PSYCHOMETRIKA—VOL. 49, NO. 2, 155–173
JUNE 1984

THE EFFECT OF SAMPLING ERROR ON CONVERGENCE, IMPROPER SOLUTIONS, AND GOODNESS-OF-FIT INDICES FOR MAXIMUM LIKELIHOOD CONFIRMATORY FACTOR ANALYSIS

JAMES C. ANDERSON

DEPARTMENT OF MARKETING ADMINISTRATION
THE UNIVERSITY OF TEXAS AT AUSTIN

DAVID W. GERBING

BAYLOR UNIVERSITY

A Monte Carlo study assessed the effect of sampling error and model characteristics on the occurrence of nonconvergent solutions, improper solutions and the distribution of goodness-of-fit indices in maximum likelihood confirmatory factor analysis. Nonconvergent and improper solutions occurred more frequently for smaller sample sizes and for models with fewer indicators of each factor. Effects of practical significance due to sample size, the number of indicators per factor and the number of factors were found for GFI, AGFI, and RMR, whereas no practical effects were found for the probability values associated with the chi-square likelihood ratio test.

Key words: Confirmatory factory analysis, LISREL, Monte Carlo, Maximum likelihood.

The development of confirmatory analyses for covariance structures (Bentler, 1983; Joreskog, 1969, 1970, 1971, 1978) has provided considerable means to test and modify theories. Given a converged and proper solution (Joreskog, 1966, 1967), a researcher would assess how well a proposed model accounted for the data with one or more goodness-of-fit indices. However, little is known about the effects of sampling error (and model characteristics) upon the occurrence of nonconvergent solutions and improper solutions and the various goodness-of-fit indices. In this paper we address these two issues with a Monte Carlo study.

The confirmatory factor analysis model (Joreskog and Sorbom, 1981) can be given as:

$$\mathbf{x} = \Lambda\xi + \delta \tag{1}$$

where $\mathbf{x}$ is a vector of k observed measures, ξ is a vector of m underlying factors such that $m < k$, Λ is a $k \times m$ matrix of weights or factor loadings relating the observed measures to the underlying factors, and δ is a vector of k variables which represent random measurement error and indicator specificity. For convenience, let $E(\mathbf{x}) = E(\xi) = E(\delta) = \mathbf{0}$. It is assumed by this model that $E(\xi\delta') = E(\delta\delta') = \mathbf{0}$. The variance-covariance matrix for $\mathbf{x}$, defined as Σ is:

$$\Sigma = \Lambda\Phi\Lambda' + \Theta_\delta \tag{2}$$

James Anderson is now at the J. L. Kellogg Graduate School of Management, Northwestern University.
The authors gratefully acknowledge the comments and suggestions of Kenneth Land and the reviewers, and the assistance of A. Narayanan with the analysis. Support for this research was provided by the Graduate School of Business and the University Research Institute of the University of Texas at Austin.
Request for reprints should be sent to James C. Anderson, J. L. Kellogg Graduate School of Management, Northwestern University, 2001 Sheridan Rd, Evanston, IL, 60201, (312) 492-3522.

0033-3123/84/0600-5098$00.75/0

where Φ is the $m \times m$ covariance matrix of ξ and Θ_δ is the diagonal $k \times k$ covariance matrix of δ. If the observations of $\mathbf{x}$ are independent and have a multivariate normal distribution, full information maximum likelihood (FIML) estimates of the model parameters are provided by the LISREL program (Joreskog and Sorbom, 1978, 1981).

Nonconvergence and Improper Solutions

Solutions are *nonconvergent* when an estimation algorithm, within a set number of iterations, is unable to arrive at values which meet prescribed criteria (see Joreskog, 1967). Although the computational method of Joreskog and Sorbom (1978, 1981) has better convergence properties than alternative methods (Joreskog, 1966, 1967, 1969), nonconvergent solutions can still occur. A primary cause of nonconvergence is a pattern of observed correlations that is fundamentally incongruent with the specified model. For example, because of sampling error, a negative correlation may be obtained between two indicators in a properly specified model even though the corresponding population value is positive.

Solutions are *improper* when one or more of the unique variances of Θ_δ is less than a positive, arbitrarily small, prescribed number such as .005 (Joreskog, 1967). Variances can even become negative since the only constraint imposed by the LISREL program is that $\hat{\Sigma}$ is positive definite (Joreskog and Sorbom, 1981). van Driel (1978) has discussed three causes of improper solutions: Sampling fluctuations combined with true values of Θ_δ close to zero, lack of a factor analytic model which would fit the data, and indefiniteness of the model. Jackson and Chan (1980) have stated that because of problems presented by improper solutions (e.g., nonunique parameter estimates), "extreme caution" should be exercised when interpreting such results.

Goodness-of-Fit Indices

The first goodness-of-fit index to be widely used was the probability associated with the chi-square likelihood ratio test (Lawley and Maxwell, 1971), which is $N - 1$ times the minimum value of the fitting function obtained for the specified model (Joreskog and Sorbom, 1978, 1981). This fitting function, F, is:

$$F = \log|\Sigma| + \text{tr}(\mathbf{S}\Sigma^{-1}) - \log|\mathbf{S}| - k \tag{3}$$

where $\mathbf{S}$ is the sample variance-covariance matrix, and Σ and k are defined as in (1) and (2). Under the assumption of multivariate normal $\mathbf{x}$, $(N-1)F$ is asymptotically distributed as chi-square. "If the model is correct and the sample size sufficiently large, the χ^2-measure is the likelihood ratio test statistic for testing the model against the alternative that Σ is unconstrained" (Joreskog and Sorbom, 1981, p. I.38).

