STRUCTURAL EQUATION MODELING, *14*(3), 361–390

The Model-Size Effect on Traditional and Modified Tests of Covariance Structures

Walter Herzog
University of St. Gallen, Switzerland

Anne Boomsma
University of Groningen, The Netherlands

Sven Reinecke
University of St. Gallen, Switzerland

According to Kenny and McCoach (2003), chi-square tests of structural equation models produce inflated Type I error rates when the degrees of freedom increase. So far, the amount of this bias in large models has not been quantified. In a Monte Carlo study of confirmatory factor models with a range of 48 to 960 degrees of freedom it was found that the traditional maximum likelihood ratio statistic, T_{ML}, overestimates nominal Type I error rates up to 70% under conditions of multivariate normality. Some alternative statistics for the correction of model-size effects were also investigated: the scaled Satorra–Bentler statistic, T_{SC}; the adjusted Satorra–Bentler statistic, T_{AD} (Satorra & Bentler, 1988, 1994); corresponding Bartlett corrections, T_{MLb}, T_{SCb}, and T_{ADb} (Bartlett, 1950); and corresponding Swain corrections, T_{MLs}, T_{SCs}, and T_{ADs} (Swain, 1975). The empirical findings indicate that the model test statistic T_{MLs} should be applied when large structural equation models are analyzed and the observed variables have (approximately) a multivariate normal distribution.

In the practice of structural equation modeling (SEM) one can observe that an increasing number of large models are estimated; that is, models with lots of indicators and latent variables, and consequently in most cases many degrees of

Correspondence should be addressed to Walter Herzog, Institute of Marketing and Retailing, Dufourstrasse 40a, CH-9000 St. Gallen, Switzerland. E-mail: walter.herzog@unisg.ch

freedom. This may raise a number of problems. First, it is not always possible and it is often too expensive to get large sample sizes needed to estimate such big models. Second, the distribution of the large number of observed variables involved can rarely be approximated by a multivariate normal density. Third, the combination of large models, relatively small sample sizes, and nonnormal data appears to be accountable for the inflated Type I error rates of the traditional maximum likelihood ratio test statistic, T_{ML}, for global model fit (see, e.g., Hoogland, 1999). The apparent consequence—which can be verified from the literature—is that in applied SEM, researchers increasingly rely on alternative fit measures rather than T_{ML}. Decisions and conclusions regarding model fit are frequently based on more popular statistics and fit indexes, applying partly subjective cutoff criteria. A brief outline of the goals of our study follows.

It is argued that the effect of model size, measured by the number of degrees of freedom d (cf. Kenny & McCoach, 2003), and its interaction with sample size requires more attention in applied research, because (a) the model-size effect makes investigators more reluctant to report p values of model fit statistics in their studies—even if of no single use—and (b) other popular statistics (e.g., the Tucker–Lewis index [TLI], and the root mean square error of approximation [RMSEA]) are affected by the inflated values of T_{ML} as well. Because relatively little is known about the effects of model size on familiar model test statistics, the first aim of our study is to quantify the impact of large model size on the finite sampling distribution of T_{ML} in SEM. In general, for the evaluation of model-size effects on model test statistics Type I error rates are of specific, although not of single importance.

Although not very obvious at first glance, a family of chi-square corrections introduced by Satorra and Bentler (1988, 1994) might be one promising approach to handle the model-size effect. Two of them are the *scaled* (mean-corrected) statistic, T_{SC}, and the *adjusted* (mean- and variance-corrected) statistic, T_{AD} (Satorra & Bentler, 1994, p. 407f), based on theoretical work by Bartlett (1937) and Satterthwaite (1941), respectively, and a classical paper by Box (1954). It is well known that these corrections have first and foremost been developed to make T_{ML} robust against effects of nonnormality. It should be noted, however, that Satorra and Bentler (2001) suggested (in their abstract) that their corrections might also work for small samples and large models, relative to distribution-free estimation methods, that is. In addition, the studies by Fouladi (2000) and Nevitt and Hancock (2004) provided empirical evidence that, relative to T_{ML}, these corrections might also improve small-sample performance even when the normality assumption is not violated at all. As large models need large sample sizes for the asymptotic properties of test statistics to hold (Muthén, 1993, p. 228), it is reasonable to assume that these statistics will also perform well in large models. Unfortunately, little is known about the finite-sample behavior of T_{SC} and T_{AD} in large models and about the interaction of sample-size and model-size

effects. Therefore, our second aim is to check whether it is beneficial (focusing on Type I error rates as well as on complete distribution functions) to favor T_{SC} or T_{AD} over T_{ML} for the test of large models even under conditions of multivariate normality. In this study we do not consider analyses of nonnormal data because, as a baseline, a detailed investigation of the effect of increasing d under the normality assumption is needed first. Once more, we included the Satorra–Bentler statistics in our research design, not because of their well-known performance for the nonnormal case (e.g., Hu, Bentler, & Kano, 1992), but because they seem to be promising for correcting model-size effects under normality conditions as well.

Another straightforward approach to attack the problem of model size is to compute the corresponding Bartlett corrections of the three model fit statistics, T_{MLb}, T_{SCb}, and T_{ADb}, as proposed by Fouladi (2000) and more recently by Nevitt and Hancock (2004). Although Bartlett (1950) developed his type of corrections for exploratory factor modeling, these researchers found an acceptable performance under conditions of small sample size for general SEM as well. Because of the dependency of sample-size requirements on model size, as mentioned earlier, it is expected that these corrections might also work in large models. Because their behavior in large models is not precisely known, it is investigated whether these statistics turn out to be adequate corrections of model-size effects. Hence, our third aim is to investigate the Type I error rates produced by T_{MLb}, T_{SCb}, and T_{ADb}, and to compare them to those of T_{ML}, T_{SC}, and T_{AD}, respectively, in large models under conditions of multivariate normality.

A less well-known correction of T_{ML} has been developed by Swain (1975). According to Browne (1982), this approach "seem[s] to result in an improvement of the approximation of the chi-squared distribution" (p. 98). With the exception of the Monte Carlo study by Fouladi (2000), to our knowledge the finite-sample behavior of this statistic is undocumented. Fouladi found a good performance of the statistic, especially for small sample sizes. For similar reasons as for the Bartlett corrections, it could be claimed that the corresponding Swain corrections T_{MLs}, T_{SCs}, and T_{ADs} might yield better Type I error rates compared to those of T_{ML}, T_{SC}, and T_{AD}. Therefore, the fourth aim of this study is to investigate the performance of the Swain corrections in large models under multivariate normality.

In summary, the purpose of our study is (a) to investigate the bias in Type I error rates produced by T_{ML}; (b) to compare the results of T_{ML} with those of T_{SC} and T_{AD}; (c) to evaluate the performance of T_{MLb}, T_{SCb}, and T_{ADb}; and (d) to check whether the behavior of T_{MLs}, T_{SCs}, and T_{ADs} is appropriate for testing covariance structure models with many degrees of freedom when multivariate normality assumptions hold.

Before we turn to the next section, it is emphasized that a careful investigation of T_{ML}, T_{SC}, and T_{AD} in large models was demanded by several researchers (e.g., Hoogland, 1999; Kenny & McCoach, 2003; Muthén, 1993, p. 228; Muthén & Satorra, 1995). To our present knowledge, no systematic Monte Carlo study of the behavior of chi-square statistics in very large models exists, although the investigation of such models "will probably result in findings that are more disappointing regarding the chi-square statistic" (Hoogland, 1999, p. 51). As indicated before, an exception is a study on some fit measures (RMSEA, TLI, and the comparative fit index [CFI]) by Kenny and McCoach (2003). Two remarks on this first investigation of the behavior of fit statistics in large models can be made. First, the study aimed at two measures (CFI and TLI) with rather subjective cutoff criteria for model fit evaluation, not at the regular chi-square statistic for overall model fit. Second, in applied research, model decision criteria for the RMSEA are mainly based on practical experience (Browne & Cudeck, 1992, p. 239), which is not undisputable: Jöreskog (2005) favored a p value for the test of close fit associated with the RMSEA of at least 0.50.

The article is structured as follows. First, the test statistics under study are defined and the corresponding asymptotic theory is presented briefly. Second, research hypotheses are developed based on findings of previous simulation studies; that is, expectations regarding the behavior of the test statistics under study are formulated. Third, based on results from a Monte Carlo research design, the expectations are tested and consequences for applied research are deduced. The practical implications of our findings are further exemplified by correcting the fit of a large structural equation model that was published recently. Finally, some limitations of this study and directions of future research are briefly mentioned.

TEST STATISTICS AND THEIR ASYMPTOTIC DISTRIBUTION

In this section, all test statistics under study are defined and the asymptotic theory underlying their distribution is summarized.

