Chapter - 3
The Model-Size Effect on Traditional and Modified Tests of Covariance Structures
Chapter - 3
The Model-Size Effect on Traditional and Modified Tests of Covariance Structures

3.1 Introduction.

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 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 non normal data appears to be accountable for the inflated Type I error rates of the traditional maximum likelihood ratio test statistic, TML, 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 TML. 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 TML 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 TML 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, TSC, and the adjusted (mean- and variance-corrected) statistic, TAD (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 TML 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 TML, 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 TSC and TAD 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 TSC or TAD over TML 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 sample size 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 non normal 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, TMLb, TSCb, and TADb, 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 TMLb, TSCb, and TADb, and to compare them to those of TML, TSC, and TAD, respectively, in large models under conditions of multivariate normality.

A less well-known correction of TML 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 TMLs, TSCs, and TADs might yield better Type I error rates compared to those of TML, TSC, and TAD. 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 TML; (b) to compare the results of TML with those of TSC and TAD; (c) to evaluate the performance of TMLb, TSCb, and TADb; and (d) to check whether the behavior of TMLs, TSCs, and TADs 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 TML, TSC, and TAD in large models was demanded by several researchers (e.g., Hoogland, 1999; Kenny & McCoach, 2003; Muthén, 1993, p. 98).
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.
3.2 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 $z(p \times 1)$ with an empirical sample covariance matrix $S(p \times p)$ based on $N = n + 1$ independent observations, and a population model of underlying relations among these variables with covariance structure $\Sigma(\theta)(p \times p)$, where $\Sigma(t \times 1)$ is the vector of independent model parameters to be estimated. If the observed variables $z$ follow a multivariate normal distribution, the sample covariance matrix $S$ 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 $S$, is equivalent to minimizing the function

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

which is a discrepancy function as defined by Browne (1984, p. 64); log denotes the natural logarithm here. The parameter vector $\hat{\theta}$, defining the minimum of $F_{ML}$ $[S; \Sigma(\theta)]$, contains the so-called maximum likelihood estimates of $\theta$. Asymptotically, as $N$ goes to infinity, the maximum likelihood estimates are normally distributed with expectation vector $E(\hat{\theta}) = \theta$, and asymptotic covariance matrix $\text{acov}(\hat{\theta}, \hat{\theta}) = \mathbf{I}^1(\theta)$, 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 \( \Sigma (p \times p) \) denote the population covariance matrix of the \( p \) observed variables \( z \), \( \Sigma (\theta_j) \) 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 \( H_0 : \Sigma = \Sigma (\theta_0) \); that is, that the postulated model holds, with the corresponding log-likelihood function, evaluated at \( \theta_0 = \hat{\theta}_0 \),

\[
\log L_0 = \log L(\Sigma (\hat{\theta}_0); S) = -\frac{n}{2} \{ \log [\Sigma (\hat{\theta}_0)] + \text{tr} [\Sigma^{-1} (\hat{\theta}_0)] \} + \log c, \quad (2)
\]

\[
\log L_1 = \log L(\Omega; S) = -\frac{n}{2} [\log |S| + \text{tr} (SS^{-1})] + \log c \quad (3)
\]

\[
= -\frac{n}{2} (\log |S| + p) + \log c
\]

It can then be shown that under \( H_0 \), the distribution of the likelihood ratio statistic, defined as

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

converges with increasing sample size \( n \) to a chi-square distribution with \( 2p \) degrees of freedom (Wilks, 1938); the likelihood criterion \( L_0 = L_1 \) 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.

Hence, the likelihood ratio test statistic can be used to test whether the proposed model 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, and a large sample size, mainly).
Satorra–Bentler Statistics

Because non normal 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, TWLSN, 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 TML.

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

\[ T_{SC} = \frac{d}{\text{tr}(A)} T_{ML}, \quad (5) \]

where matrix 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 z is elliptical, the scaling factor \( d=\text{tr}.A \) in Equation 5 provides an estimate of the common relative kurtosis of z (Satorra & Bentler, 1994, p. 407), which implies a correction for non normality.