The associated degrees of freedom (df) for χ^2 are:

$$\text{df} = [(\tfrac{1}{2})k(k+1)] - t \tag{4}$$

where t is the number of independent parameters that are estimated. The probability level for the obtained chi-square value is the probability of obtaining a larger chi-square given that the model is correct; so small chi-square values with corresponding large probability levels indicate good fit. Thus the probability value associated with the chi-square likelihood ratio (PRCHI) is used as the goodness-of-fit index, rather than the chi-square value per se, as judgments of fit are based upon this associated probability value.

It has been noted (Bentler and Bonett, 1980; Joreskog, 1978) that this chi-square statistic, being partly a function of N, is sensitive to sample size. In very large samples, residuals of no practical significance can lead to statistical rejection of a model, whereas in very small samples less appropriate models can be judged as providing "adequate fit."

Joreskog and Sorbom (1981, pp. I.38–39) have pointed out that the chi-square measure is better thought of as a goodness-of-fit measure rather than a test statistic because assumptions underlying the test (e.g., multivariate normality, large sample size) are seldom fully met in practice.

Joreskog and Sorbom (1981, pp. I.40–41) recently proposed three additional overall goodness-of-fit indices: the goodness-of-fit index (GFI), GFI adjusted for degrees of freedom (AGFI), and the root mean square residual (RMR). The goodness-of-fit index for full information maximum likelihood (FIML) is defined as:

$$\text{GFI} = 1 - \frac{\text{tr}\,(\hat{\Sigma}^{-1}\mathbf{S} - I)^2}{\text{tr}\,(\hat{\Sigma}^{-1}\mathbf{S})^2} \tag{5}$$

where $\hat{\Sigma}$ is the fitted variance-covariance matrix. The goodness-of-fit index adjusted for degrees of freedom is:

$$\text{AGFI} = 1 - [k(k+1)/2d](1 - \text{GFI}) \tag{6}$$

where d is the degrees of freedom for the model. Joreskog and Sorbom have given the following as properties of GFI (and AGFI): a) it should be between zero and one (although theoretically it can become negative), b) it indicates the relative amount of variance and covariance explained by a model, c) it is independent of sample size (unlike chi-square) and d) it is relatively robust against violations of normality. Unfortunately, the sampling distributions of GFI and AGFI are not known, limiting their usefulness in practice.

The root mean square residual (RMR) is defined by Joreskog and Sorbom (1981) as:

$$\text{RMR} = \left[2 \sum_{i=1}^{k} \sum_{j=1}^{i} (s_{ij} - \hat{\sigma}_{ij})^2 / k(k+1) \right]^{1/2} \tag{7}$$

where $\hat{\sigma}_{ij}$ are the elements of the fitted variance-covariance matrix, $\hat{\Sigma}$. Note that in this definition, and in the implementation of LISREL V, the diagonal residuals are counted twice. This goodness-of-fit index provides a measure of the average magnitude of the residuals. Joreskog and Sorbom have noted that RMR should be judged relative to the sizes of the elements in S.

The final goodness-of-fit index included in this study was a reliability coefficient for maximum likelihood factor analysis proposed by Tucker and Lewis (1973). This reliability coefficient (RHO) can be defined as:

$$\text{RHO} = \frac{M_0 - M_k}{M_0 - E(M_k)}. \tag{8}$$

Computational equations for the components of RHO are provided by Burt (1973, pp. 148–149):

$$M_0 = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} s_{ij}^2}{k(k+1)/2} \tag{9a}$$

$$M_k = \frac{F}{\text{df}} \tag{9b}$$

$$E(M_k) = 1/N. \tag{9c}$$

This reliability coefficient represents the proportion of covariation to be explained that is accounted for by a specified model. Burt has suggested that RHO is less sensitive to large

sample sizes than PRCHI because only covariation rather than total variation is considered in the computation of RHO. While the values of RHO will typically be between 0 and 1, it is possible with confirmatory models to obtain values of RHO outside these bounds (Burt 1973).

Bentler and Bonett (1980) have provided an alternate expression for RHO:

$$\text{RHO} = \frac{Q_0 - Q_k}{Q_0 - 1} \tag{10}$$

where, letting v represent the chi-square statistic,

$$Q_0 = v_0/\text{df}_0 \tag{11a}$$

The term Q_0 represents the ratio of the chi-square statistic to its degrees of freedom for Bentler and Bonett's (1980) null model, which here corresponds to a model with no common factors: "the severely restricted model specifying that the variables are mutually independent" (p. 596). A similar term can be defined for the model with k common factors:

$$Q_k = v_k/\text{df}_k \tag{11b}$$

According to this interpretation, "The index ρ represents an index of increment in fit obtained by using k common factors rather than none" (p. 599).

The equivalence of the formulas given by Burt (1973) and Bentler and Bonett (1980) follows from an asymptotic identity presented by Tucker and Lewis (1973, p. 4).

$$F = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} g_{ij}^2 \tag{12}$$

where g_{ij} is a partial correlation between variables i and j with the common factors partialed out. So, when there are no common factors, as in the null model, (10) becomes the same as (8) when both the numerator and denominator are multiplied by $1/N$.

