .70 & .75 & .80 & .80 & .00 & .00 & .00 & .00 & .00 \\ .70 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .70 & .00 & .70 & .70 & .75 & .80 & .80 \end{bmatrix}$$

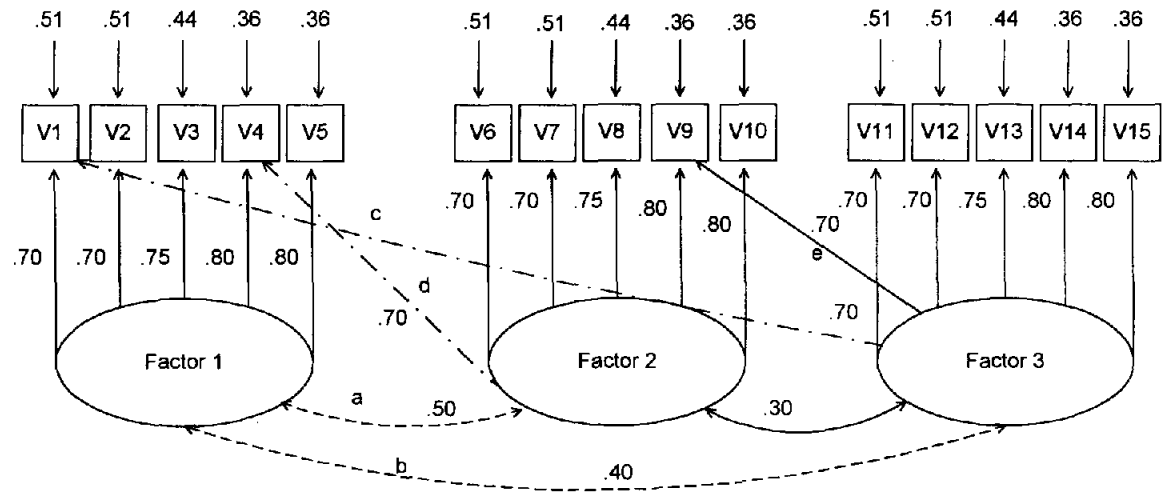

Figure 1. Structures of true-population and misspecified models used in this study. Solid lines (except solid line e) represent parameters that exist in the simple true-population model and both simple misspecified models, 1 and 2; dashed line a represents the parameter that exists in the simple true-population model but was omitted from both simple misspecified models, 1 and 2; dashed line b represents the parameter that exists in the simple true-population model but was omitted from simple Misspecified Model 2 only. Solid lines (including solid line e) and dashed lines (a and b) represent parameters that exist in the complex true-population model and both complex misspecified models, 1 and 2; dashed and dotted line c represents the parameter that exists in the complex true-population model but was omitted from both complex misspecified models, 1 and 2; dashed and dotted line d represents the parameter that exists in the complex true-population model but was omitted from Complex Misspecified Model 2 only. V = observed variable.

uncorrelated but dependent on each other. Because of the dependence, asymptotic robustness of normal-theory statistics was not to be expected under Distributional Conditions 5–7. To provide some idea about the degree of nonnormality of the factors and unique variates in Distributional Conditions 5–7 after the division, the empirical univariate kurtoses of the latent variables were computed across 5,000 × 200 = 1,000,000 observations. In Distributional Condition 5, the empirical kurtoses for the factors were 5.1, 6.0, and 5.5. The empirical kurtoses for the unique variates were 4.9, 6.0, 4.7, 4.5, 4.9, 6.1, 5.7, 5.2, 4.3, 4.8, 5.9, 4.8, 5.1, 4.8, and 5.1. In Distributional Condition 6, the empirical kurtoses for the factors were 5.1, 6.0, and 5.5. The empirical kurtoses for the unique variates were 2.6, 7.5, 10.4, 14.0, 19.3, 3.2, 9.5, 11.6, 15.1, 19.9, 4.4, 8.2, 14.2, 19.2, and 28.3. In Distributional Condition 7, the empirical kurtoses for the factors were 2.5, 18.0, and 2.14. The empirical kurtoses for the unique variates were 2.6, 7.5, 10.4, 14.0, 19.3, 3.2, 9.5, 11.6, 15.1, 19.9, 4.4, 8.2, 14.2, 19.2, and 28.3. Note that the empirical kurtoses for factors and unique variates in Distributional Conditions 1–4 were very close to the true kurtoses specified in these distributional conditions. By means of modified simulation procedures in EQS (Bentler & Wu, 1995b) and SAS program (SAS Institute, 1993), the various fit indices based on ML, GLS, and ADF estimation methods were computed in each sample.[3]

Specification of Models and Procedure

For each type of model (i.e., simple or complex), one true-population model and two misspecified models were used to examine the degree of sensitivity to model misspecification of various fit indices.

True-population model. The performance of four types of fit indices, derived from ML, GLS, and ADF estimation methods, were examined under the above-mentioned seven distributional conditions. A sample size was drawn from the population, and the model was estimated in that sample. The results were saved, and the process was repeated for 200 replications. This process was repeated for sample sizes 150, 250, 500, 1,000, 2,500, and 5,000. In all, there were 7 (distributions) × 6 (sample sizes) × 200 (replications) = 8,400 samples. The fit indices based on ML, GLS, and ADF methods were calculated for each of these

[3] BL86, BL89, RNI, gamma hat, CAK, CK, Mc, CN, and RMSEA were computed by SAS programs.

samples. This procedure was conducted for simple and complex models separately.

Misspecified models. Although both underparameterized and overparameterized models were considered as incorrectly specified models, our study only examined the sensitivity of fit indices to underparameterization. For a simple model, the covariances among the three factors in the correctly specified population model (true-population model) were nonzero (see Figure 1). The covariance between Factors 1 and 2 (Covariance a in Figure 1) was fixed to zero for Simple Misspecified Model 1. The covariances between Factors 1 and 2, as well as between Factors 1 and 3 (Covariances a and b) were fixed to zero for Simple Misspecified Model 2. For a complex model, three observed variables loaded on two factors in the true-population model: (a) The first observed variable loaded on Factors 1 and 3, (b) the fourth observed variable loaded on Factors 1 and 2, and (c) the ninth observed variables loaded on Factors 2 and 3 (see Figure 1). Complex Misspecified Model 1, the first observed variable loaded only on Factor 1 (Omitted Path c), whereas the rest of the model specification remained the same as the complex true-population model. In Complex Misspecified Model 2, the first and fourth observed variables loaded only on Factor 1: Omitted Paths c and d.

Using the design parameters specified in either the simple or complex true-population model, a sample size was drawn from the population, and each of the misspecified models was estimated in that sample. That is, the data for a given sample size were generated based on the structure specified by a true-population (correct) model, and then the goodness-of-fit between a misspecified model and the generated data was tested. For each misspecified model, there were 7 (distributions) × 6 (sample sizes) × 200 (replications) = 8,400 samples. The fit indices based on ML, GLS, and ADF methods were calculated for each of these samples.

Results

The adequacy of the simulation procedure and the characteristics specified in each distributional condition were verified by Hu et al. (1992), and thus are not discussed here. The overall mean distances (OMDs) between observed fit index values and the corresponding expected fit index values for the true-population models were calculated for each fit index and are tabulated in Table 2.[4] Separate correlation matrices among fit indices derived from ML, GLS, and ADF methods also were obtained, to determine empirically which subset of fit indices might have similar characteristics. Results are shown in Table 3. A series of analyses of variance (ANOVAs) were conducted for each fit index obtained for the simple and complex models. The η^2s, indicating the proportion of variance in each fit index accounted for by each predictor variable or interaction term, are presented in Tables 4 through 9. Note that the η^2 reported in this article is equivalent to R^2 (Hays, 1988, p. 369) and was calculated by dividing the Type 3 sum of squares for a given predictor or interaction term by the corrected total sum of squares (i.e., corrected total variance).[5] In addition, a statistical summary of the mean value and standard deviation of each fit index across the 200 replications and the empirical rejection frequency (for all but CAK and CK) based on rules of thumb were tabulated by distribution, sample size, and estimation method. Tables for the statistical summary for all fit indices are included in our technical report (Hu & Bentler, 1997).

[4] The overall coefficient of variation, which is defined as the mean of a distribution divided by its standard deviation, also was calculated for each fit index derived from ML, GLS, and ADF estimation methods. The conclusions regarding the performance of fit indices based on the mean distance and coefficients of variation were similar. However, the overall mean distance provided a much better index when compared across fit indices with different expected values (i.e., 0 and 1) for a true-population model and thus is reported in this article.