Likelihood Ratio Statistic

Consider p random variables $\mathbf{z}$ $(p \times 1)$ with an empirical sample covariance matrix $\mathbf{S}$ $(p \times p)$ based on $N = n + 1$ independent observations, and a population model of underlying relations among these variables with covariance structure $\mathbf{\Sigma}(\boldsymbol{\theta})$ $(p \times p)$, where $\boldsymbol{\theta}$ $(t \times 1)$ is the vector of independent model parameters to be estimated. If the observed variables $\mathbf{z}$ follow a multivariate normal distribution, the sample covariance matrix $\mathbf{S}$ based on independently and

identically distributed observations has a Wishart distribution (Anderson, 1958). The maximization of the corresponding log-likelihood function, conditional on the sample covariance matrix $\mathbf{S}$, is equivalent to minimizing the function

$$F_{ML}[\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})] = \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \operatorname{tr}[\mathbf{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}] - \log|\mathbf{S}| - p, \tag{1}$$

which is a discrepancy function as defined by Browne (1984, p. 64); log denotes the natural logarithm here. The parameter vector $\hat{\boldsymbol{\theta}}$, defining the minimum of $F_{ML}[\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})]$, contains the so-called maximum likelihood estimates of $\boldsymbol{\theta}$. Asymptotically, as N goes to infinity, the maximum likelihood estimates are normally distributed with expectation vector $E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$, and asymptotic covariance matrix $\operatorname{acov}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}') = \boldsymbol{I}^{-1}(\boldsymbol{\theta})$, the inverted Fisher information matrix of order $(t \times t)$, which can be estimated (cf. Bollen, 1989, p. 109), yielding estimates of the standard errors of the t parameter estimates as well as estimated covariances between those parameter estimates.

Let $\boldsymbol{\Sigma}$ $(p \times p)$ denote the population covariance matrix of the p observed variables $\mathbf{z}$, $\boldsymbol{\Sigma}(\boldsymbol{\theta}_j)$ the population covariance matrix implied by a postulated model M_j, and let c be an "irrelevant constant" (Bollen, 1989, p. 263). One can then test the null hypothesis $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$; that is, that the postulated model holds, with the corresponding log-likelihood function, evaluated at $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}_0$,

$$\log L_0 = \log L[\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_0); \mathbf{S}] = -\frac{n}{2}\{\log|\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_0)| + \operatorname{tr}[\mathbf{S}\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}}_0)]\} + \log c, \tag{2}$$

against the alternative hypothesis $H_1 : \boldsymbol{\Sigma} = \boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is any positive definite matrix, and by definition $n = N - 1$. If $\boldsymbol{\Omega}$ is set equal to the sample covariance matrix $\mathbf{S}$, it follows that the log-likelihood function under H_1 can be written as

$$\begin{aligned} \log L_1 = \log L(\boldsymbol{\Omega}; \mathbf{S}) &= -\frac{n}{2}[\log|\mathbf{S}| + \operatorname{tr}(\mathbf{S}\mathbf{S}^{-1})] + \log c \\ &= -\frac{n}{2}(\log|\mathbf{S}| + p) + \log c \end{aligned} \tag{3}$$

(for details, see, e.g., Anderson, 1958; Bollen, 1989, p. 263ff.). It can then be shown that under H_0, the distribution of the likelihood ratio statistic, defined as

$$T_{ML} \equiv -2\log\frac{L_0}{L_1} = -2\log\frac{L[\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_0); \mathbf{S}]}{L(\boldsymbol{\Omega}; \mathbf{S})} = nF_{ML}[\mathbf{S}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_0)], \tag{4}$$

converges with increasing sample size $N = n + 1$ to a chi-square distribution with $d = p(p+1)/2 - t$ degrees of freedom (Wilks, 1938); the likelihood criterion $\lambda = L_0/L_1$ in Equation 4 was introduced by Neyman and Pearson (1928). From Equations 1 and 4 it follows that the likelihood ratio test statistic, T_{ML},

is by definition n times the minimum of the maximum likelihood discrepancy function evaluated at $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}_0$. Hence, the likelihood ratio test statistic can be used to test whether the proposed model $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ is implausible at a given level of significance. In practice, the behavior of this statistic depends, of course, on its robustness against violations of underlying assumptions (independent observations, multivariate normality with covariance structure $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$, and a large sample size, mainly).

Satorra–Bentler Statistics

Because nonnormal data are very common in practice, Satorra and Bentler (1988, 1994) introduced two corrections to a family of model test statistics, aimed to yield distributional behavior that more closely follows the chi-square reference distribution that is used in structural equation model testing. Relative to distribution-free methods, these statistics can be useful when the sample size is small or the estimated model is large (Satorra & Bentler, 2001, p. 507). The corrections can, in principle, be applied to a family of test statistics, including the normal theory weighted least square model test statistic, $T_{WLS_{\mathcal{N}}}$, as it is used in the LISREL program (see Jöreskog, Sörbom, Du Toit, & Du Toit, 2001, Appendix A). In this study, we only apply it to T_{ML}.

The mean-corrected, *scaled* statistic (Satorra & Bentler, 1988, 1994, p. 407) is defined as

$$T_{SC} \equiv \frac{d}{\mathrm{tr}(\mathbf{A})} T_{ML}, \tag{5}$$

where matrix $\mathbf{A}$ is a slightly complicated function of a matrix of first-order derivatives of the ML-discrepancy function to the parameters to be estimated and an estimate of the asymptotic covariance matrix of sample covariances (cf. Muthén, 2004, Equation 105). If the distribution of $\mathbf{z}$ is elliptical, the scaling factor $d/\mathrm{tr}(\mathbf{A})$ in Equation 5 provides an estimate of the common relative kurtosis of $\mathbf{z}$ (Satorra & Bentler, 1994, p. 407), which implies a correction for nonnormality.

As usual, the test statistic T_{SC} is evaluated as having (approximately) a chi-square distribution with $d = p(p+1)/2 - t$ degrees of freedom. For certain distributions of the observed variables, for example, elliptical ones, the asymptotic distribution of T_{SC} is exactly chi-square with d degrees of freedom. In principle, however, the correction of T_{ML} involves a scaling to the correct mean, so that for general distributions asymptotically the first moment of the distribution of T_{SC} is matched to the number of degrees of freedom d. Under conditions of multivariate normality, T_{SC} has asymptotically an exact chi-square distribution with d degrees of freedom, because a multivariate normal density is also elliptical.

Furthermore, Satorra and Bentler (1988, 1994, p. 408) used a procedure developed by Satterthwaite (1941, 1946) to correct not only for the mean but for the variance of T_{ML} as well. This is possible by an adjustment of the number of degrees of freedom to d', which is the integer closest to a function of the matrix $\mathbf{A}$ (cf. Muthén, 2004, Equation 110): by definition

$$d' = \texttt{int}\left\{\frac{[\text{tr}(\mathbf{A})]^2}{\text{tr}(\mathbf{A}^2)}\right\}. \tag{6}$$

It should be noted that the value of d' may vary from sample to sample. Substituting d' for d in Equation 5, we get (cf. Muthén, 2004, Equation 108):

$$T_{AD} \equiv \frac{d'}{\text{tr}(\mathbf{A})}\, T_{ML}, \tag{7}$$

which is the *adjusted* chi-square test statistic; adjusted for mean and variance that is.

Again, for general distributions of observed variables, T_{AD} has asymptotically not an exact chi-square distribution with d' degrees of freedom, but it matches the first- and second-order moment of that distribution (Satorra & Bentler, 1994, p. 408). For multivariate normal observations, T_{AD} has asymptotically an exact chi-square distribution with d' degrees of freedom.

It should be stressed that if distributional assumptions or conditions for asymptotic robustness hold, both corrections of T_{ML} discussed in this section are "automatically inactive (asymptotically)" (Satorra & Bentler, 1994, p. 414). Notice, however, the adverb in parentheses: *asymptotically*. It has to be reemphasized, that T_{ML} also follows a chi-square distribution only asymptotically.

Bartlett-Corrected Statistics

For exploratory factor analysis models (more specifically, for principal components models) Bartlett (1950, 1954) developed a correction of the chi-square test statistic for small sample sizes. In general, Bartlett's correction consists of multiplying $-2\log\lambda = n\ F_{ML}[\mathbf{S}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_0)]$, where λ is the likelihood ratio criterion of Neyman and Pearson (1928), by a scale factor that results in a statistic having the same moments as χ^2, ignoring quantities of order n^{-2} (cf. Lawley, 1956). As pointed out by Lawley (1956), this scaling device was first employed by Bartlett (1937).

From Equation 9, it can be seen that Bartlett's correction for unrestricted factor models is a function of the number of latent variables k, the number

of observed variables p, and the sample size $N = n + 1$. Fouladi (2000) and Nevitt and Hancock (2004) studied the Bartlett correction for the analysis of general structural equation models, and applied it to the three model test statistics discussed so far, T_{ML}, T_{SC}, and T_{AD}. The corresponding Bartlett corrections for these statistics are defined as

$$T_{MLb} \equiv b\, T_{ML}\ , T_{SCb} \equiv b\, T_{SC}\ ,\ \text{and}\ T_{ADb} \equiv b\, T_{AD}, \tag{8}$$

respectively, where

$$b = 1 - \frac{4k + 2p + 5}{6n}. \tag{9}$$

It follows from Equations 8 and 9 that asymptotically the distribution of the Bartlett-corrected statistics matches the asymptotic distributions of T_{ML}, T_{SC}, and T_{AD}, respectively. The specific form of Equation 9 was derived by Bartlett (1950, Equation 3) from expansion of a moment generating function. Independently, Box (1949) derived approximations of chi-square statistics for tests on correlation matrices identical to those of Bartlett.