Past Research

Three studies are particularly relevant to the present study. The first is a Monte Carlo experiment on unrestricted or exploratory maximum likelihood factor analysis by Gweke and Singleton (1980). The distribution (based upon 100 replications) of the chi-square likelihood ratio statistic was studied for zero, one and two factor models employing five indicators and 10, 30, 150, or 300 observations. When the assumptions underlying the model were met, the likelihood ratios for a two factor model were chi-square distributed for sample sizes of thirty or more. In contrast, when one or two unique variances for the two factor model were zero, the distributions of the likelihood ratios for all sample sizes were significantly different from the theoretical chi-square distribution, with the null hypothesis being rejected more frequently than expected.

Boomsma (1982) has conducted a Monte Carlo study on confirmatory maximum likelihood factor analysis. Two factor models were studied where factor correlation (.0 or .3), factor loading pattern (.4 and .6; .6 and .8; and .9), and number of indicators per factor (3 or 4) were varied for sample sizes of 25, 50, 100, 200, or 400 (with 300 replications of each). Nonconvergence occurred most frequently for sample sizes of 25 and 50 (ranging from 0 to 55% of solutions), particularly for models with the .4 and .6 loading pattern. The presence of improper solutions decreased as 1) sample size increased, 2) the number of indicators in the model was eight rather than six, and 3) the population values of the unique variances were further from zero. Unfortunately, improper solutions were included with the proper solutions in the analysis, and as Boomsma (1982, p. 158) stated,

"Throughout this study the reader should realize, especially with small samples, that the results are more or less 'biased' by these negative estimated uniquenesses." The chi-square goodness-of-fit statistic deviated considerably from the theoretical chi-square distribution for sample sizes of 25 and 50. Related to this, there was an overall tendency for the test statistic to be too large (particularly for sample sizes of 25 and 50), thereby inflating the Type I error rate. In discussing the results, Boomsma concluded that it is "dangerous" to use confirmatory maximum likelihood factor analysis with sample sizes less than 100 (particularly for models with relatively low factor loadings) and for a "low risk of drawing wrong conclusions" a sample size of at least 200 should be employed.

A small Monte Carlo study on confirmatory maximum likelihood factor analysis has been done by Bearden, Sharma, and Teel (1982). A two-factor model and a four-factor model with three indicators per factor were studied for sample sizes of 25, 50, 100, 500, 1000, 2500, 5000, and 10000 (with 100 replications of each). Unfortunately, the usefulness of this study's findings are limited in that the characteristics of the two population covariance/correlation matrices employed are seldom found in practice. The within-factor indicator correlations range from .72 to .86 with 8 of the 18 correlations .8 or greater. As indicative of the atypical nature of these matrices, neither nonconvergent *nor* improper solutions were obtained, despite using a sample size of 25. For the two-factor model, the sample distribution of the chi-square likelihood ratio was found to not be significantly different from the theoretical chi-square distribution for any of the sample sizes, whereas for the four-factor model, the distribution of this statistic was significantly different from chi-square for sample sizes smaller than 500.

Although this past research has provided some information, clearly more research is needed on the effects of sampling error upon the occurrence of nonconvergent solutions, improper solutions, and the different goodness-of-fit indices. While PRCHI has received some study, to our knowledge no work has addressed the other goodness-of-fit indices—GFI, AGFI, RMR and RHO. Furthermore, a thorough study would require the model characteristics of number of indicators per factor, indicator loadings, number of factors, and factor correlations to be systematically varied along with sample size within a single design. Again, to our knowledge this has not been done. For these reasons, a large-scale Monte Carlo study was undertaken.

Method

Design

A primary challenge of the study was to select a manageable design that addressed a reasonably diverse set of factor models and sample sizes. The levels of each factor were chosen so as to represent the range of values typically encountered in practice.

Fifty was chosen as the smallest sample size in that it approximates the lower bound for confirmatory factor analysis. Small increments to levels of 75, 100, and 150 were used to obtain more detailed information on the performance of the goodness-of-fit indices at the lower range of sample sizes. To approximate a relatively large sample size, a level of 300 observations was included.

The minimum number of indicators per factor for multiple indicator measurement models is two, which was the lower bound in this study. Because three and four indicators per factor are commonly found in these models, they were the two other levels used. Five or more indicators per factor have seldom appeared in the literature.

One level of loadings was .9 for all indicators, which approximates the upper bound of indicator reliability obtained in practice. A loading of .6 for all indicators was selected to represent a moderate loading. For substantive interest, a level of mixed loadings was

TABLE 1

Independent Variables and Goodness-of-Fit Indices

Independent Variables		
Code	Variable	Levels
S	Sample size	50, 75, 100, 150, or 300
I	Indicators/factors	2, 3, or 4
L	Loadings	all .6, all .9, or mixed: .4, .8; .4, .6, .8; or .4, .6, .6, .8
F	Factors	2, 3, or 4
C	Correlations of Factors	.3 or .5

Indices	
Code	Index
GFI	Goodness-of-fit
AGFI	Adjusted goodness-of-fit
RMR	Root mean square residual
PRCHI	Probability of chi-square given null hypothesis
RHO	Reliability coefficient

also included where the loadings averaged .6. The values of this level for two indicators were .4 and .8; for three indicators .4, .6, .8; and for four indicators .4, .6, .6, .8.[1]

The number of factors in the model was either two, three or four. Two was included as a lower bound, whereas three and four factors per model would be encountered more often. The levels of factor correlation chosen for study were low and moderate, .3 and .5 respectively. As some correlation between factors is usually posited in a confirmatory factor analysis, a level of zero correlation between factors was not included.

This 5*3*3*3*2 design is summarized in the upper half of Table 1. The five goodness-of-fit indices that were studied are given in the lower half of Table 1.

