Can Test Statistics in Covariance Structure Analysis Be Trusted?

Li-tze Hu and P. M. Bentler
University of California, Los Angeles

Yutaka Kano
University of Osaka Prefecture
Osaka, Japan

Covariance structure analysis uses χ^2 goodness-of-fit test statistics whose adequacy is not known. Scientific conclusions based on models may be distorted when researchers violate sample size, variate independence, and distributional assumptions. The behavior of 6 test statistics is evaluated with a Monte Carlo confirmatory factor analysis study. The tests performed dramatically differently under 7 distributional conditions at 6 sample sizes. Two normal-theory tests worked well under some conditions but completely broke down under other conditions. A test that permits homogeneous nonzero kurtoses performed variably. A test that permits heterogeneous marginal kurtoses performed better. A distribution-free test performed spectacularly badly in all conditions at all but the largest sample sizes. The Satorra-Bentler scaled test statistic performed best overall.

Estimation methods in covariance structure analysis are traditionally developed under an assumption of multivariate normality (e.g., Bollen, 1989; Browne, 1974; Jöreskog, 1969). This assumption is usually violated in practice. For example, Micceri (1989) reported that among 440 large-sample achievement and psychometric measures taken from journal articles, research projects, and tests, all were significantly nonnormally distributed. Yet, normal-theory methods such as maximum likelihood (ML) and generalized least squares (GLS) are frequently applied even when normality assumptions are not tenable. A violation of the multivariate normality assumption can seriously invalidate statistical hypothesis testing (Browne, 1982, 1984; Harlow, 1985). As a result, a normal-theory test statistic may not adequately reflect the quality of a covariance structure model under such a violation. Asymptotic (large-sample) distribution-free methods, for which normality assumptions need not be made, therefore have been developed (Bentler & Dijkstra, 1985; Bentler, Lee, & Weng, 1987; Browne, 1982, 1984; Chamberlain, 1982) and made routinely available (Bentler, 1989; Jöreskog & Sörbom, 1988). Test statistics for the fit of a covariance structure model that are based on this theory are insensitive to the distribution of the observations when the sample size is large. Despite the preferable theoretical properties of these asymptotically distribution-free (ADF) methods, their wide application has been hampered because of the need for computing the fourth-order moments of the measured variables. These fourth-order moments reflect whether distributions have heavier or lighter tails as compared with normal distributions and are computationally expensive to obtain and unstable as estimators. In fact, empirical studies using Monte Carlo procedures have raised some questions about the relevance of ADF theory for practical data analysis because the basic goodness-of-fit test for model adequacy under arbitrary distributions may behave quite poorly, that is, not close to a theoretical χ^2 variate, as expected, when sample size is relatively small or model degrees of freedom are large (e.g., Chou, Bentler, & Satorra, 1991; Muthén & Kaplan, 1985, 1990; Tanaka, 1984). The limits of adequate performance are hardly known, but enough questions have been raised to again peak interest in simpler estimators that involve less computation.

A recently developed theory offers hope for the appropriate use of normal-theory methods even under violation of the normality assumption. On the basis of initial work by Amemiya (1985) and Browne (1985), the asymptotic robustness of normal-theory methods has been extensively studied (Amemiya & Anderson, 1990; Anderson & Amemiya, 1988; Browne, 1987; Browne & Shapiro, 1988; Mooijaart & Bentler, 1991; Satorra & Bentler, 1990, 1991; Shapiro, 1987). This literature has appeared only in statistical journals, so that it is hardly known to psychological researchers. The main point of this technical literature is to determine conditions under which models with nonnormally distributed variables can still be correctly described and evaluated by use of normal-theory-based methods such as ML or GLS. It is difficult to summarize verbally and succinctly this technical literature, but it has been shown that asymptotic optimality and correct standard errors of factor loadings can be obtained under normal-theory methods when the common factors are not normally distributed and the unique factors have a multivariate normal distribution and hence the observed variables are also nonnormal. For example, Anderson and Amemiya (1988) and Amemiya and Anderson (1990) have

This article was supported in part by United States Public Health Service Grants DA01070 and DA00017 and by the National Research Center on Asian-American Mental Health, National Institute of Mental Health, Grant R01 MH44331.

Correspondence concerning this article should be addressed to P. M. Bentler, Department of Psychology, 405 Hilgard Avenue, University of California, Los Angeles, California 90024-1563.

Psychological Bulletin, 1992, Vol. 112, No. 2, 351-362
 0033-2909/92/$3.00

asymptotic robustness

found that the asymptotic χ^2 goodness-of-fit test in factor analysis can be insensitive to violations of the assumption of multivariate normality of both common and unique factors, if all factors are independently distributed and the elements of the covariance matrices of common factors are all free parameters. With an additional condition of the existence of the fourth-order moments of both unique and common factors, Browne and Shapiro (1988) and Mooijaart and Bentler (1991) also demonstrated the robustness of normal-theory methods in the analysis of a general class of linear latent variate models. Satorra and Bentler (1990, 1991) obtained similar results for a wider range of discrepancy functions, estimators, and test statistics. In these results, the standard errors that are based on normal theory of some parameters, usually the variances of nonnormal variables, need correction, but the relevant computation is minor compared with that required by the distribution-free methods. Thus, asymptotic robustness theory promises to extend the range of applicability of the computationally simpler ML and GLS estimators to situations in which the more difficult distribution-free methods might seem to be needed.

In practice, the applied researcher may be tempted to use a normal-theory method in data analysis with nonnormal variables, justifying such a choice on the basis of asymptotic robustness theory. However, it is not at all certain that this theory can be invoked in practice, because nothing is known about the robustness of the asymptotic robustness theory, that is, whether asymptotic robustness theory can be applied when its assumptions such as large sample size and independence of latent variates may not hold. Adequate procedures to evaluate whether latent factors or errors not only are uncorrelated but furthermore are independent of each other do not exist. Note that independence is a much stronger condition than uncorrelatedness and that these concepts are equivalent only when variables are normally distributed.

Estimation methods that are based on distributional assumptions more general than normal, but more restricted than arbitrary, also have been developed. Browne (1982, 1984) introduced multivariate elliptical theory to covariance structure analysis. Elliptical distributions are, like the normal, symmetric, but they have tails that can be identical to those of a normal distribution as well as heavier or lighter. Browne's work was followed up by Bentler (1983; Bentler & Berkane, 1986) and Shapiro and Browne (1987), and computer implementations (e.g., ERLS in EQS) have been available for some years (Bentler, 1989). In these distributions, only one additional parameter beyond the usual normal-theory parameters is needed to yield asymptotically optimal estimators and simple χ^2 goodness-of-fit tests. This extra parameter is a kurtosis parameter reflecting the assumed common kurtosis of the variables, that is, the extent to which the distribution of variables is heavier tailed or lighter tailed as compared with the normal. Normal distributions are a special case that have no excess kurtosis. Computations are particularly simple when the model meets a scale invariance condition in which the model continues to hold when all variables are multiplied by a constant (Shapiro & Browne, 1987). Although elliptical theory is by now quite old, little is known about the robustness of elliptical theory statistics to violation of assumptions. One might expect that because normal theory is a special case of elliptical theory and elliptical methods reduce to normal-theory methods, elliptical methods should perform at least as well as or substantially better than normal-theory methods. However, this does not appear to be the case. Harlow (1985) found that elliptical χ^2 tests could be more misleading than normal-theory statistics, but this work has not been followed up.

A recent extension of elliptical distribution theory by Kano, Berkane, and Bentler (1990) has revealed that a simple adjustment of the weight matrix of normal theory, using univariate (marginal) kurtosis estimates, results in an asymptotically efficient estimator of structural parameters within the class of estimators that minimize a general discrepancy function. This method, called here HK (heterogeneous kurtosis) theory, is hardly more difficult computationally than elliptical theory but applies to a wider class of multivariate distributions that is allowed to have heterogeneous kurtosis parameters. That is, although these distributions are assumed to be symmetric, they need not be equally heavy- (or light-) tailed for all variables. Elliptical and normal-theory statistics are special cases that occur when the variables have homogeneous kurtoses or no excess kurtosis, respectively. Thus, one might expect that HK theory should perform at least as well as normal or elliptical methods.

An attractive feature of HK theory is that fourth-order moments of the measured variables do not need to be computed as they do in ADF theory, because these moments are just a function of the variances and covariances and the univariate kurtoses. As a result, the HK method can be used on models that are based on a substantially larger number of measured variables. For example, whereas ADF methods cannot be implemented in practice with 30, 40, or more variables because of the large size of a matrix that is required, this is not a limitation of the HK method. Except for an illustration given in the initial report, nothing is known about the performance of HK theory under violation of its assumptions, or even when its assumptions are met, as when the data are normal or elliptical.