Swain-Corrected Statistics

As we have emphasized, the Bartlett correction in Equation 9 is the appropriate small-sample correction for exploratory or unrestricted factor models only. For general covariance structure models, Bartlett's correction is strictly speaking not appropriate. In fact, for each class of models a specific multiplier or correction factor would be needed. Because this is quite troublesome for applied researchers, Swain (1975) developed four small-sample corrections of T_{ML} for general covariance structure models. We only study the one that seemed most promising among those four; see also Browne (1982, p. 98), who claimed that Swain used "heuristic arguments" in proposing these correction factors. It should be noted in advance that Swain (1975) is very cautious about the applicability of the corrections he proposed: "For any particular model the worth of the forms suggested [correction factors of the form $1 - k_1/n + O(n^{-2})$, where k_1 is a function of p and d] would, of course, have to be carefully evaluated before routine application" (p. 78).

From their basic derivations it is clear that both Bartlett and Swain corrections should be considered as multiplying or scale factors of $nF_{ML}[\mathbf{S}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_0)]$, not as multipliers of just the discrepancy function $F_{ML}[\mathbf{S}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_0)]$. Hence, it would be improper to suggest that these corrections can or should be interpreted as a modification of just the sample size.

For the special case of maximum likelihood estimation of structural equation models that are invariant under a constant scaling factor (cf. Browne, 1982, p. 77), the most promising small-sample correction of T_{ML} introduced by Swain (1975) is defined as

$$s = 1 - \frac{p(2p^2 + 3p - 1) - q(2q^2 + 3q - 1)}{12dn}, \tag{10}$$

where

$$q = \frac{\sqrt{1 + 4p(p + 1) - 8d} - 1}{2}, \tag{11}$$

p is the number of observed variables, d is the number of degrees of freedom, and $N = n + 1$ is the sample size, as before. Equations 10 and 11 correspond to Swain's (1975) Equations 4.14 and 4.10. The Swain corrections for the three test statistics T_{ML}, T_{SC}, and T_{AD} are now, respectively, defined as

$$T_{MLs} \equiv s\,T_{ML}\ , T_{SCs} \equiv s\,T_{SC}\ ,\ \text{and}\ T_{ADs} \equiv s\,T_{AD}. \tag{12}$$

From Equation 10 it can be seen that Swain's correction is a function of p, d, and N. Because $d = p(p+1)/2-t$, Equations 10 and 11 can also be written as a function of t instead of d, along with p and N, of course (cf. Browne, 1982, p. 98).

It follows from Equations 10 and 12 that asymptotically the distributions of the Swain-corrected statistics match those of T_{ML}, T_{SC}, and T_{AD}, respectively.

EXPECTATIONS OF FINITE SAMPLE BEHAVIOR

In this section we discuss the expected finite sample performance of the nine statistics for global model fit in large models, T_{ML}, T_{SC}, T_{AD}, T_{MLb}, T_{SCb}, T_{ADb}, T_{MLs}, T_{SCs}, and T_{ADs}, as defined previously. Statistical theory does not yield clear guidelines as to the choice among these statistics, nor does it help unequivocally to come up with proper, theory-based expectations about the issue under investigation (cf. Bentler & Yuan, 1999). In our case, the design of the study has two main factors, model size and sample size: The number of latent variables in the factor models ranges from 4 to 16, with three indicators for each latent variable, and the sample sizes are 200, 400, and 800 (details of the design are reported in the next section). In general it can be expected that the behavior of the model test statistics will improve with increasing sample size (consistent estimators, the functioning of asymptotic theory) for any given model size.

Generally, it is also expected that the statistics will show improved behavior with decreasing model size for a given sample size. There exists empirical evidence and arguments for this claim. First, the results of a meta-analysis by Hoogland (1999, section 3.3) show that the performance of the chi-square model statistics improves with a decreasing number of degrees of freedom d. Second, there are several rules of thumb in the literature indicating that one might need a specific minimal number of observations for each observed variable or for each model parameter to be estimated. Such recommendations suggest that if the number of observed or latent variables increases, more observations are needed to obtain proper estimates. As to the comparison of the test statistics under study, statistical theory is not providing solid predictions for their finite sample behavior, but in most cases it is possible to contrive expectations about the results of our investigations from the findings of previous simulation studies.

Likelihood Ratio Statistic

Under conditions of multivariate normality, for test statistic T_{ML} Hoogland (1999) found a trend to an overrejection of true models for $N < 400$, and this tendency increased as models got larger. This finding is supported by other simulation studies with various designs (Curran, Bollen, Paxton, Kirby, & Chen, 2002; Hau & Marsh, 2004; Kenny & McCoach, 2003; Marsh, Hau, Balla, & Grayson, 1998). We therefore expect that the empirical rejection rates will be inflated more or less seriously for very large models.

Scaled Satorra–Bentler Statistic

The studies by Hu, Bentler, and Kano (1992), Curran, West, and Finch (1996), Bentler and Yuan (1999), Hoogland (1999), Nevitt and Hancock (2001), and Hau and Marsh (2004) revealed that the test statistic T_{SC} produces even higher rejection rates than T_{ML} when multivariate normal variables are analyzed, and this liberal tendency increased with model size as well. Therefore, we expect that T_{SC} will perform worse than T_{ML} in large models under conditions of normality. The explanation for this expected tendency could very well be that T_{SC} requires the estimation of the asymptotic covariance matrix of sample covariances, which involves estimation of fourth-order moments and the computation of the inverse of often huge matrices.

Adjusted Satorra–Bentler Statistic

There is not a great deal of information about the finite sample behavior of T_{AD} in the literature. In a recent Monte Carlo investigation, Asparouhov (2005) found

the adjusted chi-square statistic to have excellent Type I error rates compared to T_{ML} and T_{SC}. Fouladi (2000) conducted an extensive simulation study with 12 different test statistics and found T_{AD} to outperform all other statistics with respect to Type I error rate "under more general nonnormal distributional conditions" (p. 400; cf. p. 371, Table 1). She concluded that T_{AD} "shows the most rapid convergence to the nominal level and as such can be used with smaller samples than the other procedures" (p. 401). We therefore expect that T_{AD} will outperform T_{ML} and T_{SC} in large models.

Bartlett-Corrected Statistics

Fouladi (1999, 2000) and Nevitt and Hancock (2004) examined the performance of Bartlett corrections in the context of SEM. The results of Nevitt and Hancock, in particular, indicate that T_{MLb}, T_{SCb}, and T_{ADb} tend to underestimate the nominal levels when N decreases and when d increases. Based on this finding, it is reasonable to expect that the Bartlett corrections will clearly underestimate the nominal error levels, when the model to be analyzed is larger than the models studied by Nevitt and Hancock (2004), which ranged between $d = 85$ and $d = 196$.

Swain-Corrected Statistics

To our knowledge, the only study on the Swain correction is the Monte Carlo investigation by Fouladi (2000). For the analysis of covariance structures, she found that "the normal theory procedures with the best small sample Type I error control under conditions of extremely mild distributional nonnormality were [$\cdots$] the 0-factor Bartlett rescaling or Swain rescaling of the standard ML covariance structure analysis test statistic" (p. 400). Unfortunately, she only investigated very small models with no more than 12 variables. However, as discussed earlier in the introductory section, it seems legitimate to expect an improved performance of the Swain statistics compared to T_{ML} in large models because of its favorable small-sample properties.

Summary

In summary, it is expected that T_{AD} will perform better than T_{ML}, and that T_{ML} will be more accurate than T_{SC} for large models under conditions of multivariate normality. We do not have much information about the Bartlett and the Swain statistics, but it seems reasonable to expect an improved performance compared to T_{ML} when the number of degrees of freedom increases.

Although we formulated expectations based on empirical findings from the literature mainly, our study has a partly explorative character. Where appropriate,

published results are revalidated by our investigations, but we seek to elaborate and to generalize them to large structural equation models.

MONTE CARLO DESIGN

Sample Size Conditions

Sample sizes of 200, 400, and 800 are used. It can be problematic to investigate sample sizes of $N < 200$ because it is well known that estimates of parameters and standard errors may be biased seriously. Also, nonconvergence problems and Heywood cases are more likely to occur for such small sample sizes (Boomsma, 1982, pp. 171, 1985; Boomsma & Hoogland, 2001). In practice, getting more observations than 800 is not always possible or too expensive.

Population Models and Model Size

Most Monte Carlo studies reported in the literature examined very small population models; see, for example, Asparouhov (2005) and Fouladi (2000). As for the factor models in Hoogland's (1999) meta-analysis, d ranged from 2 to 98. For our study, it was decided to restrict the population models to confirmatory factor analysis (CFA) models, because in practice these measurement models are most widely applied.

In general, a factor model without an intercept term is defined as $\mathbf{z} = \mathbf{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta}$, where $\mathbf{z}$ $(p \times 1)$ is a vector of observed variables, $\mathbf{\Lambda}$ $(p \times k)$ a matrix of factor loadings on k common factors $\xi_1, \xi_2, \ldots, \xi_k$, and $\boldsymbol{\delta}$ $(p \times 1)$ a vector with unique scores (measurement error), where $E(\boldsymbol{\xi}) = \mathbf{0}$, $E(\boldsymbol{\delta}) = \mathbf{0}$ and $\boldsymbol{\delta}$ is uncorrelated with $\boldsymbol{\xi}$. Under the usual assumptions, the population covariance matrix of $\mathbf{z}$ has the form $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{\Psi}$, where $\mathbf{\Phi} = E(\boldsymbol{\xi}\boldsymbol{\xi}')$, and $\mathbf{\Psi} = E(\boldsymbol{\delta}\boldsymbol{\delta}')$ is a diagonal matrix with unique score or error variances.