As usual, the test statistic TSC is evaluated as having (approximately) a chi-square distribution with \( C 1/=2 \) t degrees of freedom. For certain distributions
of the observed variables, for example, elliptical ones, the asymptotic distribution of TSC is exactly chi-square with d degrees of freedom. In principle, however, the correction of TML involves a scaling to the correct mean, so that for general distributions asymptotically the first moment of the distribution of TSC is matched to the number of degrees of freedom d. Under conditions of multivariate normality, TSC 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 TML as well. This is possible by an adjustment of the number of degrees of freedom to d0, which is the integer closest to a function of the matrix A (cf. Muthén, 2004, Equation 110): by definition

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

\[ T_{AD} = \frac{d'}{\text{tr}(A)} T_{ML}. \tag{7} \]

It should be noted that the value of d0 may vary from sample to sample. Substituting d0 for d in Equation 5, we get (cf. Muthén, 2004, Equation 108): which is the adjusted chi-square test statistic; adjusted for mean and variance that is. Again, for general distributions of observed variables, TAD has asymptotically not an exact chi-square distribution with d0 degrees of freedom, but it matches the first- and second-order moment of that distribution (Satorra & Bentler, 1994, p. 408). For multivariate normal observations, TAD has asymptotically an exact chi-square distribution with d0 degrees of freedom.
It should be stressed that if distributional assumptions or conditions for asymptotic robustness hold, both corrections of TML 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 TML 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, where oe is the likelihood ratio criterion of Neyman and Pearson (1928), by a scale factor that results in a statistic having the same moments as $\chi^2$, 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 $N D n C 1$. 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, TML, TSC, and TAD. The corresponding Bartlett corrections for these statistics are defined as respectively,

$$
T_{MLb} = b T_{ML} , T_{SCb} = b T_{SC} , \text{ and } T_{ADb} = b T_{AD}.
$$

(8)
where

\[ b = 1 - \frac{4k + 2p + 5}{6n}. \]  

(9)

It follows from Equations 8 and 9 that asymptotically the distribution of the Bartlett-corrected statistics matches the asymptotic distributions of TML, TSC, and TAD, 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 TML 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 \[ k_1 = n C \circ n 2/ \], 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 \( n \text{FMLOES}^\dagger \). \( O^{\text{TM0}}/ \), not as multipliers of just the discrepancy function \( \text{FMLOES}^\dagger \). \( O^{\text{TM0}}/ \). 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 TML 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, \( n \) 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 TML, TSC, and TAD are now, respectively, defined as

\[
T_{MLs} \equiv sT_{ML}, \quad T_{SCs} \equiv sT_{SC}, \quad \text{and} \quad T_{ADs} \equiv sT_{AD}. \tag{12}
\]
From Equation 10 it can be seen that Swain’s correction is a function of $p$, $d$, and $N$. Because, 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 TML, TSC, and TAD, respectively.

### 3.3 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, TML, TSC, TAD, TMLb, TSCb, TADb, TMLs, TSCs, and TADs, 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.

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 TML 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 TSC produces even
higher rejection rates than TML when multivariate normal variables are analyzed, 
and this liberal tendency increased with model size as well. Therefore, we expect 
that TSC will perform worse than TML in large models under conditions of 
normality. The explanation for this expected tendency could very well be that TSC 
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 
TAD 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 TML and TSC. Fouladi (2000) conducted an extensive simulation 
study with 12 different test statistics and found TAD to outperform all other 
statistics with respect to Type I error rate “under more general non normal 
distributional conditions” (p. 400; cf. p. 371, Table 1). She concluded that TAD 
“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 
TAD will outperform TML and TSC 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 TMLb, TSCb, and TADb 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 D 85 and d 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 non normality were 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 TML in large models because of its favorable small-sample properties.

Summary

In summary, it is expected that TAD will perform better than TML, and that TML will be more accurate than TSC 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 TML 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.