Procedure

Fifty-four population correlation matrices corresponding to the 54 factor models (3*3*3*2) were constructed using the rules of path analysis (Heise, 1975). The IMSL subroutine GGNSM in conjunction with the IMSL subroutine GGUBS (IMSL, 1980) was used to construct a multivariate normal population based upon the corresponding population correlation matrix and to generate samples of size S_i from which sample covariance matrices were computed. The GGUBS subroutine is a uniform random number generator whose accuracy has been supported (Learmouth and Lewis, 1973). Multivariate normal populations are constructed from the output of GGUBS using a set of exact transforms

[1] A necessary confound was introduced between the mixed level of loadings and the number of indicators. This level was included because of substantive interest in the effects of mixed loadings. Analysis of the mean and variance effects for the goodness-of-fit indices without the level of mixed loadings revealed that this confound was of no consequence.

embodied in the subroutine GGNSM. An initial random number is supplied as a seed which is then replaced by an internally generated random number on each of the subsequent calls to GGNSM. A different random number was used as the seed for each sample generation.

The maximum likelihood confirmatory factor analyses were performed on a CDC Cyber 170/750 computer using the LISREL IV program (Joreskog and Sorbom, 1978), adapted to run n times successively. Test runs were made using each population covariance matrix and the corresponding LISREL control statements. The respective population parameter values were specified as the start values for each analysis. A FORTRAN program was written to extract the information from the input covariance matrix and the LISREL output needed for the goodness-of-fit indices. The PRCHI index was taken directly from the output, whereas GFI, AGFI, and RMR were calculated using the formulas given by Joreskog and Sorbom (1981, pp. I40–I42). RHO was calculated using the formulas given by Burt (1973, pp. 148–149). Computational results from test data for GFI, AGFI, and RMR were checked against the corresponding results obtained from a separate matrix computational program, whereas the computational results for RHO were checked by hand.

One hundred proper and converged solutions were obtained (as replications) for each of the 270 cells in the design. The operational definition of nonconvergence employed was that of the LISREL program—failure to reach the convergence criteria (see Joreskog, 1967, p. 460) after 250 iterations. Improper solutions were not included for two reasons: The considerable problems of interpretability (and possible bias) discussed earlier, and the desire to generalize the study's findings to solutions most often obtained in practice.

Analyses

To assess the effects of sampling error and model characteristics upon the occurrence of nonconvergent and improper solutions, two qualitative dependent variables were formed. For convergence (CONVER), improper and "good" (converged and proper) solutions were coded 1 whereas nonconvergent solutions were coded 0. For proper versus improper solutions (PROPER), good solutions were coded 1, whereas improper solutions were coded 0. Log-linear/logit models were fit to the data using maximum likelihood estimation. The proportion of weighted variation explained by each model (Goodman, 1971), analogous to the coefficient of multiple determination (R^2), was calculated in addition to the usual likelihood ratio chi-square (χ^2_L) statistic.

To assess the effects of sampling error and model characteristics upon the goodness-of-fit indices, separate analyses of variance were run for each index. Given the number of observations ($N = 27000$), a criterion of practical significance of effects accounting for three or more percent of the variance was adopted in place of traditional statistical significance testing. The ω^2 statistic (Hays, 1973) was computed for this purpose. To test for differences in the variance of each index across the design, an analysis of variance (fitting main effects and two-way interactions) was performed using the $\ln s^2$ of each cell (Games, Keselman, and Clinch, 1979). Again, the practical significance criterion of three or more percent of the variance explained was adopted. Summary statistics consisting of means, standard deviations, and 5th or 95th percentile values were computed, and are presented for each mean and variance effect of practical significance.

Finally, to determine whether or not the sample distributions of the chi-square likelihood ratios were significantly from the theoretical chi-square distribution, Kolmogorov-Smirnov tests (Siegel, 1956) were computed for each of the 270 cells using the associated probability values (PRCHI) (see Geweke and Singleton, 1980). The number of solutions in each cell with probability values less than 0.05 and 0.10 was also calculated.

Results

Nonconvergence and Improper Solutions

To obtain 27000 good solutions, it was necessary to generate 43410 solutions, of which 5609 (12.9%) were nonconvergent and 10801 (24.9%) were improper. Although the presence of a relatively large number of zero cells (for both CONVER and PROPER) is desirable from a consideration of the overall research purpose, their presence made it difficult to estimate the higher-order log-linear/logit models. As a result, only main effects models could be stably estimated. However, the main effects models provided good explanation of the data, accounting for 96.3% of the weighted variation to be explained in CONVER and 86.1% of the weighted variation to be explained in PROPER. The source tables for CONVER and PROPER are presented in Table 2. The proportion of nonconvergent solutions that occurred in obtaining 100 good solutions per cell are summarized in Table 3 whereas the proportion of improper solutions are summarized in Table 4.

Examination of the parameter estimates for CONVER indicated that the proportion of convergent solutions increased as: Sample size increased, the number of indicators per factor increased, loadings varied from mixed to all .6 to all .9, and factor correlations increased. The proportion of convergent solutions decreased as the number of factors increased, although this finding is of little practical significance as indicated by its relatively small chi-square value. Examination of the parameter estimates for PROPER revealed a similar pattern of findings as obtained with CONVER, with the important differences between the two discussed below.

