Psychological Methods
1996, Vol. 1, No. 1, 16–29

1082-989X/96/$3.00

The Robustness of Test Statistics to Nonnormality and Specification Error in Confirmatory Factor Analysis

Patrick J. Curran
University of California, Los Angeles

Stephen G. West
Arizona State University

John F. Finch
Texas A&M University

Monte Carlo computer simulations were used to investigate the performance of three χ^2 test statistics in confirmatory factor analysis (CFA). Normal theory maximum likelihood χ^2 (ML), Browne's asymptotic distribution free χ^2 (ADF), and the Satorra-Bentler rescaled χ^2 (SB) were examined under varying conditions of sample size, model specification, and multivariate distribution. For properly specified models, ML and SB showed no evidence of bias under normal distributions across all sample sizes, whereas ADF was biased at all but the largest sample sizes. ML was increasingly overestimated with increasing nonnormality, but both SB (at all sample sizes) and ADF (only at large sample sizes) showed no evidence of bias. For misspecified models, ML was again inflated with increasing nonnormality, but both SB and ADF were underestimated with increasing nonnormality. It appears that the power of the SB and ADF test statistics to detect a model misspecification is attenuated given nonnormally distributed data.

Confirmatory factor analysis (CFA) has become an increasingly popular method of investigating the structure of data sets in psychology. In contrast to traditional exploratory factor analysis that does not place strong a priori restrictions on the structure of the model being tested, CFA requires the investigator to specify both the number of factors and the specific pattern of loadings of each of the measured variables on the underlying set of factors. In typical simple CFA models, each measured variable is hypothesized to load on only one factor, and positive, negative, or zero (orthogonal) correlations are specified between the factors. Such models can provide strong evidence about the convergent and discriminant validity of a set of measured variables and allow tests among a set of theories of measurement structure. More complicated CFA models may specify more complex patterns of factor loadings, correlations among errors or specific factors, or both. In all cases, CFA models set restrictions on the factor loadings, the correlations between factors, and the correlations between errors of measurement that permit tests of the fit of the hypothesized model to the data.

There are two general classes of assumptions that underlie the statistical methods used to estimate CFA models: distributional and structural (Satorra, 1990). Normal theory maximum likelihood (ML) estimation has been used to analyze the majority of CFA models. ML makes the *distributional* assumption that the measured variables have a multivariate normal distribution in the pop-

Patrick J. Curran, Graduate School of Education, University of California, Los Angeles; Stephen G. West, Department of Psychology, Arizona State University; John F. Finch, Department of Psychology, Texas A&M University.

We thank Leona Aiken, Peter Bentler, David Kaplan, David MacKinnon, Bengt Muthén, Mark Reiser, and Albert Satorra for their valuable input. This work was partially supported by National Institute on Alcohol Abuse and Alcoholism Postdoctoral Fellowship F32 AA05402-01 and Grant P50MH39246 during the writing of this article.

Correspondence concerning this article should be addressed to Patrick J. Curran, Graduate School of Education, University of California, Los Angeles, 2318 Moore Hall, Los Angeles, California 90095-1521. Electronic mail may be sent via Internet to idhbpjc@mvs.oac.ucla.edu.

ulation. However, the majority of data collected in behavioral research do not follow univariate normal distributions, let alone a multivariate normal distribution (Micceri, 1989). Indeed, in some important areas of research such as drug use, child abuse, and psychopathology, it would not be reasonable to even expect that the observed data would follow a normal distribution in the population. In addition to the distributional assumption, ML (and all methods of estimation) makes the *structural* assumption that the structure tested in the sample accurately reflects the structure that exists in the population. If the sample structure does not adequately conform to the corresponding population structure, severe distortions in all aspects of the final solution can result.

Although the chi-square test statistic can be used to measure the extent of the violation of the structural assumption (Lawley & Maxwell, 1971), the accuracy of this test statistic can be compromised given violation of the distributional assumption (Satorra, 1990). Violations of both the distributional and structural assumptions are common (and often unavoidable) in practice and can potentially lead to seriously misleading results. It is thus important to fully understand the effects of the multivariate nonnormality and specification error on maximum likelihood estimation and other alternative estimators used in CFA.

Methods of Estimation

By far the most common method used to estimate confirmatory factor models is normal theory ML. Nearly all of the major software packages use ML as the standard default estimator (e.g., EQS, Bentler, 1989; LISREL, Jöreskog & Sörbom, 1993; PROC CALIS, SAS Institute, Inc., 1990; RAMONA, Browne, Mels, & Coward, 1994). Under the assumptions of multivariate normality, proper specification of the model, and a sufficiently large sample size (N), ML provides asymptotically (large sample) unbiased, consistent, and efficient parameter estimates and standard errors (Bollen, 1989). An important advantage of ML is that it allows for a formal statistical test of model fit. $(N - 1)$ multiplied by the minimum of the ML fit function is distributed as a large sample chi-square with $1/2(p)(p + 1) - t$ degrees of freedom, where p is the number of observed variables and t is the number of freely estimated parameters (Bollen, 1989).

One potential limitation of ML estimation is the strong assumption of multivariate normality. Given the presence of non-zero third- and (particularly) fourth-order moments (skewness and kurtosis, respectively),[1] the resulting ML parameter estimates are consistent but not efficient, and the minimum of the ML fit function is no longer distributed as a large sample central chi-square. Instead, $(N - 1)$ multiplied by the minimum of the ML fit function generally produces an inflated (positively biased) estimate of the referenced chi-square distribution (Browne, 1982; Satorra, 1991). Hence, using the normal theory chi-square statistic as a measure of model fit under conditions of nonnormality will lead to an inflated Type I error rate for model rejection. Consequently, in practice a researcher may mistakenly reject or opportunistically modify a model because the distribution of the observed variables is not multivariate normal rather than because the model itself is not correct (see MacCallum, 1986; MacCallum, Roznowski, & Necowitz, 1990).