[5] We calculated η^2 values to determine the relative contribution of each main effect and interaction term. Given the very large sample size, significance tests would not be informative. Although our mixed-model ANOVA designs included a repeated measure (i.e., model misspecification or estimation method), we always used the total variance as the denominator in our calculations, so that all effects were in a common metric and are therefore directly comparable. This approach can underestimate the effect sizes for the repeated measures effects in mixed-model designs (Dodd & Schultz, 1973), and alternative approaches have been suggested (e.g., Dodd & Schultz, 1973; Dwyer, 1974; Kirk, 1995; Vaughan & Corballis, 1969); however, these approaches make comparison of between- and within-subjects estimates difficult because they are in different metrics. In our study, the error components were extremely small, and the sample size was very large, so that any advantage of using one of these alternative approaches would be negligible (see Sechrest & Yeaton, 1982).

Table 2
Overall Mean Distances Between Observed Fit-Index Values and the Corresponding True Values for Each Fit Index Under Simple and Complex True-Population Models

	Simple model			Complex model		
Fit index	ML	GLS	ADF	ML	GLS	ADF
NFI	.058	.237	.187	.047	.227	.175
BL86	.069	.284	.223	.058	.281	.216
TLI	.035	.132	.125	.029	.131	.115
BL89	.028	.102	.101	*.023*	.096	.090
RNI	.029	.110	.105	*.023*	.105	.093
CFI	.029	.106	.105	*.023*	.101	.093
GFI	.054	.050	.058	.052	.048	.054
AGFI	.075	.069	.079	.074	.069	.077
Gamma hat	*.026*	*.016*	*.046*	.025	*.016*	*.042*
CAK	.660	.585	.869	.663	.591	.832
CK	*.681*	*.606*	*.890*	*.687*	*.614*	*.855*
Mc	.092	.059	.156	.088	.057	.141
SRMR	.038	.053	.110	.035	.049	.114
RMSEA	.035	.028	.047	.034	.028	.045

Note. Mean distance = $\sqrt{\{[\Sigma(\text{observed fit-index value} - \text{true fit-index value})^2]/(\text{no. observed fit indexes})\}}$. ML = maximum likelihood; GLS = generalized least squares; ADF = asymptotic distribution-free method; NFI = normed fixed index; TLI = Tucker–Lewis Index (1973); BL86 = fit index by Bollen (1986); BL89 = fit index by Bollen (1989); RNI = relative noncentrality index; CFI = comparative fit index; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index; CAK = a rescaled version of Akaike's information criterion; CK = cross-validation index; Mc = McDonald's centrality index; CN = critical *N*; SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation. Smallest value in each column is italicized. CN methods were not applicable.

Overall Mean Distance

The OMDs between observed fit-index values and the corresponding expected fit-index values for the simple and complex true-population models were calculated for each fit index derived from ML, GLS, and ADF estimation methods. For example, the mean distance for ML-based NFI of the simple true-population model was equal to the square root of $\{[\Sigma(\text{observed fit-index value} - 1)^2]/8{,}400\}$. The smaller the mean distance, the better the fit index. The purpose for calculating the OMD was to gauge how likely and how much each fit index might depart from its true value under a correct model. Theoretically, these fit indices would equal their true values under correct models, and thus any departure from their values would indicate instability resulting from small sample size or violation of other underlying assumptions. For example, TLI or RNI would behave as a normed fit index asymptotically, but it could fall outside the 0–1 range when sample size was small or other underlying assumptions were violated. Thus, the OMD was a fair criterion for comparing the performance of fit indices under true-population (correct) models, although one might argue that it was an unfair comparison because the ranges of fit indices differ (in fact, this only occurs under some unusual conditions such as small sample size). Table 2 contains the OMDs between the observed fit-index values and the corresponding expected fit-index values. Overall, the values of the ML-based TLI, BL89, RNI, CFI, gamma hat, SRMR, and RMSEA were much closer to their corresponding true values than the other ML-based fit indices. The values of the GLS- or ADF-based GFI, gamma hat, and RMSEA as well as the GLS-based Mc and SRMR also were closer to their corresponding true values than the other GLS- or ADF-based fit indices. The distances for CAK and CK were always unacceptable.

Similarities in Performance of Fit Indices

Separate correlation matrices among fit indices derived from ML, GLS, and ADF methods for simple and complex models were obtained, to determine which fit indices might behave similarly. Each correlation matrix was calculated by collapsing across sample sizes, distributions, and model misspecifications, to determine if fit indices derived from ML, GLS, or ADF method for simple or complex models behaved similarly along three major dimensions: sample size, distribution, and model misspecification. The resulting patterns of correlations were identical;

Table 3
Overall Zero-Order Correlations Among Fit Indices Derived From ML, GLS, and ADF Methods

Fit index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
							ML method								
1. NFI[a]	—														
2. BL86[a]	0.999	—													
3. TLI[b]	0.875	0.879	—												
4. BL89[b]	0.874	0.876	0.999	—											
5. RNI[b]	0.877	0.879	0.999	1.000	—										
6. CFI[b]	0.880	0.883	0.999	0.999	0.999	—									
7. GFI[a]	0.975	0.980	0.868	0.863	0.866	0.869	—								
8. AGFI[a]	0.969	0.976	0.867	0.861	0.863	0.867	0.999	—							
9. Gamma hat[b]	0.855	0.860	0.987	0.986	0.985	0.985	0.874	0.875	—						
10. CAK[a]	−0.955	−0.959	−0.757	−0.751	−0.756	−0.760	−0.974	−0.972	−0.766	—					
11. CK[a]	−0.951	−0.954	−0.745	−0.739	−0.743	−0.748	−0.969	−0.941	−0.753	1.000	—				
12. Mc[b]	0.852	0.858	0.986	0.985	0.985	0.984	0.869	0.869	0.998	−0.758	−0.746	—			
13. CN	0.529	0.528	0.454	0.457	0.456	0.459	0.515	0.512	0.449	−0.481	−0.477	0.466	—		
14. SRMR	−0.426	−0.404	−0.384	−0.403	−0.403	−0.401	−0.325	−0.302	−0.336	0.299	0.295	−0.354	−0.451	—	
15. RMSEA[b]	−0.830	−0.835	−0.956	−0.956	−0.955	−0.952	−0.837	−0.837	−0.963	0.723	0.711	−0.974	−0.569	0.447	—
							GLS method								
1. NFI[a]	—														
2. BL86[a]	0.998	—													
3. TLI[b]	0.825	0.822	—												
4. BL89[b]	0.806	0.800	0.996	—											
5. RNI[b]	0.826	0.821	0.999	0.998	—										
6. CFI[b]	0.844	0.839	0.994	0.994	0.995	—									
7. GFI[a]	0.953	0.957	0.719	0.684	0.717	0.733	—								
8. AGFI[a]	0.947	0.954	0.714	0.675	0.709	0.725	0.999	—							
9. Gamma hat[b]	0.802	0.801	0.988	0.986	0.986	0.983	0.712	0.708	—						
10. CAK[a]	−0.884	−0.891	−0.543	−0.499	−0.540	−0.560	−0.972	−0.974	−0.532	—					
11. CK[a]	−0.875	−0.883	−0.528	−0.483	−0.524	−0.545	−0.967	−0.967	−0.516	1.000	—				
12. Mc[b]	0.800	0.799	0.989	0.986	0.986	0.982	0.709	0.705	1.000	−0.529	−0.513	—			
13. CN	0.611	0.615	0.455	0.450	0.451	0.467	0.531	0.531	0.453	−0.496	−0.489	0.460	—		
14. SRMR	−0.604	−0.582	−0.589	−0.585	−0.589	−0.590	−0.537	−0.516	−0.574	0.445	0.437	−0.580	−0.487	—	
15. RMSEA[b]	−0.779	−0.781	−0.949	−0.949	−0.949	−0.942	−0.680	−0.679	−0.961	0.510	0.494	−0.967	−0.563	0.626	—