While it was not possible to fit the two-way or higher-order interaction log-linear models for the six-way contingency tables, a separate (though non-independent) assessment of each two-way interaction effect was possible. Three-way contingency tables were formed (e.g., CONVER**S***I*), and the chi-square difference values between the main effects model (e.g., CONVER**S* and CONVER**I*) and the two-way interaction (saturated) model (e.g., CONVER**S***I*) were examined. For CONVER, a two-way interaction effect of considerable size was *S***I* ($\chi^2_{L(8)} = 175.0$, $p < 0.0001$). It was found that as sample size increases, fewer indicators per factor are needed for convergence.

For PROPER, three two-way interaction effects were of considerable size: *I***L* ($\chi^2_{L(4)} = 1588.5$, $p < 0.0001$), *L***C* ($\chi^2_{L(4)} = 487.7$, $p < 0.0001$), and *S***I* ($\chi^2_{L(8)} = 238.7$, $p < 0.0001$). It was found from these analyses that with two indicators per factor, loadings

TABLE 2

The Effects Of Sampling Error and Model Characteristics

Upon Nonconvergent and Improper Solutions

		CONVER	PROPER
Source	df	χ^2	χ^2
Constant	1	3384.4	7197.8
S	4	1527.0	2221.5
I	2	1915.3	5610.2
L	2	2203.1	1075.6
F	2	38.9	146.1
C	1	1198.2	1288.4
Main effects model	258	461.6	2482.4

Note. All effects significant beyond the .0001 level.

TABLE 3

Proportion of Nonconvergent Solutions Across the Design

I=2						
	L=.4,.8		L=.6,.6		L=.9,.9	
S	C=.3	C=.5	C=.3	C=.5	C=.3	C=.5
50	.58	.31	.53	.23	.13	.01
75	.48	.22	.44	.10	.05	.00
100	.40	.11	.31	.06	.02	.00
150	.25	.04	.20	.01	.01	.00
300	.01	.00	.04	.00	.00	.00
I=3						
	L=.4,.6,.8		L=.6,.6,.6		L=.9,.9,.9	
S	C=.3	C=.5	C=.3	C=.5	C=.3	C=.5
50	.18	.03	.12	.02	.00	.00
75	.05	.01	.02	.00	.00	.00
100	.02	.00	.01	.00	.00	.00
150	.00	.00	.00	.00	.00	.00
300	.00	.00	.00	.00	.00	.00
I=4						
	L=.4,.6,.6,.8		L=.6,.6,.6,.6		L=.9,.9,.9,.9	
S	C=.3	C=.5	C=.3	C=.5	C=.3	C=.5
50	.01	.01	.02	.02	.00	.00
75	.00	.00	.00	.00	.00	.00
100	.00	.01	.00	.00	.00	.00
150	.00	.00	.00	.00	.00	.00
300	.00	.00	.00	.00	.00	.00

Note. Summing across the three levels of F.

of .9 give the largest proportion of improper solutions, whereas with three or four indicators per factor, .9 loadings give the smallest proportion (no improper solutions occurred in obtaining the 6000 solutions). In addition, when factor correlations are .3, loadings of .9 yield the largest proportion of improper solutions; whereas when factor correlations are .5, the proportion of improper solutions for loadings of .9 markedly decreases with the largest proportion occurring for mixed loadings. Lastly, as sample size increases, fewer indicators per factor are required to obtain proper solutions.

Goodness-of-Fit Indices

Mean and variance differences. The information on each of the indices across the 270 cells can be sumarized by examining, for each index, mean differences attributable to the design with $\omega^2 \geq 3.0\%$. Table 5a lists for each index all effects which account for 3.0% or more of the variance in its respective univariate analysis.

A general result is that all mean differences which met the criterion of practical significance are either main effects or two-way interactions. As can be seen from the Table, sample size, the number of indicators per factor, and the number of factors in the model strongly influence the mean values of GFI, AGFI, and RMR. In addition, factor

TABLE 4

Proportion of Improper Solutions Across the Design

			I=2			
	L=.4,.8		L=.6,.6		L=.9,.9	
S	C=.3	C=.5	C=.3	C=.5	C=.3	C=.5
50	.62	.58	.57	.38	.86	.53
75	.60	.50	.49	.29	.76	.36
100	.58	.54	.46	.19	.69	.22
150	.55	.42	.37	.10	.50	.11
300	.43	.25	.17	.00	.21	.00
			I=3			
	L=.4,.6,.8		L=.6,.6,.6		L=.9,.9,.9	
S	C=.3	C=.5	C=.3	C=.5	C=.3	C=.5
50	.53	.38	.25	.07	.00	.00
75	.40	.22	.08	.03	.00	.00
100	.30	.13	.03	.00	.00	.00
150	.16	.04	.00	.00	.00	.00
300	.03	.00	.00	.00	.00	.00
			I=4			
	L=.4,.6,.6,.8		L=.6,.6,.6,.6		L=.9,.9,.9,.9	
S	C=.3	C=.5	C=.3	C=.5	C=.3	C=.5
50	.25	.13	.06	.01	.00	.00
75	.10	.06	.00	.01	.00	.00
100	.08	.00	.00	.00	.00	.00
150	.01	.01	.00	.00	.00	.00
300	.00	.00	.00	.00	.00	.00

Note. Summing across the three levels of F.

loadings exert a considerable effect upon the values of RMR. Note that factor correlation (with levels of .3 and .5) was found not to have any effects of practical significance for any of the goodness-of-fit indices. Considering the indices, the design factors account for the most variance in GFI (86.8%) followed by RMR (74.9%) and AGFI (66.6%). The corresponding ANOVA tables for each univariate analysis where effects with $\omega^2 \geq 3.0\%$ were found are presented in Table 6.