Several different approaches have been proposed to address the problems with ML estimation under conditions of multivariate nonnormality. One example is the development of alternative methods of estimation that do not assume multivariate normality. One such estimator that is currently available in structural modeling programs such as EQS (Bentler, 1989), LISCOMP (Muthén, 1987), LISREL (Jöreskog & Sörbom, 1993), and RAMONA (Browne et al., 1994) is Browne's (1982, 1984) asymptotic distribution free (ADF) method of estimation. The derivation of the ADF estimator was not based on the assumption of multivariate normality so that variables possessing non-zero kurtoses theoretically pose no special problems for estimation. ADF provides asymptotically consistent and efficient parameter estimates and standard errors, and $(N - 1)$ times the minimum of the fit function is distributed as a large sample chi-square (Browne, 1984). One practical disadvantage of ADF is that it is computationally

[1] The multivariate normal distribution is actually characterized by skewness equal to 0 and kurtosis equal to 3. However, it is common practice to subtract the constant value of 3 from the kurtosis estimate so that the normal distribution is characterized by zero skewness and zero kurtosis. We will similarly refer to the normal distribution as defined by zero skewness and zero kurtosis.

very demanding. Browne (1984) anticipated that models with greater than 20 variables could not be feasibly estimated with ADF. A second possible disadvantage is that initial findings suggest that this estimator performs poorly at the small to moderate sample sizes that typify much of psychological research.

A second approach that has been developed for computing a more accurate test statistic under conditions of nonnormality is to adjust the normal theory ML chi-square estimate for the presence of non-zero kurtosis. Because the normal theory chi-square does not follow the expected chi-square distribution under conditions of nonnormality, the normal theory chi-square must be corrected, or rescaled, to provide a statistic that more closely approximates the referenced chi-square distribution (Browne, 1982, 1984). One variant of the rescaled test statistic that is currently only available in EQS (Bentler, 1989) is the Satorra-Bentler chi-square (SB χ^2; Satorra, 1990, 1991; Satorra & Bentler, 1988). The SB χ^2 corrects the normal theory chi-square by a constant k, a scalar value that is a function of the model implied residual weight matrix, the observed multivariate kurtosis, and the model degrees of freedom. The greater the degree of observed multivariate kurtosis, the greater downward adjustment that is made to the inflated normal theory chi-square.

Review of Monte Carlo Studies

Muthén and Kaplan (1985) studied a properly specified four indicator single-factor model under five distributions ranging from normal to severely nonnormal for one sample size (1,000). For univariate skewness greater than 2.0, the ML χ^2 was clearly inflated whereas the ADF χ^2 remained consistent. Muthén and Kaplan (1992) extended these findings by adding more complex model specifications, an additional sample size (500), and increasing the number of replications to 1,000 per condition. The normal theory chi-square was extremely sensitive to both nonnormality and model complexity (defined as the number of parameters estimated in the model). The ADF χ^2 appeared to be very sensitive to model complexity, with extreme inflation of the model chi-square as the tested model became increasingly complex. The ADF χ^2 was also particularly inflated at the smaller sample size.

Satorra and Bentler (1988) performed a Monte Carlo simulation using a properly specified four indicator single-factor model to evaluate the behavior of the SB χ^2 test statistic. The unique variances of the four indicators were calculated with univariate skewness of 0 and a homogenous univariate kurtosis of 3.7. The models were estimated using ML, unweighted least squares (ULS), and ADF, based on 1,000 replications of a single sample size of 300. The normal theory ML χ^2 and the SB χ^2 performed similarly to one another. On average, the ML χ^2 slightly underestimated the expected value of the model chi-square while the SB χ^2 slightly overestimated the expected value. However, the ML χ^2 had a larger variance than did the SB χ^2. The ADF χ^2 resulted in the highest average value, although it also attained the lowest variance.

Chou, Bentler, and Satorra (1991) similarly used a Monte Carlo simulation to examine the ML, ADF, and SB χ^2 test statistics for a properly specified model under varying conditions of normality. A two-factor six indicator CFA model was replicated 100 times per condition based on two sample sizes (200 and 400) and six multivariate distributions. Two versions of the model were estimated, one in which all of the necessary parameters were freely estimated, and one in which the factor loadings were fixed to the population values. Consistent with previous research, the ML χ^2 was inflated under nonnormal conditions. The SB χ^2 outperformed both the ML and ADF χ^2 test statistics in nearly all conditions.

Finally, Hu, Bentler, and Kano (1992) performed a major simulation study based on a three-factor confirmatory factor model with five indicators per factor. Six sample sizes were used (ranging from 150 to 5,000) with 200 replications per condition. Seven different symmetric distributions were considered, ranging from normal to severely nonnormal (high kurtosis). The normal theory estimators (maximum likelihood and generalized least squares) provided inflated chi-square values as nonnormality increased. The ADF test statistic was relatively unaffected by distribution but was only reliable at the largest sample size (5,000). Finally, the SB χ^2 performed the best of all test statistics, although models were rejected at a higher frequency than was expected at small sample sizes.

In summary, Monte Carlo simulation studies have consistently supported the theoretical prediction that the normal theory ML χ^2 test statistic is significantly inflated as a function of multivariate nonnormality. The ADF χ^2 is theoretically asymptotically robust to multivariate nonnormality, but its behavior at smaller (and more realistic) sample sizes is poor. The SB χ^2 has not been fully examined, but initial findings indicate that it outperforms both the ML and ADF test statistics under nonnormal distributions, although it does tend to overreject models at smaller sample sizes.

Although the ramifications of violating the distributional assumptions in CFA is becoming better understood, much less is known about violations of the structural assumption. Recent work has addressed the effects of model misspecification on the computation of parameter estimates and standard errors (Kaplan, 1988, 1989) as well as post hoc model modification (MacCallum, 1986) and the need for alternative indices of fit (MacCallum, 1990). However, little is known about the behavior of chi-square test statistics under simultaneous violations of both the distributional and the structural assumptions. Indeed, we are not aware of a single empirical study that has examined the ADF and SB χ^2 test statistics under these two conditions. Given the low probability that the structure tested in a sample precisely conforms to the structure that exists in the population, it is critical that a better understanding be gained of the behavior of the test statistics under these more realistic conditions.

The Present Study

A series of Monte Carlo computer simulations were used to study the effects of sample size, multivariate nonnormality, and model specification on the computation of three chi-square test statistics that are currently widely available to the practicing researcher: ML, SB, and ADF.[2] Four specifications of an oblique three-factor model with three indicators per factor were considered. The first two models were *correctly specified* such that the structure estimated in the sample precisely corresponded to the structure that existed in the population. The second two models were *misspecified* such that the structure tested in the sample did *not* correspond to the structure that existed in the population.

Method

Model Specification

Four specifications of an oblique three-factor model with three indicators per factor were examined. The basic confirmatory factor model is presented in Figure 1. The population parameters consisted of factor loadings (each $\lambda = .70$), uniquenesses (each $\Theta_\delta = .51$), interfactor correlations (each $\phi = .30$), and factor variances (all set to 1.0).

