Performance of Bootstrapping Approaches to Model Test Statistics and Parameter Standard Error Estimation in Structural Equation Modeling

Jonathan Nevitt and Gregory R. Hancock

Department of Measurement, Statistics, and Evaluation
University of Maryland, College Park

Though the common default maximum likelihood estimator used in structural equation modeling is predicated on the assumption of multivariate normality, applied researchers often find themselves with data clearly violating this assumption and without sufficient sample size to utilize distribution-free estimation methods. Fortunately, promising alternatives are being integrated into popular software packages. Bootstrap resampling, which is offered in AMOS (Arbuckle, 1997), is one potential solution for estimating model test statistic $p$ values and parameter standard errors under nonnormal data conditions. This study is an evaluation of the bootstrap method under varied conditions of nonnormality, sample size, model specification, and number of bootstrap samples drawn from the resampling space. Accuracy of the test statistic $p$ values is evaluated in terms of model rejection rates, whereas accuracy of bootstrap standard error estimates takes the form of bias and variability of the standard error estimates themselves.

For a system of $p$ measured variables, let $\Sigma_0$ represent the true population covariance matrix underlying the variables of interest. Then, a covariance structure model represents the elements of $\Sigma_0$ as functions of model parameters with null hypothesis $H_0$: $\Sigma_0 = \Sigma(\theta)$, in which $\theta$ is a vector of $q$ model parameters. An hypothesized model may be fit to a $p \times p$ sample covariance matrix ($S$), and for any vector of model parameter estimates ($\hat{\theta}$) the hypothesized model can be used to evaluate the model implied covariance matrix, $\Sigma(\theta) = \hat{\Sigma}$. The goal in parameter estimation is to
obtain a vector of parameter estimates such that $\theta$ is as close to $\mathbf{S}$ as possible. The disparity between $\hat{\Sigma}$ and $\mathbf{S}$ is measured by a discrepancy function—the maximum likelihood (ML) function is the most commonly employed discrepancy function in structural equation modeling (SEM) and is defined as $F_{ML} = \ln|\hat{\Sigma}| - \ln|\mathbf{S}| + \text{tr}(\mathbf{S}^{-1} - \hat{\Sigma}^{-1}) - p$ (see, e.g., Bollen, 1989).

The popularity of ML estimation stems from ML’s desirable properties: ML yields unbiased, consistent, and efficient parameter estimates and provides a model test statistic $T_{ML} = (n - 1)F_{ML}$, for a sample of size $n$ evaluated at the minimum value of $F_{ML}$ for assessing the adequacy of an hypothesized model. These properties inherent in ML are asymptotic results (i.e., as the sample size increases toward infinity) that are derived from the theoretical behavior of ML under the assumption of multivariate normality. Thus, given the null hypothesis, a system of $p$ measured variables (yielding $p^* = \frac{p(p+1)}{2}$ unique variances and covariances), and $q$ model parameters to be estimated, asymptotic theory establishes that $T_{ML}$ follows a central chi-square distribution with $p^* - q$ degrees of freedom ($df$).

Whereas ML estimation rests on the assumption of multivariate normality and is based on large-sample theory, in practice researchers are commonly faced with relatively small samples clearly from nonnormal populations (Mickey, 1989). Thus, there has been considerable interest in evaluating the robustness of ML estimators and other estimation methods with respect to violations of distributional assumptions (Anderson & Gerbing, 1984; Boomsma, 1983; Browne, 1982; Chou, Bentler, & Satorra, 1991; Curran, West, & Finch, 1996; Finch, West, & MacKinnon, 1997; Harlow, 1985; Hu, Bentler, & Kano, 1992). Nonnormality appears to have little impact on model parameters estimated via ML (i.e., parameters remain relatively unbiased). However, research has demonstrated that $T_{ML}$ and parameter standard errors under ML may be substantially affected when the data are nonnormal. Specifically, under certain nonnormal conditions (i.e., with heavy-tailed leptokurtic distributions), $T_{ML}$ tends to inflate whereas parameter standard errors become attenuated (for reviews see Chou & Bentler, 1995; West, Finch, & Curran, 1995).

Fundamentally different approaches have been developed to address the problems associated with ML estimation under nonnormal conditions. Browne (1982, 1984) advanced an asymptotically distribution free (ADF) estimation method that relaxes distributional assumptions and yields a model test statistic, $T_{ADF}$. At large sample sizes ($n \geq 5,000$) $T_{ADF}$ operates as expected, yielding observed Type I error rates at the nominal level (Chou et al., 1991; Curran et al., 1996; Hu et al., 1992). However, with large models and small to moderate sample sizes, ADF has been shown to be problematic; it tends to yield high rates of nonconvergence and improper solutions when minimizing the ADF fit function (a problem much less frequently encountered when minimizing $F_{ML}$). Moreover, $T_{ADF}$ appears to reject true models too often, yielding Type I error rates as high as 68% under some sampling conditions (Curran et al., 1996).
Another approach is to adjust $T_{ML}$ and ML standard errors to account for the presence of nonzero kurtosis in the sample data. The adjustment is a rescaling of $T_{ML}$ to yield a test statistic that more closely approximates the referenced chi-square distribution (Browne, 1982, 1984). Satorra and Bentler (1988, 1994) introduced a rescaled test statistic ($T_{ML-SB}$) that has been incorporated into the EQS program (Bentler, 1996). Empirical research has demonstrated that model complexity and sample size have less effect on $T_{ML-SB}$ as compared to the ADF model test statistic (Chou & Bentler, 1995; Chou et al., 1991; Hu et al., 1992). Similar in principle to the rescaling of $T_{ML}$, Browne (1982, 1984) also formulated a scaling correction to ML standard errors for nonnormal data conditions. A variant of this correction procedure (Bentler & Dijkstra, 1985) is currently available in EQS and yields what are referred to as “robust” standard errors. The correction involves applying a scaling constant to the covariance matrix of the parameter estimates. Robust standard error estimates are then obtained by taking the square root of the elements along the main diagonal of the scaled covariance matrix.

A third approach to managing nonnormality in SEM is bootstrap resampling (i.e., establishing an empirical sampling distribution associated with a statistic of interest by repeatedly sampling from the original “parent” sample data). Efron (1979) pioneered the bootstrap in a landmark article in the late 1970s, which led to a host of publications exploring the method. With specific regard to latent variable models, recent bootstrapping investigations have surfaced within the context of exploratory and confirmatory factor analysis (CFA; Beran & Srivastava, 1985; Bollen & Stine, 1988, 1990, 1992; Boomsma, 1986; Chatterjee, 1984; Ichikawa & Konishi, 1995; Stine, 1989; Yung & Bentler, 1994, 1996). Additionally, Yung and Bentler (1996) considered the bootstrap’s potential for obtaining robust statistics in SEM. They noted that because the primary statistical concern in SEM centers on the sampling properties of parameter estimates and the model fit statistic, bootstrap methods may be a viable alternative to normal theory methods. In fact, the AMOS program (Arbuckle, 1997) is the first to offer bootstrap-derived robust statistics as an alternative to normal theory hypothesis testing methods, providing both standard errors and an adjusted model test statistic $p$ value.

When estimating standard errors using the bootstrap, a completely nonparametric approach is taken; that is, the resampling scheme for the bootstrap does not depend on any assumption regarding the distributional form of the population or on any covariance structure model for the data. For a model parameter of interest, the bootstrap-estimated standard error is calculated as the standard deviation of the parameter estimates for that model parameter across the number of bootstrap samples drawn ($B$).