Table 3 *(continued)*

Fit index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
							ADF method								
1. NFI[a]	—														
2. BL86[a]	0.998	—													
3. TLI[a]	0.956	0.955	—												
4. BL89[a]	0.953	0.946	0.998	—											
5. RNI[a]	0.958	0.952	0.999	1.000	—										
6. CFI[a]	0.958	0.951	0.998	0.999	1.000	—									
7. GFI[c]	0.742	0.759	0.672	0.646	0.657	0.657	—								
8. AGFI[c]	0.727	0.747	0.657	0.630	0.640	0.640	0.999	—							
9. Gamma hat[b]	0.552	0.542	0.548	0.553	0.554	0.554	0.227	0.214	—						
10. CAK[b]	−0.461	−0.458	−0.404	−0.401	−0.407	−0.406	−0.208	−0.199	−0.957	—					
11. CK[b]	−0.457	−0.454	−0.399	−0.396	−0.402	−0.401	−0.206	−0.197	−0.955	1.000	—				
12. Mc[b]	0.581	0.573	0.576	0.579	0.581	0.580	0.246	0.233	0.988	−0.959	−0.957	—			
13. CN	0.623	0.632	0.528	0.517	0.520	0.518	0.427	0.423	0.402	−0.412	−0.411	0.437	—		
14. SRMR	−0.702	−0.705	−0.647	−0.637	−0.643	−0.643	−0.653	−0.643	−0.453	0.424	0.423	−0.481	−0.558	—	
15. RMSEA[b]	−0.680	−0.677	−0.668	−0.666	−0.668	−0.667	−0.337	−0.326	−0.953	0.914	0.912	−0.966	−0.583	0.581	—

Note. ML = maximum likelihood; GLS = generalized least squares; ADF = asymptotic distribution-free; NFI = normed fit index; BL86 = fit index by Bollen (1986); TLI = Tucker–Lewis Index (1973); BL89 = fit index by Bollen (1989); RNI = relative noncentrality index; CFI = comparative fit index; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index; CAK = a rescaled version of Akaike's of formation criterion; CK = cross-validation index; Mc = McDonald's centrality index; CN = critical *N*; SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation. Fit indices having the same superscript belong to the same cluster of correlated indices. Higher intercorrelations between fit indices are underscored.

thus, we further calculated separate overall correlation matrices across simple and complex models for ML, GLS, and ADF methods. Table 3 contains the correlations. Inspection of the correlation matrix for the ML-based fit indices revealed that there were two major clusters of correlated fit indices. NFI, BL86, GFI, AGFI, CAK, and CK were clustered with high correlations. Another cluster of high intercorrelations included TLI, BL89, RNI, CFI, Mc, and RMSEA. CN and SRMR were found to be least similar to the other ML-based fit indices. The same pattern was observed for the GLS-based fit indices. Finally, three clusters of ADF-based fit indices were observed in the correlation matrix. The first cluster included NFI, BL86, TLI, BL89, RNI, and CFI. The second cluster included CAK, CK, gamma hat, Mc, and RMSEA. The last cluster included GFI and AGFI. As with ML and GLS, CN and SRMR seemed to be less similar to the other ADF-based fit indices.

Sensitivity to Underparameterized Model Misspecification and Effects of Sample Size and Distribution

Our preliminary analyses indicated that values of most fit indices vary across different estimation methods; thus, we performed a series of ANOVAs separately for fit indices based on ML, GLS, and ADF methods, to determine if different patterns of effects of model misspecification, sample size, and distribution existed among the three estimation methods. Specifically, to examine the potential additive or multiplicative effects of model misspecification (i.e., sensitivity to underparameterized model misspecification) to the effect of sample size and distribution on fit indices, we performed a series of 6 × 7 × 3 (Sample Size × Distribution × Model Misspecification) ANOVAs on each of the ML-, GLS-, and ADF-based fit indices. Separate analyses were performed for simple and complex models, to determine if different types of model misspecification (i.e., models with misspecified factor covariance[s] and models with misspecified factor loadings) exerted differential effects on fit indices derived from ML, GLS, and ADF methods. The larger the amount of variance accounted for by model misspecification and the smaller the amount of variance accounted for by sample size and distribution, the better the fit index was considered to be. Tables 4 through 6 display the η^2 for each predictor variable and interaction term derived from the ANOVA performed on each fit index.

Analyses for simple models. For the ML- and GLS-based fit indices derived for simple models (see Tables 4 and 5), an extremely large proportion of variance in SRMR (η^2s = .914 and .859, respectively) and a moderate proportion of variance in TLI, BL89, RNI, CFI, gamma hat, Mc, and RMSEA were accounted for by model misspecification (η^2s ranged from .309 to .487). Inspection of the cell means suggested that the mean values of these fit indices derived from the two simple misspecified models were substantially different from those derived from the simple true-population model. Thus, these fit indices, especially SRMR, were more sensitive to simple misspecified models than the rest of the other fit indices. Model misspecification accounted for a substantial amount of variance (η^2 = .608) in the ADF-based SRMR and a moderate amount of variance (η^2s ranged from .389 to .516) in the ADF-based NFI, BL86, TLI, BL89, RNI, and CFI; thus, these ADF-based fit indices were more sensitive to simple misspecified models than the other fit indices (see Table 6).

Furthermore, sample size accounted for a substantial amount of variance (η^2s ranged from .605 to .882) in the ML- and GLS-based NFI, BL86, GFI, AGFI, CAK, and CK, after controlling for the effects of distribution, model misspecification, and their interaction terms. Distribution accounted for a relatively small proportion of variance in any of the ML- and GLS-based indices. Sample size accounted for a large proportion of variance (η^2s ranged from .674 to .877) in the ADF-based gamma hat, CAK, CK, Mc, and RMSEA. Sample size also accounted for a moderate proportion of variance (η^2s = .343) in the ADF-based CN. Distribution exerted a moderate effect on the ADF-based GFI and AGFI (η^2s = .373 and .382, respectively). Also, a moderate interaction effect between sample size and model misspecification on the ML-, GLS-, and ADF-based CN (η^2s ranged from .340 to .390) indicated that the sample-size effect was more substantial for the simple true-population model than for the two complex misspecified models.

Analyses for complex models. For the ML- and GLS-based fit indices derived for complex models (see Tables 4 and 5), a relatively large proportion of variance in TLI, BL89, RNI, CFI, gamma hat, Mc, and RMSEA (η^2s ranged from .699 to .766) was accounted for by model misspecification. A moderate amount of variance in ML- and GLS-based NFI and BL86 and the ML-based GFI and AGFI (η^2s ranged from .454 to .549) was accounted for by model misspecification. Model misspecification accounted for a

Table 4

η^2 Derived From a 6 × 7 × 3 Analysis of Variance (Sample Size × Distribution × Model Misspecification) Performed Separately on Maximum-Likelihood-Based Fit Indices of Simple and Complex Models

	Sample size		Distribution		Misspecification		Sample size × distribution		Sample size × misspecification		Distribution × misspecification		Sample size × distribution × misspecification	
Fit index	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex
NFI[a]	.635	.319	.054	.034	.151	.548	.035	.020	.003	.001	.004	.001	.014	.000
BL86[a]	.635	.326	.059	.038	.143	.534	.037	.022	.003	.000	.004	.000	.014	.000
TLI[b]	.115	.039	.158	.062	.315	.748	.115	.040	.007	.000	.011	.000	.040	.000
BL89[b]	.111	.035	.152	.056	.330	.763	.112	.038	.007	.000	.011	.001	.039	.000
RNI[b]	.114	.038	.151	.057	.326	.759	.113	.039	.007	.000	.011	.001	.039	.000
CFI[b]	.122	.040	.149	.055	.321	.759	.112	.038	.007	.000	.010	.001	.040	.000
GFI[a]	.646	.370	.067	.051	.101	.471	.046	.028	.004	.003	.005	.000	.017	.000
AGFI[a]	.645	.375	.071	.059	.094	.454	.049	.030	.004	.004	.005	.000	.017	.000
Gamma hat[b]	.122	.044	.159	.066	.309	.743	.121	.044	.007	.000	.012	.001	.041	.000
CAK[a]	.801	.597	.037	.030	.061	.301	.030	.022	.002	.000	.002	.000	.009	.000
CK[a]	.813	.617	.034	.029	.057	.286	.028	.029	.001	.000	.002	.000	.009	.000
Mc[b]	.118	.040	.161	.064	.339	.766	.115	.040	.007	.001	.013	.002	.040	.001
CN	.240	.201	.027	.024	.221	.256	.030	.029	.343	.357	.042	.042	.061	.059
SRMR	.020	.151	.002	.063	.914	.653	.001	.010	.010	.026	.001	.005	.000	.001
RMSEA[b]	.091	.038	.161	.077	.466	.763	.061	.025	.017	.009	.044	.021	.023	.002

Note. η^2 = the proportion of variance accounted for by each predictor variable or interaction term (η^2 was calculated by dividing the Type 3 sum of squares for a given predictor or interaction term by the corrected total sum of squares). NFI = normed fit index; BL86 = fit index by Bollen (1986); TLI = Tucker–Lewis Index (1973); BL89 = fit index by Bollen (1989); RNI = relative noncentrality index; CFI = comparative fit index; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index; CAK = a rescaled version of Akaike's of formation criterion; CK = cross-validation index; Mc = McDonald's centrality index; CN = critical *N;* SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation. Fit indices having the same superscript behaved similarly.