Model 1. Model 1 was properly specified such that the model that was estimated in the sample directly corresponded to the model that existed in the population. Thus, both the sample and the population models corresponded to the solid lines presented in Figure 1.

Model 2. Model 2 contained two factor loadings that were estimated in the sample but did *not* exist in the population. Thus, in Figure 1, the double dashed lines represent the two factor loadings that linked Item 5 to Factor 3 and Item 8 to Factor 2, and the expected value of these parameters was 0. This is a misspecification of *inclusion*. Note that from the standpoint of statistical theory, estimation of parameters with an expected value of 0 in the population does not bias the sample results. Model 2 is thus considered to be a properly specified model.

Model 3. Model 3 excluded two loadings from the sample that *did* exist in the population. Thus, in Figure 1, the single dashed lines represent the two excluded factor loadings (both population λs $= .35$) that linked Item 6 to Factor 3 and Item 7 to Factor 2. The value of $\lambda = .35$ was chosen to reflect a small to moderate factor loading that might be commonly encountered in practice. This is a misspecification of *exclusion*.

Model 4. Finally, Model 4 was the combination of Models 2 and 3. Like Model 2, two factor loadings were estimated in the sample that did *not* exist in the population (the double dashed lines linking Item 5 to Factor 3 and Item 8 to Factor 2). Additionally, like Model 3, two factor loadings were excluded from the sample that *did* exist in the population (the single dashed lines that linked Item 6 to Factor 3 and Item 7 to Factor 2). This is a misspecification of both *inclusion* and *exclusion*.

[2] Note that GLS is also available in standard packages and is relatively widely used. However, GLS is a normal theory estimator that is asymptotically equivalent to ML, and previous studies (e.g., Muthén & Kaplan, 1985, 1992) have shown the behavior of ML and GLS to be very similar.

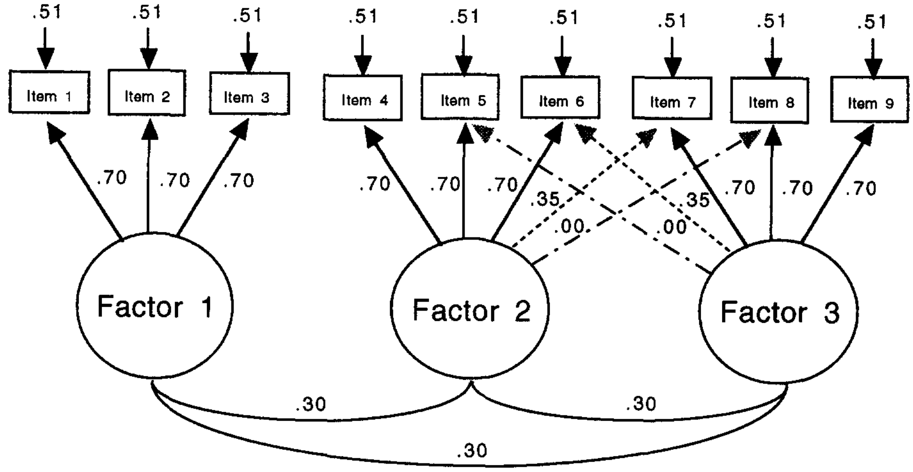

Figure 1. Nine-indicator three-factor oblique confirmatory factor analysis model with population parameter values. Solid lines represent parameters that were shared between the sample and the population; single dashed lines represent parameters that existed in the population (λ = .35) but were omitted from the sample; double dashed lines represent parameters that did not exist in the population (λ = 0) but were estimated in the sample.

Conditions

Multivariate distributions. Three population distributions were considered for all four model specifications (see Figure 2). Distribution 1 was multivariate normal with univariate skewness and kurtoses equal to 0. Distribution 2 was moderately nonnormal with univariate skewness of 2.0 and kurtoses of 7.0. Finally, Distribution 3 was severely nonnormal with univariate skewness of 3.0 and kurtoses of 21.0. These levels of nonnormality were chosen to represent moderate and severe nonnormality based on our examination of the levels of skewness and kurtoses in data sets from several community-based mental health and substance abuse studies.

Sample size. Four sample sizes were considered for all model specifications: 100, 200, 500, and 1,000.

Replications. All models were replicated 200 times per condition.

Data generation. The raw data were generated using both the PC and mainframe version of EQS (Version 3; Bentler, 1989). Details of the data generation procedure are presented in the Appendix.

Measures

Three chi-square test statistics were studied: normal theory ML, ADF, and the SB scaled χ^2. All three test statistics were computed by EQS (Version 3.0). Note that the ML and ADF statistics provided by EQS should be identical to that available through current versions of LISCOMP, LISREL, and RAMONA.

Results

Expected Value of Test Statistics

The expected values of the chi-square test statistics for Models 1 and 2 were simply the model degrees of freedom for all three estimators across all distributions and sample sizes (24.0 for Model 1 and 22.0 for Model 2). Because Models 3 and 4 were misspecified, the expected values of these test statistics could not be computed directly. Instead, the expected values were computed as large sample empirical estimates that differed as a function of method of estimation, multivariate distribution, and sample size. Further details regarding the computations of these estimates are presented in the Appendix.

Monte Carlo Results

Tables 1, 2, 3, and 4 present the mean observed value, the expected value, the percentage of bias,

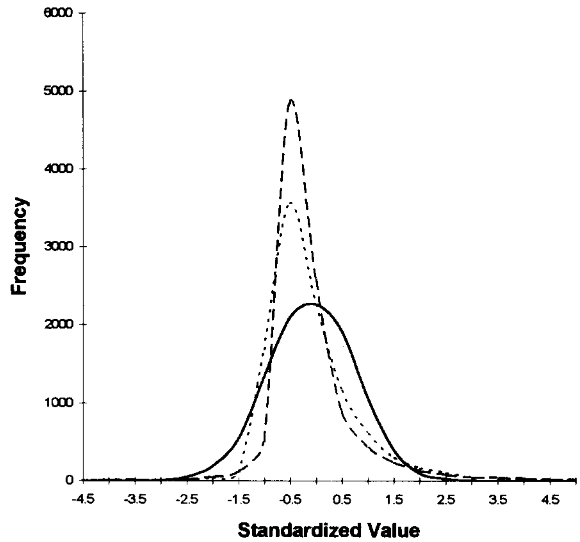

Figure 2. Plots of normal (solid line), moderately nonnormal (dotted line), and severely nonnormal (dashed line) empirical distributions based on a random sample of $N = 10{,}000$.

and the percentage of models rejected at $p < .05$ for the ML, SB, and ADF χ^2 test statistics for all four model specifications.[3] For the correctly specified models (Models 1 and 2), the expected rejection rate was 5%; and given 200 replications and $\alpha = .05$, the 95% confidence interval for the percentage of rejected models defined an approximate upper and lower bound of 2% and 8%, respectively. Rejection rates for the obtained chi-square values falling within these bounds are consistent with the null hypothesis that the estimator is unbiased. Model rejection rates were not as meaningful for misspecified models (Models 3 and 4), so relative bias was computed (the observed value minus the expected value divided by the expected value). Bias in excess of 10% was considered significant (Kaplan, 1989).