3.4 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, non convergence 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 \( \mathbf{D} \mathbf{Y} \mathbf{b} \), where \( \mathbf{y} \) is a vector of observed variables, \( \mathbf{Y} \) matrix of factor loadings on \( k \) common factors \( \mathbf{Y}_1; \mathbf{Y}_2; \ldots; \mathbf{Y}_k \), and \( \mathbf{b} \) a vector with unique scores (measurement error), where uncorrelated with \( \mathbf{Y} \). Under the usual assumptions, the population covariance matrix of \( z \) has the form where 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 D 1,200 replications was used. Although 300 replications would have been a “reasonable trade off between precision, and the amount of 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 non normal data in future research. The population covariance matrix of these normal variables is defined by the population factor structure of the models under study: \( \hat{\Sigma} \), \( j = 1; 2; \ldots; 7 \). Both the generation of the sample data and the estimation of the models was performed using the Mplus 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 Mplus. For

\[
\begin{array}{cccccc}
 k & p & p^* & t & d & \text{Seed} \\
\hline
4 & 12 & 78 & 30 & 48 & \text{77703570} & \text{49330250} & \text{71578326} \\
6 & 18 & 111 & 51 & 120 & \text{83444508} & \text{39023988} & \text{68738111} \\
8 & 24 & 300 & 76 & 224 & \text{16159776} & \text{44724671} & \text{97116941} \\
10 & 30 & 465 & 105 & 360 & \text{71034416} & \text{06466931} & \text{85864123} \\
12 & 36 & 666 & 138 & 528 & \text{56460497} & \text{38267030} & \text{98682926} \\
14 & 42 & 903 & 175 & 728 & \text{64459199} & \text{07380304} & \text{07013316} \\
16 & 48 & 1176 & 216 & 960 & \text{48795874} & \text{79583898} & \text{23965379} \\
\end{array}
\]

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 \( \Sigma \); \( t \) the number of parameters to be estimated; \( d = p^* - t \) the number of degrees of freedom.
the mean-adjusted and mean- and variance-adjusted estimation of the chi-square statistic, the estimation option in Mplus 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 d0. 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.

3.5 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 ‘D 0:05 defined as O’2 OE0:035; 0:068 , implying that acceptable difference rates in the tables are within the range 1:5%;C1:8%.

Likelihood ratio statistic. The quantile bias of this statistic reduces with increasing sample size and decreasing model size. It can be seen that TML 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 TML 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 non normal variables. Curran et al. (1996), for example, reported empirical rejection rates of 48\% for the nominal 5\% Type I error rate when severely non normal 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.

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

**Scaled Satorra–Bentler statistic.** Like for TML, the finite sample bias of the test statistic TSC 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 TSC is slightly worse compared to that of TML. 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
It follows that the use of TSC is no option for the evaluation of large models.

### Table 2

**Empirical Minus the 5% Nominal Type I Error Rates of Nine Model Fit Statistics**

*for N = 200 (NR = 1,200)*

<table>
<thead>
<tr>
<th>k</th>
<th>TML</th>
<th>TSC</th>
<th>TAD</th>
<th>TMLB</th>
<th>TSCB</th>
<th>TADD</th>
<th>TMLS</th>
<th>TSCS</th>
<th>TADS</th>
<th>(\bar{d}(\hat{d}))</th>
<th>N: t</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3.2</td>
<td>3.8</td>
<td>1.1</td>
<td>-3</td>
<td>-3.4</td>
<td>-1.4</td>
<td>2.0</td>
<td>-2</td>
<td>-524</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>4.9</td>
<td>6.3</td>
<td>-6</td>
<td>-3.5</td>
<td>-3.5</td>
<td>-3.5</td>
<td>1.2</td>
<td>-3.1</td>
<td>170</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>9.7</td>
<td>13.2</td>
<td>-5</td>
<td>-7</td>
<td>-4.6</td>
<td>-2.7</td>
<td>27</td>
<td>-3.5</td>
<td>224 (98)</td>
<td>2.6</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>20.3</td>
<td>24.9</td>
<td>-5</td>
<td>-2.9</td>
<td>-1.7</td>
<td>-4.7</td>
<td>.8</td>
<td>32</td>
<td>-4.4</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>33.3</td>
<td>38.9</td>
<td>-3.3</td>
<td>-2.4</td>
<td>-4.9</td>
<td>2.5</td>
<td>46</td>
<td>-4.7</td>
<td>528 (136)</td>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>50.9</td>
<td>57.1</td>
<td>-3.8</td>
<td>-3.4</td>
<td>-5.0</td>
<td>2.8</td>
<td>43</td>
<td>-5.0</td>
<td>728 (149)</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>70.7</td>
<td>75.4</td>
<td>-4.5</td>
<td>-4.0</td>
<td>-5.0</td>
<td>3.2</td>
<td>69</td>
<td>-5.0</td>
<td>960 (158)</td>
<td>0.9</td>
<td></td>
</tr>
</tbody>
</table>