To study a variety of model sizes, the number of factors k was set at 4, 6, 8, 10, 12, 14, and 16. Each factor has three indicators, so the number of observed variables p ranges from 12 to 48. To achieve identifiable models, the variance of each latent construct was fixed to the value of one. Furthermore, the population factor loadings were set to 0.70 and the error variance to 0.51 for each indicator. The correlation between each pair of factors was set to 0.30. Table 1 gives an overview of characteristics of the seven factor models.

Number of Replications

A total number of $NR = 1{,}200$ replications was used. Although 300 replications would have been a "reasonable trade off between precision, and the amount of

TABLE 1
Overview of Factor Models of the Monte Carlo Design and Seed Values for Data Generation

					Seed		
k	p	p^*	t	d	$N = 200$	$N = 400$	$N = 800$
4	12	78	30	48	77703570	49330350	71578326
6	18	171	51	120	83444508	39023988	68738111
8	24	300	76	224	16159776	44724671	97116941
10	30	465	105	360	71034416	06466931	85864123
12	36	666	138	528	56460497	36267030	98682926
14	42	903	175	728	64459199	07380304	07013316
16	48	1176	216	960	48795874	79583898	23965379

Note. k is the number of factors; $p = 3k$ the number of observed variables; $p^* = p(p+1)/2$ the number of independent elements of $\mathbf{S}$; t the number of parameters to be estimated; $d = p^* - t$ the number of degrees of freedom.

information to be handled" (Hoogland, 1999, p. 59), it was decided to use four times as many replications to lower the standard error of percentages presented in Tables 2, 3, and 4 (see next section). For example, under the null hypothesis that the nominal value of a 5% significance level holds, the standard error of the percentages reported in the cells of these tables equals 0.629%, where it would have been twice as large if only 300 replications had been used.

Data Generation and Model Estimation

Multinormal variables were generated to isolate the effect of model size (and sample size) on the test statistics, and to set a normal baseline for comparison with nonnormal data in future research. The population covariance matrix of these normal variables is defined by the population factor structure of the models under study: $\mathbf{\Sigma}(\boldsymbol{\theta}_j)$, $j = 1, 2, \ldots, 7$. Both the generation of the sample data and the estimation of the models was performed using the M*plus* software program (Version 3.11; Muthén & Muthén, 2004). The seed values for the pseudo-random draws of samples from the multivariate normal population distributions for each cell in the design are listed in Table 1. The starting values for the model parameter estimates were fixed at their population values.

The factor models were estimated using the primary estimation setting of maximum likelihood (ML) in M*plus*. For the mean-adjusted and mean- and variance-adjusted estimation of the chi-square statistic, the estimation option in M*plus* was MLM and MLMV, respectively, which are both maximum likelihood

procedures. For the statistical analyses of the generated model estimates, R software (Version 2.1.1) was used (see, e.g., Venables & Smith, 2005).

Statistics

The sampling distributions of the nine test statistics based on the 1,200 replications were observed. First, the empirical rejection rates on the 5% Type I error level were inspected. A tolerable rejection rate is defined here as one that falls in the two-sided 99% adjusted Wald confidence interval estimate, calculated as [3.5, 6.8]; see Agresti and Coull (1998). If the observed rejection rate falls outside this interval, it is concluded that the population rejection rate differs from 0.05; that is, rejecting the null hypothesis that the population rejection rate equals 0.05, using a 1% significance level. A 99% interval estimate was chosen because of the large number of replications, hence slightly reducing the power of the test compared to a 95% interval estimate.

Second, by means of a one-sample Kolmogorov–Smirnov test (e.g., Birnbaum, 1952) it was tested at a 1% significance level whether the empirical sampling distributions of the fit statistics follow the proper theoretical chi-square distribution. Because the value of the number of degrees of freedom for AD-based test statistics varies over sample covariance matrices, the rounded mean value over 1,200 replications was used as the number of degrees of freedom of the theoretical chi-square distribution. In Tables 2 through 7, this rounded mean value is shown in brackets in column 12; in all cases it was equal to the median value of d'. In addition, selected PP and QQ plots (percentile-percentile and quantile-quantile plots), were used to illustrate the findings, so as to provide a visual reply to the question: How do the deviations from the theoretical chi-square distributions look?

Information about the discrepancies between empirical and theoretical distributions of test statistics, by means of both Kolmogorov–Smirnov tests and PP and QQ plots, is reported here for two reasons. First, 5% Type I error rates are quite arbitrary; sometimes 1% or 10% significance levels might be preferred. Second, in applied research p values of estimated model fit statistics are reported quite often, especially if in favor of the postulated model. If we had confined ourselves to rejection rate behavior at a 5% significance level, not only would it be difficult to generalize results to other significance levels, but also, and more important, no information about the empirical distribution function of the statistics as compared to the theoretical chi-square distribution would have been obtained.

In the statistical analyses, all 1,200 replications were used for all cells in the design, because no convergence problems and no improper solutions occurred in model estimation.

FINDINGS AND RECOMMENDATIONS

In this section, we first focus on the empirical rejection rates of the nine test statistics for model fit and compare them with the rejection rates predicted by asymptotic theory. Second, the sampling distributions of the test statistics are compared to the theoretical chi-square distributions by means of a one-sample Kolmogorov–Smirnov test. Third, the findings are further visualized by means of PP and QQ plots of the empirical sampling distributions of the test statistics. Finally, based on the results of these analyses, recommendations are formulated for the use of appropriate model test statistics in applied research when large models are at stake. In addition, the implications of our findings are briefly illustrated by correcting the fit of a recently published applied model.

Type I Error Rates

The empirical rejection rates were computed across the 1,200 replications. The *differences* of these rejection rates to the nominal 5% value are summarized in Table 2 ($N = 200$), Table 3 ($N = 400$), and Table 4 ($N = 800$). Values larger than zero indicate that the population model is rejected too frequently, whereas values smaller than zero indicate that the corresponding statistic is too conservative. The boldfaced numbers in these tables indicate acceptable rejection rates, for nominal $\alpha = 0.05$ defined as $\hat{\alpha} \in [0.035, 0.068]$, implying that acceptable *difference rates* in the tables are within the range $[-1.5\%, +1.8\%]$.

Likelihood ratio statistic. The quantile bias of this statistic reduces with increasing sample size and decreasing model size. It can be seen that T_{ML} performs extremely badly. In fact, the rejection rate is not acceptable for all model sizes for a sample size of $N = 200$ and $N = 400$. This latter finding is in line with research findings of Boomsma (1983, Table 4.4.16, Model 4CM), who analyzed a very similar model. The amount of this bias is considerable: For the largest model with $d = 960$ and $N = 200$ the progressive bias is 70.7%. Furthermore, the performance is not even acceptable for $N = 800$ when models with six or more factors are analyzed.

As a consequence of these findings, it is not recommendable to employ T_{ML} for the test of large models. Although the effect of increasing degrees of freedom has been reported frequently, the amount of the bias detected here is quite alarming. The effect of increasing degrees of freedom seems to be comparable to the effect of testing models with nonnormal variables. Curran et al. (1996), for example, reported empirical rejection rates of 48% for the nominal 5% Type I error rate when severely nonnormal variables (univariate kurtoses of 21.0 and skewnesses of 3.0) were analyzed (Curran et al., 1996, p. 22, Table 1). The rejection rate bias in our study is similar to the bias reported by these authors.

TABLE 2
Empirical Minus the 5% Nominal Type I Error Rates of Nine Model Fit Statistics for $N = 200$ ($NR = 1{,}200$)

k	T_{ML}	T_{SC}	T_{AD}	T_{MLb}	T_{SCb}	T_{ADb}	T_{MLs}	T_{SCs}	T_{ADs}	d ($\overline{d}'$)[a]	$N{:}t$
4	3.2	3.8	**1.1**	**.3**	**1.4**	**−1.0**	**1.4**	2.0	**−.2**	48 (36)	6.7
6	4.9	6.3	**−.6**	**−.8**	**−.5**	−3.5	**.4**	**1.2**	−3.1	120 (69)	3.9
8	9.7	13.2	**−.5**	−1.7	**−.7**	−4.6	**.8**	2.7	−3.5	224 (98)	2.6
10	20.3	24.9	**−.5**	−2.9	−1.7	−4.7	**.8**	3.2	−4.4	360 (120)	1.9
12	33.3	38.9	**.8**	−3.3	−2.4	−4.9	2.5	4.6	−4.7	528 (136)	1.4
14	50.9	57.1	**1.2**	−3.8	−3.4	−5.0	2.8	4.3	−5.0	728 (149)	1.1
16	70.7	76.4	4.2	−4.3	−4.0	−5.0	3.2	6.9	−5.0	960 (158)	.9

[a] $\overline{d}'$ denotes the rounded mean of d' for T_{AD}, T_{ADb}, and T_{ADs} over 1,200 replications.

Note. Values in the range [−1.5, 1.8] are defined as acceptable and are thus printed in bold face.