When normal-, elliptical, or heterogeneous kurtosis theory distributional assumptions are false, statistics that are based on these assumptions can be corrected using a method developed by Satorra and Bentler (1988a, 1988b) and further studied by Kano (1990). In their approach, a scaling correction is computed on the basis of the model, estimation method, and sample fourth-order moments, and the given test statistic is divided by this correction factor. The correction factor has no impact when the distributional assumption is correct. This approach, which is a type of Bartlett correction to the χ^2 statistic, has not been evaluated widely even though it has been available in the EQS program (Bentler, 1989) for some years. Preliminary indications are that this corrected test statistic, here called the SCALED statistic, can perform as well as, or perhaps better than the ADF method under violation of distributional assumptions (Chou et al. 1991). However, nothing definitive is known.

The purpose of this study is to evaluate the performance of six goodness-of-fit test statistics obtained from a variety of estimators under violation of assumptions, that is, the empirical robustness of these statistics. In all cases, under an assumed distribution and a hypothesized model $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ for the population covariance matrix $\boldsymbol{\Sigma}$, these statistics have an asymptotic χ^2 distribution that describes the mean, variance, and tail performance of the statistics. Three ways of violating theoretical con-

ditions, chosen for their relevance to data analysis practice and recent theoretical results, are investigated. First, distributional assumptions are violated. Second, assumed independence conditions are violated. Third, asymptotic sample size requirements are violated. The effects of these conditions on normal-theory maximum likelihood (ML) and generalized least squares (GLS), elliptical theory (ERLS), heterogeneous kurtosis (HK), asymptotic distribution-free (ADF), and scaling-corrected (SCALED) test statistics (T_{ML}, T_{GLS}, T_{ERLS}, T_{HK}, T_{ADF}, T_{SCALED}) are studied in an extensive Monte Carlo sampling experiment. Technical definitions for these statistics are given in the Appendix.

Method

The confirmatory factor model $\mathbf{x} = \Lambda\xi + \epsilon$ is used to generate measured variables $\mathbf{x}$ under various conditions on the common factors ξ and unique variates ("errors") ϵ. In the usual approach, factors and errors are assumed to be normally distributed, factors are allowed to correlate with covariance matrix $\mathcal{E}(\xi\xi') = \Phi$, errors are uncorrelated with factors, that is $\mathcal{E}(\xi\epsilon') = 0$, and various error variates are uncorrelated and have a diagonal covariance matrix $\mathcal{E}(\epsilon\epsilon') = \Psi$. As a result, $\Sigma = \Sigma(\theta) = \Lambda\Phi\Lambda' + \Psi$, and the elements of θ are the unknown parameters in Λ, Φ, and Ψ. In one condition, factors and errors are created to be multivariate normally distributed, so that the latent variates that are uncorrelated in the factor model are also independent of each other. Additional conditions are created in which the factors or errors (or both factors and errors) are not normally distributed; in some of these conditions, factor/error variates that are uncorrelated under the model also are independent, but in other conditions, variables that are uncorrelated as assumed by the factor model are not also independent. After generation of the population covariance matrix Σ under the assumed conditions, random samples of a given size from the population are taken. In each sample, the parameters of the factor model are estimated using the methods ML, GLS, ERLS, HK, ADF, and SCALED as described above and in the Appendix, and the associated test statistics T_{ML}, T_{GLS}, T_{ERLS}, T_{HK}, T_{ADF}, and T_{SCALED} are computed. Results for each sample are saved. The performance of the test statistics across the sampling replications at a given sample size represents the main data of the study.

In particular, the confirmatory factor-analytic model studied is based on 15 observed variables with three common factors. The factor-loading pattern in Λ (15×3) is a simple one in which a variable is influenced by one, and only one, common factor and the three common factors are allowed to correlate. The factor-loading matrix (transposed) Λ' has the following structure:

$$\begin{bmatrix} 0.70 & 0.70 & 0.75 & 0.80 & 0.80 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.70 & 0.70 & 0.75 & 0.80 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}$$

$$\begin{bmatrix} 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.80 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.70 & 0.70 & 0.75 & 0.80 & 0.80 \end{bmatrix}.$$

Variances of the factors are 1.0, and the covariances among the three factors are taken to be 0.30, 0.40, and 0.50. The unique variances are taken as values that would yield unit-variance measured variables under normality. In estimation, the factor loading of the last indicator of each factor is fixed for identification at 0.80, and the remaining nonzero parameters are free to be estimated. The behavior of the various test statistics $T = (n-1)\hat{F}$ (for variously defined functions F) is observed at sample sizes of 150, 250, 500, 1,000, 2,500, and 5,000. In each condition at each sample size, 200 replications (samples) are drawn from the population. The various estimators and goodness-of-fit tests are computed in each sample using a modification of the simulation feature of EQS (Bentler, 1989). A statistical summary of the mean value and standard deviation of T across the 200 replications, and the empirical rejection rate at the $\alpha = .05$ level on the basis of the assumed χ^2 distribution, is used to compare the various methods.[1]

Table 1 contains the experimental design and the expectations (or the first-order moments) of the asymptotic distributions of the goodness-of-fit test statistics T in each condition. Under the modeling conditions, the expectations in Table 1 would be very close to the corresponding empirical mean values of the test statistics, one of the statistical summaries in our experiment, if the sample size and number of replications are large enough. In some conditions, these are the expected values of a central χ^2 variate, that is, the degrees of freedom, but in other cases, these are other values as based on predictions from the current literature and the rationale given below. Seven conditions are examined, as shown in the rows of the table. These conditions correspond to various distributional specifications on the common and unique (error) factors. In Condition 1, all factors are normally distributed, with no excess kurtosis. In Conditions 2 and 3, all factors are nonnormally distributed. In these conditions, the true kurtoses, using the formula $\sigma_{iiii}/\sigma_{ii}^2 - 3$, for the nonnormal factors in the various conditions are -1.0, 2.0, and 5.0, and the true kurtoses of the unique variates for Conditions 2–4 with nonnormal errors are -1.0, 0.5, 2.5, 4.5, 6.5, -1.0, 1.0, 3.0, 5.0, 7.0, -0.5, 1.5, 3.5, 5.5, and 7.5. In Conditions 1–4, all the factors and unique variates are independently distributed regardless of whether they are normally distributed. Additionally, in Condition 2, all elements in the factor covariance matrix Φ are fixed at their true values. Thus, although Conditions 3–4 are designed to be consistent with asymptotic robustness theory, the fixed covariance matrix invalidates this theory, and asymptotic robustness of normal-theory test statistics would not be expected in Condition 2. In Conditions 5–7, the factors and error variates are divided by a random variable $Z = [\chi^2_{(5)}]^{1/2}/\sqrt{3}$ that is distributed independently of the original common and unique factors. The division by $\sqrt{3}$ is made so that $\mathcal{E}(Z^{-2}) = 1$, that is, the variances and covariances of the factors remain unchanged by the division (Kano, 1990), but the kurtoses of the factors and errors become modified. A consequence of the division by a random variable is that the factors and errors are dependent, even though they remain uncorrelated. Because of the dependency, asymptotic robustness of normal-theory statistics is not to be expected under Conditions 5–7.

Under the model $\Sigma(\theta)$, as can be seen in the table, in many conditions the anticipated means of the asymptotic goodness-of-fit statistics are the degrees of freedom. The degrees of freedom for the model in all conditions but the second is 87, whereas the degrees of freedom for the model in Condition 2 is 93. Under other conditions, the predicted means of the test statistics depend on the kurtoses of the variables as well. These values were computed using the known relation of normal to elliptical theory as well as the results of Kano (1990). In all cases, the expectations of Table 1 are based on the assumption of the correctness of the model $\Sigma = \Sigma(\theta)$ and the assumption of infinite sample size. As noted above, in the simulation, the correctness of the model is maintained in all sampling situations and conditions, but sample size is varied in a standard range.

The simplest predictions are made for Condition 1 (row 1) and the ADF and SCALED statistics (last column) of Table 1. In row 1, all variables are normally distributed, and hence $\mathcal{E}(T) = 87$ for all testing conditions. The anticipated means of T_{ADF} and T_{SCALED} are the degrees of freedom, regardless of the seven types of distributions and conditions that are considered. The results of T_{ML} and T_{GLS} (in the column

[1] The number of replications was chosen as a compromise between practicality and a very large number. Each condition took several weeks of central processing unit time on DECstation 3100 and VAXstation 3100 workstations.