Table 5
η^2 Derived From a 6 × 7 × 3 Analysis of Variance (Sample Size × Distribution × Model Misspecification) Performed Separately on Generalized-Least-Squares-Based Fit Indices of Simple and Complex Models

	Sample size		Distribution		Misspecification		Sample size × distribution		Sample size × misspecification		Distribution × misspecification		Sample size × distribution × misspecification	
Fit index	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex
NFI[a]	.605	.293	.054	.058	.217	.549	.013	.005	.017	.022	.022	.038	.019	.001
BL86[a]	.619	.313	.047	.043	.207	.548	.013	.007	.018	.024	.020	.030	.020	.001
TLI[b]	.036	.015	.153	.080	.442	.749	.073	.027	.011	.001	.051	.048	.037	.003
BL89[b]	.022	.005	.154	.090	.466	.745	.065	.023	.014	.004	.058	.061	.035	.001
RNI[b]	.034	.014	.152	.088	.447	.737	.069	.025	.011	.001	.056	.059	.036	.002
CFI[b]	.048	.019	.152	.088	.445	.743	.064	.022	.017	.003	.054	.059	.036	.001
GFI[a]	.730	.529	.044	.054	.106	.331	.026	.020	.003	.002	.011	.020	.013	.000
AGFI[a]	.736	.547	.043	.048	.099	.320	.028	.022	.003	.002	.010	.015	.013	.000
Gamma hat[b]	.044	.011	.150	.105	.381	.699	.104	.044	.011	.005	.045	.040	.058	.000
CAK[a]	.873	.775	.018	.023	.045	.156	.014	.011	.002	.001	.006	.010	.008	.000
CK[a]	.882	.792	.017	.021	.041	.144	.013	.011	.002	.001	.005	.009	.007	.000
Mc[b]	.041	.010	.151	.103	.393	.705	.101	.044	.011	.005	.046	.038	.056	.001
CN	.247	.224	.027	.022	.213	.238	.033	.028	.340	.352	.040	.038	.059	.060
SRMR	.046	.122	.012	.126	.859	.588	.003	.009	.006	.006	.004	.053	.000	.000
RMSEA[b]	.031	.013	.151	.101	.487	.708	.055	.026	.024	.015	.075	.050	.031	.002

Note. η^2 = the proportion of variance accounted for by each predictor variable or interaction term (η^2 was calculated by dividing the Type 3 sum of squares for a given predictor or interaction term by the corrected total sum of squares). NFI = normed fit index; BL86 = fit index by Bollen (1986); TLI = Tucker–Lewis Index (1973); BL89 = fit index by Bollen (1989); RNI = relative noncentrality index; CFI = comparative fit index; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index; CAK = a rescaled version of Akaike's of formation criterion; CK = cross-validation index; Mc = McDonald's centrality index; CN = critical *N;* SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation. Fit indices having the same superscript behaved similarly.

Table 6

η^2 Derived From a 6 × 7 × 3 Analysis of Variance (Sample Size × Distribution × Model Misspecification) Performed Separately on Asymptotic-Distribution-Free-Based Fit Indices of Simple and Complex Models

	Sample size		Distribution		Misspecification		Sample size × distribution		Sample size × misspecification		Distribution × misspecification		Sample size × distribution × misspecification	
Fit index	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex
NFI[a]	.252	.088	.124	.102	.391	.593	.011	.007	.032	.061	.042	.040	.008	.004
BL86[a]	.266	.095	.108	.085	.389	.602	.011	.007	.035	.067	.039	.032	.008	.004
TLI[a]	.106	.014	.107	.077	.516	.683	.004	.008	.034	.059	.047	.034	.005	.004
BL89[a]	.093	.010	.128	.096	.510	.665	.004	.008	.033	.057	.054	.044	.005	.004
RNI[a]	.100	.013	.125	.094	.506	.665	.004	.008	.032	.055	.052	.043	.005	.003
CFI[a]	.100	.013	.126	.095	.505	.667	.004	.008	.031	.054	.053	.043	.005	.003
GFI[c]	.129	.048	.373	.409	.202	.328	.089	.014	.019	.041	.024	.053	.060	.017
AGFI[c]	.130	.049	.382	.422	.193	.315	.090	.013	.020	.042	.023	.052	.061	.018
Gamma hat[b]	.710	.578	.051	.081	.081	.171	.023	.036	.007	.007	.014	.028	.002	.005
CAK[b]	.871	.818	.018	.030	.025	.054	.013	.021	.004	.005	.006	.012	.002	.005
CK[b]	.877	.827	.017	.028	.024	.052	.013	.020	.004	.005	.005	.012	.002	.005
Mc[b]	.738	.600	.049	.077	.093	.197	.012	.019	.003	.002	.014	.026	.002	.001
CN	.343	.310	.002	.002	.221	.237	.002	.001	.390	.401	.001	.001	.001	.001
SRMR	.137	.169	.129	.212	.608	.392	.018	.028	.021	.017	.005	.042	.002	.008
RMSEA[b]	.674	.541	.039	.055	.184	.300	.006	.010	.007	.011	.012	.018	.002	.002

Note. η^2 = the proportion of variance accounted for by each predictor variable or interaction term (η^2 was calculated by dividing the Type 3 sum of squares for a given predictor or interaction term by the corrected total sum of squares). NFI = normed fit index; BL86 = fit index by Bollen (1986); TLI = Tucker–Lewis Index (1973); BL89 = fit index by Bollen (1989); RNI = relative noncentrality index; CFI = comparative fit index; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index; CAK = a rescaled version of Akaike's of formation criterion; CK = cross-validation index; Mc = McDonald's centrality index; CN = critical *N;* SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation. Fit indices having the same superscript behaved similarly.

small-to-moderate amount of variance in the GLS-based GFI and AGFI (η^2s = .331 and .320, respectively). It accounted for a moderate to relatively large amount of variance in the ML- and GLS-based SRMR (η^2s = .653 and .588, respectively). Model misspecification accounted for a moderate to relatively large amount of variance (η^2s ranged from 5.93 to .667) in the ADF-based NFI, BL86, TLI, BL89, RNI, and CFI (see Table 6). Overall, all types of fit indices (except SRMR) seemed more sensitive in detecting the complex misspecified models (i.e., models with misspecified factor loading[s]) than the simple misspecified models (i.e., models with misspecified factor covariance[s]).[6] SRMR was more sensitive in detecting the simple than the complex misspecified models, although the ability to detect complex misspecified models for the ML- and GLS-based SRMR remained reasonably high.

Sample size accounted for a small-to-large proportion of variance in the ML- and GLS-based NFI, BL86, GFI, AGFI, CAK, and CK (η^2s ranged from .293 to .792). Sample size also accounted for a substantial amount of variance in the ADF-based gamma hat, CAK, CK, Mc, and RMSEA (η^2s ranged from .541 to .827). Distribution accounted only for a moderate amount of variance in the ADF-based GFI and AGFI (η^2s = .409 and .422, respectively). A moderate interaction effect between sample size and model misspecification on the ML-, GLS-, and ADF-based CN (η^2s ranged from .352 to .401) also was observed, indicating that the sample-size effect was more substantial for the complex true-population model than for the two complex misspecified models.

Effects of Estimation Method, Distribution, and Sample Size on Fit Indices

To determine the importance of the additive and multiplicative effects of sample size, distribution, and estimation method on fit indices, we conducted a series of ANOVAs on fit indices derived from each of the simple and complex true-population models and misspecified models. These analyses were performed separately for simple and complex true-population models and misspecified models, to determine if the effect of estimation method after controlling for the effects of sample size and distribution varied as a function of model quality, as reported by Sugawara and MacCallum (1993). The results for simple and complex models were similar and hence are discussed together. Tables 7 through 9 contain the proportion of variance in each fit index accounted for by sample size, distribution, estimation method, and various interaction terms derived from each ANOVA. Note that the smaller the effects of sample size, distribution, and estimation method, the better was the fit index.

Analyses for simple and complex true-population models. The 6 × 7 × 3 (Sample Size × Distribution × Estimation Method) ANOVAs performed on the fit indices derived for the two types of true-population models revealed that sample size accounted for a substantial amount of variance in each of the following fit indices (see Table 7): NFI, BL86, GFI, AGFI, CAK, CK, and CN (η^2s ranged from .480 to .888). A small-to-moderate amount of variance was observed also for the other fit indices. The interaction between sample size and estimation method accounted for relatively small amounts of variance in NFI, BL86, TLI, BL89, RNI, CFI, gamma hat, Mc, and RMSEA (η^2s ranged from .102 to .266). Inspection of cell means revealed that NFI, BL86, TLI, BL89, RNI, and CFI behaved differently across estimation methods at small sample sizes, but they behaved consistently across estimation methods at large sample sizes. Gamma hat, Mc, and RMSEA also behaved less consistently across estimation methods at small sample sizes. In addition, distribution accounted for a relatively small proportion of variance in TLI, BL89, RNI, CFI, GFI, AGFI, and RMSEA (η^2s ranged from .116 to .160). Estimation method accounted for a small proportion of variance in NFI and BL86 (η^2s ranged from .242 to .264).