TABLE 3
Empirical Minus the 5% Nominal Type I Error Rates of Nine Model Fit Statistics for $N = 400$ ($NR = 1{,}200$)

k	T_{ML}	T_{SC}	T_{AD}	T_{MLb}	T_{SCb}	T_{ADb}	T_{MLs}	T_{SCs}	T_{ADs}	d ($\overline{d}'$)	$N{:}t$
4	2.6	3.1	**1.6**	**1.2**	**1.7**		**1.5**	2.0	**.7**	48 (41)	13.3
6	3.1	3.8	**.7**	**.5**	**1.1**	−1.6	**1.3**	1.9	**−1.1**	120 (88)	7.8
8	3.6	4.5	**−1.5**	−1.8	**−1.0**	−3.6	**−1.3**	**.3**	−3.2	224 (136)	5.3
10	6.5	8.3	**−.9**	**−1.1**	**−.7**	−4.0		**1.3**	−3.3	360 (179)	3.8
12	11.4	14.3	**−1.0**	−2.0	**−1.1**	−4.8	**.2**	**1.3**	−4.6	528 (215)	2.9
14	21.0	22.0	−1.9	−2.8	−2.2	−5.0	**1.4**	2.9	−4.7	728 (245)	2.3
16	26.0	29.7	−1.7	−3.4	−2.8	−5.0	**.8**	2.1	−4.6	960 (268)	1.9

Note. Blank cell indicates that the empirical error rate equals the nominal rate of 5%. Values in the range [−1.5, 1.8] are defined as acceptable and are thus printed in bold face.

TABLE 4
Empirical Minus the 5% Nominal Type I Error Rates of Nine Model Fit Statistics for $N = 800$ ($NR = 1{,}200$)

k	T_{ML}	T_{SC}	T_{AD}	T_{MLb}	T_{SCb}	T_{ADb}	T_{MLs}	T_{SCs}	T_{ADs}	d ($\overline{d}'$)	$N{:}t$
4	**1.4**	**1.7**	**1.1**	**1.0**	**1.3**	**.7**	**1.1**	**1.3**	**.7**	48 (44)	26.7
6	2.2	2.7	**.7**	**.6**	**1.0**	**−.6**	**1.2**	**1.6**	**−.4**	120 (101)	15.7
8	3.1	3.0	**.8**	**.8**	**1.3**	**−1.5**	**1.7**	2.1	**−.5**	224 (169)	10.5
10	1.9	2.6	**−1.1**	**−1.0**	**−.7**	−3.3	**−.2**	**−.1**	−2.6	360 (238)	7.6
12	5.6	6.1	**−1.1**	**−1.0**	**−.8**	−3.6	**.7**	**1.8**	−2.9	528 (305)	5.8
14	5.7	6.6	−1.7	−1.8	−1.6	−4.4	**−.1**	**.3**	−3.8	728 (365)	4.6
16	8.8	10.9	−2.1	−2.3	−1.7	−4.7	**−.1**	**.8**	−4.2	960 (418)	3.7

Note. Values in the range [−1.5, 1.8] are defined as acceptable and are thus printed in bold face.

Therefore, one could argue that, in both theoretical and applied research, the issue of model size should deserve similar attention as the robustness against nonnormality.

Scaled Satorra–Bentler statistic. Like for T_{ML}, the finite sample bias of the test statistic T_{SC} reduces with increasing sample size and decreasing model size. As expected, and therefore consistent with the results of simulation studies mentioned earlier, the performance of T_{SC} is slightly worse compared to that of T_{ML}. For nearly all investigated sample sizes, the rejection rates are not acceptable. For $N = 200$ and 16 factors, the bias in the empirical rejection rates is 76.4%. It follows that the use of T_{SC} is no option for the evaluation of large models.

Adjusted Satorra–Bentler statistic. For T_{AD} with $N = 200$, there is a slight tendency of a reduced finite sample bias when model size decreases, but this tendency is much weaker compared to that of T_{ML} and T_{SC}. For $N = 400$ and $N = 800$, T_{AD} slightly underestimates nominal Type I error levels when the model size increases. Overall, however, the results indicate that T_{AD} clearly outperforms T_{ML} and T_{SC} for all models under study. The rejection rates on the 5% error level are nearly perfect for $N = 200$ and models with up to 14 factors. Therefore, our study revalidates the finding of Fouladi (2000) that test statistic T_{AD} has excellent Type I error control. The reason for the good performance of T_{AD} seems to be Satterthwaite's (1941, 1946) variance correction, which adjusts the tail of the distribution of T_{ML} adequately.

In general, our expectations with respect to the behavior of the mean- and variance-adjusted test statistic T_{AD} are not refuted. Recall that Fouladi (2000) found that T_{AD} outperforms 12 other statistics with respect to Type I error control under various distributional conditions and for different models. Therefore, T_{AD} seems to be relatively robust against model size, small sample size, and nonnormality. Nevitt and Hancock (2004) seem to be disinclined to recommend this statistic, because it slightly underestimates the nominal Type I error rates when nonnormal variables are analyzed. Their conclusions challenge those of Fouladi (2000); more research on this issue is therefore necessary. Nevertheless, after inspection of the empirical rejection rates, it seems legitimate to use T_{AD} with approximately normal data, but a more final judgment will be postponed after inspection of the Kolmogorov–Smirnov test results.

Bartlett-corrected statistics. All Bartlett statistics underestimate the nominal rejection rates with increasing model size. Where most statistics are progressive (i.e., the null hypothesis is rejected too often, or the rejection rates are too high) for $N = 200$, the Bartlett corrections show a conservative trend (i.e., the null hypothesis is "conserved" too often, the rejection rates are too low). This

is consistent with our expectation based on the results of Nevitt and Hancock (2004). Compared to T_{AD}, the statistics T_{MLb}, T_{SCb}, and T_{ADb} are slightly more influenced by model size. Interestingly, T_{SCb} performs better than T_{MLb}. It seems that the progressive tendency of T_{SC} dominates for smaller model sizes, whereas a general conservative effect of the Bartlett corrections dominates when the models get larger. Based on the empirical rejection rate performance only, we are slightly hesitant to recommend the use of Bartlett statistics, because these statistics are too conservative and do not reveal an adequate Type I error control, at least not for large models and small sample sizes.

Swain-corrected statistics. The results indicate that T_{MLs} is less affected by model size compared to T_{MLb}. The statistic T_{MLs} has appropriate rejection rates for $N = 200$ up to 10 factors. Compared to all other statistics, T_{MLs} is less influenced by the model-size effect, especially when the sample size is 400 or 800. T_{SCs} performs equally well compared to T_{SCb}. T_{ADs} is clearly too conservative. Thus, it seems legitimate to use T_{MLs} in applied research, but again, a more final judgment will be formulated after looking at the results of the Kolmogorov–Smirnov test.

Intermediate conclusion. To summarize the results presented so far, we conclude that (a) T_{MLs}, (b) T_{AD}, and (c) T_{SCs} or T_{SCb}—in that order—yield the best 5% Type I error control in large models.

Kolmogorov–Smirnov Tests

To check whether the empirical sampling distributions of the test statistics, $F_{NR}(x)$, deviate significantly from their reference chi-square distribution, $F_d(x)$, with d degrees of freedom, the one-sample Kolmogorov–Smirnov test statistic $D_{NR} = \sup_x[|F_{NR}(x) - F_d(x)|]$ was computed. The D_{NR} values are presented in Table 5 ($N = 200$), Table 6 ($N = 400$), and Table 7 ($N = 800$). In the evaluation of test results we applied a two-sided 1% significance level. In our case, with $NR = 1{,}200$ replications, the critical value of the D_{NR} statistic at that 1% level equals $1.63/\sqrt{1{,}200} = 0.047$ (Massey, 1951). Nonsignificant D_{NR} values, indicating closeness of fit, are boldfaced in the tables.

For the smallest sample size $N = 200$, T_{MLs} clearly outperforms all other statistics for large models. Although significant deviations for the larger models are reported, the relatively good performance of T_{MLs} compared to the other statistics under study is obvious. The statistic T_{AD} does not perform well, although it produced Type I error rates close to those of T_{MLs}. When the sample size increases to $N = 400$, T_{SCb} is the second best statistic. For $N = 800$, T_{MLs} and T_{SCs} are the best performing statistics regarding their expected distributional match.

TABLE 5
The D_{NR} Values of the One-Sample Kolmogorov–Smirnov Test of Nine Model Fit Statistics for $N = 200$ ($NR = 1{,}200$)

k	T_{ML}	T_{SC}	T_{AD}	T_{MLb}	T_{SCb}	T_{ADb}	T_{MLs}	T_{SCs}	T_{ADs}	d ($\overline{d}'$)	$N{:}t$
4	.087	.110	.139	**.022**	**.025**	.070	**.029**	**.044**	.085	48 (36)	6.7
6	.138	.167	.203	.060	**.037**	.078	**.013**	**.043**	.111	120 (69)	3.9
8	.253	.295	.292	.068	**.027**	.100	.054	.097	.151	224 (98)	2.6
10	.368	.414	.367	.133	.076	.151	.057	.116	.178	360 (120)	1.9
12	.482	.528	.443	.195	.141	.186	.060	.124	.213	528 (136)	1.4
14	.626	.668	.516	.284	.205	.275	.099	.148	.230	728 (149)	1.1
16	.761	.800	.598	.362	.283	.301	.104	.189	.264	960 (158)	.9

Note. The critical value of $D_{1,200}$ at a two-sided 1% significance level equals 0.047. Values in the range [.000, .047] are defined as acceptable and are thus printed in bold face.