Table 1
Asymptotic Expected Goodness-of-Fit Statistics Under Seven Conditions

Distribution		Method of estimation			
Factor	Unique factor	ML & GLS	ERLS	HK	ADF & SCALED
1. Normal	Normal	87	87	87	87
2. Non[a]	Non	—	—	—	93
3. Non	Non	87	$87/\kappa_3$	—	87
4. Normal	Non	87	$87/\kappa_4$	—	87
5. Normal/Z^b	Normal/Z^b	87*3	87	87	87
6. Normal/Z^b	Non/Z^b	87*3	$87*3/\kappa_6$	—	87
7. Non/Z^b	Non/Z^b	87*3	$87*3/\kappa_7$	—	87

Note. ML = maximum likelihood, GLS = generalized least squares, ERLS = elliptical theory, HK = heterogenous kurtosis, ADF = asymptotic distribution free, SCALED = Satorra-Bentler scaling corrected, Non = nonnormal distribution.
[a] Factor covariance matrix Φ is taken as known and fixed. [b] Common and unique factors are divided by the same variate $Z = [\chi^2_{(5)}]^{1/2}/\sqrt{3}$.

marked ML & GLS) depend on the independence of factors from errors, as noted above. When they are independent, as in Conditions 1 and 3–4, asymptotic robustness conditions apply, and $\mathcal{E}(T_{ML}) = \mathcal{E}(T_{GLS}) = 87$, but the fixed factor covariance matrix in Condition 2 invalidates robustness theory, and no prediction is made there. Lack of robustness for normal-theory methods in Conditions 5–7 leads to expectations of $261 = 87 \times 3$ as the mean value of the normal-theory test statistics, which is obtained from a chi-squared variable multiplied by $\mathcal{E}(Z^{-4})/\mathcal{E}(Z^{-2})^2 = 3$ (Kano, 1990). The predicted mean of the statistic T_{ERLS} of elliptical theory is given by the expected value under the ML method divided by $(\kappa + 1) = \mathcal{E}[(x-\mu)'\Sigma^{-1}(x-\mu)]^2/p(p+2)$, which is the rescaled Mardia kurtosis parameter for multivariate distributions. Note that κ_3, κ_4, κ_6, and κ_7 in Table 1 are the specific values of the Mardia kurtosis defined for different distributions. They have the following relationships: $3 \times \kappa_3 = \kappa_7$ and $3 \times \kappa_4 = \kappa_6$. Predictions are made for the HK theory only under normality (Condition 1) and elliptical (Condition 5) distributions. Of course, HK should yield the same result as the normal-theory statistic in Condition 1, and because the distribution in Condition 5 is elliptical, ERLS and HK methods are expected to work correctly.[2]

To summarize, the calculated sample mean of the test statistic across the 200 replications in each condition should approximate the degrees of freedom (87 or 93) if the given statistic were χ^2 distributed. Otherwise, on the basis of our analysis, these means would be close to the alternative anticipated means shown in Table 1. We also tabulated the sample standard deviation of the test statistic across the 200 replications. Under the model and a χ^2 distribution, the variance of a χ^2 variate is twice the degrees of freedom, which is approximately the value that should be observed empirically across the 200 samples in each condition if χ^2 is indeed the appropriate reference distribution for that condition. In addition, we tabulated the frequencies of rejection of the model within each condition. If each statistic were performing as a χ^2 variate, the expected number of rejections at the 5% level would be 10.

Results

The simulation procedure produced the distributional characteristics desired. In each condition, computed across all cases in all replication samples, the means of all factors, errors, and measured variables were essentially zero (with a maximum deviation of 0.002), and standard deviations of latent and measured variables were similarly close to 1.0. Skewnesses of all latent and observed variables were close to zero as well, with a maximum deviation of 0.09. To provide some idea about the degree of nonnormality of the factors and errors, we present in Table 2 the empirical univariate kurtoses of the latent generating variables computed across $5{,}000 \times 200 = 1{,}000{,}000$ observations. As intended, the kurtoses of the factors in Conditions 1 and 4 were close to zero, and the kurtoses in Conditions 2 and 3 were close to −1, 2, and 5. The kurtoses of all variables were similar in Condition 5, the elliptical condition. Kurtoses of the factors were large and homogeneous in Condition 6 and larger on the average and more heterogeneous in Condition 7. Kurtoses of the error variates in Conditions 2–4 were very close to the magnitude intended, as given above. Kurtoses of the errors were largest in Conditions 6 and 7. Finally, it may be instructive to indicate how kurtoses of the factor and error variates translated into the range of kurtoses on the measured variables. These ranges, in Conditions 1–7, were (1) −0.010 to 0.010, (2) −0.502 to 3.098, (3) −0.502 to 3.098, (4) −0.262 to 0.989, (5) 4.658 to 6.827, (6) 4.635 to 9.659, (7) 3.930 to 20.013.

Tables 3 through 9 contain summaries of the results of the simulation, one table per condition. All of the tables are organized the same way. The columns of each table give the sample size used for a particular set of 200 replications from the population. At each sample size, a sample was drawn, and each of the six methods shown in the rows of the table (ML, GLS, ERLS, HK, ADF, SCALED) was applied to estimate the parameters of the model and compute the resulting test statistic T; this process was repeated 200 times. For each estimation method, the resulting T statistics were used to compute (a) the mean of the 200 T statistics, (b) the standard deviation of the 200 T statistics, and (c) the frequency of rejecting the null hypothesis at the .05 level. These are the three entries in each cell of each table. As noted in the columns of each table, these procedures were repeated at sample sizes of 150, 250, 500, 1,000, 2,500, and 5,000. When converged solutions were not obtained in each of the 200 replications of a given cell of the table, the statistics reported in the table are based on the results for those replications that did converge. The ADF method at the smallest sample size provided the only consistent lack of convergence.

The results of Condition 1 are easiest to understand, because it is the baseline condition in which the factors and errors, and hence measured variables, are multivariate normally distributed. The results are tabulated in Table 3. Asymptotically, each estimation method should yield a mean test statistic T of about 87, a standard deviation of $13.19 = \sqrt{174}$ ($174 = 2 \times 87$), and $10 = .05 \times 200$ rejections. The last column of the table gives the results for $n = 5{,}000$, which is as close to asymptotic sample size as was considered in this study. For all six estimation methods, the mean goodness-of-fit statistic T was quite close to 87; the

[2] A technical reviewer questioned whether there is a theoretical result showing that the moments of the test statistics converge to those of the limiting distribution. In general, this occurs under conditions that the statistics are uniformly integrable. Usually this is easily verified, but when the statistics are defined by an implicit function, as in this article, it is hard to do, though we believe that uniform integrability is met in our situation.

Table 2
Univariate Kurtoses Across Replications of Factors and Errors Under Seven Conditions

	Condition						
Variate	1	2	3	4	5	6	7
F1	0.005	−0.997	−0.997	0.004	5.087	5.087	2.521
F2	−0.002	2.038	2.038	0.001	6.031	6.031	17.992
F3	0.004	5.126	5.126	0.001	5.545	5.545	21.354
E1	−0.007	−1.001	−1.001	−1.001	4.877	2.609	2.609
E2	−0.008	0.511	0.511	0.511	5.963	7.474	7.474
E3	−0.000	2.505	2.505	2.505	4.672	10.383	10.383
E4	−0.010	4.461	4.461	4.461	4.522	13.977	13.977
E5	−0.008	6.283	6.283	6.283	4.914	19.343	19.343
E6	−0.002	−0.999	−0.999	−0.999	6.052	3.163	3.163
E7	0.005	1.005	1.005	1.005	5.653	9.477	9.477
E8	−0.005	3.036	3.036	3.036	5.216	11.616	11.616
E9	−0.002	5.013	5.013	5.013	4.321	15.050	15.050
E10	0.010	6.957	6.957	6.957	4.753	19.941	19.941
E11	0.001	−0.500	−0.500	−0.500	5.932	4.368	4.368
E12	−0.004	1.521	1.521	1.521	4.755	8.171	8.171
E13	0.001	3.467	3.467	3.467	5.114	14.154	14.154
E14	−0.002	5.584	5.584	5.584	4.755	19.168	19.168
E15	−0.003	7.770	7.770	7.770	5.083	28.349	28.349

standard deviations were a bit smaller than 13.19, and the rejections were just slightly below 10. Clearly, however, these results are very close to the theoretical values, indicating that the Monte Carlo procedure as implemented in the computer program was working correctly. In addition, note the following features. As can be seen in the first row, the ML method worked well when sample sizes were equal or greater than 500, but the rejection frequency was higher than nominal at smaller sample sizes. Similar results have been reported by Boomsma (1983). The GLS method performed better than ML at the smaller sample sizes, though at the smallest sample size models were rejected too infrequently. The ERLS and HK methods seemed to perform equally well, and a bit better than ML at the two smallest sample sizes. The ADF method yielded unacceptably high rejection rates and test statistics at all sample sizes up to 1,000, with performance being completely unacceptable at n = 250 or below, where almost all true models were rejected. At n = 150, only 191 replications yielded converged solutions, all of which rejected the null hypothesis. In contrast, the SCALED statistic, which was based on the ML statistic, performed about the same as the ML statistic itself.