Within the context of exploratory factor analysis, Chatterjee (1984) was the first to propose the use of bootstrap standard errors; subsequently, Ichikawa and Konishi (1995) conducted a full simulation study. They showed bootstrap-estimated standard errors are less biased than unadjusted ML estimates un-
der nonnormality; however, with normal data their results suggested the bootstrap did not perform as well as ML. Additionally, they found for samples of size \( n = 150 \) the bootstrap did not work well, consistently overestimating standard errors. These problems dissipated at sample sizes of \( n = 300 \). Similar work proceeded concurrently within the SEM paradigm, starting with Boomsma’s (1986) simulation evidence indicating a tendency for bootstrap standard errors within covariance structure analysis to be larger than ML standard errors under skewed data conditions. Stine (1989) and Bollen and Stine (1990) extended the use of the bootstrap to estimate the standard errors of the estimates of standardized regression coefficients, as well as of direct, indirect, and total effects. Additionally Bollen and Stine and Yung and Bentler (1996), using examples from existing data sets, provided promising evidence for the performance of the bootstrap in SEM.

For testing model fit, Bollen and Stine (1992), in work similar to that of Beran and Srivastava (1985), proposed a bootstrap method for adjusting the \( p \) value associated with \( T_{\text{ML}} \). In general, to obtain adjusted \( p \) values under the bootstrap resampling approach \( T_{\text{ML}} \) is referred to an empirical sampling distribution of the test statistic generated via bootstrap samples drawn from the original parent sample data. The bootstrap-adjusted \( p \) value is calculated as the proportion of bootstrap model test statistics that exceed the value of \( T_{\text{ML}} \) obtained from the original parent sample. Bollen and Stine noted naïve bootstrapping of \( T_{\text{ML}} \) (i.e., completely nonparametric resampling from the original sample data) for SEM models is inaccurate because the distribution of bootstrapped model test statistics follows a noncentral chi-square distribution instead of a central chi-square distribution. To adjust for this inaccuracy, Bollen and Stine formulated a transformation on the original data that forces the resampling space to satisfy the null hypothesis (i.e., making the model-implied covariance matrix the true underlying covariance matrix in the population). The transformation is of the form

\[
Z = YS^{-\frac{1}{2}}\hat{\Sigma}^{-\frac{1}{2}}
\]

in which \( Y \) is the original data matrix from the parent sample. They demonstrated analytically that \( T_{\text{ML}} \) values from bootstrap samples drawn from the transformed data matrix \( Z \) have an expectation equal to the model \( df \). In addition, they showed empirically that these values’ distribution is a reasonable approximation to a central chi-square distribution. Finally, Yung and Bentler (1996) also provide some evidence of the usefulness of the bootstrap for adjusting model test statistic \( p \) values. However, this evidence is limited to a single data set and is primarily a comparison of alternative methods for obtaining bootstrap samples.

In sum, research to date has shown the bootstrap to be a potential alternative for obtaining robust statistics in SEM. However, issues clearly remain to be addressed. First, there is only minimal evidence supporting the accuracy of the boot-
strap for estimating standard errors and model test statistic $p$ value adjustment in SEM. Yung and Bentler (1996) appropriately cautioned embracing the bootstrap with blind faith and advocated a critical evaluation of the empirical behavior of these methods. Similarly, West et al. (1995) noted that there have been no large simulation studies in SEM investigating the accuracy of the bootstrap under varied experimental conditions. As a consequence, little is currently known about the performance of the bootstrap with respect to adjusting $p$ values and estimating standard errors. Second, minimum sample size requirements for the original parent sample that defines the resampling space are rather unclear. The failure of the bootstrap with relatively small sample sizes (Ichikawa & Konishi, 1995; Yung & Bentler, 1994) suggests that the bootstrap may not be an appropriate method under such conditions. These results point to the need for a systematic examination of sample size with respect to the bootstrap in SEM. Finally, no investigations have been conducted that address the minimum $B$ required to yield accurate estimates of standard errors and adjusted $T_{ML}p$ values in SEM. Yung and Bentler (1996) noted that an ideal bootstrapping paradigm would use $B = n^n$ (i.e., all possible samples of size $n$ drawn from the resampling space). Unfortunately, this collection would (currently) prove too large for practical implementation and would contain improper samples with singular covariance matrices not suitable for fitting a covariance structure model. In practice, $B$ is set to a large number (say, hundreds) to generate an approximate empirical sampling distribution of a statistic of interest; however, no guidelines exist to establish an appropriate minimum number of bootstrap samples for obtaining accurate results with respect to standard errors and model test statistics in SEM.

Based on previous findings, we anticipate results in this study to show inflation in the unadjusted $T_{ML}$ and attenuation in ML standard errors under the leptokurtic nonnormal conditions investigated here. With respect to bootstrap-estimated standard errors, results from Boomsma (1986) and Ichikawa and Konishi (1995) lead to an expectation that bootstrap-estimated standard errors will resist suppression under nonnormal conditions at sufficient sample sizes but may become too large at smaller sample sizes. Unfortunately, no theory or empirical evidence exists to drive expectations with respect to the relative performance of bootstrap-estimated standard errors as compared to EQS-robust standard errors. For bootstrap-adjusted $T_{ML}p$ values, evidence provided by Bollen and Stine (1992) leads to anticipation that under nonnormal conditions the bootstrap adjustment may yield more appropriate $p$ values than the unadjusted $T_{ML}$. However, no strong theory exists to provide expectations regarding the relative performance of the bootstrap-adjusted $T_{ML}p$ value against $p$ values obtained from other robust methods such as the scaled $T_{ML-SB}$.

The purpose of this study is to provide a systematic investigation of these issues surrounding the bootstrap in SEM. Specifically, a Monte Carlo simulation is used to evaluate the bootstrap with respect to $T_{ML}p$ value adjustment and estimation of parameter standard errors under varying data distributions, sample sizes, and
model specifications, as well as using differing numbers of bootstrap samples in the investigation.

**METHOD**

Model specifications, distributional forms, sample sizes, and the number of replications per population condition established in this study are the same as those used by Curran et al. (1996). We chose to carefully replicate their population conditions so that our results for the bootstrap (with respect to assessing model fit) may be directly compared to the results reported by Curran et al. for the ADF model test statistic and $T_{ML-SB}$. The $T_{ML-SB}$ model fit statistic was also collected in this investigation to replicate the results found by Curran et al.

**Model Specifications**

The base underlying population model in this study is an oblique CFA model with three factors, each factor having three indicator variables. Population parameter values are such that all factor variances are set to 1.0, all factor covariances and correlations are set to .30, all factor loadings are set to .70, and all error variances are set to .51, thereby yielding unit variance for the variables.