\(\bar{d}\) 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)*

<table>
<thead>
<tr>
<th>k</th>
<th>TML</th>
<th>TSC</th>
<th>TAD</th>
<th>TMLB</th>
<th>TSCB</th>
<th>TADD</th>
<th>TMLS</th>
<th>TSCS</th>
<th>TADS</th>
<th>(\bar{d}(\hat{d}))</th>
<th>N: t</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2.6</td>
<td>3.1</td>
<td>1.6</td>
<td>1.2</td>
<td>1.7</td>
<td>1.5</td>
<td>2.0</td>
<td>.7</td>
<td>48 (41)</td>
<td>13.3</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>3.1</td>
<td>3.8</td>
<td>.7</td>
<td>.5</td>
<td>1.1</td>
<td>-1.6</td>
<td>1.3</td>
<td>1.9</td>
<td>-1.1</td>
<td>7.8</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3.6</td>
<td>4.5</td>
<td>-1.5</td>
<td>-1.8</td>
<td>-1.6</td>
<td>-3.6</td>
<td>-1.3</td>
<td>.3</td>
<td>-3.2</td>
<td>5.3</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>6.5</td>
<td>8.3</td>
<td>-9</td>
<td>-1.1</td>
<td>-7</td>
<td>-4.0</td>
<td>1.3</td>
<td>3.3</td>
<td>360 (179)</td>
<td>3.8</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>11.4</td>
<td>14.3</td>
<td>-1.0</td>
<td>-2.0</td>
<td>-1.1</td>
<td>-4.8</td>
<td>.2</td>
<td>1.3</td>
<td>-4.6</td>
<td>2.9</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>21.0</td>
<td>22.0</td>
<td>-1.9</td>
<td>-2.8</td>
<td>-2.2</td>
<td>-5.0</td>
<td>1.4</td>
<td>2.9</td>
<td>-4.7</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>26.0</td>
<td>28.7</td>
<td>-1.7</td>
<td>-3.4</td>
<td>-2.8</td>
<td>-5.0</td>
<td>.8</td>
<td>2.1</td>
<td>-4.6</td>
<td>1.9</td>
<td></td>
</tr>
</tbody>
</table>

**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)*

<table>
<thead>
<tr>
<th>k</th>
<th>TML</th>
<th>TSC</th>
<th>TAD</th>
<th>TMLB</th>
<th>TSCB</th>
<th>TADD</th>
<th>TMLS</th>
<th>TSCS</th>
<th>TADS</th>
<th>(\bar{d}(\hat{d}))</th>
<th>N: t</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1.4</td>
<td>1.7</td>
<td>1.1</td>
<td>1.0</td>
<td>1.3</td>
<td>1.1</td>
<td>1.3</td>
<td>.7</td>
<td>48 (44)</td>
<td>26.7</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2.2</td>
<td>2.7</td>
<td>.7</td>
<td>.6</td>
<td>1.6</td>
<td>-2.6</td>
<td>1.2</td>
<td>1.6</td>
<td>-4</td>
<td>15.7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2.1</td>
<td>3.0</td>
<td>.8</td>
<td>1.3</td>
<td>-3.5</td>
<td>1.7</td>
<td>2.1</td>
<td>1.6</td>
<td>-5</td>
<td>10.5</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1.9</td>
<td>3.6</td>
<td>-1.1</td>
<td>-1.0</td>
<td>-3.3</td>
<td>-2</td>
<td>1.8</td>
<td>-1.1</td>
<td>-2.6</td>
<td>7.6</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>5.6</td>
<td>6.1</td>
<td>-1.1</td>
<td>-1.0</td>
<td>-8</td>
<td>3.6</td>
<td>.7</td>
<td>1.8</td>
<td>-2.9</td>
<td>5.8</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>5.7</td>
<td>6.6</td>
<td>-1.7</td>
<td>-1.6</td>
<td>-4.4</td>
<td>-1</td>
<td>1.3</td>
<td>-3.8</td>
<td>728 (265)</td>
<td>4.6</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>8.8</td>
<td>10.9</td>
<td>-2.1</td>
<td>-2.3</td>
<td>-1.7</td>
<td>-4.7</td>
<td>.8</td>
<td>4.2</td>
<td>960 (418)</td>
<td>3.7</td>
<td></td>
</tr>
</tbody>
</table>