The results for Condition 2 are presented in Table 4. The mathematical–statistical asymptotic robustness theory does not predict robustness when all the elements in Φ are fixed at true values rather than free to be estimated. However, the behaviors of all the methods in Condition 2 were similar to those in Conditions 3 and 4, discussed next. Thus, the main surprise here was that normal-theory methods showed asymptotic robustness under a condition in which it was not expected.

For Conditions 3 and 4, where asymptotic robustness for normal-theory statistics had been predicted, ML and GLS indeed performed well, as is shown in Tables 5 and 6. That is, at sample sizes of 2,500 and 5,000, the statistics yielded their expected behavior. As with the results under normality, GLS performed somewhat better than ML at the smaller sample sizes, with ML tending to reject models somewhat too frequently. ERLS and HK tended to overcorrect, that is, to accept models too readily. Although ADF performed well at n = 5,000, at smaller sample sizes, it rejected models far too frequently. At n = 150, about 5% of the replications did not yield converged solutions. In all the converged solutions, the true model was rejected. Conditions 3 and 4 thus represent situations in which ML and GLS methods indeed perform better than ADF methods at all but the largest sample size. The SCALED test statistic performed better than ADF at all but the largest sample size, where it performed equally well. It performed about the same as the ML statistic, though with marginally greater rejection of true models at the smaller sample sizes.

In summary, when the latent common and unique factors were independently distributed, regardless of the form of their distributions (Conditions 1–4), the anticipated asymptotic robustness properties were retained for normal-theory methods if the sample size was relatively large. Asymptotic robustness, however, could not be guaranteed at smaller sample sizes with ML. ERLS slightly, and HK somewhat more, overcorrected the test statistics when some or all the latent variates were nonnormal (Conditions 2–4). The ADF method was very sensitive to sample size; except under normality, it did not even perform acceptably with a sample size as large as 2,500. The SCALED statistic outperformed ADF at all but the largest sample sizes.

When the factors and errors were dependent on each other, that is, in Conditions 5–7, as summarized in Tables 7–9, asymptotic robustness theory was not relevant, and the results indicate that empirical robustness completely broke down. Specifically, the normal-theory methods (ML, GLS) essentially always rejected the true model even at the largest sample sizes. The ERLS method performed substantially better than the normal-theory methods under all these conditions, though it did not perform perfectly under the elliptical distribution condition, Condition 5 (Table 7), where it retained a tendency to reject models too frequently at even the largest sample size. On the other hand, it performed quite well at the largest sample sizes in Conditions 6 and 7, which are not elliptical. The HK method performed considerably better than all but the

Table 3
Summary of Simulation Results for Condition 1

Method	n 150	250	500	1,000	2,500	5,000
ML						
M	92.674	90.540	87.771	86.166	87.136	86.583
SD	13.175	14.622	12.617	12.450	12.232	11.728
Freq	20	21	9	5	4	8
GLS						
M	85.491	86.546	85.355	85.214	86.779	86.328
SD	12.440	13.667	12.212	12.490	12.154	11.578
Freq	5	12	5	5	3	7
ERLS						
M	89.887	88.905	87.250	85.812	86.986	86.573
SD	12.363	14.496	12.558	12.241	12.205	11.794
Freq	11	18	7	5	5	9
HK						
M	87.747	88.031	86.180	85.579	86.882	86.373
SD	13.441	14.103	12.634	12.575	12.084	11.531
Freq	8	15	5	7	2	7
ADF						
M	229.118	144.930	109.333	96.077	91.096	88.447
SD	48.426	27.487	17.613	15.017	13.222	12.057
Freq	191/191	184	100	32	12	9
SCALED						
M	94.469	91.540	88.251	86.398	87.251	86.654
SD	13.485	14.732	12.644	12.426	12.246	11.750
Freq	23	22	8	5	6	9

Note. ML = maximum likelihood, GLS = generalized least squares, ERLS = elliptical theory, HK = heterogenous kurtosis, ADF = asymptotic distribution free, SCALED = Satorra-Bentler scaling corrected, Freq = frequency of rejection of the null hypothesis [critical value from $\chi^2_{(p^*-q)}$ at $\alpha = .05$]. All numbers are based on the number of converged replications. Usually this is 200, but when two numbers are given in an entry in the freq row, the first gives the frequency of rejection of the null hypothesis, and the second gives the number of converged replications.

SCALED statistic across the different sample sizes, though it retained its tendency to accept models too frequently. As under independence of factors and errors, when the sample sizes were smaller than 2,500, ADF consistently yielded test statistics that were too large and rejection rates that were too high. However, at $n = 5{,}000$, the ADF method performed as expected. The SCALED statistic performed better than the ADF method at all sample sizes below 2,500, though it tended to overreject models at the largest sample size.

These results can be summarized as follows. Under conditions of dependency among latent factors and unique variates, normal-theory methods cannot be trusted, HK works substantially better, ADF works only at very large sample sizes, and across all sample sizes, the Satorra-Bentler SCALED statistic performs at closest to nominal levels of all the methods considered.

Discussion

This study has several important implications for practice, especially with regard to the ADF method, normal-theory methods, and the SCALED statistic. Regarding the ADF method, our results provide support for and extend the cautions raised, for example, by Harlow, Chou, and Bentler (1986), Muthén and Kaplan (1990), and Chou et al. (1991). ADF theory originally was introduced as a general-purpose solution to the problem of nonnormal distributions of variables in structural modeling (Browne, 1982; Chamberlain, 1982); that is, it was expected that the ADF method would work well for any arbitrary distribution. The results of the present study indicate that this expectation is correct asymptotically. T_{ADF} does perform as a χ^2 variate when the sample size is about 5,000 cases under the modeling conditions studied. However, applications in practice typically have substantially smaller sample sizes, for which our results are even more pessimistic than those reported by Muthén and Kaplan (1990). In more than half of the conditions, a sample size of 2,500 was not large enough to yield the number of model rejections expected on the basis of the assumed χ^2 distribution, and in none of the conditions was a sample size of 1,000 large enough to yield a nominal rejection rate. At sample sizes of 250 or less, the ADF statistic, when it could be computed, yielded model rejections from 93%–99.9% of the time: It almost never correctly diagnosed that a true model was being evaluated. This spectacularly poor performance occurred even with multivariate normal data that would be expected to show the best behavior of the statistic. Thus, clearly ADF is not a general panacea to the problem of nonnormal variables in structural modeling. In fact, like the normal-

Table 4
Summary of Simulation Results for Condition 2

Method	n 150	250	500	1,000	2,500	5,000
ML						
M	98.786	98.489	94.852	96.065	96.416	95.317
SD	14.837	13.989	14.526	15.373	15.121	14.421
Freq	21	25	14	19	14	10
GLS						
M	94.005	94.795	93.120	95.224	96.020	95.050
SD	14.353	13.834	14.446	15.481	15.188	14.433
Freq	12	14	17	21	14	11
ERLS						
M	88.818	88.674	84.789	85.324	85.404	84.409
SD	13.602	12.928	13.297	14.055	13.483	12.723
Freq	6	6	2	4	4	2
HK						
M	81.646	78.678	74.843	74.902	73.873	72.828
SD	14.704	13.484	13.758	14.047	12.844	11.696
Freq	4	1	0	1	0	0
ADF						
M	282.501	166.790	119.898	106.837	99.099	95.619
SD	69.391	30.284	20.245	18.449	15.395	14.167
Freq	195/195	196	105	55	25	13
SCALED						
M	100.007	98.415	94.063	95.123	95.199	94.047
SD	15.039	13.944	14.754	15.569	14.968	14.236
Freq	25	20	14	18	13	9

Note. ML = maximum likelihood, GLS = generalized least squares, ERLS = elliptical theory, HK = heterogenous kurtosis, ADF = asymptotic distribution free, SCALED = Satorra-Bentler scaling corrected, Freq = frequency of rejection of the null hypothesis [critical value from $\chi^2_{(p^*-q)}$ at $\alpha = .05$]. All numbers are based on the number of converged replications. Usually this is 200, but when two numbers are given in an entry in the freq row, the first gives the frequency of rejection of the null hypothesis, and the second gives the number of converged replications.