Two model specifications for fitting sample data were examined in this investigation: a correctly specified model and a misspecified model. For the correctly specified model, simulated samples of data were drawn from the base population model as described previously and then fit in AMOS and EQS using the specification for the base population model. For generating simulated samples of data for the improperly specified model, a variant of the base population model was established to include two population cross-loadings: $\lambda_{72} = \lambda_{63} = .35$. Simulated samples of data drawn from this new population model were then fit in AMOS and EQS to the base population model specification, which omitted the variable-factor cross-loadings, thus creating errors of exclusion.1

When fitting sample data to models, model identification was established by estimating the three factor variances and fixing one factor loading to 1.0 for each factor ($\lambda_{11}, \lambda_{42}, \lambda_{93}$). This approach to model identification was chosen (i.e., rather than fixing the factor variances to 1.0 and estimating all factor loadings) to ensure stability of the parameter estimates in the bootstrap samples. As noted by Arbuckle (1997) and illustrated by Hancock and Nevitt (1999), if model identification is

---

1A population analysis was conducted to determine the power of the model misspecification under multivariate normal data conditions (see Bollen, 1989, pp. 338–349 for a review) yielding power of .892, .996, 1.000, and 1.000 at sample sizes of $n = 100, 200, 500$, and 1,000, respectively.
achieved by fixing the factor variances, then the criterion for minimizing the model fit function may yield parameter estimates that are unique only up to a sign change. Although the choice of approach to model identification is irrelevant in most applied settings, it has great importance with respect to bootstrap resampling. In bootstrapping, if the signs of some of the parameter estimates are arbitrary, these estimates could potentially vary from bootstrap sample to bootstrap sample (some positive and some negative), thereby causing the resulting bootstrap standard errors to become artificially inflated. To avoid this problem, we fixed a factor loading and estimated the factor variance to establish identification for each factor in the CFA models.

Distributional Forms and Data Generation

Three multivariate distributions were established through the manipulation of univariate skewness and kurtosis. All manifest variables were drawn from the same univariate distribution for each data condition. Distribution 1 is multivariate normal with univariate skewness and kurtosis both equal to 0. Distribution 2 represents a moderate departure from normality with univariate skewness of 2.0 and kurtosis of 7.0. Distribution 3 is severely nonnormal (i.e., extremely leptokurtic) with univariate skewness of 3.0 and kurtosis 21.0. Curran et al. (1996) reported that these levels of nonnormality are reflective of real data distributions found in applied research.

Simulated raw data matrices were generated in GAUSS (Aptech Systems, 1996) to achieve the desired levels of univariate skewness, kurtosis, and covariance structure. Multivariate normal and nonnormal data were generated via the algorithm developed by Vale and Maurelli (1983), which is a multivariate extension of the method for simulating nonnormal univariate data proposed by Fleishman (1978). Programming used in this study to generate simulated data has been tested and verified for accuracy; the programs were scrutinized externally and are now available to the research community (Nevitt & Hancock, 1999).

Design

Three conditions were manipulated in this investigation: model specification (two model types), distributional form of the population (three distributional forms), and sample size (n = 100, 200, 500, and 1,000). The three manipulated conditions were completely crossed to yield 24 population conditions. Two hundred random sam-

---

2 Note that we define normality, as is commonly done in practice, by using a shifted kurtosis value of 0 rather than a value of 3.
amples (i.e., raw data matrices) were analyzed in each of the population conditions. To investigate the accuracy of the bootstrap with respect to the number of bootstrap samples drawn, each simulated data set was repeatedly modeled in AMOS using $B = 250, 500, 1,000,$ and $2,000$ bootstrap samples drawn from the original sample. All bootstrap samples were mutually independent of one another (i.e., none of the bootstrap samples drawn for a particular value of $B$ were used as samples for any other value of $B$). Additionally, bootstrap samples drawn to obtain the adjusted $T_{ML} p$ values were completely independent of the bootstrap samples drawn to obtain estimated bootstrap standard errors.3

Model Fittings and Data Collection

For each simulated data matrix, models were fit in AMOS (Arbuckle, 1997) to obtain $T_{ML}$ with its associated $p$ value, ML parameter standard errors, bootstrap-adjusted $T_{ML} p$ values, and bootstrap-estimated standard errors. Bootstrap samples drawn from each simulated data matrix were sampled with replacement and were of the same size as the original simulated data matrix. The AMOS program automatically discards unusable bootstrap samples and continues resampling until the target number of usable bootstrap samples has been achieved.

In addition to modeling each simulated data matrix in AMOS, each data matrix was also analyzed within EQS 5.4 (Bentler, 1996). Raw simulated data were input into both AMOS and EQS because bootstrapping and rescaling (for $T_{ML-SB}$ and EQS-robust standard errors) require raw data; the associated sample covariance matrix, rather than the correlation matrix, was modeled in each program. Start values for modeling simulated data were established using the default values provided by AMOS and EQS. The maximum number of iterations to convergence for each model fitting was set to 200. This maximum was established for modeling each simulated parent data set in AMOS and EQS and was also established for modeling bootstrap samples in AMOS. Any simulated data matrix that failed to converge or yielded an improper solution in either AMOS or EQS was discarded and replaced with a replicate yielding a convergent proper solution.

From EQS, the $T_{ML}$ and $T_{ML-SB}$ test statistics with associated $p$ values and ML and robust standard errors were collected. For each simulated data matrix, the $T_{ML}$ test statistic obtained from EQS was carefully screened against the $T_{ML}$ obtained from AMOS, with roughly 98% of the replications yielding ML test statistics that were within .01 of one another. Data matrices that yielded $T_{ML}$ values from the two programs that differed by more than .01 were discarded and replaced. A total of 58

---

3 For each of the 200 replications per population condition, each data set was modeled independently in AMOS a total of eight times—four times using $B = 250, 500, 1,000,$ and $2,000$ to obtain adjusted $p$ values, and four times using $B = 250, 500, 1,000,$ and $2,000$ to obtain estimated bootstrap standard errors.
data matrices were replaced out of the 4,800 total data matrices, 41 of which were from the nonnormal data conditions under the misspecified model.

**Summary Measures**

In the first part of this study, $T_{ML}$ and its bootstrap alternatives are evaluated in terms of their model rejection rates (i.e., the proportion of replications in which the target model is rejected at the $\alpha = .05$ level of significance). For the properly specified model we consider Type I error rates, and for the misspecified model we consider the power to reject an incorrectly specified model. Model rejection rates are considered here, rather than bias in the model test statistics, because the AMOS program only provides a bootstrap adjusted $T_{ML}$ $p$ value for analysis, unlike EQS, which provides $T_{ML-SB}$ that can be compared against an expected value for calculating estimator bias.

In the second part of the study, parameter standard error estimates are assessed in terms of bias and standard error variability. Bias is an assessment of standard errors relative to a true standard error. Two different approaches were taken to obtain or estimate the true standard error for each model parameter under each population condition. For conditions in which the underlying distributional form was multivariate normal, true standard errors were obtained via a population analysis, modeling the population covariance matrix and specifying the sample size for that research condition. For conditions in which the distributional form was not multivariate normal, true standard errors for model parameters could not be obtained using a population analysis because ML estimation assumes a multivariate normal distribution in the population. Instead, true standard errors were approximated empirically using a Monte Carlo simulation, independent of the samples drawn for the main part of our investigation. In this case, true standard errors were estimated using the standard deviation of 2,000 parameter estimates drawn from samples from the original population covariance matrix under a given population condition.