Variance differences across the design with $\omega^2 \geq 3.0\%$, as indicated by analyses of variance of $\ln s^2$, are presented in Table 5b. Considering the indices, variance effects for GFI, AGFI, RMR and RHO are all well explained by the design factors with sample size having by far the strongest effects. Note that although no practically significant mean effects were found for RHO, strong differences in variances across the design occurred.

Useful descriptive statistics for each index are the mean, standard deviation and 5th or 95th percentile. The 5th percentile was used for indices in which large values indicate a good fit whereas the 95th percentile was used for indices in which 0.00 indicates a perfect fit. These reported percentile values provide the index value of interest for evaluating the fit of a model in practice, given that the null hypothesis is true that the proposed model fits the data, and assuming an underlying multivariate normal distribution. The calculated

TABLE 5

Effects with $\omega^2 \geq 3.0\%$

for Goodness-of-Fit Indices

Table 5a: Mean Effects

	Index				
Effect	GFI	AGFI	RMR	PRCHI	RHO
Sample size	28.3	41.7	24.7		
Indicators	28.7	10.8	21.9		
Factors	17.7	10.0	8.9		
Loadings			19.4		
S*I	8.3	4.1			
S*F	3.8				
Total	86.8	66.6	74.9		

Table 5b: Log of Variance Effects

	Index				
Effect	GFI	AGFI	RMR	PRCHI	RHO
Sample size	68.6	69.6	50.5	3.8	31.1
Indicators	11.2	13.1	3.1	3.6	19.7
Factors	4.8	6.7	9.7		4.4
Loadings			21.1		31.2
S*F				3.7	
I*F	5.4		5.3		
Total	90.0	89.4	89.7	11.1	86.4

TABLE 6

Analysis of Variance for Goodness-of-Fit Indices

		GFI		AGFI	
Effect	df	SS	F	SS	F
Between	269	55.42	1009.21	106.93	278.58
Sample size	4	17.24	21117.98	60.57	10611.65
Indicators	2	17.48	42813.45	15.67	5490.42
Factors	2	10.78	26400.44	14.58	5107.91
S*I	8	5.07	3102.60	5.89	516.27
S*F	8	2.33	1428.92		
Within	26730	5.46		38.14	
Total	26999	60.87		145.07	

	RMR	
Effect	df	SS	F
Between	269	12.96	514.97
Sample size	4	3.82	10211.76
Factors	2	1.38	7393.46
Loadings	2	2.99	15992.68
Within	26730	2.50	
Total	26999	15.46	

index percentile values are correspondent with a 95% confidence interval, the boundary values of which are typically constructed from a sample mean and standard deviation.

These descriptive statistics for each index across the 270 cells are summarized according to the pattern of mean and variance effects with $\omega^2 \geq 3.0\%$. At one extreme, if no design effects for an index are of practical significance, then the design can be summed across and only a single mean, standard deviation and percentile value need to be reported. At the other extreme, if a five-way interaction was of practical importance, then the descriptive statistics would need to be reported for all 270 cells. The descriptive statistics were computed by summing across as many design factors as possible given the $\omega^2 \geq 3.0\%$ criterion. These statistics for the main effects are listed in Table 7; the interaction effect statistics appear in Table 8.

A primary finding is that GFI and AGFI tend to increase as sample size increases, although the presence of *S*I* and *S*F* interaction effects moderate this pattern. As the number of indicators per factor or the number of factors increases, the values of GFI and AGFI decrease, particularly for smaller sample sizes. Thus the smallest values are found with four indicators per factor or four factors in the model, and a sample size of 50. As an example, the GFI grand mean is .95 but the GFI mean value for four indicators and a sample size of 50 is .85, as shown in Tables 7 and 8. Similarly, when there are four factors and a sample size of 50, the GFI mean value is .87.

The RMR index dramatically decreases as an inverse function of sample size, from .060 for a sample size of 50 to .025 for a sample size of 300. The RMR mean value for four

TABLE 7

Summary Statistics on Indices for Main Effects with $\omega^2 \geq 3.0\%$

	Index								
Effect	GFI			AGFI			RMR		
	Mean	SD	5pcnt	Mean	SD	5pcnt	Mean	SD	95pcnt
Overall	.949	.047	.850	.903	.073	.758	.043	.024	.087
Sample Size									
50	.909	.064	.791	.828	.084	.697	.060	.028	.105
75	.936	.045	.853	.880	.060	.785	.050	.023	.086
100	.952	.034	.887	.908	.045	.838	.043	.020	.074
150	.965	.026	.919	.932	.051	.881	.036	.016	.061
300	.983	.012	.960	.968	.016	.942	.025	.012	.043
Indicators									
2	.980	.022	.939	.933	.068	.812	.027	.019	.064
3	.948	.038	.869	.902	.063	.775	.048	.021	.087
4	.918	.054	.805	.874	.076	.723	.054	.023	.096
Factors									
2	.974	.027	.918	.933	.062	.817	.033	.024	.077
3	.947	.044	.860	.900	.072	.762	.045	.022	.087
4	.925	.054	.813	.876	.074	.729	.051	.023	.093
Loadings									
mixed							.051	.024	.094
.6							.050	.024	.091
.9							.028	.015	.055

indicators (.054) is twice the size of the mean value for two indicators (.027). High factor loadings (.9) also yield smaller mean values (.028) than moderate (.050) or mixed (.051) loadings, and the RMR mean values increase with the number of factors in the model.