Model Specification 1. Table 1 presents the results for Model 1. Recall that Model 1 was properly specified such that the model estimated in the sample directly corresponded to the model that existed in the population. The expected value for all three test statistics was $E(\chi^2) = 24.0$.

Under multivariate normality, the ML χ^2 rejected the expected number of models across all sample sizes (approximately 5%). Consistent with both theory and previous simulation research, the ML χ^2 became increasingly positively biased as the distribution became increasingly nonnormal. This inflation was exacerbated with increasing sample size. For example, nearly half of the correctly specified models were rejected for $N = 1{,}000$ under the severely nonnormal condition. Under multivariate normality, the ADF χ^2 was inflated at small sample sizes, for example, rejecting 43% of the correctly specified models at $N = 100$. The performance of the ADF improved with increasing sample size, but even under multivariate normality at $N = 1{,}000$, 10% of the correct models were rejected. The ADF χ^2 was also positively biased with increasing nonnormality, but this bias was attenuated with increasing sample size. Finally, the SB χ^2 was very well behaved at nearly all sample sizes across all distributions. For example, at a sample size of $N = 200$ under severe nonnormality, the SB χ^2 rejected 7% of the properly specified models (compared to 25% for ADF and 36% for ML). Under these conditions, the performance of the SB χ^2 represented a distinct improvement over the ML χ^2 under conditions of nonnormality. Interestingly, the SB and ADF performed similarly at samples of $N = 500$ and $N = 1{,}000$.

Model Specification 2. The results from Model 2 are presented in Table 2. Recall that Model 2 estimated two factor loadings in the sample that did not exist in the population. Because the error is the addition of two truly nonexistent parameters, this can be considered a properly specified model, and the expected value of the model chi-square was equal to the model degrees of freedom for all estimators across all sample sizes, $E(\chi^2) = 22.0$.

Overall, the results from Model 2 closely followed those of Model 1. Under multivariate normality, the rejection rates for both the ML and SB were slightly higher than expected at $N = 100$ but were unbiased at $N = 200$ and greater. Under multivariate normality, the ADF again rejected a very high number of models at the two smaller sample sizes but was unbiased at the two larger sample sizes. As with Model 1, the ML χ^2 was

[3] All improper solutions (nonconverged solutions and solutions that converged but resulted in out-of-bound parameters, e.g., Heywood cases) were dropped from subsequent analyses. Collapsing across all conditions, 90% of the replications were proper for ML and 83% were proper for ADF.

Table 1
Observed Chi-Square, Expected Chi-Square, Percentage Bias, and Percentage of Rejected Models for Model Specification 1

		Normal				Moderately nonnormal				Severely nonnormal			
Size	χ^2	Observed χ^2	Expected χ^2	% Bias	% Reject	Observed χ^2	Expected χ^2	% Bias	% Reject	Observed χ^2	Expected χ^2	% Bias	% Reject
100	ML	25.01	24.0	4.0	5.5	29.35	24.0	22.0	20.0	33.54	24.0	40.0	30.0
	SB	25.87	24.0	8.0	7.5	26.06	24.0	9.0	8.5	27.26	24.0	14.0	13.0
	ADF	36.44	24.0	52.0	43.0	38.04	24.0	59.0	49.0	44.82	24.0	87.0	68.0
200	ML	24.78	24.0	3.0	6.5	30.15	24.0	26.0	25.0	34.40	24.0	43.0	36.0
	SB	25.22	24.0	5.0	8.5	25.44	24.0	6.0	8.0	25.80	24.0	8.0	6.5
	ADF	29.19	24.0	22.0	19.0	29.27	24.0	22.0	19.0	31.29	24.0	30.0	25.0
500	ML	23.94	24.0	0.0	3.5	31.26	24.0	30.0	24.0	35.55	24.0	48.0	40.0
	SB	24.10	24.0	0.0	5.0	25.44	24.0	6.0	6.9	24.85	24.0	4.0	8.5
	ADF	25.92	24.0	8.0	11.0	26.42	24.0	10.0	6.7	26.83	24.0	12.0	8.5
1000	ML	25.05	24.0	4.0	7.0	30.78	24.0	28.0	24.0	37.40	24.0	56.0	48.0
	SB	25.16	24.0	5.0	8.0	24.77	24.0	3.0	7.5	25.01	24.0	4.0	7.0
	ADF	25.79	24.0	7.0	9.5	25.36	24.0	6.0	7.5	25.47	24.0	6.0	7.2

Note. Univariate skewness and kurtoses were (0,0), (2,7), and (3,21) for normal, moderately nonnormal, and severely nonnormal distributions, respectively. ML = maximum likelihood; SB = Satorra-Bentler rescaled; ADF = asymptotic distribution free.

increasingly positively biased with increasing nonnormality, and this inflation was exacerbated with increasing sample size. In comparison, the SB χ^2 showed minimal bias with increasing nonnormality, although the observed rejection rates at the smallest sample size were slightly larger than expected. Even under severe nonnormality, the SB χ^2 again showed little evidence of bias, especially at sample sizes of $N = 200$ or greater. Finally, the ADF χ^2 was positively biased with increasing nonnormality at the smaller sample sizes but was unbiased at sample sizes of $N = 500$ and $N = 1{,}000$, even under severe nonnormality.