To verify the accuracy of estimating true standard errors via Monte Carlo simulation, we also estimated true standard errors using the Monte Carlo approach for the multivariate normal conditions. For the multivariate normal conditions, a comparison of true standard errors obtained via population analysis against estimated true standard errors via Monte Carlo simulation showed reasonable agreement between the two methods. True standard errors are presented in Appendixes A and B for the two methods for the correctly specified model, using the factor covariance $\phi_{21}$ and the loading $\lambda_{21}$ parameters as exemplars. True standard errors for the two methods tended to deviate from one another only as a function of sample size, with decreasing sample size generally yielding increasing discrepancy between the two methods, as expected. The discrepancy between Monte Carlo and population true standard errors was never more than 2.5% of the population standard error for $\phi_{21}$, and never more than 8.5% for $\lambda_{21}$.
The accuracy of parameter standard errors is evaluated using percentages of bias relative to an estimator’s appropriate true standard error. Relative bias percentages are computed on each case for each standard error estimate as

\[ \% \text{bias} = \left( \frac{\hat{\theta} - \theta}{\theta} \right) \times 100\% \]

with \( \hat{\theta} \) representing an estimated standard error for a given model parameter, and \( \theta \) representing the relevant true parameter standard error. Examining raw bias in the standard errors proved difficult because true standard errors differed substantially across population conditions. Inspecting percentages of relative bias, rather than raw bias, places the performance of standard error estimates on a more common metric. Moreover, Muthén, Kaplan, and Hollis (1987) provided a benchmark by which to assess standard error relative bias, suggesting that relative bias values less than 10% of the true value can be considered negligible in SEM. The data are summarized via the means and standard deviations of the relative bias percentages, examining standard error central tendency and variability across the 200 replications for each estimation method under each study condition.

Convergence Rates and Unusable Bootstrap Samples

All simulated data matrices yielded converged model fittings in both AMOS and EQS. Simulated data matrices that yielded improper solutions in either program often led to a discrepancy in the \( T_{\text{ML}} \) test statistic between AMOS and EQS that exceeded 0.01 and were thus discarded and replaced. In total, 2.4% of the simulated data sets were discarded and replaced; approximately 71% of these discarded data sets were from the misspecified model under nonnormal distributions. Bootstrap samples that did not converge to a solution within 200 iterations were considered unusable and were discarded automatically by the AMOS program. For diagnostic purposes, the number of unusable bootstrap samples was monitored for the \( B = 2,000 \) bootstrap sample standard error estimator. The frequency of unusable bootstrap samples appeared to be a function of sample size, distributional form, and model specification, as would be expected. For the larger sample sizes that we investigated (\( n = 500 \) and \( n = 1,000 \)) there were no bootstrap samples that were unusable. The most extreme levels of unusable bootstrap samples were found in the \( n = 100 \), nonnormal, and misspecified model conditions. The largest percentage of unusable bootstrap samples under these conditions was 8.3%. Again, all unusable bootstrap samples were automatically discarded and replaced by the AMOS program, which continues to draw bootstrap samples until the target number of usable bootstrap samples has been reached.
RESULTS

Model Rejection Rates

For evaluating the performance of the model test statistics under the properly specified model, a quantitative measure of robustness as suggested by Bradley (1978) was utilized. Using Bradley’s liberal criterion, an estimator is considered robust if it yields an empirical model rejection rate within the interval \([.5\alpha, 1.5\alpha]\). Using \(\alpha = .05\), this interval for robustness of rejection rates is \([.025, .075]\). Note that this interval is actually narrower than an expected 95% confidence interval, which evaluates to \([.0198, .0802]\) given the 200 replications per condition in this investigation and \(\alpha = .05\) (i.e., 
\[
.05 \pm 1.96\left(\frac{0.05 \times 0.95}{200}\right)^{1/2}
\].

For both the properly specified and misspecified models, Table 1 presents model rejection rates based on \(T_{\text{ML}}, T_{\text{ML-SB}},\) and the Bollen and Stine bootstrap adjusted \(p\) values. The ML test statistic obtained from EQS is reported in the table, rather than \(T_{\text{ML}}\) from AMOS, because the corresponding \(T_{\text{ML-SB}}\) is a scaled form of the EQS \(T_{\text{ML}}\) test statistic. Model rejection rates for \(T_{\text{ML}}\) from AMOS never differed from parallel EQS \(T_{\text{ML}}\) results by more than .5%. For the properly specified model, rejection rates falling outside the robustness interval are shown in boldface type.

Results in Table 1 for the ML estimator are consistent with findings in previous research (Chou & Bentler, 1995; Curran et al., 1996). Under the multivariate normal distribution and properly specified model conditions, rejection rates for \(T_{\text{ML}}\) are within the criterion for robustness even at the smallest sample size. With departures from multivariate normality, however, \(T_{\text{ML}}\) is not robust even under the largest sample sizes, with percentages of model rejections ranging from about 20% to about 40%. Model rejection rates in Table 1 associated with \(T_{\text{ML-SB}}\) match up very well with parallel results from Curran et al (1996). Under the multivariate normal distribution model, rejection rates for \(T_{\text{ML-SB}}\) are within the robustness interval at \(n \geq 200\), and only marginally above the .075 upper bound at \(n = 100\). Under nonnormal conditions \(T_{\text{ML-SB}}\) maintains control of Type I error rates given adequate sample size; the estimator appears to be robust at \(n \geq 200\) under the moderately nonnormal distribution and at \(n \geq 500\) under the severely nonnormal distribution. On the other hand, the Bollen and Stine bootstrap adjusted \(p\) values in Table 1 reflect model rejection rates under the properly specified model that are within the robustness interval under nearly every condition, even with the combination of extreme departures from multivariate normality and the smallest sample sizes.

To understand better the trends in Table 1, logistic regression analyses were conducted on the cells in the properly specified model and on the cells in the misspecified model. Two sets of such analyses were conducted, the first examining the issue of number of bootstrap replications (results from this analysis are de-
# TABLE 1
Proportion of Model Rejections (Using 200 Replications) for $T_{ML}$, $T_{ML-SB}$, and Bootstrap Estimators

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$n$</th>
<th>$T_{ML}$</th>
<th>$T_{ML-SB}$</th>
<th>Bootstrap</th>
<th>$T_{ML}$</th>
<th>$T_{ML-SB}$</th>
<th>Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>250</td>
<td>500</td>
<td>1000</td>
<td>2000</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>.055</td>
<td>.080</td>
<td>.035 .035</td>
<td>.035 .030</td>
<td>.500 .520</td>
<td>.305 .315</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>.070</td>
<td>.070</td>
<td>.030 .040</td>
<td>.040 .040</td>
<td>.830 .850</td>
<td>.785 .790</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>.060</td>
<td>.065</td>
<td>.050 .050</td>
<td>.050 .050</td>
<td>1.000 1.000</td>
<td>1.000 1.000</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>.075</td>
<td>.070</td>
<td>.055 .060</td>
<td>.065 .070</td>
<td>1.000 1.000</td>
<td>1.000 1.000</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>.240</td>
<td>.100</td>
<td>.035 .035</td>
<td>.040 .035</td>
<td>.700 .400</td>
<td>.235 .225</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>.200</td>
<td>.065</td>
<td>.045 .040</td>
<td>.040 .035</td>
<td>1.000 .990</td>
<td>.990 .990</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>.220</td>
<td>.040</td>
<td>.045 .040</td>
<td>.040 .040</td>
<td>1.000 1.000</td>
<td>1.000 1.000</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>.300</td>
<td>.120</td>
<td>.025 .025</td>
<td>.030 .035</td>
<td>.770 .490</td>
<td>.165 .170</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>.400</td>
<td>.110</td>
<td>.060 .055</td>
<td>.055 .055</td>
<td>.930 .680</td>
<td>.515 .500</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>.360</td>
<td>.040</td>
<td>.025 .020</td>
<td>.025 .030</td>
<td>1.000 .990</td>
<td>.945 .940</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>.370</td>
<td>.035</td>
<td>.025</td>
<td>.035 .035</td>
<td>1.000 1.000</td>
<td>1.000 1.000</td>
</tr>
</tbody>
</table>