TABLE 6
The D_{NR} Values of the One-Sample Kolmogorov–Smirnov Test of Nine Model Fit Statistics for $N = 400$ ($NR = 1{,}200$)

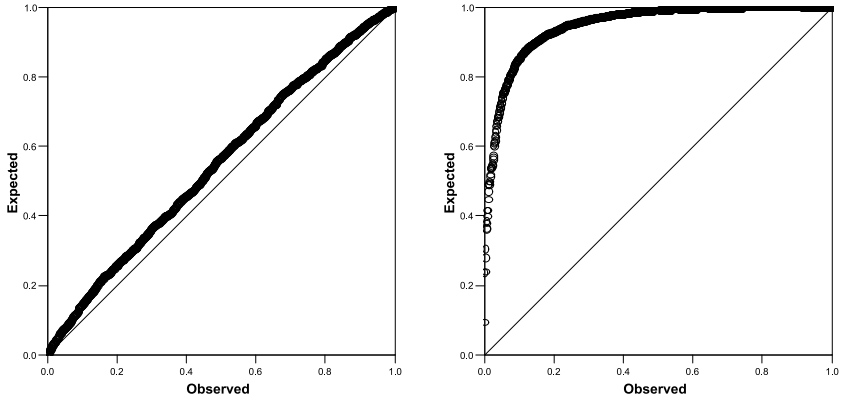
k	T_{ML}	T_{SC}	T_{AD}	T_{MLb}	T_{SCb}	T_{ADb}	T_{MLs}	T_{SCs}	T_{ADs}	d ($\overline{d}'$)	$N{:}t$
4	.084	.089	.102	**.033**	**.044**	.050	**.044**	.054	.060	48 (41)	13.3
6	.086	.102	.092	**.038**	**.030**	**.033**	**.021**	**.037**	**.042**	120 (88)	7.8
8	.145	.169	.176	**.031**	**.016**	.076	**.038**	.063	.093	224 (136)	5.3
10	.186	.211	.212	.070	**.044**	.105	**.036**	.059	.109	360 (179)	3.8
12	.260	.292	.291	.103	.070	.121	**.034**	.065	.151	528 (215)	2.9
14	.351	.385	.332	.118	.085	.164	.055	.092	.157	728 (245)	2.3
16	.428	.463	.399	.184	.138	.199	**.047**	.093	.190	960 (268)	1.9

Note. Values in the range [.000, .047] are defined as acceptable and are thus printed in bold face.

TABLE 7
The D_{NR} Values of the One-Sample Kolmogorov–Smirnov Test of Nine Model Fit Statistics for $N = 800$ ($NR = 1{,}200$)

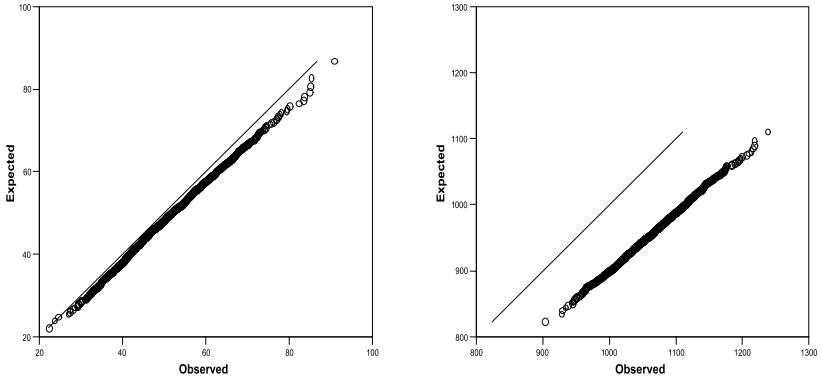
k	T_{ML}	T_{SC}	T_{AD}	T_{MLb}	T_{SCb}	T_{ADb}	T_{MLs}	T_{SCs}	T_{ADs}	d ($\overline{d}'$)	$N{:}t$
4	.048	.055	.074	**.030**	**.025**	**.043**	**.025**	**.029**	.048	48 (44)	26.7
6	**.044**	**.047**	.064	**.026**	**.023**	**.031**	**.020**	**.023**	**.039**	120 (101)	15.7
8	.096	.104	.109	**.018**	**.023**	**.047**	**.037**	**.046**	.061	224 (169)	10.5
10	.087	.096	.126	.062	.053	.072	**.023**	**.022**	.074	360 (238)	7.6
12	.135	.157	.159	.063	.054	.073	**.024**	**.040**	.086	528 (305)	5.8
14	.175	.192	.208	.072	.055	.109	**.027**	**.044**	.108	728 (365)	4.6
16	.235	.257	.268	.090	.065	.130	**.037**	.056	.143	960 (418)	3.7

Note. Values in the range [.000, .047] are defined as acceptable and are thus printed in bold face.



(a) $d = 48$ (b) $d = 960$

FIGURE 1 PP plots for T_{ML} ($N = 200$; $NR = 1{,}200$).

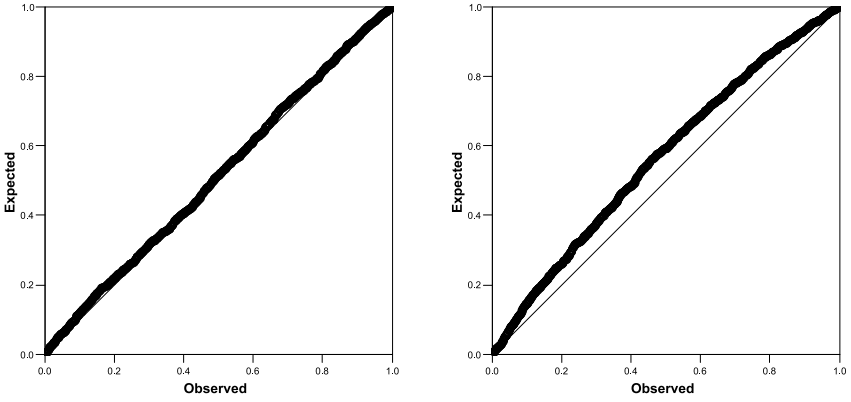


(a) $d = 48$ (b) $d = 960$

FIGURE 2 QQ plots for T_{ML} ($N = 200$; $NR = 1{,}200$).

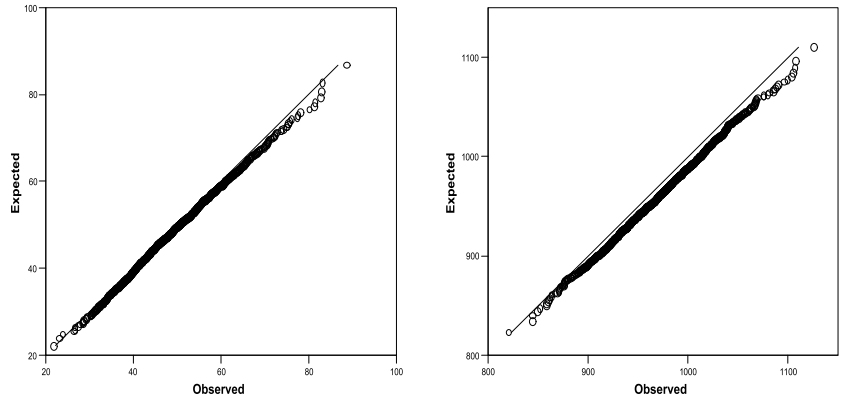
PP Plots and QQ Plots

Graphical comparisons of the sampling distributions of the statistics to their reference chi-square distributions are provided to visualize information from Tables 2 through 7. Both PP plots and QQ plots are shown because PP plots are more sensitive to deviations in the middle of a distribution, whereas QQ plots are more sensitive to deviations in its tails (Gnanadesikan, 1977). The plots for T_{ML} (Figures 1 and 2) are included because T_{ML} serves here as the reference statistic to illustrate the potential benefits of using T_{MLs} (Figures 3 and 4). In addition, Figures 5 and 6 demonstrate the extremely bad distributional



(a) $d = 48$ (b) $d = 960$

FIGURE 3 PP plots for T_{MLs} ($N = 200$; $NR = 1{,}200$).

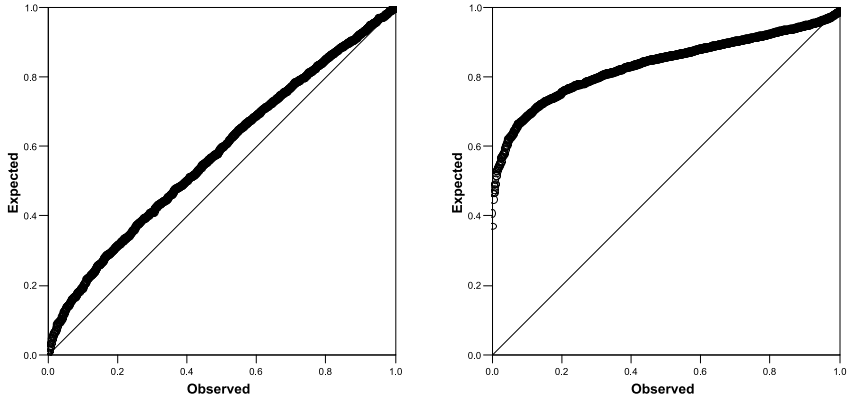


(a) $d = 48$ (b) $d = 960$

FIGURE 4 QQ plots for T_{MLs} ($N = 200$; $NR = 1{,}200$).