Model Specification 3. Model Specification 3 excluded two factor loadings in the sample ($\lambda = .35$) that truly existed in the population. These results are presented in Table 3. Recall that due to the exclusion of existing parameters, there was a different expected value for each test statistic. Also, because the model was misspecified in the

Table 2
Observed Chi-Square, Expected Chi-Square, Percentage Bias, and Percentage of Rejected Models for Model Specification 2

		Normal				Moderately nonnormal				Severely nonnormal			
Size	χ^2	Observed χ^2	Expected χ^2	% Bias	% Reject	Observed χ^2	Expected χ^2	% Bias	% Reject	Observed χ^2	Expected χ^2	% Bias	% Reject
100	ML	23.42	22.0	6.0	9.6	26.89	22.0	22.0	22.2	29.82	22.0	36.0	34.4
	SB	24.19	22.0	9.0	12.1	23.80	22.0	8.0	8.5	25.07	22.0	14.0	11.1
	ADF	31.0	22.0	41.0	30.5	34.95	22.0	59.0	40.8	46.45	22.0	111.0	63.2
200	ML	22.48	22.0	2.0	6.0	27.70	22.0	26.0	26.5	31.77	22.0	44.0	36.7
	SB	22.86	22.0	3.0	7.0	23.75	22.0	8.0	8.5	24.10	22.0	10.0	8.5
	ADF	26.43	22.0	20.0	17.5	26.06	22.0	18.0	14.5	28.52	22.0	30.0	22.0
500	ML	21.89	22.0	0.0	6.0	26.68	22.0	21.0	19.0	31.86	22.0	45.0	32.5
	SB	22.02	22.0	0.0	7.0	21.90	22.0	0.0	5.5	23.33	22.0	6.0	6.5
	ADF	23.13	22.0	5.0	7.0	23.04	22.0	5.0	8.0	24.02	22.0	9.0	5.5
1000	ML	22.25	22.0	0.0	5.0	26.05	22.0	18.0	14.5	33.37	22.0	52.0	41.5
	SB	22.31	22.0	0.0	3.5	21.14	22.0	4.0	4.0	22.74	22.0	3.0	7.0
	ADF	22.09	22.0	0.0	6.0	23.32	22.0	6.0	9.5	23.41	22.0	6.0	7.5

Note. Univariate skewness and kurtoses were (0,0), (2,7), and (3,21) for normal, moderately nonnormal, and severely nonnormal distributions, respectively. ML = maximum likelihood; SB = Satorra-Bentler rescaled; ADF = asymptotic distribution free.

Table 3
Observed Chi-Square, Expected Chi-Square, Percentage Bias, and Percentage of Rejected Models for Model Specification 3

		Normal				Moderately nonnormal				Severely nonnormal			
Size	χ^2	Observed χ^2	Expected χ^2	% Bias	% Reject	Observed χ^2	Expected χ^2	% Bias	% Reject	Observed χ^2	Expected χ^2	% Bias	% Reject
100	ML	38.45	37.62	2.0	54.3	46.50	37.75	23.0	78.9	52.87	37.77	40.0	81.4
	SB	39.63	37.65	5.0	57.9	37.63	33.99	11.0	49.1	38.10	30.85	24.0	47.1
	ADF	52.04	33.74	54.0	80.3	60.88	29.68	105.0	86.3	81.31	27.29	198.0	95.3
200	ML	51.07	51.38	0.0	88.1	59.72	51.66	16.0	93.7	68.58	51.68	33.0	95.4
	SB	51.65	51.44	0.0	89.9	47.04	44.09	7.0	82.4	44.66	37.77	18.0	78.9
	ADF	50.17	43.59	15.0	85.6	47.93	35.42	36.0	84.8	47.16	30.61	54.0	75.2
500	ML	92.27	92.66	0.0	100.0	99.39	93.37	6.0	100.0	109.87	93.42	18.0	100.0
	SB	92.52	92.80	0.0	100.0	75.04	74.40	1.0	100.0	63.46	58.52	8.0	97.2
	ADF	76.89	73.11	5.0	100.0	60.10	52.63	14.0	99.5	53.67	40.58	32.0	96.7
1000	ML	161.46	161.35	0.0	100.0	171.07	162.88	5.0	100.0	180.90	162.98	11.0	100.0
	SB	161.66	161.74	0.0	100.0	126.23	124.89	2.0	100.0	101.25	93.11	9.0	100.0
	ADF	127.71	122.32	4.0	100.0	90.23	81.31	11.0	100.0	76.10	57.20	33.0	100.0

Note. Univariate skewness and kurtoses were (0,0), (2,7), and (3,21) for normal, moderately nonnormal, and severely nonnormal distributions, respectively. ML = maximum likelihood; SB = Satorra-Bentler rescaled; ADF = asymptotic distribution free.

sample, the percentage of rejected models was no longer a meaningful guide with which to judge the behavior of the test statistics. Thus, the following results will now be presented in terms of the relative bias in the test statistics.[4]

Under multivariate normality, the expected values for the ML and SB χ^2 were nearly identical across all four sample sizes. This is further support that for normal distributions, no scaling correction is required for the ML χ^2, and the SB χ^2 thus simplifies to the ML χ^2. Additionally, neither the ML or SB test statistic showed appreciable bias under normality across all four sample sizes. In comparison, the empirical estimate of the expected value of the ADF χ^2 was smaller than that of the ML or SB test statistics. Recall that the expected value for all three test statistics were equal for the properly specified models. The lower expected value of the ADF for misspecified models even under multivariate normality suggests that this test statistic may have less power to reject the null hypothesis compared with the ML or SB χ^2. Unlike the ML and SB, the ADF was significantly positively biased at the two smaller sample sizes. For example, at $N = 100$ the average observed ADF χ^2 was 54% larger than the expected value. This bias dropped to 15% at $N = 200$ and was negligible at the two larger sample sizes.

The findings become more complicated given nonnormal distributions. The expected value of the ML χ^2 was the same across all three distributions. As with the previous models, the ML χ^2 showed increasing levels of positive bias with increasing nonnormality. A particularly interesting finding pertained to the expected values of the SB and ADF test statistics under nonnormality. Both the expected and the observed values of the SB and ADF test statistics decreased with increasing nonnormality. For example, at sample size $N = 200$, the expected value for the SB χ^2 was approximately 51 under normality, 44 under moderate nonnormality, and 38 under severe nonnormality. The ADF test statistic showed a similar pattern. The direct interpretation of this finding is that it is increasingly difficult to detect a misspecification within the model given the added variability due to the nonnormal distribution of the data. Thus, the power of SB and ADF test statistics decreased with increasing nonnormality.

Under moderate nonnormality, the SB χ^2 was slightly biased at $N = 100$ (11%) but was unbiased at sample sizes of $N = 200$ and above. In comparison, also under moderate nonnormality, the ADF χ^2 showed extreme bias at the smaller sample sizes

[4] Models 1 and 2 could have similarly been evaluated using the percentage of bias, and the same conclusions would have been drawn. The percentage of rejected models was chosen instead for Models 1 and 2 given the more direct interpretability of the findings.