**Note.** Distribution 1 is multivariate normal, Distribution 2 is moderately nonnormal, and Distribution 3 is severely nonnormal. Model rejection rates for the properly specified model that lie outside the [.025, .075] robustness interval are shown in boldface type.
scribed hereafter but not tabled), and the second focusing on differences among the
three estimation methods across the study conditions (Table 2 gives numerical re-
sults for this analysis). Consider first the four bootstrap replication conditions,
which are crossed with three distribution and four sample size conditions. Sepa-
rately for the properly specified and misspecified models, a logistic regression
analysis was conducted using as the dependent variable the retain/reject (0/1) deci-
sion for all 9,600 outcomes across the 48 relevant cells. In each analysis the categ-
orical predictor was distribution (normal, moderately nonnormal, and severely
nonnormal), whereas the sample size variable (100, 200, 500, and 1,000) and the
number of bootstrap replications (250, 500, 1,000, and 2,000) were both treated as
continuous predictors within the analysis. It should be noted that, because of the
inability of SPSS 9.0 (SPSS, Inc., 1999) to accommodate repeated measures in lo-
gistic regression and loglinear modeling subroutines, the repeated-measure boot-
strap variable was treated as a between-cases variable. However, due to the
enormous sample size, the anticipated loss in power as a result of treating the data
as completely between-cases should be infinitesimal. Thus, for each analysis (i.e.,
for the properly specified model and for the misspecified model), there were 200
replications for three distributions crossed with four sample sizes crossed with
four bootstrap methods (200 × 3 × 4 × 4 = 9,600 cases).

The key element of interest in the analyses focusing on the number of bootstrap
replications is the role the bootstrap variable plays in predicting the model rejec-
tions for both the properly specified and misspecified models. In both cases, a like-
lihood ratio criterion was used to forward select predictors (and possibly their two-
and three-way interactions) into the regression model. For the properly specified
model, only the distribution by sample size interaction was selected as a viable
predictor, whereas for the misspecified model the selected predictors were sample
size, distribution by sample size interaction, and distribution, in order, respec-
tively. Notice that in neither case did the number of bootstrap replications have a

<table>
<thead>
<tr>
<th>TABLE 2</th>
<th>Predictive Contribution of Forward Selected Design Variables to Model Retentions and Rejections (n = 7200 Cases)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \Delta G^2 )</td>
</tr>
<tr>
<td>Properly specified model</td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>444.201</td>
</tr>
<tr>
<td>Method × Distribution</td>
<td>224.657</td>
</tr>
<tr>
<td>Misspecified model</td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>2500.494</td>
</tr>
<tr>
<td>Method</td>
<td>292.552</td>
</tr>
<tr>
<td>Method × Distribution</td>
<td>90.918</td>
</tr>
<tr>
<td>n × Distribution</td>
<td>34.660</td>
</tr>
<tr>
<td>Distribution</td>
<td>7.103</td>
</tr>
</tbody>
</table>
predictive impact, alone or interactively. Although one cannot prove a null condition, this result at least supports the notion that the number of bootstrap replications beyond $B = 250$ seems to be irrelevant in terms of model rejections. An inspection of Table 1 bears this out. The results for the Bollen and Stine bootstrap $p$ values in Table 1 indicate only small and apparently trivial differences when comparing the model rejection rates for the varying number of bootstrap resamplings. This pattern of consistency in the performance across the levels of $B$ is seen under both model specifications and under all distributional forms and sample sizes. Based on the results of the logistic regression analysis then, the small reported differences between the bootstrap estimators in Table 1 are interpreted as being attributable to sampling variability, rather than due to any systematic effect of $B$ on model rejection decisions.

The next phase of logistic regression analyses examined differences in model rejection rates as a function of estimation method, distribution, and sample size, for the correctly specified model and the misspecified model. Again, the dependent variable was the retain/reject (0/1) decision for all of the outcomes in the relevant cells. The two categorical predictors were distribution (normal, moderately nonnormal, and severely nonnormal) and method (ML, Satorra-Bentler, and bootstrap), and the sample size variable (100, 200, 500, 1,000) was treated as a continuous predictor within the logistic regression analyses. Regarding the second categorical predictor, method, only the $B = 2,000$ bootstrap data were used as they were deemed representative of the other bootstrap conditions given the prior results. Also, as before, the method repeated-measure variable was treated as a between-cases variable with no anticipated loss of interpretability. Thus, for the analysis of each model, there were 200 replications for three distributions crossed with three methods crossed with four sample sizes ($200 \times 3 \times 3 \times 4 = 7,200$ cases).

For both the properly specified model and the misspecified model, the likelihood ratio criterion was used to forward select variables (and possibly their two- and three-way interactions) into the regression model. Results are presented in Table 2, including $\Delta G^2$ (i.e., change in $–2\text{log likelihood}$) and $\Delta df$, the $p$ value for $\Delta G^2$, the Nagelkerke-adjusted $R^2$ (Nagelkerke, 1991), and $\Delta R^2$. Because of the tremendous statistical power to detect even miniscule effects, the $R^2$ and $\Delta R^2$ measures are offered to facilitate an assessment of the practical value of any main effect and interaction predictors. It should be noted, however, that because so many models are retained in the case of the properly specified model and because so many models are rejected in the case of the misspecified model, the $R^2$ values themselves are not be terribly high. Thus, it is the $\Delta R^2$ values that are most useful.

For the properly specified model, the only two statistically significant ($p < .05$) predictors are the method main effect and interaction of method with distribution. An inspection of the model rejection rates in Table 1 corroborates this finding. Notice in Table 1 that model rejection rates under the properly specified model are
generally largest for $T_{\text{ML}}$, smallest for the bootstrap, with $T_{\text{ML-SB}}$ yielding rejection rates intermediate to those for $T_{\text{ML}}$ and the bootstrap. This pattern of results becomes more pronounced with increasing departures from multivariate normality, a manifestation of the interactive nature of method with distribution. Also, model rejection rates in Table 1 reflect the lack of significance of the other effects and interactions in the logistic regression analysis. Note in Table 1 that, for the properly specified model, rejection rates are mostly constant across sample sizes and distributional forms, with no patterns in the table indicating the presence of any higher order interactions.

Turning to the misspecified model, several statistically significant predictors emerged from the logistic regression analysis. Foremost among them is $n$, which would be expected given that increases in sample size lead to increases in power, as reflected in greater rejection rates. An inspection of Table 1 quickly bears this out. As with the correctly specified model, the next strongest predictors are method and the interaction of method by distribution. Again, this is not unexpected because as evidenced in Table 1, $T_{\text{ML}}$ yields overall the largest rates of model rejections, with the bootstrap yielding the smallest and $T_{\text{ML-SB}}$ intermediate to $T_{\text{ML}}$ and the bootstrap. Clearly this pattern of model rejection rates becomes more pronounced with increasing departures from normality, again indicating the significant interaction between method and distributional form of the data. It should be noted, however, that model rejection results for the misspecified model under nonnormal conditions must be evaluated with caution. The apparent advantage in power exhibited by ML is largely an artifact of the inflation of $T_{\text{ML}}$ in the presence of nonnormality, as reflected in ML’s extremely liberal Type I error rates with the properly specified model. Lastly, although the sample size by distribution interaction and overall distribution effects are reported as statistically significant, notice also in Table 2 that their contribution to the change in $R^2$ is extremely small. Again this result is consistent with model rejection rates reported in Table 1; for the misspecified model, rejection rates change only minimally across the three distributional forms, a pattern of results that appears to be consistent across sample sizes within each distribution.