**Note.** Values in the range \([-1.5, 1.8]\) are defined as acceptable and are thus printed in bold face.
Adjusted Satorra–Bentler statistic. For TAD 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 TML and TSC. For \( N = 400 \) and \( N = 800 \), TAD slightly underestimates nominal Type I error levels when the model size increases. Overall, however, the results indicate that TAD clearly outperforms TML and TSC 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 TAD has excellent Type I error control. The reason for the good performance of TAD seems to be Satterthwaite’s (1941, 1946) variance correction, which adjust the tail of the distribution of TML adequately.

In general, our expectations with respect to the behavior of the mean- and variance-adjusted test statistic TAD are not refuted. Recall that Fouladi (2000) found that TAD outperforms 12 other statistics with respect to Type I error control under various distributional conditions and for different models. Therefore, TAD 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 non normal 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 TAD 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 \( N \geq 200 \), 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 TAD, the statistics TMLb, TSCb, and TADb are slightly more influenced by model size. Interestingly, TSCb performs better than TMLb. It seems that the progressive tendency of TSC 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 TMLs is less affected by model size compared to TMLb. The statistic TMLs has appropriate rejection rates for \( N \geq 200 \) up to 10 factors. Compared to all other statistics, TMLs is less influenced by the model-size effect, especially when the sample size is 400 or 800. TSCs performs equally well compared to TSCb. TADs is clearly too conservative. Thus, it seems legitimate to use TMLs 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) TMLs, (b) TAD, and (c) TSCs or TSCb 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, FNR.x/, deviate significantly from their reference chi-square distribution, F with d degrees of freedom, the one-sample Kolmogorov–Smirnov test statistic DNR supxOE was computed. The DNR 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 DNR statistic at that 1% level equals 1:63/p1,200 = 0:047 (Massey, 1951). Nonsignificant DNR values, indicating closeness of fit, are boldfaced in the tables.

For the smallest sample size N = 200, TMLs clearly outperforms all other statistics for large models. Although significant deviations for the larger models are reported, the relatively good performance of TMLs compared to the other statistics under study is obvious. The statistic TAD does not perform well, although it produced Type I error rates close to those of TMLs. When the sample size increases to N = 400, TSCb is the second best statistic. For N = 800, TMLs and TSCs 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 $N = 200$ ($NR = 1,200$)