Analysis for simple and complex misspecified models 1 and 2. A series of 6 × 7 × 3 (Sample Size × Distribution × Estimation Method) ANOVAs were conducted on the fit indices derived from the simple and complex misspecified models. The results were similar for all the misspecified models; however, the effect of estimation method was slightly increased as the degree of model misspecification increased (see Tables 8 and 9). Sample size was found to account for a relatively small proportion of variance in NFI and BL86 (η^2s ranged from .144 to .206) and a moderate-to-substantial amount of variance in GFI, AGFI, gamma hat, CAK, CK, Mc, CN, and RMSEA (η^2s

[6] Results from a five-way ANOVA (Sample Size × Distribution × Model Misspecification × Estimation Method × Model Type) revealed that there were moderate-to-substantial interaction effects between model misspecification and model type (simple vs. complex model) for all fit indices but CN.

Table 7
η^2 Derived From a 6 × 7 × 3 Analysis of Variance (Sample Size × Distribution × Estimation Method) Performed Separately on Each Fit Index of the Simple or Complex True-Population Model

	Sample size		Distribution		Method		Sample size × distribution		Sample size × method		Distribution × method		Sample size × distribution × method	
Fit index	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex
NFI	.521	.481	.042	.040	.242	.264	.008	.008	.126	.139	.008	.009	.003	.003
BL86	.519	.480	.044	.043	.242	.263	.008	.008	.126	.139	.008	.010	.003	.003
TLI	.237	.213	.131	.134	.075	.079	.063	.063	.111	.102	.101	.115	.045	.050
BL89	.240	.217	.128	.131	.078	.081	.057	.057	.120	.112	.100	.115	.042	.046
RNI	.236	.213	.128	.131	.076	.079	.061	.061	.112	.103	.101	.116	.047	.052
CFI	.306	.283	.116	.119	.095	.104	.050	.049	.114	.107	.083	.096	.034	.037
GFI	.628	.620	.155	.153	.008	.006	.047	.045	.035	.043	.025	.023	.008	.009
AGFI	.621	.613	.160	.158	.008	.006	.048	.047	.034	.043	.025	.023	.008	.009
Gamma hat	.361	.354	.056	.062	.056	.049	.028	.030	.266	.244	.049	.054	.040	.044
CAK	.866	.879	.010	.010	.013	.009	.005	.005	.059	.046	.010	.009	.009	.009
CK	.874	.888	.010	.010	.012	.009	.005	.005	.056	.043	.009	.009	.009	.008
Mc	.368	.359	.064	.071	.055	.047	.030	.033	.261	.239	.053	.057	.039	.042
CN	.814	.815	.033	.032	.007	.007	.046	.046	.017	.017	.017	.016	.022	.022
SRMR	.415	.336	.111	.101	.185	.196	.025	.023	.096	.104	.086	.093	.018	.021
RMSEA	.398	.390	.142	.148	.031	.026	.020	.021	.186	.173	.086	.091	.019	.020

Note. η^2 = the proportion of variance accounted for by each predictor variable or interaction term (η^2 was calculated by dividing the Type 3 sum of squares for a given predictor or interaction term by the corrected total sum of squares). NFI = normed fit index; BL86 = fit index by Bollen (1986); TLI = Tucker–Lewis Index (1973); BL89 = fit index by Bollen (1989); RNI = relative noncentrality index; CFI = comparative fit index; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index; CAK = a rescaled version of Akaike's of formation criterion; CK = cross-validation index; Mc = McDonald's centrality index; CN = critical *N;* SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation.

Table 8
η^2 Derived From a 6 × 7 × 3 Analysis of Variance (Sample Size × Distribution × Estimation Method) Performed Separately on Each Fit Index of the Simple or Complex Misspecified Model 1

	Sample size		Distribution		Method		Sample size × distribution		Sample size × method		Distribution × method		Sample size × distribution × method	
Fit index	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex
NFI	.202	.148	.043	.052	.592	.662	.003	.002	.059	.059	.020	.022	.011	.005
BL86	.206	.151	.035	.041	.599	.673	.003	.002	.061	.061	.016	.018	.011	.006
TLI	.035	.022	.081	.092	.589	.658	.019	.012	.008	.005	.053	.049	.033	.024
BL89	.025	.012	.095	.112	.586	.644	.018	.010	.008	.002	.061	.060	.030	.024
RNI	.034	.021	.092	.107	.577	.641	.019	.011	.008	.005	.058	.056	.031	.024
CFI	.035	.021	.092	.107	.580	.642	.018	.011	.008	.005	.058	.056	.031	.023
GFI	.349	.311	.160	.241	.083	.053	.021	.012	.069	.091	.084	.146	.101	.022
AGFI	.346	.306	.163	.247	.082	.053	.022	.012	.069	.090	.084	.146	.101	.022
Gamma hat	.311	.273	.027	.036	.066	.139	.038	.020	.298	.251	.070	.091	.039	.047
CAK	.795	.813	.006	.008	.019	.031	.010	.005	.087	.066	.019	.023	.014	.017
CK	.806	.825	.006	.007	.018	.030	.009	.005	.082	.062	.018	.022	.014	.016
Mc	.320	.272	.030	.038	.066	.160	.043	.020	.293	.250	.075	.094	.031	.034
CN	.647	.423	.019	.028	.044	.159	.039	.039	.072	.153	.072	.083	.049	.056
SRMR	.104	.088	.104	.116	.449	.524	.013	.015	.038	.026	.076	.109	.008	.018
RMSEA	.291	.251	.032	.032	.061	.181	.056	.025	.288	.252	.092	.109	.023	.020

Note. η^2 = the proportion of variance accounted for by each predictor variable or interaction term (η^2 was calculated by dividing the Type 3 sum of squares for a given predictor or interaction term by the corrected total sum of squares). NFI = normed fit index; BL86 = fit index by Bollen (1986); TLI = Tucker–Lewis Index (1973); BL89 = fit index by Bollen (1989); RNI = relative noncentrality index; CFI = comparative fit index; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index; CAK = a rescaled version of Akaike's of formation criterion; CK = cross-validation index; Mc = McDonald's centrality index; CN = critical *N*; SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation.

Table 9

η^2 Derived From a 6 × 7 × 3 Analysis of Variance (Sample Size × Distribution × Estimation Method) Performed Separately on Each Fit Index of the Simple or Complex Misspecified Model 2

	Sample size		Distribution		Method		Sample size × distribution		Sample size × method		Distribution × method		Sample size × distribution × method	
Fit index	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex
NFI	.144	.021	.061	.083	.642	.784	.007	.001	.046	.028	.036	.040	.005	.003
BL86	.148	.023	.051	.060	.652	.812	.007	.001	.047	.030	.031	.032	.005	.004
TLI	.020	.001	.102	.096	.621	.764	.021	.002	.006	.013	.068	.050	.018	.006
BL89	.014	.004	.119	.129	.608	.719	.019	.002	.006	.010	.078	.064	.016	.006
RNI	.019	.001	.115	.123	.605	.730	.020	.002	.005	.012	.075	.061	.017	.006
CFI	.019	.001	.115	.123	.608	.732	.020	.002	.005	.013	.075	.061	.017	.006
GFI	.292	.113	.168	.263	.122	.096	.054	.003	.064	.128	.112	.234	.074	.035
AGFI	.291	.112	.170	.264	.122	.097	.054	.003	.063	.127	.112	.235	.075	.036
Gamma hat	.268	.191	.035	.067	.081	.226	.028	.016	.345	.222	.089	.114	.034	.040
CAK	.766	.694	.010	.022	.026	.070	.009	.008	.113	.084	.027	.042	.015	.026
CK	.778	.709	.009	.021	.025	.067	.008	.007	.108	.080	.026	.040	.014	.025
MC	.275	.185	.038	.067	.083	.267	.032	.013	.342	.220	.097	.118	.026	.020
CN	.603	.332	.027	.049	.054	.195	.041	.044	.086	.152	.094	.116	.059	.055
SRMR	.072	.047	.078	.129	.521	.605	.009	.006	.025	.013	.070	.102	.005	.009
RMSEA	.249	.183	.037	.049	.079	.263	.041	.016	.339	.227	.111	.133	.022	.017