Factor Covariance Standard Errors

Four types of parameters exist in the models under investigation: factor covariance, loading, factor variance, and error variance. Although the standard error behavior for all parameter types was monitored in the full investigation, only results regarding factor covariances and loadings are presented. For the other types of parameters, factor variances and error variances, applied settings rarely find their parameter value estimation or significance testing (requiring standard errors) of
substantive interest. For this reason, analysis of the standard errors associated with these parameters is not reported here.

Data were collected for two of the three factor covariances in the population model, $\phi_{21}$ and $\phi_{32}$, which would be expected to behave identically in the properly specified model but not necessarily so in the misspecified model. However, inspection of summary results associated with the standard errors for these covariances show the two covariances behave nearly identically to each other and almost identically across both model specifications. Because of this consistency across models and covariances, the results for $\phi_{21}$ under the properly specified model are presented here to characterize the general behavior of the covariances in the two types of models.

Table 3 presents the mean and standard deviation relative bias percentages in the $\phi_{21}$ standard errors for the ML, EQS-robust, and bootstrap estimation methods. As with $T_{ML}$, the ML standard errors obtained from EQS are reported in the table rather than those from AMOS because the corresponding EQS-robust standard errors are a scaled form of the EQS ML standard errors. Companion ML standard errors obtained from AMOS were carefully monitored and yielded nearly identical summary patterns as compared to the EQS ML standard errors.

The relative bias percentages in Table 3 reveal some noteworthy tendencies in the standard error estimators. Under the multivariate normal condition mean bias is quite low for ML and EQS-robust standard errors ($< 1.2\%$) at all sample sizes, whereas bootstrap standard errors under normality yield a pattern of decreasing mean bias with increasing sample size. At the smallest sample size of $n = 100$, bootstrap standard errors yielded positive average relative bias (i.e., a tendency to be inflated) with bias percentages of 10\% to 11\%; mean relative bias in the bootstrap standard errors drops down to less than 1\% at the largest sample size of $n = 1,000$.

Under nonnormal conditions ML yields large negative mean bias in the $\phi_{21}$ standard error, with bias percentages in Table 3 for ML ranging from 13\% to 33\%. This result is consistent with previous research that has demonstrated that ML standard errors become attenuated with departures from multivariate normality (see West et al., 1995, for a review). In contrast to ML, EQS-robust and bootstrap standard errors exhibit considerably smaller percentages of average relative bias under nonnormal conditions. EQS-robust standard errors appear to resist attenuation under nonnormality at $n \geq 500$; mean relative bias is as high as 17\% under the moderately nonnormal distribution and smaller sample sizes, and around 20\% at small $n$ with severely nonnormal data. On the other hand, bootstrap standard errors remain relatively unbiased across all nonnormal data conditions and sample sizes (with a notable exception at $n = 200$ and severely nonnormal data).\footnote{Anomalous results in the patterns of means and standard deviations in Tables 3 and 4 were addressed by carefully screening the raw standard error data for outliers and influential observations. Some large positive relative bias percentages were found for all estimation methods under various study conditions, mostly without pattern but with a few exceptions. More large bias percentages tended to be present un-}
when comparing the bootstrap standard errors against one another, one finds very little difference in mean relative bias percentages across the four levels of $B$. Thus, from the perspective of average bias, there appears to be no real advantage to drawing numbers of bootstrap samples beyond $B = 250$.

With respect to the variability of the standard errors associated with $\phi_{21}$, the standard deviations, for all estimators, generally tend to decrease with increasing sample size under normal and nonnormal conditions. Such results indicate that standard error estimates tend to become relatively more stable as $n$ increases, as expected. Comparatively, the ML estimator yields smaller standard error variability than either EQS-robust or bootstrap standard errors, a pattern of results that holds across the three distributional forms; however, recall that for nonnormal conditions ML yields considerably attenuated standard error estimates. Also, under every study conditions EQS-robust standard errors exhibited less variability than bootstrap standard error estimates. Finally, as with the mean relative bias percentages, comparing the bootstrap standard errors against one another shows little difference in the variability of the standard errors with varying values of $B$.

Variable-Factor Loading Standard Errors

Data were collected for the variable-factor loadings $\lambda_{21}$, $\lambda_{52}$, and $\lambda_{62}$. As with the factor covariances, patterns of results across the two models for each of the factor loadings were quite similar, as were the patterns of results from one factor loading to another. Thus, for simplicity, only the results for the $\lambda_{21}$ loading standard errors under the properly specified model are presented here.

Table 4 presents the means and standard deviations for relative bias percentages in the standard errors for the $\lambda_{21}$ parameter. Under the normal distribution ML and EQS-robust standard errors show only marginal average bias (<2.5%). For the bootstrap standard errors, mean bias is negligible at $n \geq 200$ but becomes large and positive, jumping to about 70% at the $n = 100$ sample size.

Like the factor covariance parameter, notice again that increasing departures from multivariate normality lead to increased negative bias in the ML standard errors for the loading parameter. ML standard errors show about 30% attenuation under the moderately nonnormal condition and about 50% attenuation under the severely nonnormal condition. Unlike ML, EQS-robust and bootstrap standard error estimates appear to show some resistance to standard error suppression, given
<table>
<thead>
<tr>
<th>Distribution</th>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>1.13</td>
<td>19.21</td>
<td>-0.37</td>
<td>21.30</td>
<td>11.22</td>
<td>24.44</td>
<td>10.38</td>
<td>24.04</td>
<td>10.20</td>
<td>23.77</td>
<td>10.29</td>
<td>23.83</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.72</td>
<td>14.54</td>
<td>-0.52</td>
<td>15.91</td>
<td>3.28</td>
<td>17.69</td>
<td>3.55</td>
<td>17.71</td>
<td>3.58</td>
<td>17.67</td>
<td>3.38</td>
<td>17.53</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.91</td>
<td>8.40</td>
<td>1.12</td>
<td>10.19</td>
<td>2.44</td>
<td>11.43</td>
<td>2.89</td>
<td>10.87</td>
<td>2.92</td>
<td>10.68</td>
<td>2.58</td>
<td>10.49</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>-0.43</td>
<td>6.59</td>
<td>-0.70</td>
<td>7.54</td>
<td>-0.14</td>
<td>9.49</td>
<td>0.17</td>
<td>9.38</td>
<td>-0.08</td>
<td>8.90</td>
<td>-0.12</td>
<td>8.14</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>-25.49</td>
<td>23.45</td>
<td>-17.19</td>
<td>47.75</td>
<td>-6.75</td>
<td>52.38</td>
<td>-7.16</td>
<td>52.25</td>
<td>-7.34</td>
<td>52.10</td>
<td>-7.05</td>
<td>52.53</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>-18.77</td>
<td>15.88</td>
<td>-11.03</td>
<td>29.35</td>
<td>-7.18</td>
<td>32.24</td>
<td>-6.88</td>
<td>33.20</td>
<td>-6.76</td>
<td>33.32</td>
<td>-6.64</td>
<td>34.11</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>-17.49</td>
<td>10.81</td>
<td>-4.60</td>
<td>23.18</td>
<td>-2.30</td>
<td>24.76</td>
<td>-2.39</td>
<td>24.06</td>
<td>-2.20</td>
<td>(24.13</td>
<td>-2.32</td>
<td>24.31</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>-13.28</td>
<td>7.50</td>
<td>3.02</td>
<td>16.99</td>
<td>4.30</td>
<td>17.13</td>
<td>4.19</td>
<td>17.19</td>
<td>4.11</td>
<td>17.53</td>
<td>3.96</td>
<td>17.64</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>-32.54</td>
<td>25.90</td>
<td>-19.74</td>
<td>56.90</td>
<td>-9.64</td>
<td>63.06</td>
<td>-9.44</td>
<td>63.27</td>
<td>-9.32</td>
<td>63.12</td>
<td>-9.05</td>
<td>63.58</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>-33.44</td>
<td>18.07</td>
<td>-21.31</td>
<td>38.01</td>
<td>-15.91</td>
<td>40.81</td>
<td>-16.26</td>
<td>40.46</td>
<td>-16.41</td>
<td>40.17</td>
<td>-16.44</td>
<td>40.33</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>-28.10</td>
<td>13.37</td>
<td>-5.00</td>
<td>54.69</td>
<td>-2.92</td>
<td>60.29</td>
<td>-2.58</td>
<td>59.24</td>
<td>-2.56</td>
<td>59.93</td>
<td>-2.36</td>
<td>59.62</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>-28.99</td>
<td>10.00</td>
<td>-5.39</td>
<td>25.91</td>
<td>-4.24</td>
<td>27.56</td>
<td>-4.22</td>
<td>27.26</td>
<td>-4.11</td>
<td>27.16</td>
<td>-4.07</td>
<td>27.11</td>
</tr>
</tbody>
</table>