<table>
<thead>
<tr>
<th>$k$</th>
<th>$T_{ML}$</th>
<th>$T_{SC}$</th>
<th>$T_{AD}$</th>
<th>$T_{MLb}$</th>
<th>$T_{SCb}$</th>
<th>$T_{ADb}$</th>
<th>$d$</th>
<th>$N$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>.087</td>
<td>.110</td>
<td>.139</td>
<td>.022</td>
<td>.025</td>
<td>.070</td>
<td>.039</td>
<td>48</td>
<td>(36)</td>
</tr>
<tr>
<td>6</td>
<td>.138</td>
<td>.167</td>
<td>.203</td>
<td>.060</td>
<td>.037</td>
<td>.078</td>
<td>.013</td>
<td>111</td>
<td>120</td>
</tr>
<tr>
<td>8</td>
<td>.253</td>
<td>.295</td>
<td>.292</td>
<td>.068</td>
<td>.027</td>
<td>.100</td>
<td>.054</td>
<td>151</td>
<td>224</td>
</tr>
<tr>
<td>10</td>
<td>.368</td>
<td>.414</td>
<td>.362</td>
<td>.135</td>
<td>.076</td>
<td>.151</td>
<td>.057</td>
<td>178</td>
<td>360</td>
</tr>
<tr>
<td>12</td>
<td>.482</td>
<td>.528</td>
<td>.443</td>
<td>.195</td>
<td>.141</td>
<td>.186</td>
<td>.060</td>
<td>213</td>
<td>528</td>
</tr>
<tr>
<td>14</td>
<td>.626</td>
<td>.668</td>
<td>.516</td>
<td>.284</td>
<td>.205</td>
<td>.275</td>
<td>.099</td>
<td>230</td>
<td>728</td>
</tr>
<tr>
<td>16</td>
<td>.761</td>
<td>.800</td>
<td>.598</td>
<td>.362</td>
<td>.283</td>
<td>.301</td>
<td>.104</td>
<td>264</td>
<td>960</td>
</tr>
</tbody>
</table>

*Note.* The critical value of $D_{1,200}$ at a two-sided 1% significance level equals 0.047. Values in the range $[0.00, 0.047]$ 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 $N = 400$ ($NR = 1,200$)

<table>
<thead>
<tr>
<th>$k$</th>
<th>$T_{ML}$</th>
<th>$T_{SC}$</th>
<th>$T_{AD}$</th>
<th>$T_{MLb}$</th>
<th>$T_{SCb}$</th>
<th>$T_{ADb}$</th>
<th>$d$</th>
<th>$N$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>.086</td>
<td>.089</td>
<td>.102</td>
<td>.033</td>
<td>.044</td>
<td>.050</td>
<td>.044</td>
<td>48</td>
<td>(41)</td>
</tr>
<tr>
<td>6</td>
<td>.086</td>
<td>.102</td>
<td>.092</td>
<td>.038</td>
<td>.030</td>
<td>.033</td>
<td>.021</td>
<td>442</td>
<td>120</td>
</tr>
<tr>
<td>8</td>
<td>.145</td>
<td>.169</td>
<td>.176</td>
<td>.031</td>
<td>.016</td>
<td>.076</td>
<td>.038</td>
<td>109</td>
<td>360</td>
</tr>
<tr>
<td>10</td>
<td>.186</td>
<td>.211</td>
<td>.212</td>
<td>.070</td>
<td>.044</td>
<td>.105</td>
<td>.036</td>
<td>151</td>
<td>360</td>
</tr>
<tr>
<td>12</td>
<td>.260</td>
<td>.292</td>
<td>.291</td>
<td>.103</td>
<td>.070</td>
<td>.121</td>
<td>.034</td>
<td>190</td>
<td>360</td>
</tr>
<tr>
<td>14</td>
<td>.351</td>
<td>.385</td>
<td>.352</td>
<td>.118</td>
<td>.085</td>
<td>.164</td>
<td>.055</td>
<td>245</td>
<td>360</td>
</tr>
<tr>
<td>16</td>
<td>.428</td>
<td>.463</td>
<td>.399</td>
<td>.184</td>
<td>.138</td>
<td>.199</td>
<td>.047</td>
<td>268</td>
<td>960</td>
</tr>
</tbody>
</table>

*Note.* Values in the range $[0.00, 0.047]$ 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 $N = 800$ ($NR = 1,200$)