Note. η^2 = the proportion of variance accounted for by each predictor variable or interaction term (η^2 was calculated by dividing the Type 3 sum of squares for a given predictor or interaction term by the corrected total sum of squares). NFI = normed fit index; BL86 = fit index by Bollen (1986); TLI = Tucker–Lewis Index (1973); BL89 = fit index by Bollen (1989); RNI = relative noncentrality index; CFI = comparative fit index; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index; CAK = a rescaled version of Akaike's of formation criterion; CK = cross-validation index; Mc = McDonald's centrality index; CN = critical *N;* SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation.

ranged from .268 to .825) under simple misspecified models 1 and 2, as well as complex misspecified model 1. Sample size accounted only for a moderate-to-large proportion of variance in CAK, CK, and CN (η^2s ranged from .332 to .709) under complex misspecified model 2. A small proportion of variance in GFI and AGFI also was accounted for by distribution (η^2s ranged from .153 to .264). Estimation method had a moderate-to-substantial effect on NFI, BL86, TLI, BL89, RNI, CFI, and SRMR (η^2s ranged from .292 to .673) derived from simple and complex misspecified models. A relatively small estimation-method effect (η^2s ranged from .226 to .263) was observed for gamma hat, Mc, and RMSEA derived from complex misspecified model 2. Furthermore, there were also relatively small-to-moderate interaction effects between sample size and estimation method (η^2s ranged from .222 to .345) on gamma hat, Mc, and RMSEA derived from simple and complex misspecified models. Inspection of cell means revealed that these three fit indices behaved less consistently at small sample sizes than at large sample sizes. Under the complex misspecified model 2, there were a small distribution effect and a small interaction effect between distribution and estimation method on GFI and AGFI. Inspection of cell means suggested that GFI and AGFI derived from complex misspecified model 2 behaved less consistently across estimation methods under Distributional Conditions 1, 3, and 4. Finally, inspection of Tables 7 through 9 yielded a systematic decrease in the magnitude of estimation-method effect as a result of a decrease in quality of models.[7]

Discussion

Our findings suggest that the performance of fit indices is complex and that additional research with a wider class of models and conditions is needed, to provide final answers on the relative merits of many of these indices. In spite of this complexity, there are enough clear-cut results from this study to permit us to make some very specific recommendations for practice. We do this in a sequential manner, first making suggestions about which indices not to use, then concluding with suggestions about indices to use. A good fit index should have a large model misspecification effect accompanied with trivial effects of sample size, distribution, and estimation method. Summary tables and detailed description of various sources of effects on fit indices are presented in our technical report (Hu & Bentler, 1997).

Recommendations for the Selection of Fit Indices in Practice

CAK and CK are not sensitive to model misspecification, estimation method, or distribution but are extremely sensitive to sample size. We do not recommend their use.

CN is not sensitive to model misspecification, estimation method, or distribution but is very sensitive to sample size. We do not recommend its use.

NFI and BL86 are not sensitive to simple model misspecification but are moderately sensitive to complex model misspecification. Although a slight effect of estimation method under true-population models and a substantial estimation-method effect under misspecified models were observed for NFI and BL86, they are not sensitive to distribution. ML- and GLS-based NFI and BL86 are sensitive to sample sizes. The ADF-based NFI and BL86 are less sensitive to sample size, but they substantially underestimate true-population values. We do not recommend their use.

GFI and AGFI are not sensitive to model misspecification and estimation method. ML- and GLS-based GFI and AGFI are not sensitive to distribution but are sensitive to sample size. ADF-based GFI and AGFI are sensitive to distribution but are not sensitive to sample size. We do not recommend their use.

TLI, BL89, RNI, and CFI are moderately sensitive to simple model misspecification but are very sensitive to complex model misspecification. They are not influenced by estimation method under true-population models but are substantially influenced by estimation method under misspecified models. These fit indices are less sensitive to distribution and sample size. We recommend these fit indices be used in general; however, ML-based TLI, BL89, RNI, and CFI are more preferable when sample size is small (e.g., $N \leq 250$), because the GLS- and ADF-based TLI, BL89, RNI, and CFI underestimate their true-population values and have much larger variances than those based on ML at small sample size.

ML- and GLS-based gamma hat, Mc, and RMSEA are moderately sensitive to simple model misspecifi-

[7] Four-way ANOVAs (Sample Size × Distribution × Model Misspecification × Estimation Method) revealed that there are substantial interaction effects between model misspecification and estimation method for Type 1, Type 2, and Type 3 incremental fit indices and SRMR.

cation and are very sensitive to complex model misspecification. These fit indices based on the ADF method are less sensitive to both simple and complex model misspecification. Estimation method exerts an effect on gamma hat, Mc, and RMSEA at small sample sizes but exerts no effect at large sample sizes. ML- and GLS-based gamma hat, Mc, and RMSEA are less sensitive to distribution and sample size. The fit indices based on the ADF method are not sensitive to distribution but are very sensitive to sample size. ML- and GLS-based gamma hat, Mc, and RMSEA performed equally well, and we recommended their use. However, we do not recommend that the ADF-based gamma hat, Mc, and RMSEA be used in practice.

Among all the fit indices studied, SRMR is most sensitive to simple model misspecification and is moderately sensitive to complex model misspecification. SRMR is not sensitive to estimation method under true-population models but is sensitive to estimation method under misspecified models. SRMR is less sensitive to distribution and sample size. At small sample sizes, GLS-based SRMR has a slight tendency to overestimate true-population values, and ADF-based SRMR substantially overestimates true-population values. We recommend the ML-, GLS-, and ADF-based SRMR be used in general, but ML-based SRMR is preferable when sample size is small (e.g., $N \leq 250$). The average absolute standardized residual computed by EQS, not studied here, has an identical rationale and should perform the same as SRMR.

On the basis of these results, with ML and GLS methods, we recommend a two-index presentation strategy for researchers. This would include definitely using SRMR and supplementing this with one of the following indices: TLI, BL89, RNI, CFI, gamma hat, Mc, or RMSEA. By using cutoff criteria for both SRMR and one of the supplemented indices, researchers should be able to identify models with underparameterized factor covariance(s), underparameterized factor loading(s), or a combination of both types of underparameterization. These alternative indices perform interchangeably in all distributional conditions (see Table 3) except when sample size is small (e.g., $N \leq 250$). At small sample size, (a) the range of TLI (or NNFI) tends to be large (e.g., Bentler, 1990); (b) Mc tends to depart substantially from its true-population values; and (c) RMSEA tends to overreject substantially true-population models. Therefore a cautious interpretation of model acceptability based on any of these three fit indices is recommended when sample size is small. Note that Marsh et al. (1996) have proposed a normed version of TLI, to reduce the variance of TLI, and have suggested that the normed version of TLI may be more preferable when sample size is small.

With the ADF method, we recommend the definite use of SRMR, supplemented with one of the following indices: TLI, BL89, RNI, or CFI. However, we do not recommend the use of any ADF-based fit indices when sample size is small, because they depart substantially from their true-population values and tend to overreject their true-population models (see also Hu et al., 1992). Better results may be observed with new approaches that attempt to improve ADF estimation in small samples.[8]

Finally, most of the fit indices (except gamma hat, Mc, and RMSEA, which perform equally well under ML and GLS methods) obtained from ML perform much better (less likely to be influenced by various sources of irrelevant effects and less likely to depart from their true-population values) than those obtained from GLS and ADF and should be preferred indicators for model selection and evaluation.

Other General Observations

The ability to discriminate well-fitting from badly fitting models for the ML-, GLS-, and ADF-based SRMR is substantially superior to that of any other fit index under simple misspecified models, but it is slightly less sensitive to complex model misspecification than several above-mentioned fit indices. One possible explanation for this finding is that the loadings of the observed indications on a given factor become biased due to the misspecification of the co-

[8] Under the ADF method, there was a substantial sample-size effect on the three noncentrality-based absolute-fit indices. Because these absolute-fit indices rely very heavily on the quality of the ADF chi-square statistic and because this statistic simply cannot be trusted at smaller sample sizes (e.g., Bentler & Dudgeon, 1996; Hu et al., 1992), we are optimistic that the finite sample improvements in the ADF tests made, for example, by Yuan and Bentler (1997) will remove this performance problem in the near future. In general, these indices also have good sensitivity to model misspecification. This does break down with ADF estimation, and it is possible that this breakdown also will be prevented with the Yuan–Bentler ADF test. Future work will have to evaluate this suggestion.

variance between two factors and thus the average of squared residuals is more likely to capture this type of misspecification as a result of a greater number of biased parameter estimates obtained. Our findings are consistent with La Du and Tanaka's (1989) findings that ML-based NFI is more sensitive to the underparameterized model misspecification than the ML- and GLS-based GFI. However, in contrast to the results of Maiti and Mukherjee (1991), we have found GFI to be quite insensitive to various types of underparameterized model misspecification. Because they found GFI to be sensitive as their newly proposed indices of structural closeness (ISC), we suspect that ISC also would not have performed well in our study. However, ISC possesses, under some circumstances, an excellent property of going to an extremely small value under extreme misspecification, which they call *specificity*. Certainly this feature, and the ISC indices, require further evaluation under conditions of extreme model misfit.