**Note.** Values in the table are for the correctly specified model. Distribution 1 is multivariate normal, Distribution 2 is moderately nonnormal, and Distribution 3 is severely nonnormal.
### TABLE 4
Mean and Standard Deviation Relative Bias Percentages (Using 200 Replications) in Standard Errors for Parameter $\lambda_{21}$

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$n$</th>
<th>$M$</th>
<th>$SD$</th>
<th>$M$</th>
<th>$SD$</th>
<th>$M$</th>
<th>$SD$</th>
<th>$M$</th>
<th>$SD$</th>
<th>$M$</th>
<th>$SD$</th>
<th>$M$</th>
<th>$SD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>1.72</td>
<td>32.33</td>
<td>–0.68</td>
<td>34.26</td>
<td>66.49</td>
<td>168.34</td>
<td>68.86</td>
<td>169.90</td>
<td>69.03</td>
<td>159.06</td>
<td>73.01</td>
<td>154.97</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>–0.41</td>
<td>18.19</td>
<td>–2.32</td>
<td>19.20</td>
<td>7.35</td>
<td>25.50</td>
<td>7.42</td>
<td>25.04</td>
<td>7.54</td>
<td>24.80</td>
<td>7.51</td>
<td>24.46</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>–0.35</td>
<td>12.55</td>
<td>–0.94</td>
<td>12.90</td>
<td>2.27</td>
<td>15.69</td>
<td>2.48</td>
<td>15.11</td>
<td>2.43</td>
<td>14.70</td>
<td>2.36</td>
<td>14.65</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>1.09</td>
<td>9.43</td>
<td>0.91</td>
<td>9.94</td>
<td>2.35</td>
<td>11.64</td>
<td>2.10</td>
<td>11.23</td>
<td>2.19</td>
<td>10.79</td>
<td>2.24</td>
<td>10.80</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>–29.54</td>
<td>35.49</td>
<td>–15.03</td>
<td>44.63</td>
<td>78.68</td>
<td>218.66</td>
<td>88.36</td>
<td>240.81</td>
<td>93.33</td>
<td>239.78</td>
<td>93.23</td>
<td>236.43</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>–32.38</td>
<td>18.27</td>
<td>–10.09</td>
<td>26.46</td>
<td>3.31</td>
<td>35.96</td>
<td>3.89</td>
<td>36.07</td>
<td>3.84</td>
<td>35.80</td>
<td>4.61</td>
<td>38.25</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>–30.81</td>
<td>10.80</td>
<td>–4.13</td>
<td>19.78</td>
<td>–0.93</td>
<td>20.84</td>
<td>–0.59</td>
<td>20.75</td>
<td>–0.64</td>
<td>20.55</td>
<td>–0.85</td>
<td>20.30</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>–30.91</td>
<td>8.28</td>
<td>–2.75</td>
<td>15.41</td>
<td>–1.24</td>
<td>16.61</td>
<td>–1.20</td>
<td>16.16</td>
<td>–1.05</td>
<td>15.99</td>
<td>–1.05</td>
<td>16.00</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>–46.19</td>
<td>31.17</td>
<td>–24.17</td>
<td>43.25</td>
<td>45.62</td>
<td>137.68</td>
<td>48.69</td>
<td>138.26</td>
<td>56.54</td>
<td>159.93</td>
<td>56.66</td>
<td>157.67</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>–49.21</td>
<td>17.24</td>
<td>–21.09</td>
<td>33.42</td>
<td>–2.83</td>
<td>49.91</td>
<td>–2.76</td>
<td>49.81</td>
<td>–2.91</td>
<td>49.13</td>
<td>–3.16</td>
<td>48.29</td>
</tr>
</tbody>
</table>

**Note.** Values in the table are for the correctly specified model. Distribution 1 is multivariate normal, Distribution 2 is moderately nonnormal, and Distribution 3 is severely nonnormal.
sufficient sample size. Under the moderately nonnormal distribution EQS-robust standard errors yield mean relative bias percentages of less than 10% at sample sizes of \( n \geq 200 \); under severely nonnormal data conditions standard error attenuation is only controlled at the \( n = 1,000 \) sample size. Bootstrap standard errors appear to resist reduction in standard error estimates under both moderately and severely nonnormal distributions at sample sizes \( n \geq 200 \) (with a notable exception at \( n = 500 \) and severely nonnormal data). At the \( n = 100 \) sample size, bootstrap standard errors inflate with average relative bias percentages large and positive at this sample size condition. Finally, as with the factor covariance parameter, mean bias percentages for the bootstrap standard errors for the loading parameter show no apparent advantage to using more than \( B = 250 \) bootstrap resamplings.

Standard error variability in the \( \lambda_{21} \) parameter, as assessed via the standard deviation of the relative bias percentages, shows a systematic decreasing of standard error variability with increasing sample size. Comparing the standard error estimators against each other, the relative variability in the ML standard errors tends to be the smallest, with larger standard deviations seen in the EQS-robust standard errors and even larger variability exhibited by bootstrap standard errors. This pattern exists across the three distributional forms and appears to be the most pronounced at small sample sizes. At the smallest sample size of \( n = 100 \), the standard deviations for bootstrap standard error estimates are always above 100%, implying that expected fluctuations in standard errors from sample to sample exceed the value of the standard errors themselves. Lastly, one again sees very little difference in the variability of the resampled loading standard errors with increasing numbers of bootstrap samples.