<table>
<thead>
<tr>
<th>$k$</th>
<th>$T_{ML}$</th>
<th>$T_{SC}$</th>
<th>$T_{AD}$</th>
<th>$T_{MLb}$</th>
<th>$T_{SCb}$</th>
<th>$T_{ADb}$</th>
<th>$d$</th>
<th>$N$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>.048</td>
<td>.055</td>
<td>.074</td>
<td>.030</td>
<td>.025</td>
<td>.043</td>
<td>.025</td>
<td>48</td>
<td>(44)</td>
</tr>
<tr>
<td>6</td>
<td>.096</td>
<td>.104</td>
<td>.109</td>
<td>.018</td>
<td>.023</td>
<td>.047</td>
<td>.037</td>
<td>061</td>
<td>224</td>
</tr>
<tr>
<td>8</td>
<td>.107</td>
<td>.096</td>
<td>.126</td>
<td>.052</td>
<td>.053</td>
<td>.072</td>
<td>.023</td>
<td>022</td>
<td>360</td>
</tr>
<tr>
<td>10</td>
<td>.135</td>
<td>.157</td>
<td>.159</td>
<td>.063</td>
<td>.054</td>
<td>.073</td>
<td>.024</td>
<td>040</td>
<td>528</td>
</tr>
<tr>
<td>12</td>
<td>.175</td>
<td>.192</td>
<td>.208</td>
<td>.072</td>
<td>.055</td>
<td>.109</td>
<td>.027</td>
<td>044</td>
<td>728</td>
</tr>
<tr>
<td>14</td>
<td>.235</td>
<td>.257</td>
<td>.268</td>
<td>.090</td>
<td>.065</td>
<td>.130</td>
<td>.057</td>
<td>143</td>
<td>960</td>
</tr>
</tbody>
</table>

*Note.* Values in the range $[0.00, 0.047]$ are defined as acceptable and are thus printed in bold face.
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 TML (Figures 1 and 2) are included because TML serves here as the reference statistic.
to illustrate the potential benefits of using TMLs (Figures 3 and 4). In addition, Figures 5 and 6 demonstrate the extremely bad distributional performance of TAD: 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 $N = 200$. 

![PP plots for $T_{MLS}$ ($N = 200$; $NR = 1200$).](image1)

![QQ plots for $T_{MLS}$ ($N = 200$; $NR = 1200$).](image2)
When comparing Figures 1 and 2 to Figures 3 and 4, the disastrous results for TML clearly emerge. Overall, TMLs has a very close approximation to the reference chi-square distribution. Therefore, we reconfirm our recommendation to use this correction of TML in applied research when large structural equation models are analyzed.

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

**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 TMLs. Therefore, we recommend using this statistic when many (approximately) multi normal 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 chisquare value for smaller models or larger sample sizes. From that perspective it would make sense to apply the correction quite generally

Software

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 non central 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 The calculation of TMLs is quite easy once the value of TML 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 TMLs are also easily computed with computer software, for example with the function `pchisq(x,d)`, where \( x \) D TMLs, 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 TMLs 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 TMLs, the value of TML 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 \( TML = 1,307 \) 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 TMLs 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 TMLs.
3.6 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 TML, 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 TMLs for TML 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 TML 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 TML. 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 TML 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 TML to be approximately chi-square. Even for our smallest model and our largest sample size (d = 48, t = 30, N D = 800), with a N: t ratio of 26.7:1, the Kolmogorov–Smirnov test for TML indicates a significant deviation from the chi-square reference distribution (see Table 7). For our second smallest model (d = 120, t = 51, and N D = 800), a N: t ratio of 15.7:1 is not large enough for proper Type I error behavior of TML at the 5% significance level (see Table 4). Also, in contrast to Fouladi (2000, p. 401), we would not conclude that TAD 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 TAD 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 TSCb, this effect is masked by the slightly more liberal tendency of TSC compared to TML. 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 _ 2:1”

The most salient conclusion of our study is that overall the Swain-corrected statistic TMLs performs best. The results in Tables 2 through 7 validate the (strong) conclusion that for the models under study, apart from single smallsample fluctuations, TMLs is robust against large model size if N: t 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 TMLs 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 non convergent 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.

3.7 Limitations and Future Work

- It is well known that non normality 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 non normality.

- 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 TMLs 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 TML 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 TML and TSC 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 (1999) 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 TML (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 TML, 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 non normal variables are analyzed. A comparison of the performance of the Bayesian approach to that of TMLs for large models would be intriguing.

3.8 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).