A major effort in prior research on fit indices has been to examine sensitivity of fit indices to sample size. Virtually all of this research has been conducted under the true models (e.g., Anderson & Gerbing, 1984; Anderson et al., 1985; Bollen, 1986, 1989, 1990; Marsh et al., 1988). To test the generality of previous findings, we examined the effect of sample size on fit indices under both true-population and misspecified models. The means of the empirical sampling distributions for Type 2 and Type 3 incremental indices varied with sample size to a lesser extent than was found for Type 1 incremental fit indices. In keeping with the findings of Marsh et al. (1988), Type 1 incremental fit indices tended to underestimate their asymptotic values and overreject true models at small sample sizes. This was especially true for indices obtained from GLS and ADF. Obviously, Type 1 incremental indices are influenced by the badness of the null model as well as the goodness of fit of the target model. Among the absolute-fit indices, GFI, AGFI, CAK, and CK derived from ML and GLS methods, as well as CAK, CK, and the noncentrality-based absolute-fit indices derived from the ADF method, were substantially influenced by sample size. The quality of models does not have a substantial effect on the relationship between the sample size and the mean values of most of the fit indices studied here (CN is the only exception). The pattern of association between the mean values of all three types of fit indices and sample size for the two misspecified models are quite similar to that for the true-population model. Our results on absolute indices are mixed. The Type 2 and Type 3 incremental fit indices and the noncentrality-based absolute-fit indices, in general, outperform the Type 1 incremental and the rest of the absolute-fit indices. The underestimation of perfect fit by the fit indices studied here, which is evident at the smaller sample sizes, becomes trivially small at the two largest sample sizes (i.e., 2,500 and 5,000). This is consistent with the theoretically predicted asymptotic properties and has been noted previously in several other studies (e.g., Bearden et al., 1982; Bentler, 1990; La Du & Tanaka, 1989).

Our findings on the effect of estimation method on all three types of incremental fit indices are more optimistic than those of Sugawara and MacCallum (1993). Sugawara and MacCallum have reported that values of incremental fit indices such as NFI, BL86, BL89, and TLI varied substantially across estimation methods and that this phenomenon held for both poor- and well-fitting methods. However, our results indicated that Type 2 and Type 3 incremental as well as absolute-fit indices behave relatively consistently across the three estimation methods under both types of true-population models (especially when sample size is relatively large), although Type 1 incremental fit indices seem to behave less consistently across estimation methods under both true-population and misspecified models. These inconsistent findings may be due to the differences in the range of sample sizes and quality of models used in each of the studies, for example, (a) small sample-size-to-model-size ratios and (b) the use of good-fitting models instead of true-population models by Sugawara and MacCallum.

Under both simple and complex misspecified models, all three types of incremental fit indices behave less consistently across ML, GLS, and ADF methods. These findings are consistent with those of Sugawara and MacCallum (1993). Sugawara and MacCallum have suggested that the effect of estimation methods on fit is tied closely to the nature of the weight matrices used by the methods. According to them, incremental fit indices, which use the discrepancy function value for the null model in their calculation, tend to behave erratically across estimation methods, because the discrepancy function values for a null model vary as a function of the weight matrices defined in various estimation methods. They also suggest that this phenomenon will occur even for a model that is quite consistent with the observed data. Our findings suggest that their proposition cannot be generalized to various situations (e.g., when there is dependence

among latent variates or when a true-population model is analyzed). For example, Type 2 and Type 3 incremental fit indices for the true-population model behave consistently at moderate or large sample sizes under the independence condition. It seems that when more information is used for deriving a fit-index value, the influence of weight matrices (and hence estimation methods) on the performance of incremental fit indices (e.g., Type 2 and Type 3 incremental fit indices) decreases. This is evident from our findings that Type 2 and Type 3 incremental fit indices behave much more consistently across estimation methods than Type 1 incremental fit indices. This is especially true when the sample size is large, the model is correctly specified, and the conditions for asymptotic robustness theory are satisfied. In addition, estimation method has no effect on GFI, AGFI, CAK, and CK derived from simple and complex true-population and misspecified models. Estimation method has no effect on CN under simple models, but it exerts small effect on CN under complex models when sample size is small, especially when there is dependence among latent variates. Estimation method has a relatively small effect on SRMR under both simple and complex true-population models, whereas it has a moderate-to-large effect on SRMR under both types of misspecified models. Thus, Sugawara and MacCallum's suggestion that nonincremental fit indices tend to behave much more consistently across estimation than do incremental fit indices is only partially supported, and the differential performance among three types of incremental fit indices need to be emphasized. Furthermore, the interaction effect between sample size and estimation method on the noncentrality-based absolute-fit indices (i.e., gamma hat, Mc, and RMSEA) seems to suggest that difference of weight matrices used for various estimation methods by itself does not provide sufficient rationale for explaining the inconsistent performance of various fit indices across estimation methods. One of the plausible explanations to this unexpected finding may be that the difference between a sample test statistic T and its degrees of freedom provides a biased estimate of the corresponding population noncentrality parameter when sample size is small.

The quality of models (degrees of model misspecification) seems to be related to the inconsistent performance of all fit indices, although this relationship is much less substantial for GFI, AGFI, CAK, and CK. In general, they tend to perform less consistently across estimation methods under the misspecified models than under the true-population model. All the fit indices behave more consistently across estimation methods under the true-population model than under the two misspecified models. In keeping with Sugawara and MacCallum's (1993) findings, the extent of consistent performance across estimation methods for the absolute-fit indices depends on the quality of models. One relevant and interesting question is how the extent of model misspecification may affect the performance of the noncentrality-based Type 3 incremental and absolute-fit indices. As suggested, a test statistic T can be approximated in large samples by the noncentral χ^2 (df, λ) distribution with true or not extremely misspecified models and distributional assumptions. It is likely that the degree of model misspecification will influence the performance of these noncentrality-based fit indices more than it will affect the other types of fit indices because of the violation of assumption underlying the noncentrality-based fit indices (i.e., they may not be distributed as a noncentral chi-square variate under extremely misspecified models). Future research needs to further address this issue.

The only important remaining issue is the cutoff value for these indices. Considering any model with a fit index above .9 as acceptable (Bentler & Bonett, 1980), and one with an index below this value as unacceptable, we have evaluated the rejection rates for most of the fit indices, except CAK, CK, CN, SRMR, and RMSEA. A cutoff value of 200 was used for CN (cf., Hoelter, 1983). A cutoff value of .05 was used for SRMR and RMSEA. Steiger (1989), Browne and Mels (1990), and Browne and Cudeck (1993) have recommended that values of RMSEA less than .05 be considered as indicative of close fit. Browne and Cudeck have also suggested that values in the range of .05 to .08 indicate fair fit and that values greater than .10 indicate poor fit. MacCallum, Browne, and Sugawara (1996) consider values in the range of .08 to .10 to indicate mediocre fit.

Although it is difficult to designate a specific cutoff value for each fit index because it does not work equally well with various types of fit indices, sample sizes, estimators, or distributions, our results suggest a cutoff value close to .95 for the ML-based TLI, BL89, CFI, RNI, and gamma hat; a cutoff value close to .90 for Mc; a cutoff value close to .08 for SRMR; and a cutoff value close to .06 for RMSEA, before one can conclude that there is a relatively good fit between the hypothesized model and the observed data. Furthermore, the proposed two-index presentation strategy

(i.e., the use of the ML-based SRMR, supplemented by either TLI, BL89, RNI, CFI, gamma hat, Mc, or RMSEA) and the proposed cutoff values for the recommended fit indices are required to reject reasonable proportions of various types of true-population and misspecified models. Finally, the ML-based TLI, Mc, and RMSEA tend to overreject true-population models at small sample sizes ($N \leq 250$), and are less preferable when sample size is small. Note that different cutoff values under various conditions (e.g., various sample sizes) are required for GLS- and ADF-based fit indices and, hence, no cutoff values for GLS- and ADF-based fit indices are recommended here. We present a detailed discussion on the selection of cutoff values for the ML-based fit indices elsewhere (Hu & Bentler, 1997, 1999).