Table 5
Summary of Simulation Results for Condition 3

	n					
Method	150	250	500	1,000	2,500	5,000
ML						
M	91.752	90.800	87.571	88.531	88.766	87.517
SD	14.481	13.401	14.227	14.620	14.056	13.559
Freq	20	20	14	16	10	10
GLS						
M	85.072	86.138	85.450	87.366	88.320	87.228
SD	12.342	12.274	13.353	14.239	14.043	13.441
Freq	8	5	8	14	12	9
ERLS						
M	82.491	81.682	78.242	78.582	78.549	76.547
SD	13.255	12.487	13.197	13.408	12.486	11.763
Freq	8	4	3	5	4	2
HK						
M	73.854	71.818	69.112	69.216	68.499	67.493
SD	12.733	12.242	12.757	12.978	11.899	11.077
Freq	3	0	0	1	0	0
ADF						
M	216.807	143.761	108.541	97.816	91.983	88.899
SD	39.495	25.163	18.668	16.441	14.311	13.467
Freq	188/188	185	89	44	18	11
SCALED						
M	93.777	92.066	88.160	88.914	88.896	87.568
SD	14.403	13.427	14.372	14.786	14.083	13.551
Freq	25	25	16	18	12	10

Note. ML = maximum likelihood, GLS = generalized least squares, ERLS = elliptical theory, HK = heterogenous kurtosis, ADF = asymptotic distribution free, SCALED = Satorra-Bentler scaling corrected, Freq = frequency of rejection of the null hypothesis [critical value from $\chi^2_{(p^*-q)}$ at $\alpha = .05$]. All numbers are based on the number of converged replications. Usually this is 200, but when two numbers are given in an entry in the freq row, the first gives the frequency of rejection of the null hypothesis, and the second gives the number of converged replications.

theory methods ML and GLS, under some conditions it will yield completely misleading results. Typically, models that are true would be rejected far too frequently when using the ADF method.

An important theoretical question for future research involves development of a detailed mathematical explanation of why the ADF method seems to break down so easily. Here, we provide a possible explanation. Consider Condition 1, the normal case, for instance. According to asymptotic theory, ADF should work as well as the GLS method. Yet this does not happen. Obviously, a reason for this is that the required sample size for asymptotic theory to be relevant to the behavior of a test statistic must depend on the particular method of estimation under consideration, with ADF requiring much larger samples than GLS. Because the only distinction between the two methods is whether the actual fourth-order moments or their expression in terms of second-order moments under normality is used in the test statistic, differential behavior must be traced to this difference. More specifically, GLS uses the weight matrix with typical element

$$s_{ik}s_{jl} + s_{il}s_{jk}, \tag{1}$$

which is based on sample covariances s_{ij} only, whereas ADF uses the expression

$$s_{ijkl} - s_{ij}s_{kl}, \tag{2}$$

which requires estimating the fourth-order moments s_{ijkl} (see Appendix for a precise definition). Although under normality, both Equations 1 and 2 are estimators of the population parameters $\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$ and converge in probability to these parameters as the sample size goes to infinity, their variability is different. In fact, because Equation 1 is the ML estimator of $\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$ and is efficient (under no covariance structure), it follows that elementwise,

$$Var(s_{ik}s_{jl} + s_{il}s_{jk}) \leq Var(s_{ijkl} - s_{ij}s_{kl}). \tag{3}$$

In addition, the large covariance matrices of order $p(p+1)(p+2)(p+3)/24$ made up of these elements will have the same inequality as in Equation 3, that is,

$$MatCov\{s_{ik}s_{jl} + s_{il}s_{jk}\} \leq MatCov\{s_{ijkl} - s_{ij}s_{kl}\},$$

denoting that the difference is a nonnegative definite matrix. The direct assessment of their variances also will show these inequalities.

In summary, because the distinction between GLS and ADF is based on the differences in the weight matrices, as shown above, this difference must yield the differential performance of these methods. Although both weight matrices meet the re-

Table 6
Summary of Simulation Results for Condition 4

	n					
Method	150	250	500	1,000	2,500	5,000
ML						
M	91.818	90.890	87.634	88.496	88.546	87.619
SD	14.203	13.464	14.158	14.483	14.059	13.490
Freq	19	19	13	18	11	9
GLS						
M	85.107	86.228	85.546	87.318	88.087	87.333
SD	12.178	12.376	13.325	14.106	14.018	13.355
Freq	8	8	9	14	10	8
ERLS						
M	83.352	82.803	79.448	79.824	79.721	78.876
SD	13.048	12.613	13.141	13.370	12.706	12.153
Freq	7	5	3	6	4	1
HK						
M	81.055	80.381	78.060	78.622	78.422	77.614
SD	12.376	12.368	12.691	13.002	12.678	11.900
Freq	7	2	2	3	4	1
ADF						
M	220.571	144.184	108.728	97.737	91.791	89.007
SD	41.290	25.157	18.915	16.435	14.310	13.406
Freq	193/193	188	88	46	19	11
SCALED						
M	93.742	92.091	88.164	88.827	88.668	87.664
SD	14.164	13.506	14.320	14.647	14.105	13.473
Freq	25	24	16	17	11	9

Note. ML = maximum likelihood, GLS = generalized least squares, ERLS = elliptical theory, HK = heterogenous kurtosis, ADF = asymptotic distribution free, SCALED = Satorra-Bentler scaling corrected, Freq = frequency of rejection of the null hypothesis [critical value from $\chi^2_{(p^*-q)}$ at $\alpha = .05$]. All numbers are based on the number of converged replications. Usually this is 200, but when two numbers are given in an entry in the freq row, the first gives the frequency of rejection of the null hypothesis, and the second gives the number of converged replications.

Table 7
Summary of Simulation Results for Condition 5

	n					
Method	150	250	500	1,000	2,500	5,000
ML						
M	170.213	179.175	195.297	202.506	219.422	228.736
SD	44.950	48.089	63.894	58.175	69.067	92.457
Freq	197	197	200	200	200	200
GLS						
M	129.819	146.658	168.294	184.205	207.902	219.653
SD	20.560	25.472	31.581	38.030	52.392	62.771
Freq	166	187	199	200	200	200
ERLS						
M	114.569	108.920	104.427	98.028	94.218	90.800
SD	31.183	25.501	24.478	20.129	18.189	17.692
Freq	94	83	57	48	34	19
HK						
M	76.931	80.550	83.142	84.540	88.097	87.504
SD	14.488	13.748	13.664	14.671	16.727	16.330
Freq	3	3	3	8	22	15
ADF						
M	208.613	138.331	109.313	97.438	92.535	89.812
SD	38.930	22.694	15.733	12.593	13.981	13.419
Freq	170/170	181	90	28	17	12
SCALED						
M	91.578	88.805	87.587	86.409	86.650	85.620
SD	12.411	12.735	11.540	12.563	13.109	12.724
Freq	16	11	4	6	9	5

Note. ML = maximum likelihood, GLS = generalized least squares, ERLS = elliptical theory, HK = heterogenous kurtosis, ADF = asymptotic distribution free, SCALED = Satorra-Bentler scaling corrected, Freq = frequency of rejection of the null hypothesis [critical value from $\chi^2_{(p^*-q)}$ at $\alpha = .05$]. All numbers are based on the number of converged replications. Usually this is 200, but when two numbers are given in an entry in the freq row, the first gives the frequency of rejection of the null hypothesis, and the second gives the number of converged replications.

quirement for the associated T statistics to converge to an asymptotic χ^2 variate, the speed of convergence must depend on the variability of the weight matrices. Thus, we would expect that variability of the ADF weight matrix would be excessive when sample size is small. Further research can evaluate this suggestion empirically.

Under conditions in which latent common factors and unique factors were distributed independently of each other (Conditions 1–4), which is a basic requirement of asymptotic robustness theory, normal-theory methods (ML, GLS) outperformed ADF at all but the very largest sample sizes. This occurred even when the data were quite nonnormally distributed and even under a condition in which asymptotic robustness theory had not been shown to hold (Condition 2). Thus, conditions certainly exist in which nonnormal data are better analyzed for model adequacy using normal-theory methods than those specifically developed for nonnormal data. Note, however, that the word *asymptotic* in asymptotic robustness theory requires careful attention. At the smaller sample sizes, this theory also had some limitations. For ML, sample sizes of 2,500 were needed before the rejection rate approached nominal levels. GLS performed much better, performing near nominal levels at even the smallest sample sizes with only a few exceptions. As yet, there is no theory to explain the differential behaviors of ML and GLS or to explain the robustness observed in Condition 2, in which the common factor covariance matrix was fixed rather than free to be estimated.