Table 4
Observed Chi-Square, Expected Chi-Square, Percentage Bias, and Percentage of Rejected Models for Model Specification 4

		Normal				Moderately nonnormal				Severely nonnormal			
Size	χ^2	Observed χ^2	Expected χ^2	% Bias	% Reject	Observed χ^2	Expected χ^2	% Bias	% Reject	Observed χ^2	Expected χ^2	% Bias	% Reject
100	ML	34.03	32.52	5.0	48.7	40.37	32.52	24.0	63.4	45.15	32.45	39.0	78.6
	SB	34.97	32.56	7.0	51.8	33.94	29.76	14.0	45.5	34.09	27.33	25.0	50.3
	ADF	47.99	30.66	56.0	82.1	48.67	27.19	79.0	85.4	63.25	25.06	152.0	91.8
200	ML	43.75	43.14	1.0	81.1	49.84	43.14	16.0	91.0	58.14	43.01	35.0	88.1
	SB	44.34	43.23	3.0	81.6	39.55	37.59	5.0	67.9	38.48	32.71	18.0	60.4
	ADF	46.42	39.41	18.0	85.5	41.84	32.43	29.0	73.4	46.05	28.16	64.0	84.5
500	ML	75.29	75.04	0.0	100.0	81.35	75.04	8.0	100.0	91.71	74.68	23.0	100.0
	SB	75.62	75.23	0.0	100.0	62.09	61.09	2.0	97.0	55.94	48.85	15.0	92.6
	ADF	69.62	65.66	6.0	100.0	53.84	48.15	12.0	94.2	49.13	37.45	31.0	93.2
1000	ML	128.71	128.20	0.0	100.0	133.86	128.20	4.0	100.0	144.56	127.46	13.0	100.0
	SB	129.16	128.56	0.0	100.0	100.48	100.26	0.0	100.0	83.44	75.76	10.0	100.0
	ADF	111.39	109.41	2.0	100.0	80.91	74.35	9.0	100.0	68.25	52.92	30.0	100.0

Note. Univariate skewness and kurtoses were (0,0), (2,7), and (3,21) for normal, moderately nonnormal, and severely nonnormal distributions, respectively. ML = maximum likelihood; SB = Satorra-Bentler rescaled; ADF = asymptotic distribution free.

and remained biased even at N = 1,000 (11%). Under severe nonnormality, the SB χ^2 showed substantial bias at the two smaller sample sizes (e.g., 18% at N = 200) but was only moderately biased at the larger sample sizes (e.g., 9% at N = 1,000). The ADF showed very high levels of relative bias across all four sample sizes under severe nonnormality and was overestimated by 33% even at the largest sample size N = 1,000.

Model Specification 4. Model Specification 4 contained both errors of inclusion and exclusion. Two cross-loadings existed in the population that were not estimated in the sample (λ = .35), and two cross-loadings were estimated in the sample that did not exist in the population (λ = 0). These results are presented in Table 4.

The findings from Model Specification 4 followed the same general pattern as was observed for Model Specification 3. The primary difference was that the expected values and rejection rates in Model 4 were lower compared with those of Model 3. This result may initially appear counterintuitive given that Model 4 combined errors of both inclusion and exclusion. However, unlike Model 3, Model 4 contained the simultaneous estimation of the two truly nonexistent paths and the exclusion of the two truly existent paths. The mean estimated factor loadings for the two additional paths in Model 2 (where no other paths were excluded) was λ = 0 (the population expected value). However, the mean factor loadings for these same two additional paths in Model 4 was λ = −.40. Thus, these additional free parameters served to "absorb" the misspecification, and Model 4 resulted in a better fit to the data than did Model 3.

As with Model 3, under multivariate normality, the expected values of the ML and SB χ^2 were equal to one another whereas the expected value of the ADF χ^2 was smaller. Neither the ML or SB χ^2 showed any significant bias under the normal distribution across any of the four sample sizes. The largest bias was for the SB χ^2 at N = 100 (7%), but the magnitude of bias dropped to near 0 at sample sizes of N = 200 and above. In comparison, the ADF χ^2 was again significantly overestimated at the two smaller sample sizes but was unbiased at sample sizes of N = 500 and above.

The expected value of the ML χ^2 was again equal across distributions, and the ML χ^2 was increasingly positively biased with increasing nonnormality. Like Model 3, the expected values for the SB and ADF χ^2 decreased with increasing nonnormality. The SB χ^2 showed increasing positive bias with increasing nonnormality. This bias became negligible at N = 200 under moderate nonnormality (5%) but was still slightly biased even at N = 1,000 under severe nonnormality (10%). Finally, the ADF χ^2 was again strongly biased, with increasing nonnormality even at the largest sample size.

Discussion

The first two models were theoretically properly specified. Model 1 was estimated in the sample precisely as it existed in the population, whereas Model 2 included two parameters in the sample that did not exist in the population. The findings for the ML χ^2 from Models 1 and 2 closely replicate both previous theoretical predictions and empirical findings. For example, the ML χ^2 showed no evidence of bias across all sample sizes under multivariate normal distributions but was significantly inflated with increasing nonnormality. Thus, a correct model was significantly more likely to be erroneously rejected based on the ML χ^2 given departures from a multivariate normal distribution (thus resulting in an increased Type I error rate).

The ADF and SB test statistics have been proposed as alternatives to the normal theory ML test statistic when the observed data do not meet the multivariate normality assumption. Consistent with previous research on properly specified models, the ADF χ^2 was substantially inflated at smaller sample sizes, even under multivariate normal distributions. Although this small sample size inflation was exacerbated with increasing nonnormality, the ADF was unbiased at sample sizes of $N = 500$ and above, regardless of distribution. The SB χ^2 performed quite well across nearly all sample sizes and all distributions and showed no evidence of bias even under severely nonnormal distributions at sample sizes of $N = 200$ or more. These are very heartening findings for the practicing researcher who encounters nonnormal data as a way of life (e.g., in the study of adolescent substance use or psychopathology). Not only was the SB χ^2 accurate under even severely nonnormal distributions, but the SB χ^2 simplified to the ML χ^2 under conditions of multivariate normality. Assuming a properly specified model, the SB χ^2 appears to be a very useful measure of fit given moderately sized samples and nonnormal data.