Considering the log of variance (ln s^2) effects, sample size was found to have a predominant influence upon GFI, AGFI, RMR and RHO, with its effects in the expected direction. As an illustration of this, the standard deviation of the GFI values for a sample size of 50 is .064, whereas for a sample size of 300 the standard deviation is just .012. Corresponding to this marked decrease, the fifth percentile value for GFI increases from .791 to .960 (as sample size increases from 50 to 300). A consistent pattern of results can be observed in Table 7 for AGFI, RMR, and RHO.

As noted earlier, strong variance effects were found for RHO, although the mean values were practically constant, varying in Table 7 from .989 to 1.035. The effect of outliers on the distribution, and in particular, the variance of RHO, was considerable. To better understand this effect, further analyses of RHO were done: the outlying observations were identified whose absolute values were greater than or equal to 50, and the marginal means and standard deviations were recomputed for a related index, RHO*, where these outliers were not present. Eleven observations out of the 27000 were found to surpass this criterion value. The values of these observations for the five goodness-of-fit indices are presented in Table 9, and the comparison values for RHO* appear in Table 7.

TABLE 7 (continued)

Summary Statistics on Indices for Main Effects with $\omega^2 \geq 3.0\%$

	Index							
Effect	PRCHI			RHO			RHO*	
	Mean	SD	5pcnt	Mean	SD	5pcnt	Mean	SD
Overall	.471	.292	.035	1.014	2.836	.823	1.017	.809
Sample Size								
50	.442	.299	.020	.989	5.788	.615	1.029	1.690
75	.464	.292	.031	1.050	2.579	.794	1.026	.582
100	.475	.293	.038	1.019	.251	.858	1.019	.251
150	.485	.287	.044	1.009	.107	.902	1.009	.107
300	.488	.288	.049	1.001	.056	.949	1.001	.056
Indicators								
2	.536	.282	.075	1.030	4.254	.724	1.066	1.366
3	.461	.290	.034	1.000	.295	.841	1.000	.295
4	.416	.292	.020	1.010	2.440	.864	.984	.093
Factors								
2				1.028	2.131	.768	1.035	1.118
3				1.017	4.386	.830	1.020	.596
4				.996	.599	.849	.996	.599
Loadings								
mixed				1.035	3.079	.760	1.014	1.081
.6				1.008	3.828	.759	1.039	.890
.9				.998	.044	.946	.998	.044

Note. RHO* corresponds to RHO where the 11 outliners (11/27000 or .04%) with absolute values larger than or equal to 50 have been removed.

TABLE 8

Summary Statistics on Indices

for Interaction Effects with $\omega^2 \geq 3.0\%$

Effect	Index					
	GFI			AGFI		
	Mean	SD	5pcnt	Mean	SD	5pcnt
S*I						
50, 2	.967	.026	.921	.886	.083	.754
50, 3	.908	.042	.842	.826	.061	.733
50, 4	.851	.054	.771	.771	.065	.673
75, 2	.977	.019	.944	.922	.057	.835
75, 3	.932	.035	.872	.873	.051	.784
75, 4	.899	.038	.840	.844	.045	.776
100, 2	.982	.014	.956	.940	.042	.866
100, 3	.950	.024	.911	.906	.035	.850
100, 4	.922	.030	.875	.880	.036	.826
150, 2	.982	.026	.918	.940	.080	.773
150, 3	.966	.016	.939	.936	.024	.897
150, 4	.947	.021	.914	.918	.025	.881
300, 2	.994	.005	.984	.978	.016	.952
300, 3	.982	.008	.968	.967	.012	.948
300, 4	.972	.011	.955	.958	.013	.937
	GFI			PRCHI		
S*F	Mean	SD	5pcnt	Mean	SD	5pcnt
50, 2	.952	.038	.886	.496	.287	.049
50, 3	.903	.058	.803	.443	.297	.024
50, 4	.870	.063	.778	.387	.304	.009
75, 2	.967	.026	.922	.505	.287	.051
75, 3	.936	.036	.879	.474	.291	.037
75, 4	.904	.046	.837	.415	.292	.018
100, 2	.975	.020	.939	.500	.286	.058
100, 3	.951	.028	.908	.482	.295	.039
100, 4	.928	.035	.875	.445	.295	.026
150, 2	.983	.013	.960	.504	.283	.060
150, 3	.963	.024	.930	.483	.289	.041
150, 4	.948	.025	.911	.467	.287	.037
300, 2	.991	.007	.979	.501	.286	.050
300, 3	.983	.010	.966	.486	.287	.051
300, 4	.974	.012	.955	.478	.290	.040

It can be seen from Table 9 that the observed values for the other indices compare favorably to their respective mean values in Table 7. As an example, the observation with the largest absolute value of RHO (-295.67) has a GFI value of .990 and an RMR value of .028. As can be seen from a comparison of RHO and RHO* in Table 7, removing the 11 outliers from consideration results in a change in both the size and pattern of the standard deviations. For instance, the standard deviations of RHO values as a function of the number of indicators per factor are 4.254, 0.295 and 2.440 for two, three, and four indicators respectively. This peculiar pattern of standard deviations, however, is due to a single outlying observation with four indicators per factor which has a RHO value of 232.27 (as another comparison, the next largest value of RHO for the 99 other observations in that cell is 1.182). Removal from consideration of this one observation from the 9000 observations with four indicators per factor lowers the standard deviation to .093, a

TABLE 8 (continued)

Summary Statistics on Indices

for Interaction Effects with $\omega^2 \geq 3.0\%$

	Index					
	GFI			RMR		
I*F	Mean	SD	5pcnt	Mean	SD	5pcnt
2,2	.996	.007	.984	.012	.012	.036
2,3	.978	.020	.946	.031	.016	.062
2,4	.966	.022	.922	.039	.018	.073
3,2	.973	.019	.935	.040	.019	.076
3,3	.947	.031	.888	.050	.021	.089
3,4	.923	.042	.847	.054	.021	.092
4,2	.952	.029	.897	.048	.021	.088
4,3	.916	.049	.823	.054	.022	.095
4,4	.886	.057	.786	.059	.023	.101

decrease of 2.347 (96.2%). Similar changes can be observed for variance effects due to sample size and the number of factors in a model.