DISCUSSION AND CONCLUSIONS

Results in this investigation replicate previous findings, as well as expand our understanding of the bootstrap as applied to SEM. As expected based on the literature (e.g., Bentler & Chou, 1987), this study shows that, under violations of multivariate normality, normal theory ML estimation yields inflated model test statistics for correctly specified models as well as attenuated parameter standard errors. Additionally, our results for bootstrap standard errors are consistent with the findings of Ichikawa and Konishi (1995). Under multivariate normality, ML outperformed the bootstrap yielding standard errors that exhibited less bias and variability than bootstrap standard errors. Also consistent with the results of Ichikawa and Konishi is the finding that bootstrap standard errors fail under small sample sizes.

New findings from this study center on the relative performance of the bootstrap for parameter standard error and model assessment and are discussed in terms of behavior under nonnormal conditions as bootstrapping methods would not likely be selected if one’s data met standard distributional assumptions. Re-
Regarding standard errors, we first reiterate that increasing the number of bootstrapped samples beyond $B = 250$ did not appear to afford any change in quality of the bootstrapped standard error estimator; even fewer bootstrapped samples may work as well. Second, findings seem to indicate that using bootstrap methods is unwise with the smallest sample size of $n = 100$; standard error bias and variability can become highly inflated, and many bootstrap samples can prove unusable. Third, for sample sizes of $n \geq 200$, bias information would seem to favor the bootstrap over ML estimation, and to some extent over EQS-robust standard errors, when data are derived from nonnormal distributions.

Before championing the bootstrap, however, one must also consider the variability in the resampled statistics. Small bias in the long run is of little use if individual results behave erratically. For the covariance parameter standard error under nonnormal data and $n \geq 200$, bootstrap methods yielded a worst case standard deviation of 60% (at $n = 500$ and severely nonnormal data), implying that a typical fluctuation in standard error would be 60% of the value of the standard error itself, with larger fluctuations possible. A best-case scenario, with moderately nonnormal data conditions and $n = 1,000$, leads one to find a typical fluctuation in bootstrap standard errors to be 17% of the value of the standard error itself, though larger fluctuations could certainly occur. As for variability in the estimation of a loading parameter standard error using bootstrap-based methods, the worst case of $n = 200$ (other than $n = 100$) under severe nonnormality yields standard error variability that is roughly half the parameter standard error itself with a reported standard deviation of about 50%. The best case, on the other hand, with $n = 1,000$ under moderate nonnormality, yields an associated standard deviation of 16%.

We now consider assessing the overall fit of the model under nonnormal conditions. Given that the prior discussion regarding parameter standard error estimation recommended against the $n = 100$ case, results for the bootstrap under nonnormal conditions with $n \geq 200$ are evaluated against the results for $T_{ML-SB}$ and against those reported by Curran et al. (1996) for the ADF estimator. Curran et al. showed Type I error rates associated with the ADF test statistic become intolerably high at $n \leq 500$ with reported error rates under the $n = 200$ sample size of 19% and 25% for the moderately nonnormal and severely nonnormal distributions, respectively. At the same sample size condition, error rates for $T_{ML-SB}$ in this study remained controlled under moderately nonnormal data but were inflated under the severely nonnormal distribution. Type I error rates for the bootstrap are notably lower than those for $T_{ML-SB}$ with severely nonnormal data.

Results here for the bootstrap suggest the resampling-based method may be conservative in its control over model rejections, thus having an impact on the statistical power associated with the method. Indeed, this appears to be the case, as evidenced in the proportion of model rejections for the improperly specified model. The comparison between the bootstrap and $T_{ML-SB}$ to robust model assessment in SEM illustrates the often-unavoidable tradeoff between control over
Type I error and statistical power. Although TML-SB demonstrates greater power to reject an improperly specified model, it also has the tendency to over-reject a correctly specified model. The bootstrap in its adjustment to model \( p \) values appears to exert greater control over model rejections than TML-SB, but at the expense of the power to reject a misspecified model. It is perhaps unrealistic to think that practitioners will, in the context of their specific models, conduct an a priori cost-benefit analysis before seeking out the SEM software that offers the robust method with the desired emphasis on error control. Rather, it may be more practical simply to suggest that researchers be aware of the robust methods’ error control propensities in the programs to which they have access and interpret results with appropriate caution.

To sum, the results of this investigation for bootstrap standard errors have led to the apparent recommendation that use of the bootstrap with samples of size \( n = 100 \) is unwise. However, it must be emphasized that such results emanate from the specific model investigated here, a nine-variable, three-factor CFA model with 21 parameters requiring estimation. Bootstrapping standard errors with sample sizes of \( n = 100 \) (or even smaller) may work fine with less complex models, whereas bootstrapping may fail with samples of sizes \( n = 200 \) (or even larger) with more complicated models. In this study, the problematic \( n = 100 \) case has a ratio of observations to parameters barely below 5:1, a minimum loosely recommended for normal theory estimation methods (Bentler, 1996). Further, bootstrap standard errors were seen to lack stability with a sample size of \( n = 200 \); this case has a corresponding ratio of sample size to parameters of just under 10:1, a ratio that may be a practical lower bound for parameter estimation under arbitrary distributions (Bentler, 1996). As bootstrap procedures become more popular and integrated into more SEM software packages, future research is certainly warranted employing models of wide-ranging complexity to home in on model-based minimum sample size recommendations.

REFERENCES


---

**APPENDIX A**

Population and Estimated True Standard Errors (Via Monte Carlo Simulation Using 2000 Replications) for Parameters $\phi_{21}$ and $\lambda_{21}$ Under the Multivariate Normal Distribution and Correctly Specified Model

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\phi_{21}$ Population</th>
<th>$\phi_{21}$ Monte Carlo</th>
<th>$\lambda_{21}$ Population</th>
<th>$\lambda_{21}$ Monte Carlo</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>.07211623</td>
<td>.07389620</td>
<td>.20107951</td>
<td>.21719432</td>
</tr>
<tr>
<td>200</td>
<td>.05086559</td>
<td>.05064775</td>
<td>.14182698</td>
<td>.15377571</td>
</tr>
<tr>
<td>500</td>
<td>.03212183</td>
<td>.03171152</td>
<td>.08956434</td>
<td>.08980573</td>
</tr>
<tr>
<td>1000</td>
<td>.02270219</td>
<td>.02250867</td>
<td>.06329985</td>
<td>.06471447</td>
</tr>
</tbody>
</table>
**APPENDIX B**


<table>
<thead>
<tr>
<th>Distribution</th>
<th>n</th>
<th>$\phi_{21}$</th>
<th>$\lambda_{21}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multivariate normal</td>
<td>100</td>
<td>.07211623</td>
<td>.20107951</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>.05086559</td>
<td>.14182698</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>.03212183</td>
<td>.08956434</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>.02270219</td>
<td>.06329985</td>
</tr>
<tr>
<td>Moderately nonnormal</td>
<td>100</td>
<td>.09346542</td>
<td>.31654237</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>.06278746</td>
<td>.21547391</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>.03845778</td>
<td>.13104480</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>.02640384</td>
<td>.09266848</td>
</tr>
<tr>
<td>Severely nonnormal</td>
<td>100</td>
<td>.10235749</td>
<td>.41877029</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>.07340406</td>
<td>.29015204</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>.04436619</td>
<td>.19171851</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>.03175627</td>
<td>.13169042</td>
</tr>
</tbody>
</table>