Whereas many of the results from Models 1 and 2 were predicted from theory and previous research, the findings from Models 3 and 4 were not. Recall that Models 3 and 4 were two variations of a misspecified model where the model estimated in the sample did not conform to the model that existed in the population. Studying the behavior of the test statistics under these conditions is of particular interest given the high likelihood that the model estimated in the sample does not precisely conform to the model that exists in the population. The results for the ML χ^2 were as expected: The ML test statistic showed no evidence of bias at any sample size under multivariate normality but was increasingly inflated given increasing nonnormality. As in Models 1 and 2, the SB χ^2 also showed no evidence of bias at any sample size given multivariate normality, and thus simplified to the ML χ^2. Interestingly, the expected value for the ADF χ^2 under model misspecification was much smaller than that of the ML and SB, even under multivariate normality. This suggests that, compared to the ML and SB, the ADF test statistic may be a less powerful test of the null hypothesis. This conclusion is tentative, and more work is needed to better understand this finding. Like Models 1 and 2, the ADF was positively biased under multivariate normal distributions at the two smaller sample sizes but showed no bias at the two larger sample sizes.

The most surprising findings related to the behavior of the SB and ADF test statistics under the simultaneous conditions of misspecification and multivariate nonnormality (Models 3 and 4). The expected values of these test statistics markedly decreased with increasing nonnormality. That is, all else being equal, the SB and ADF test statistics were less likely to detect a specification error given increasing departures from a multivariate normal distribution. The more severe the nonnormality, the greater the corresponding loss of power. This result was unexpected, and we are not aware of any previous discussions of this finding.

Although the specific reason for this loss of power is currently not known, we theorize that it is due to the inclusion of the fourth-order moments (kurtoses) in the computation of the SB and ADF test statistics, information that is ignored by the normal theory ML χ^2. Recall that a normal distribution is completely described by the first two moments, the mean and the variance. As the distribution becomes increasingly nonsymmetric, is characterized by thicker or thinner tails (compared with the normal curve), or both, additional parameters are needed to describe this more complex distribution. Because ML is a normal theory estimator, it is assumed that the fourth-order moments are equal to 0, multivariate kurtosis is ignored, and the expected value of the ML χ^2 is equal across all distributions.[5] In contrast, the ADF and SB do not assume multivariate normality, the fourth-order moments are not assumed to equal 0, and measures of multivariate kurtosis are explicitly incorporated into the computation of the test statistics. As a result, the expected values of the ADF and SB directly depend upon the particular characteristics of the multivariate distribution under consideration.

The inclusion of multivariate kurtosis into the computation of the SB and ADF test statistics provides the critical information necessary to fully describe the more complex nonnormal distribution. However, this added information resulting from the more complex distribution also reduces the ability of the SB and ADF to identify a given model misspecification. Otherwise stated, we can think of a hypothetical signal to noise ratio in which the test statistic is attempting to identify the presence of the signal (i.e., the misspecification) against the background noise (i.e., the sampling variability of the data). Compared with the normal distribution, the nonnormal distribution is characterized by additional noise (in the form of nonzero kurtosis) that makes it correspondingly more difficult to identify the presence of the signal. Thus, any particular signal is easier to detect given multivariate normality than is the very same signal given the multivariate nonnormal distributions considered here. The power of the ADF and SB test statistics (and any test statistic that incorporates information from fourth-order moments) to detect a given misspecification is thus decreased as multivariate nonnormality increases.[6] This interpretation is only speculative, and we are currently working on discerning precisely why this loss of power under nonnormality exists.

There are two important implications of these findings for the practicing researcher. First, the SB χ^2 will almost always be smaller than the ML χ^2 under conditions of multivariate nonnormality. However, the lower SB χ^2 does not necessarily imply that the model is a better fit to the data because under nonnormality there is a simultaneous decrease in the ability of the SB χ^2 to detect a model misspecification. The SB χ^2 is smaller than the ML χ^2 because of two (inseparable) reasons: a correction for the inflation to the normal theory ML χ^2 and a decrease in statistical power to detect a misspecification. The ML χ^2 and SB χ^2 should thus be interpreted with this in mind.

A second implication of these findings is that if a researcher is planning a study that will not be characterized by a multivariate normal distribution, further steps must be taken to compensate for the decreased statistical power that results as a function of the nonnormal data (i.e., plan to include additional subjects in the study). For example, the power estimation methods developed by Satorra and Saris (1985) only apply to normal theory estimators. Using this method to compute the required sample size needed to achieve a given level of statistical power will be underestimated if the hypothesized model was misspecified and tested based on data that do not follow a multivariate normal distribution.

Recommendations

On the basis of the previous results, we have several recommendations for the practicing researcher. First, we have not identified at what point the data appreciably deviate from multivariate normality. Similar to previous researchers (e.g., Muthén and Kaplan, 1985, 1992), we found significant problems arising with univariate skewness of 2.0 and kurtoses of 7.0. Further research is needed to better understand more precisely when nonnormality becomes problematic, but it seems clear that obtained univariate values approaching at least 2.0 and 7.0 for skewness and kurtoses are suspect. Second, we agree with previous researchers (e.g., Hu et al., 1992; Muthén & Kaplan, 1992) that the ADF χ^2 not be used with small sample sizes. Although we found adequate behavior at samples as small as $N = 500$, other researchers have found problems with the ADF χ^2 at samples as large as $N = 5{,}000$ when testing more complex models (Hu et al., 1992). There are some epidemiological and catchment area studies that do have these large sample sizes available, and in these cases the ADF is a promising method of estimation, particularly for smaller models. Recent research has also shown the possibility of using bootstrapping techniques to compute more stable ADF

[5] Note that although the obtained ML χ^2 values increased with increasing nonnormality, the expected ML χ^2 values were equal across distribution.

[6] We thank both Albert Satorra and Peter Bentler, whom each independently suggested this same argument as a potential explanation for the obtained results.

χ^2 estimates (Yung and Bentler, 1994); however, more work is needed to explore the utility of this approach in applied research settings.

Finally, relative to the ML χ^2 and the ADF χ^2, the SB χ^2 behaved extremely well in nearly every condition across sample size, distribution, and model specification. Additionally, the SB χ^2 had the desirable property of simplifying to the ML χ^2 under multivariate normality. We thus recommend reporting both the ML χ^2 and the SB χ^2 when nonnormal data is suspected with the clear realization that the lower SB value may be reflecting decreased power and not simply that the model is a better fit to the data based on the SB χ^2. Model fit should thus be evaluated with appropriate caution. There are a few disadvantages to using the SB χ^2 in practice. One is that the computation of the SB χ^2 requires raw data, which might pose a problem for some researchers. Second, the SB χ^2 is currently only available in EQS. This poses a practical problem for researchers who are either not trained in or do not have access to EQS.