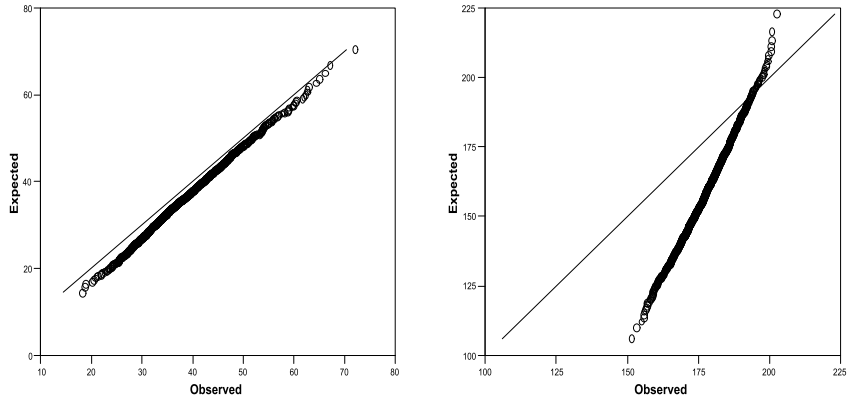
performance of T_{AD}: The 5% Type I error rate is approximately correct but the overall behavior is clearly deviant. The plots for the smallest model ($d = 48$) and the largest model ($d = 960$) are shown for the worst case scenario where $N = 200$.

When comparing Figures 1 and 2 to Figures 3 and 4, the disastrous results for T_{ML} clearly emerge. Overall, T_{MLs} has a very close approximation to the reference chi-square distribution. Therefore, we reconfirm our recommendation to use this correction of T_{ML} in applied research when large structural equation models are analyzed.



(a) $\bar{d}' = 36$ (b) $\bar{d}' = 158$

FIGURE 5 PP plots for T_{AD} ($N = 200$; $NR = 1{,}200$).



(a) $\bar{d}' = 36$ (b) $\bar{d}' = 158$

FIGURE 6 QQ plots for T_{AD} ($N = 200$; $NR = 1{,}200$).

Final Conclusion

In summary, the best performing statistic with respect to Type I error control and the approximation of the reference chi-square distribution is T_{MLs}. Therefore, we recommend using this statistic when many (approximately) multinormal distributed variables are under study in SEM. From Equations 10 through 12 it can be seen that the correction will have only a very small effect on the chi-square value for smaller models or larger sample sizes. From that perspective it would make sense to apply the correction quite generally.

Software

The calculation of T_{MLs} is quite easy once the value of T_{ML} is available, because Swain's correction factor is a simple function of known values of p, N, and d or t. The p values for the test statistic T_{MLs} are also easily computed with computer software, for example with the function `pchisq(x,d)`, where `x` $= T_{MLs}$, and `d` is the number of degrees of freedom, from freely available `R` software (cf. Venables & Smith, 2005, section 8.1). Although this is a small effort in practice (the `R`-function `swain` for the calculation of T_{MLs} and its corresponding p value can be downloaded from http://www.gmw.rug.nl/∼boomsma), we would recommend implementing the Swain correction in standard SEM software.

Example

To illustrate the effects of using T_{MLs}, the value of T_{ML} was corrected in a recently published article. Ramaswami and Singh (2003) estimated a confirmatory factor model with $N = 154$, $k = 13$, $p = 51$, $d = 1{,}147$, and $t = 179$. They reported $T_{ML} = 1{,}307$ with a p value of 0.0007, which would lead to a rejection of the model if a formal test was applied at significance levels of 5% or 10%, say. When the Swain correction is applied, the value of T_{MLs} equals 1,146 with a relatively large increase of the p value to 0.5034. Hence, the model is certainly not rejected when this Swain-corrected test of exact fit is performed. Of course, chi-square dependent statistics like the RMSEA are also affected by the model-size effect: The RMSEA test statistic for close fit would drop from 0.0302 (Ramaswami and Singh reported 0.0320) to 0.0000 when using T_{MLs}.

DISCUSSION

A Retrospective View on Applied Research

In the following we briefly discuss the consequences of our results for past applied research using large covariance structure models. Even if the estimated models in those applications were specified correctly, with variables having nearly normal distributions, we suspect that the fit of most models was underestimated. Two strategies might have been used when small p values of the chi-square model fit statistics occurred.

First, the chi-square statistic for global model fit might be neglected completely and refuge might be taken to other fit statistics (e.g., the RMSEA) or fit indexes (e.g., the TLI, the CFI, and the standardized root mean square residual, SRMR). Apart from the RMSEA, which is asymptotically based on a noncentral chi-square distribution, research on the distribution of the latter statistics is still at its beginning (e.g., Hu & Bentler, 1999; Ogasawara, 2001). The sampling

distribution of most fit indexes is just unknown. Researchers therefore rely on certain cut-off values for such indexes, that have been recommended in the literature (e.g., Hu & Bentler, 1999). These cut-off values are partly arbitrary, and moreover, the blindfolded use of such "golden rules" has proven to be inaccurate under circumstances (Kaplan, 1988; Marsh, Hau, & Wen, 2004; Saris, Den Ronden, & Satorra, 1987). More important, however, is the fact that most fit statistics and indexes are also affected by the inflated T_{ML}, because they are a function of this statistic when maximum likelihood estimation is applied. Given the results of our study, it would make sense to substitute T_{MLs} for T_{ML} when calculating these fit statistics and fit indexes. For incremental fit indexes it is not clear whether the fit statistic for the independence model needs to be adjusted similarly; these are issues in need of further research (for first results see Herzog & Boomsma, 2006).

Second, in applied (exploratory) SEM, modification indexes (Sörbom, 1989) are often used extensively, as a last resort in the search for models that cannot be rejected. In many cases, restrictions on covariances among measurement errors are removed without interpreting their meaning, or explaining why such covariances make sense from a theoretical point of view in the first place. This seems to become a common practice, although Jöreskog (1993, p. 297) and many others explicitly criticized this kind of pseudo-theory testing. Given our research findings, the reliability of such model explorations, with T_{ML} as its basis, must be questioned even further when at least 12 observed variables are analyzed with sample sizes of up to $N = 800$.

The results of our study also suggest that it is not unlikely that there may have been many studies in the past where correctly specified large models were not published, because the models were rejected due to the inflated T_{ML}. Such phenomena, also labeled "file drawer" problems (e.g., Scargle, 2000), clearly attenuate scientific progress.

The N:t Ratio Criterion

The robustness of model test statistics against model size is not unimportant, as our study shows. An obvious overall remedy to avoid the problem of inflated values of test statistics is to increase sample size N relative to the number of degrees of freedom d, or to increase N relative to the number of parameters to be estimated t, because t can in principle be interpreted as a measure of model size as well. Certain rules of thumb regarding an adequate sample size relative to the number of parameters t, the $N{:}t$ ratio, can be found in the literature. Bentler (1995), for example, recommended a ratio of at least 5:1 when T_{ML} is used and the assumption of multivariate normality holds. Although such rules of thumb are not without criticism (e.g., Jackson, 2003), we could evaluate our results also in terms of the $N{:}t$ ratio, that is, the relative sample adequacy. The

last column of Tables 2 through 7 shows the value of this ratio. We can now compare our results with earlier $N:t$ recommendations and try to formulate general guidelines in terms of relative sample adequacy for proper behavior of model test statistics. One should realize, however, that the $N:t$ ratio is a simplifying rule of thumb regarding only two of the many factors that matter in a research design.

Our results clearly show that Bentler's 5:1 rule of thumb is not sufficient for the sampling distribution of T_{ML} to be approximately chi-square. Even for our smallest model and our largest sample size ($d = 48$, $t = 30$, $N = 800$), with a $N:t$ ratio of 26.7:1, the Kolmogorov–Smirnov test for T_{ML} indicates a significant deviation from the chi-square reference distribution (see Table 7). For our second smallest model ($d = 120$, $t = 51$, and $N = 800$), a $N:t$ ratio of 15.7:1 is not large enough for proper Type I error behavior of T_{ML} at the 5% significance level (see Table 4). Also, in contrast to Fouladi (2000, p. 401), we would not conclude that T_{AD} can be applied under conditions of small $N:t$ ratios. The results in Table 7 show that a ratio of 26.7:1 is insufficient for proper behavior of T_{AD} in moderately large models when inspecting its sampling distribution as a whole, not just its 5% Type I error rates.

Earlier we discussed evidence that the Bartlett statistics suffer from an increasingly conservative trend when model size increases. This effect may be due to the fact that these corrections were originally developed for exploratory factor analyses and not for general covariance structure analyses. For T_{SCb}, this effect is masked by the slightly more liberal tendency of T_{SC} compared to T_{ML}. Thus, for the models under study here, we do not observe and cannot conclude, unlike Nevitt and Hancock (2004), that the Bartlett corrections "frequently delivered acceptable Type I error rates at $N:t \leq 2{:}1$" (p. 467).

The most salient conclusion of our study is that overall the Swain-corrected statistic T_{MLs} performs best. The results in Tables 2 through 7 validate the (strong) conclusion that for the models under study, apart from single small-sample fluctuations, T_{MLs} is robust against large model size if $N:t \geq 2{:}1$ under conditions of normality. As will be indicated in the next section, more research is needed to investigate the interaction of nonnormality and model size.