Conclusion

Our study has several strengths. First, a wide variety of fit indices, including several new indices such as gamma hat, Mc, and RMSEA, were evaluated under various conditions, such as estimation method, distribution, and sample size, often encountered in practice. Second, we studied performance of fit indices under various types of correct and misspecified models. However, there are also limitations to this study. Although a misspecified model has often been defined by a nonzero noncentrality parameter (e.g., MacCallum et al., 1996; Satorra & Saris, 1985), the rationale for model selection or misspecification remains a weak link in any simulation study, in the absence of consensus on the definition of model misspecification or systematic study of models in the literature and their likely misspecification. In our view, parsimony is a separate issue, and we did not evaluate the performance of fit indices against this criterion. Some fit indices include penalty functions for nonparsimonious models (e.g., AGFI, TLI, CAK, CK, RMSEA), whereas others do not (e.g., NFI, GFI, and CFI). Finally, our study examined the performance of fit indices only under correct and underparameterized confirmatory factor models. Further work should be performed to explore the limits of generalizability in various ways, for example, across types of structural models and overparameterized models.

On the basis of the findings from previous studies and our Monte Carlo study, we identified several critical factors that may influence the adequacy of performance of fit indices. These factors include the degree of sensitivity to model misspecification, sample size, assumptions regarding the independence of latent variates, and estimation methods. Violation of multivariate normality assumption alone seems to exert less impact on the performance of fit indices. Like chi-square statistics, fit indices are measures of the overall model fit, but it is likely that one may acquire a very good overall fit of the model while one or more areas of local misspecification may remain. Thus, although our discussion has been focused on the issues regarding overall fit indices, consideration of other aspects such as the adequacy and interpretability of parameter estimates, model complexity, and many other issues remains critical in deciding on the validity of a model.

References

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

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

Anderson, J., & Gerbing, D. W. (1984). The effects of sampling error on convergence, improper solutions and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49,* 155–173.

Anderson, J., Gerbing, D. W., & Narayanan, A. (1985). A comparison of two alternate residual goodness-of-fit indices. *Journal of the Market Research Society, 24,* 283–291.

Anderson, T. W., & Amemiya, Y. (1988). The asymptotic normal distribution of estimators in factor analysis under general conditions. *Annals of Statistics, 16,* 759–771.

Bearden, W. D., Sharma, S., & Teel, J. E. (1982). Sample size effects on chi-square and other statistics used in evaluating causal models. *Journal of Marketing Research, 19,* 425–430.

Bentler, P. M. (1983). Some contributions to efficient statistics for structural models: Specification and estimation of moment structures. *Psychometrika, 48,* 493–571.

Bentler, P. M. (1989). *EQS structural equations program manual.* Los Angeles: BMDP Statistical Software.

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

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

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

Bentler, P. M., & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology, 47,* 541–570.

Bentler, P. M., & Wu, E. J. C. (1995a). *EQS for Macintosh user's guide.* Encino, CA: Multivariate Software.

Bentler, P. M., & Wu, E. J. C. (1995b). *EQS for Windows user's guide.* Encino, CA: Multivariate Software.

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

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Research and Methods, 17,* 303–316.

Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin, 107,* 256–259.

Bollen, K A., & Liang, J. (1988). Some properties of Hoelter's CN. *Sociological Research and Methods, 16,* 492–503.

Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal, 8,* 1–24.

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). Cambridge, England: Cambridge University Press.

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. (1987). Robustness of statistical inference in factor analysis and related models. *Biometrika, 74,* 375–384.

Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean- and covariance-structure models. In G. Arminger, C. C., Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for social and behavioral science* (pp. 185–249). New York: Plenum.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24,* 445–455.

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.

Browne, M. W., & Mels, G. (1990). *RAMONA user's guide.* Unpublished report, Department of Psychology, Ohio State University, Columbus.

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

Chou, C.-P., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. Hoyle (Ed.), *Structural equation modeling: Issues, concepts, and applications* (pp. 37–55). Newbury Park, CA: Sage.

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

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

de Leeuw, J. (1983). Models and methods for the analysis of correlation coefficients. *Journal of Econometrics, 22,* 113–137.

Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling, 2,* 119–144.

Dodd, D. H., & Schultz, R. F. (1973). Computational procedures for estimating magnitude of effect for some analysis of variance designs. *Psychological Bulletin, 79,* 391–395.

Dwyer, J. H. (1974). Analysis of variance and the magnitude of effects: A general approach. *Psychological Bulletin, 81,* 731–737.

Gerbing, D. W., & Anderson, J. C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 40–65). Newbury Park, CA: Sage.

Gierl, M. J., & Mulvenon, S. (1995). *Evaluation of the application of fit indices to structural equation models in educational research: A review of literature from 1990 through 1994.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Goffin, R. D. (1993). A comparison of two new indices for the assessment of fit of structural equation models. *Multivariate Behavioral Research, 28,* 205–214.

Hays, W. L. (1988). *Statistics.* New York: Holt, Rinehart & Winston.

Hoelter, J. W. (1983). The analysis of covariance structures: Goodness-of-fit indices. *Sociological Methods and Research, 11,* 325–344.

Hoyle, R. H. (Ed.). (1994). Structural equation modeling in clinical research [Special section]. *Journal of Consulting and Clinical Psychology, 62,* 427–521.

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In

R. H. Hoyle (Ed.), *Structural equation modeling: Issues, concepts, and applications* (pp. 76–99). Newbury Park, CA: Sage.

Hu, L., & Bentler, P. M. (1997). *Selecting cutoff criteria for fit indexes for model evaluation: Conventional criteria versus new alternatives* (Technical report). Santa Cruz, CA: University of California.

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

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

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Models, assumptions, and data.* Beverly Hills, CA: Sage.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183–202.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika, 43,* 443–477.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by the method of maximum likelihood.* Chicago: National Educational Resources.

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI user's guide* (3rd ed.). Mooresville, IN: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Hillsdale, NJ: Erlbaum.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences.* Pacific Grove, CA: Brooks/Cole.

La Du, T. J., & Tanaka, S. J. (1989). The influence of sample size, estimation method, and model specification on goodness-of-fit assessments in structural equation models. *Journal of Applied Psychology, 74,* 625–636.

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.

Maiti, S. S., & Mukherjee, B. N. (1991). Two new goodness-of-fit indices for covariance matrices with linear structures. *British Journal of Mathematical and Statistical Psychology, 28,* 205–214.

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* (pp. 315–353). Mahwah, NJ: Erlbaum.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: Effects of sample size. *Psychological Bulletin, 103,* 391–411.

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

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

Mooijaart, A., & Bentler, P. M. (1991). Robustness of normal theory statistics in structural equation models. *Statistica Neerlandica, 45,* 159–171.

Mulaik, S. A., James, L. R., Van Alstine, J., Bonnett, N., Lind, S., & Stillwell, C. D. (1989). An evaluation of goodness of fit indices for structural equation models. *Psychological Bulletin, 105,* 430–445.

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

SAS Institute. (1993). *SAS/STAT user's guide.* Cary, NC: Author.

Satorra, A., & Bentler, P. M. (1990). Model conditions for asymptotic robustness in the analysis of linear relations. *Computational Statistics & Data Analysis, 10,* 235–249.

Satorra, A., & Bentler, P. M. (1991). Goodness-of-fit test under IV estimation: Asymptotic robustness of a NT test statistic. In R. Gutierrez & M. J. Valderrama (Eds.), *Applied stochastic models and data analysis* (pp. 555–567). Singapore: World Scientific.

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

Sechrest, L., & Yeaton, W. H. (1982). Magnitudes of experimental effects in social science research. *Evaluation Review, 6,* 579–600.

Sobel, M. E., & Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. In N. B. Tuma (Ed.), *Sociological methodology* (pp. 152–178). San Francisco: Jossey-Bass.

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

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

Sugawara H. M., & MacCallum, R. C. (1993). Effect of estimation method on incremental fit indexes for covariance structure models. *Applied Psychological Measurement, 17,* 365–377.

Tanaka, J. S. (1987). How big is big enough? Sample size

and goodness of fit in structural equation models with latent variables. *Child Development, 58,* 134–146.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structure equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Tanaka, J. S., & Huba, G. J. (1985). A fit index for covariance structural models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology, 42,* 233–239.

Tanaka, J. S., & Huba, G. J. (1989). A general coefficient of determination for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology, 42,* 233–239.

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

Vaughan, G. M., & Corballis, M. C. (1969). Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin, 72,* 204–213.

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

Yuan, K.-H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association, 92,* 766–773.

Received June 14, 1995
Revision received December 19, 1997
Accepted March 6, 1998 ■