References

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

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

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

Browne, M. W. (1984). Asymptotic distribution free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 127–141.

Browne, M. W., Mels, G., & Coward, M. (1994). Path Analysis: RAMONA. In L. Wilkinson & M. A. Hill (Eds.), *Systat: Advanced applications* (pp. 163–224). Evanston, IL: SYSTAT.

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

Fleishman, A. I. (1978). A method for simulating nonnormal distributions. *Psychometrika, 43,* 521–532.

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

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL-VIII: A guide to the program and applications.* Mooresville, IN: Scientific Software.

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

Kaplan, D. (1989). A study of the sampling variability of the z-values of parameter estimates from misspecified structural equation models. *Multivariate Behavioral Research, 24,* 41–57.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method.* London: Butterworth.

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

MacCallum, R. (1990). The need for alternative measures of fit in covariance structure modeling. *Multivariate Behavioral Research, 25,* 157–162.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1990). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111,* 490–504.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166.

Muthén, B. (1987). *LISCOMP: Analysis of linear structural equations with a comprehensive measurement model.* Mooresville, IN: Scientific Software.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171–189.

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

Saris, W. E., & Stronkhorst, H. (1984). *Causal modelling in nonexperimental research.* Amsterdam: Sociometric Research Foundation.

SAS Institute, Inc. (1990). *SAS/STAT user's guide: Vols. 1 and 2.* Cary, NC: Author.

Satorra, A. (1990). Robustness issues in structural equation modeling: A review of recent developments. *Quality & Quantity, 24,* 367–386.

Satorra, A. (1991, June). *Asymptotic robust inferences in the analysis of mean and covariance structure.* Barcelona, Spain: Universitat Pompeu Fabra.

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *ASA Proceedings of the Business and Economic Section,* 308–313.

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

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48,* 465–471.

Yung, Y.-F., & Bentler, P. M. (1994). Bootstrap-corrected ADF test statistics in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology, 47,* 63–84.

Appendix

Technical Details

Data Generation

EQS generates the raw data with non-zero skewness and kurtosis using the formulae developed by Fleishman (1978) in accordance with the procedures described by Vale and Maurelli (1983). The raw data were generated based upon the covariance matrix implied by the model parameters for each of the three models. The availability of the raw data was necessary for the computation of the ADF and SB χ^2 test statistics.

The samples were created using the model implied population covariance matrix $\Sigma(\theta)$. The measurement equations consisted of the population parameter values that defined the particular model. EQS generated the population covariance matrix based on these measurement equations. The sample raw data were created using a random number generator in conjunction with the characteristics of the population covariance matrix. The raw data were generated under two constraints: (a) the expected value of S should equal the population covariance matrix $\Sigma(\theta)$, and (b) the expected value of the indices of skewness and kurtosis should equal the values specified for each measured variable.

Verification of Data Generation

To verify that EQS properly generated the raw data in accordance with the desired levels of skewness and kurtoses, three sets of raw data of sample size N = 60,000 were generated. The three data sets were produced using the same procedures that created the multivariate normal, moderately nonnormal, and severely nonnormal distributions for the simulations. The large sample size provides a more accurate estimate of the coefficients of skewness and kurtosis for the generated data.

Univariate skewness and kurtoses were computed for the three samples of N = 60,000 using SAS PROC UNIVARIATE. For the normally distributed condition (skewness = 0, kurtosis = 0), the mean univariate skewness for the nine variables was .001 and the mean kurtosis was .004. For the moderately nonnormal distribution (skewness = 2.0, kurtosis = 7.0), the mean skewness was 1.973 and the mean kurtosis was 6.648. Finally, for the severely nonnormal distribution (skewness = 3.0, kurtosis = 21.0), the mean skewness was 2.986 and the mean kurtosis was 21.44. These large sample values of skewness and kurtosis closely reflected the population values. Previous published studies have also successfully utilized this same method of data generation (Chou et al., 1991; Hu et al., 1992).

Expected Value of Test Statistics

For a properly specified model, the expected value for all three chi-square test statistics is equal to the model degrees of freedom. Thus, the expected chi-square for Model 1 was 24.0 and for Model 2 was 22.0. A complication arises when computing the expected value of the test statistics for the misspecified models. Under misspecification, the expected value of the model chi-square is a combination of the model degrees of freedom plus the noncentrality parameter, λ. The value of λ is dependent on both the particular method of estimation and sample size, with the expected value of the model chi-square becoming larger with increasing sample size.

Satorra and Saris (1985) provided a method for computing the noncentrality parameter for normal theory ML for misspecified models. First, a covariance matrix is created to reflect the structure of the model as it exists in the population. Second, this covariance matrix is used to estimate the model as it is thought to exist in the sample. The chi-square value that results from this model is the corresponding noncentrality parameter λ. This value, when added to the model degrees of freedom, provides the expected value of the ML χ^2 test statistic under model misspecification (Kaplan, 1988; Saris & Stronkhorst, 1984).

The Satorra-Saris method does not apply to the ADF or SB test statistics, and it is currently not known how to compute the theoretical expected values of these test statistics for misspecified models, particularly under conditions of nonnormality. We thus computed an empirical estimate of the expected value of the SB and ADF test statistics. Three samples of N = 60,000 were generated using EQS reflecting the normal, moderately nonnormal, and severely nonnormal distributions described previously. Models 3 and 4 were then fit to these three large samples, and the minimum of the fit function for SB and ADF was obtained. This value was then scaled by the sample size (minus 1) of interest (100, 200, 500, and 1000) and was added to the model degrees of freedom to result in a large sample empirical estimate of the expected value of the chi-square test statistics under misspecification. We thank Douglas Bonett for pointing out the dependence of the noncentrality parameter on the method of estimation and Peter Bentler for suggesting the procedure to compute the empirical estimates of the population noncentrality parameters. For comparative purposes, the expected values of the test statistics for Models 1 and 2 were also computed using the large sample empirical method. All expected values for all test statistics across all conditions were very close to the corresponding model degrees of freedom. Additionally, the Satorra-Saris method was used to compute the expected value for ML for Models 3 and 4, and these values closely approximated the large sample empirical estimates. This cross-validation of estimation methods increases our confidence in the accuracy of the large sample empirical estimates of the expected values.

Received March 27, 1995
Revision received August 7, 1995
Accepted October 23, 1995 ■