However, although it seems convenient for applied researchers to have rules of thumb like $N:t$ (or $N:p$ ratios for that matter) it would be unwise to follow these guidelines blindly; compare the sincere warnings of Marsh et al. (1998) and Boomsma and Hoogland (2001, p. 142f). First, the mild requirement that for the use of T_{MLs} the $N:t$ ratio should be at least 2:1 should certainly not be interpreted as an encouragement to always stay away from large models, or to use a small number of indicators per factor, which, as a start, would increase the occurrence of nonconvergent and improper solutions. Second, easy formulated rules of thumb regarding the $N:t$ ratio also should not overshadow sample size requirements related to the stability of parameter estimates or the size of

estimated standard errors of parameter estimates, and considerations as to the power of model test statistics, either locally or globally.

Limitations and Future Work

- It is well known that nonnormality has an inflating effect on chi-square model fit statistics (cf. Boomsma, 1983). It should be investigated how well the test statistics, and in particular the Swain-corrected scaled Satorra–Bentler statistic, behave in large models under conditions of nonnormality.
- This study was confined to factor models. It seems necessary to expand the scope of structural equation models under investigation to a broader range. For these other types of models a main question is also whether and to which extent Bartlett adjustments are effective in comparison with Swain's correction.
- Another issue concerns the specific value of 0.70 of the factor loadings that was used in our study. According to the research by Hoogland (1999), the rejection rates are more accurate for smaller factor loadings. Maybe the same pattern will be observed for the test statistics from our study as well.
- The test statistic T_{MLs} deserves additional attention from a statistical power perspective. After assessing the Type I error rates, future studies should also focus on the power of this corrected test statistic in comparison with a few other promising ones. Emphasis would then turn more to Type II error rates (cf. Nevitt & Hancock, 2004).
- As mentioned earlier, the effect of the proposed corrections of T_{ML} on other fit statistics and indexes, like the RMSEA, the TLI, and the CFI, requires further attention. It needs to be investigated to which extent other fit measures are affected by corrected global test statistics (for first results see Herzog & Boomsma, 2006). The SRMR, in our view a fit measure that needs to be inspected in all circumstances, certainly is not.
- This simulation study emphasized the importance of investigating the finite sample behavior of statistics in large models. The disastrous results for T_{ML} and T_{SC} may raise questions regarding the generalizations made in many previous simulation studies. One direction of further investigation could be to revisit those studies, and to check whether reported findings generalize to larger models.
- Wakaki, Eguchi, and Fujikoshi (1990) derived a (relatively complex) Bartlett adjustment factor for the test of general covariance structures. In a first simulation study, this correction significantly improved the performance of T_{ML} (Kensuke, Takahiro, & Kazuo, 2005). Therefore, it would be of interest to compare its performance with that of the statistics presented here.

- Within the framework of Bayesian estimation of structural equation models, Lee and Song (2004) made a comparison with the classical, frequentist use of T_{ML}, and found that the Bayesian posterior predictive p values are less biased compared to the maximum likelihood p values under conditions of small sample sizes (cf. Scheines, Hoijtink, & Boomsma, 1999). They also found that the posterior predictive p values are not accurate when nonnormal variables are analyzed. A comparison of the performance of the Bayesian approach to that of T_{MLs} for large models would be intriguing.

CONCLUSION

Some years ago, Kaplan (1988) came to the conclusion that the chi-square model statistic "should be taken seriously as a means of formally testing model specification" (p. 85). For large models, it has been shown here that researchers should seriously consider corrected model test statistics if such a formal approach of model testing is being taken. Otherwise, biased inference might be an undesirable consequence. If this problem is acknowledged, and proper corrections are indeed applied, there are enough obstacles to clean inference left (cf. Jöreskog, 1993).

ACKNOWLEDGMENTS

This article is an elaboration of a paper presented at the 70th annual meeting of the Psychometric Society, July 5, 2005, Tilburg University, The Netherlands.

We would like to thank three anonymous referees and the editor for a number of useful suggestions. Further, we would like to thank Anthony J. Swain for sending his dissertation and Michael W. Browne for his help in contacting Dr. Swain. The Research Funds (GFF) of the University of St. Gallen (Switzerland) provided financial support to the first author during some time of the research project.

REFERENCES

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician, 52,* 119–126.

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis.* New York: Wiley.

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12,* 411–434.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London, Series A, 160,* 268–282.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology (Statistical Section), 3,* 77–85.

Bartlett, M. S. (1954). A note on the multiplying factors for various χ^2 approximations. *Journal of the Royal Statistical Society, Series B, 16,* 296–298.

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

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

Birnbaum, Z. W. (1952). Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size. *Journal of the American Statistical Association, 47,* 425–441.

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

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction* (Part 1, pp. 149–173). Amsterdam: North-Holland.

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality*. Unpublished doctoral dissertation, University of Groningen, The Netherlands.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika, 50,* 229–242.

Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future: A Festschrift in honor of Karl Jöreskog* (pp. 139–168). Chicago: Scientific Software International.

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika, 36,* 317–346.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics, 25,* 290–302.

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–142). Cambridge, UK: 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., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21,* 230–258.

Curran, P. J., Bollen, K. A., Paxton, P., Kirby, J., & Chen, F. (2002). The noncentral chi-square distribution in misspecified structural equation models: Finite sample results from a Monte Carlo simulation. *Multivariate Behavioral Research, 37,* 1–36.

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.

Fouladi, R. T. (1999, April). *Model fit in covariance structure analysis under small sample conditions—Modified maximum likelihood and asymptotically distribution free generalized least squares procedures.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation Modeling, 7,* 356–410.

Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. New York: Wiley.

Hau, K.-T., & Marsh, H. W. (2004). The use of item parcels in structural equation modelling: Non-normal data and small sample sizes. *British Journal of Mathematical and Statistical Psychology, 57,* 327–351.

Herzog, W., & Boomsma, A. (2006, June). *Finite sample corrections for RMSEA estimation.* Paper presented at the annual meeting of the Psychometric Society, Montreal, Canada.

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

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

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

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling, 10,* 128–141.

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

Jöreskog, K. G. (2005, May). *Structural equation modeling with LISREL 8.7*. Symposium conducted at The Gleacher Center (in cooperation with K. A. Bollen), Chicago.

Jöreskog, K. G., Sörbom, D., Du Toit, S., & Du Toit, M. (2001). *LISREL 8: New statistical features* (3rd rev. ed.). Chicago: Scientific Software International.

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research, 23,* 69–86.

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling, 10,* 333–351.

Kensuke, O., Takahiro, H., & Kazuo, S. (2005, July). *Bartlett correction of likelihood ratio statistics in structural equation modeling*. Poster presented at the annual meeting of the Psychometric Society, Tilburg, The Netherlands.

Lawley, D. N. (1956). A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika, 43,* 295–303.

Lee, S.-Y., & Song, X.-Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research, 39,* 653–686.

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

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

Massey, F. J., Jr. (1951). The Kolmogorov–Smirnov test for goodness of fit. *Journal of the American Statistical Association, 46,* 68–78.

Muthén, B. O. (1993). Goodness of fit with categorical and other nonnormal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–234). Newbury Park, CA: Sage.

Muthén, B. O. (2004). *Mplus*: *Statistical analysis with latent variables: Technical appendices*. Los Angeles: Muthén & Muthén.

Muthén, B. O., & Satorra, A. (1995). Technical aspects of Muthén's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika, 60,* 489–503.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide* (3rd ed.). Los Angeles: Muthén & Muthén.

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

Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research, 39,* 439–478.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika, 20A,* 175–240.

Ogasawara, H. (2001). Approximations to the distributions of fit indexes for misspecified structural equation models. *Structural Equation Modeling, 8,* 556–574.

Ramaswami, S. N., & Singh, J. (2003). Antecedents and consequences of merit pay fairness for industrial salespeople. *Journal of Marketing, 67,* 46–66.

Saris, W. E., Den Ronden, J., & Satorra, A. (1987). Testing structural equation models. In P. Cuttance & J. Ecob (Eds.), *Structural modeling by example: Applications in educational, sociological, and behavioral research* (pp. 202–220). Cambridge, UK: Cambridge University Press.

Satorra, A., & Bentler, P. M. (1988). *Scaling corrections for statistics in covariance structure analysis* (UCLA Statistics Series, No. 2). Los Angeles: University of California, Department of Psychology.

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

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66,* 507–514.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika, 6,* 309–316.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2,* 110–114.

Scargle, J. D. (2000). Publication bias: The "file drawer" problem in scientific inference. *Journal of Scientific Exploration, 14,* 91–106.

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika, 64,* 37–52.

Sörbom, D. (1989). Model modification. *Psychometrika, 54,* 371–384.

Swain, A. J. (1975). *Analysis of parametric structures for variance matrices*. Unpublished doctoral dissertation, University of Adelaide, Australia.

Venables, W. N., & Smith, D. M. (2005). *An introduction to* R (Version 2.1.1). Retrieved August 15, 2005 from http://www.r-project.org.

Wakaki, H., Eguchi, S., & Fujikoshi, Y. (1990). A class of tests for a general covariance structure. *Journal of Multivariate Analysis, 32,* 313–325.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics, 9,* 60–62.