On the other hand, the practitioner certainly cannot blindly trust normal-theory test statistics to yield correct results with nonnormal data. Our results show that normal-theory methods performed extremely poorly when there was a dependency among latent variates. In fact, under these conditions, even the poorly performing ADF method always outperformed the normal-theory methods, which for all practical purposes were completely useless at evaluating model adequacy at all sample sizes, because they almost always rejected the true model. These results must give serious pause to the covariance structure analysis practitioner. It would be inappropriate for a practitioner to use normal-theory methods in their analysis with the justification that "asymptotic robustness theory verifies that nonnormal data can be appropriately analyzed using ML or GLS." Without some diagnostic about the relevance of this theory to the particular model and data analysis situation, it is entirely possible that the data being analyzed comes from a data generation mechanism that does not yield asymptotic robustness for ML or GLS test statistics. It might be countered that in this study, under these conditions, ML and GLS tended to reject true models far too often and that if in practice one has a model

Table 8
Summary of Simulation Results for Condition 6

	n					
Method	150	250	500	1,000	2,500	5,000
ML						
M	167.410	175.943	191.954	199.956	215.911	225.006
SD	46.089	48.044	70.363	66.343	71.016	80.043
Freq	195	194	200	200	200	200
GLS						
M	128.573	145.037	166.022	182.371	205.124	216.938
SD	20.853	25.114	32.925	40.883	53.613	59.441
Freq	167	186	200	199	200	200
ERLS						
M	107.704	101.600	96.589	89.919	85.503	82.540
SD	30.837	24.805	24.701	20.513	17.210	15.230
Freq	68	53	33	17	14	6
HK						
M	73.319	74.607	76.005	76.721	79.052	79.570
SD	13.933	13.985	13.498	13.841	14.057	13.782
Freq	1	2/199	2	3	7	5
ADF						
M	207.585	138.225	108.673	97.149	91.814	89.218
SD	39.100	20.203	15.546	12.422	13.439	13.213
Freq	179/180	184	97	26	11	10
SCALED						
M	92.355	89.277	87.759	86.430	86.128	85.531
SD	12.600	12.581	11.534	12.656	12.757	12.557
Freq	14	13	7	7	6	5

Note. ML = maximum likelihood, GLS = generalized least squares, ERLS = elliptical theory, HK = heterogenous kurtosis, ADF = asymptotic distribution free, SCALED = Satorra-Bentler scaling corrected, Freq = frequency of rejection of the null hypothesis [critical value from $\chi^2_{(p^*-q)}$ at $\alpha = .05$]. All numbers are based on the number of converged replications. Usually this is 200, but when two numbers are given in an entry in the freq row, the first gives the frequency of rejection of the null hypothesis, and the second gives the number of converged replications.

Table 9
Summary of Simulation Results for Condition 7

	n					
Method	150	250	500	1,000	2,500	5,000
ML						
M	166.485	175.366	191.309	199.481	215.325	225.643
SD	45.844	47.955	70.507	65.989	70.836	81.001
Freq	194	195	200	199	200	200
GLS						
M	128.266	144.718	165.662	181.992	204.578	217.556
SD	20.923	25.089	33.047	40.587	53.465	60.689
Freq	162	185	199	198	200	200
ERLS						
M	106.482	100.633	95.455	88.795	84.535	81.369
SD	30.482	25.031	25.156	20.558	17.377	15.355
Freq	64	50	31	14	11	7
HK						
M	67.711	70.218	70.166	70.130	71.966	70.487
SD	13.147	14.124	14.022	13.864	14.983	14.731
Freq	0	1	3	1	3	1
ADF						
M	206.609	137.762	108.634	96.721	91.343	89.150
SD	37.517	20.131	15.645	12.365	13.433	13.360
Freq	177/177	183/199	93	27	11	10
SCALED						
M	92.032	89.154	87.539	86.359	86.114	85.625
SD	12.474	12.653	11.628	12.681	12.859	12.805
Freq	14	11	4	7	6	5

Note. ML = maximum likelihood, GLS = generalized least squares, ERLS = elliptical theory, HK = heterogenous kurtosis, ADF = asymptotic distribution free, SCALED = Satorra-Bentler scaling corrected, Freq = frequency of rejection of the null hypothesis [critical value from $\chi^2_{(p^*-q)}$ at $\alpha = .05$]. All numbers are based on the number of converged replications. Usually this is 200, but when two numbers are given in an entry in the freq row, the first gives the frequency of rejection of the null hypothesis, and the second gives the number of converged replications.

that is statistically acceptable, then this worry should be irrelevant. This line of reasoning is not correct. For example, it would no doubt have been possible to empirically modify the models being studied in each sample of this simulation through the Lagrange Multiplier statistics (Lee & Bentler, 1980; Satorra, 1989) to find parameters such as correlated error terms that would reduce a test statistic T_{ML} (or T_{GLS}) in each sample, so that the associated simulation results would have yielded closer to nominal model rejection frequencies. However, it is clear that such modifications would produce models that are incorrect, at least in the sense that they would contain parameters that are in fact zero in the population.

Elliptical and HK theory promise a methodology to correct for normality that requires only trivially heavier computations than normal-theory methods. These methods performed variably. When the latent common and error variates were independently distributed, both methods tended to accept models more frequently than expected. When these variates were dependent, ERLS tended to reject models more frequently than expected, and the HK method accepted models too often, though both performances were certainly substantially better than those obtained by the normal-theory methods, which essentially always rejected true models. The performance of the new HK method was remarkably consistent across all conditions, yielding mean test statistics T_{HK} and associated standard deviations that were reasonably close to the theoretically expected values under a χ^2 distribution, though the number of rejections was consistently smaller than nominal levels. Except at the largest sample sizes, HK theory performed better than ADF theory, though improvements in HK theory are clearly still needed. It is possible that a new implementation of this theory (Bentler, Berkane, & Kano, 1991) would yield better performance. In any case, the HK method seems to be the most promising of all methods when considering extremely large models, in which normal-theory methods can be misleading (under dependence of factors and errors) and in which alternative methods like ADF and SCALED basically cannot be implemented at all because of the size of the matrices involved.

The best performance across all conditions was shown by Satorra and Bentler's (1988a, 1988b) SCALED test statistic, which has been documented and available in EQS since 1989. Although preliminary positive research on its performance has previously been obtained (Chou et al., 1991), this is the first comprehensive study of its behavior. It performed better than ADF at all but the largest sample sizes. When considering the closeness of the empirical mean of the statistic to the expected value (87, or 93 in Condition 2), the SCALED statistic was closer than the ADF statistic 42 times out of 42 comparisons. In terms of closeness to the expected 10 rejections at $\alpha = .05$, at sample sizes up to and including 1,000 the SCALED statistic performed better than the ADF statistic in 28 out of 28 comparisons. At $n = 2{,}500$ or $n = 5{,}000$, the SCALED and ADF statistics were closer to nominal levels an equivalent number of times.

Although the SCALED statistic performed extremely well overall, it had a tendency to overreject models at smaller sample sizes. The same tendency was observed for ML as compared with GLS when these statistics performed adequately (Conditions 1–4). Because the SCALED statistic, as implemented in this study, was based on the T_{ML} statistic (using METHOD = ML, ROBUST in EQS; see Appendix), it is likely that closer to nominal rejection rates would be observed if the GLS statistic had been scaled instead (using METHOD = GLS, ROBUST). Of course the scaling correction can work with many other estimators as well. It is interesting to speculate whether the tendency of the HK method to overaccept models would be eliminated if the scaling correction were used on the T_{HK} statistic. Clearly, more work is needed in this area.

The reason for the superior performance of the SCALED over the ADF statistic is not known. An obvious hypothesis is that although both statistics rely on the sample fourth-order moments of the variables as part of their estimation, the SCALED statistic uses a matrix computed from these moments directly, but ADF requires the relevant matrix to be inverted (see Appendix). This inverse may not even exist in sufficiently small samples, and there may be accuracy problems in intermediate-size samples. It is interesting to speculate whether Satorra and Bentler's (1988a, 1988b) adjusted test statistic, which requires still heavier computations than the SCALED statistic and attempts to adjust the variance of the test statistic as well as its mean, would perform better still. This statistic was not studied here because it is not yet routinely available in public computer programs. Clearly, there remains much to learn about

these statistics. Kano (1990) provided a further in-depth study of various forms of the SCALED statistic.

Although this study evaluated the empirical behavior of a larger variety of test statistics under a more varied set of conditions than has previously been attempted, clearly any simulation such as this has its limitations. In addition to being limited to a particular confirmatory factor-analytic design with a given set of parameter values, this study only evaluated the performance of six statistics under seven conditions. More work is clearly needed, both empirical and theoretical. As noted above, there are other statistics that need to be evaluated, and additional models need to be studied, especially models with larger number of variables (such as 30–50 variables), where ADF and SCALED statistics become difficult if not impossible to implement. At a more theoretical level, corrections for improved behavior of all statistics at small sample sizes need to be developed further (Tanaka, 1987). In the case of normal-theory statistics under multivariate normality, a Bartlett correction to the χ^2 statistics has been available for exploratory factor analysis for some time (Anderson, 1984; Lawley & Maxwell, 1971). Since this study was completed, a new theory for correcting more general covariance structure models came to our attention (Wakaki, Eguchi, & Fujikoshi, 1990). It seems likely that the Wakaki et al. correction could fix the overrejection problem found for ML at small sample sizes under the independence conditions of this study, but empirical evidence is needed to evaluate this suggestion. In fact, we would expect the Wakaki et al. corrected test statistic to outperform the GLS test statistic.

In any case, the current study was comprehensive enough to generate a clear warning about the current practice of covariance structure analysis, which relies heavily on normal theory and asymptotically distribution-free statistics. Unambiguous evidence was obtained on the inadequate behavior of normal-theory test statistics under some conditions of nonnormality and about the inadequate behavior of asymptotically distribution-free covariance structure analysis at all but the largest sample sizes. It also showed that a scaled test statistic exists that can outperform these better-known statistics under a variety of conditions.

References

Amemiya, Y. (1985). *On the goodness-of-fit tests for linear structural relationships* (Tech Rep No. 10). Stanford, CA: Stanford University, Econometric Workshop.

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

Anderson, T. W. (1984). Estimating linear statistical relationships. *Annals of Statistics, 12,* 1–45.

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

Bentler, P. M. (1983). Some contributions to efficient statistics for structural models: Specification and estimation of moment structures. *Psychometrika, 48,* 493–517.

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

Bentler, P. M., & Berkane, M. (1986). The greatest lower bound to the elliptical theory kurtosis parameter. *Biometrika, 73,* 240–241.

Bentler, P. M., Berkane, M., & Kano, Y. (1991). Covariance structure analysis under a simple kurtosis model. In E. M. Keramidas (Ed.), *Computing science and statistics* (pp. 463–465). Fairfax Station, VA: Interface Foundation of North America.

Bentler, P. M., & Dijkstra, T. (1985). Efficient estimation via linearization in structural models. In P. R. Krishnaiah (Ed.), *Multivariate analysis VI* (pp. 9–42). Amsterdam: North-Holland.

Bentler, P. M., Lee, S.-Y., & Weng, J. (1987). Multiple population covariance structure analysis under arbitrary distribution theory. *Communications in Statistics-Theory, 16,* 1951–1964.

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

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

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

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

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62–83.

Browne, M. W. (1985, July). *Robustness of normal theory tests of fit of factor analysis and related models against nonnormally distributed common factors.* Paper presented at the Fourth European Meeting of the Psychometric Society and the Classification Societies, Cambridge, England.

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

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

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics, 18,* 5–46.

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

Harlow, L. L. (1985). *Behavior of some elliptical theory estimators with non-normal data in a covariance structure framework: A Monte Carlo study.* Unpublished doctoral dissertation, University of California, Los Angeles.

Harlow, L. L., Chou, C. P., & Bentler, P. M. (1986, June). *Performance of chi-square statistic with ML, ADF, and elliptical estimators for covariance structures.* Paper presented at the annual meeting of the Psychometric Society, Toronto, Ontario, Canada.

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

Jöreskog, K. G., & Sörbom, D. (1988). *LISREL 7: A guide to the program and applications.* Chicago: SPSS.

Kano, Y. (1990). *A simple adjustment of the normal theory inference for a wide class of distribution in linear latent variate models.* Technical Report, University of Osaka Prefecture, Osaka, Japan, Department of Mathematical Sciences, College of Engineering.

Kano, Y., Berkane, M., & Bentler, P. M. (1990). Covariance structure analysis with heterogenous kurtosis parameters. *Biometrika, 77,* 575–585.

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

Lee, S.-Y., & Bentler, P. M. (1980). Some asymptotic properties of constrained generalized least squares estimation in covariance structure models. *South African Statistical Journal, 14,* 121–136.

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

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

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. (1990). *A comparison of some methodologies for the factor analysis of nonnormal Likert variables: A note on the size of the model.* Technical Report, Los Angeles: University of California, Graduate School of Education.

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika, 54,* 131–151.

Satorra, A., & Bentler, P. M. (1986). Some robustness properties of goodness of fit statistics in covariance structure analysis. *American Statistical Association 1986 proceedings of the Business and Economics Sections,* (pp. 549–554). Alexandria, VA: American Statistical Association.

Satorra, A., & Bentler, P. M. (1988a). Scaling corrections for chi-square statistics in covariance structure analysis. *American Statistical Association 1988 proceedings of the Business and Economics Sections* (pp. 308–313). Alexandria, VA: American Statistical Association.

Satorra, A., & Bentler, P. M. (1988b). *Scaling corrections for statistics in covariance structure analysis.* (UCLA Statistics Series 2). Los Angeles: University of California, Department of Psychology.

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

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

Shapiro, A. (1987). Robustness properties of the MDF analysis of moment structures. *South African Statistical Journal, 21,* 39–62.

Shapiro, A., & Browne, M. (1987). Analysis of covariance structures under elliptical distributions. *Journal of the American Statistical Association, 82,* 1092–1097.

Tanaka, J. S. (1984). *Some results on the estimation of covariance structure models.* Unpublished doctoral dissertation, University of California, Los Angeles.

Tanaka, J. S. (1987). "How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development, 58,* 134–146.

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

Appendix

Test Statistics

This Appendix describes the technical definitions of the six goodness-of-fit test statistics studied in this report. Let S represent the usual unbiased estimator that is based on a sample of size n of a $p \times p$ population covariance matrix $\boldsymbol{\Sigma}$, whose elements are functions of a $q \times 1$ parameter vector $\boldsymbol{\theta}$: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. A discrepancy function $F = F[\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})]$ can be considered to be a measure of the discrepancy between S and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ evaluated at an estimator $\hat{\boldsymbol{\theta}}$. The normal-theory maximum-likelihood (ML) discrepancy function is

$$F_{\mathrm{ML}} = \log|\boldsymbol{\Sigma}| - \log|\mathbf{S}| + tr(\mathbf{S}\boldsymbol{\Sigma}^{-1}) - p.$$

At the minimum, $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ and F_{ML} takes on the value $\hat{F}_{\mathrm{ML}}$, where $T_{\mathrm{ML}} = (n-1)\hat{F}_{\mathrm{ML}}$ is distributed, under the null hypothesis, as an asymptotic goodness-of-fit χ^2 variate with $(p^* - q)$ degrees of freedom, where $p^* = p(p+1)/2$. T_{ML} can be used as a test statistic to evaluate the null hypothesis $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. The null hypothesis is rejected if T_{ML} exceeds a critical value in the χ^2 distribution at an α level of significance. T_{ML} is the ML statistic whose behavior is studied in this article, with $\alpha = .05$.

A quadratic form discrepancy function is

$$F_{QD} = [\mathbf{s} - \sigma(\boldsymbol{\theta})]'\mathbf{W}^{-1}[\mathbf{s} - \sigma(\boldsymbol{\theta})],$$

where s and $\sigma(\boldsymbol{\theta})$ are $p^* \times 1$ column vectors formed from the nonduplicated elements of S and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, respectively and W is a $p^* \times p^*$ positive-definite weight matrix. The asymptotically distribution-free (ADF) covariance structure method used in this study minimizes F_{QD} under the choice of optimal weight matrix W with typical elements

$$w_{ij,kl} = \sigma_{ijkl} - \sigma_{ij}\sigma_{kl},$$

where $\sigma_{ijkl} = \mathcal{E}(x_{ti} - \mu_i)(x_{tj} - \mu_j)(x_{tk} - \mu_k)(x_{tl} - \mu_l)$ is the fourth-order multivariate moment of variables x_i about their means μ_i and σ_{ij} is an element of $\boldsymbol{\Sigma}$. In practice, sample moment estimators $s_{ijkl} = \sum_1^n (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j)(x_{tk} - \bar{x}_k)(x_{tl} - \bar{x}_l)/n$ and $s_{ij} = \sum_1^n (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j)/(n-1)$ are used to consistently estimate σ_{ijkl} and σ_{ij}. The ADF estimator provides an asymptotically efficient estimator $\hat{\boldsymbol{\theta}}$ without the need for distributional assumptions on variables. Under the null hypothesis, the associated test statistic $T_{\mathrm{ADF}} = (n-1)\hat{F}_{QD}$ has an asymptotic χ^2 distribution that is based on $(p^* - q)$ degrees of freedom.

The fitting function F_{QD} for normal-theory generalized least squares (GLS) can be simplified to

$$F_{\mathrm{GLS}} = \tfrac{1}{2}tr\{[\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta})]\mathbf{V}^{-1}\}^2,$$

if $\mathbf{W} = 2\mathbf{K}'_p(\mathbf{V} \otimes \mathbf{V})\mathbf{K}_p$, where V is a positive definite matrix that converges to $\boldsymbol{\Sigma}$ in probability and $\mathbf{K}_p$ is a known transition matrix. At the minima of the respective functions, both T_{ML} and $T_{\mathrm{GLS}} = (n-1)\hat{F}_{\mathrm{GLS}}$ have asymptotic χ^2 distributions with $(p^* - q)$ degrees of freedom; they are asymptotically equivalent when the model is correct. Browne (1974) has shown that if V converges in probability to $\boldsymbol{\Sigma}$ (e.g., $\mathbf{V} = \mathbf{S}$ as used in this study), then GLS estimators are asymptotically equivalent to ML estimators.

Under the assumption that all marginal distributions of a multivariate distribution are symmetric and have the same relative kurtosis, elliptical theory parameter estimators and test statistics can be obtained by readjusting the statistics derived from normal-theory methods. Let $\kappa = \sigma_{iiii}/3\sigma_{ii}^2 - 1$ be the common kurtosis parameter of a distribution from the elliptical class. Multivariate normal distributions are members of this class with $\kappa = 0$. The fourth-order multivariate moments σ_{ijkl} are related to κ by

$$\sigma_{ijkl} = (\kappa + 1)(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}),$$

where σ_{ij} is an element of $\boldsymbol{\Sigma}$. As a result of this simplification, the discrepancy function for an elliptical distribution may be written as

$$F_E = \tfrac{1}{2}(\kappa + 1)^{-1}\mathrm{tr}\{[\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta})]\mathbf{V}^{-1}\}^2 - \delta\{tr[\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta})]\mathbf{V}^{-1}\}^2,$$

where as before V is any consistent estimator of $\boldsymbol{\Sigma}$ and $\delta = \kappa/[4(\kappa + 1)^2 + 2p\kappa(\kappa + 1)]$ (Bentler, 1983). The selection of V as a consistent estimator of $\boldsymbol{\Sigma}$ leads, under the model and assumptions, to an asymptotically efficient estimator of $\boldsymbol{\theta}$ with $T_E = (n-1)\hat{F}_E$ at $\hat{\boldsymbol{\theta}}$ asymptotically distrib-

Appendix continues on next page

uted as a $\chi^2_{(p^*-q)}$ variate. In this study, we chose $\mathbf{V} = \hat{\boldsymbol{\Sigma}}$ at the minimum and

$$(\hat{\kappa} + 1) = \sum_{1}^{n}[(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}})]^2/np(p + 2).$$

Because the models to be investigated are invariant with respect to a constant scaling factor, at the minimum of F_E the second term drops out (Browne, 1984), yielding $T_E = T_{\text{ERLS}}$ as used in this study.

Heterogeneous kurtosis theory (Kano, Berkane, & Bentler, 1990) defines a more general class of multivariate distributions that allows marginal distributions to have heterogeneous kurtosis parameters. The elliptical distribution is a special case of this class of distributions. Let $\kappa_i^2 = \sigma_{iiii}/3\sigma_{ii}^2$ represent a measure of excess kurtosis of the ith variable and the fourth-order moments have the structure

$$\sigma_{ijkl} = (a_{ij}\,a_{kl})\sigma_{ij}\,\sigma_{kl} + (a_{ik}a_{jl})\sigma_{ik}\sigma_{jl} + (a_{il}\,a_{jk})\sigma_{il}\sigma_{jk},$$

where $a_{ij} = (\kappa_i + \kappa_j)/2$. If the covariance structure $\boldsymbol{\Sigma}(\theta)$ is fully scale invariant and the modeling and distributional assumptions are met, the F_{QD} discrepancy function can be expressed as

$$F_{\text{HK}} = \tfrac{1}{2}\text{tr}\{[\mathbf{S} - \boldsymbol{\Sigma}(\theta)]\hat{\mathbf{C}}^{-1}\}^2,$$

where $\hat{\mathbf{C}} = \hat{\mathbf{A}}^*\hat{\boldsymbol{\Sigma}}$ and * denotes the elementwise (Hadamard) product of the two matrices of the same order. In this study, we take $\hat{\mathbf{A}} = (\hat{a}_{ij}) = (\hat{\kappa}_i + \hat{\kappa}_j)/2$ on the basis of the usual moment estimators $\hat{\kappa}_i^2 = s_{iiii}/3s_{ii}^2$, and we take $\hat{\mathbf{C}} = \hat{\mathbf{A}}^*\mathbf{S}$. Kano, Berkane, and Bentler (1990) demonstrated that the simple adjustment of the weight matrix $\hat{\mathbf{C}}$ of the normal-theory generalized-least-squares procedure (see F_{GLS} above) produces asymptotically efficient estimators. The associated test statistic $T_{\text{HK}} = (n-1)\hat{F}_{\text{HK}}$ at the minimum has an asymptotic $\chi^2_{(p^*-q)}$ distribution under the assumed model.

Satorra and Bentler (1988a, 1988b) developed two modifications of any standard goodness-of-fit statistic test T (= T_{ML}, T_{HK}, etc.), so that its distributional behavior should more closely approximate χ^2. One of these, the scaled test statistic, is available in EQS (Bentler, 1989, p. 218). See also Kano (1990). Satorra and Bentler (1986) noted that the general distribution of T is in fact not χ^2, but rather a mixture

$$T \xrightarrow{\mathcal{L}} \sum_{1}^{df} \alpha_i\tau_i,$$

where α_i is one of the df nonnull eigenvalues of the matrix $\mathbf{UV}_{ss}$, $\mathbf{V}_{ss}$ is the asymptotic covariance matrix of $\sqrt{n}[\mathbf{s} - \sigma(\boldsymbol{\theta})]$, τ_i is one of the df independent χ_1^2 variates, and, when there are no constraints on free parameters (as in this study),

$$\mathbf{U} = \mathbf{W}^{-1} - \mathbf{W}^{-1}\dot{\sigma}(\dot{\sigma}'\mathbf{W}^{-1}\dot{\sigma})^{-1}\dot{\sigma}'\mathbf{W}^{-1}$$

is the residual weight matrix under the model and the weight matrix $\mathbf{W}$ used in the estimation. The scaled statistic used in this article is based on T_{ML}, with $\mathbf{W} = 2\mathbf{K}_p'(\hat{\boldsymbol{\Sigma}} \otimes \hat{\boldsymbol{\Sigma}})\mathbf{K}_p$, the normal-theory ML weight matrix at the minimum of F_{ML} and $\dot{\sigma} = \partial F_{\text{ML}}/\partial\theta'$ evaluated at $\hat{\theta}$. The mean of the asymptotic distribution of T_{ML} is given by $tr(\mathbf{UV}_{ss})$. Then, defining the scaling estimate $k = tr(\hat{\mathbf{U}}\hat{\mathbf{V}}_{ss})/df$, where $\hat{\mathbf{U}}$ is a consistent estimator of $\mathbf{U}$ on the basis of $\hat{\theta}$ and $\hat{\mathbf{V}}_{ss}$ is the distribution-free estimator with elements $s_{ijkl} - s_{ij}s_{kl}$ (see above), the scaled ML statistic

$$\bar{T} = T_{\text{ML}}/k$$

defines Satorra and Bentler's scaling-corrected test statistic as applied in this article.

Received March 19, 1991
Revision received November 19, 1991
Accepted November 26, 1991 ■