No mean effects of practical significance were found for PRCHI. Apart from this, only small variance effects due to sample size, indicators per factor, and sample size by number of factors were found. A finding of interest for PRCHI, however, is the convergence of the fifth percentile value upon the expected probability value (.05) as sample size increases to 300 (.049).

Kolmogorov-Smirnov values. For the purpose of parsimony, an analysis of variance was performed on the Kolmogorov-Smirnov (KS) values resulting from the 100 replications within each of the 270 cells. For the main effects and the two-way interactions which were estimated, effects of practical significance (as given by ω^2) were found for sample size (17.7%), factors (12.9%), indicators (11.5%), indicators by factors (8.9%), sample size by factors (8.2%), sample size by indicators (5.5%), and indicators by loadings (3.3%). As expected, the Kolomogrov-Smirnov mean values decrease as sample size increases, and the mean values increase with the total number of indicators in the model

TABLE 9

Values of Goodness-of-Fit Indices

for Observations with |RHO| $\geq$ 50

Design					Index				
S	I	L	F	C	RHO	GFI	AGFI	RMR	PRCHI
50	2	.6	3	.3	-295.67	.990	.965	.028	.961
50	4	m	3	.3	232.27	.879	.815	.080	.738
75	2	.6	3	.3	107.84	.986	.951	.040	.759
75	2	m	2	.3	103.75	.994	.941	.029	.348
50	2	m	3	.5	-84.85	.981	.935	.036	.827
50	2	.6	2	.3	-84.23	.998	.983	.014	.681
50	2	.6	3	.3	77.85	.975	.913	.053	.716
75	2	.6	2	.5	-72.97	.999	.987	.010	.663
75	2	.6	3	.3	-60.21	.998	.991	.014	.997
50	2	m	2	.3	-57.44	.993	.926	.030	.390
75	2	.6	2	.3	54.96	.997	.969	.017	.494

TABLE 10

Correlations Between Goodness-of-Fit Indices[1]

	Index				
Index	GFI	AGFI	RMR	PRCHI	RHO
GFI	1.00	.86	-.67	.78	.03
AGFI	.91	1.00	-.68	.72	.03
RMR	-.73	-.71	1.00	-.72	-.03
PRCHI	.43	.50	-.40	1.00	.03
RHO	.01	.02	-.01	.03	1.00

Note. N = 27000

[1] Grand mean deviated below the diagonal, Within-cell deviated above the diagonal.

(the pattern of the factor and indicator main effects, and their interaction). This increase in mean values as the number of indicators per factor and the number of factors increase is strongly attenuated as sample size increases. For example, the mean values for two, three and four factors are .119, .176, and .271 for a sample size of 50 compared to .103, .096, and .095 for a sample size of 300. The mean values for two three and four indicators are .131, .164 and .270 for a sample size of 50 compared to .096, .090 and .108 for a sample size of 300. There is an overall tendency for the likelihood ratio test to become conservative when the sample distribution significantly departs from the theoretical distribution, resulting in rejection of the null hypothesis (at both the .05 and .10 levels) more often than would be expected.

Index correlations. The correlations between the indices are presented in Table 10. The usual grand mean deviated correlations are given below the diagonal. Correlations were also computed by deviating each of the 27000 values of each index from its respective cell mean as suggested by Finn (1974).

A general result is that GFI, AGFI, and RMR are all highly correlated with one another using either correlational approach. GFI and AGFI evince the highest correlation, .91 (or .86). From this table it is clear that RHO hardly correlates, whereas PRCHI has a moderate correlation with GFI, AGFI, and RMR.

Discussion

Several contributions have been made by this study. First, detailed information on the effects of sampling error and model characteristics upon the occurrence of nonconvergent, and improper solutions has been provided. Second, differences of practical significance which resulted from the design factors have been determined for the various proposed goodness-of-fit indices. Related to these differences, the expected sample values of the goodness-of-fit indices for a wide variety of factor models have been tabled. These tabled values allow researchers to compare their results with values that would be obtained from correctly specified models with multivariate normally distributed data, given a particular sample size. Finally, the relationships between the goodness-of-fit indices have been provided, as well as the values for these correlations independent of the design.

Nonconvergence and Improper Solutions

The occurrence of nonconvergent solutions and improper solutions was found to depend primarily upon sample size and the number of indicators per factor. With three or more indicators per factor, a sample size of 100 will usually be sufficient for convergence,

whereas with only two indicators per factor, problems can arise with sample sizes of 150 or lower. A related general finding is that to obtain solutions that are proper as well as convergent requires a sample size that is somewhat larger, or more indicators per factor. A sample size of 150 for models with three or more indicators per factor will usually be sufficient for a convergent and proper solution.

The loadings of the indicators upon their factors can also be a consideration, particularly in conjunction with the number of indicators per factor. As an example, loadings of .9 provided the highest proportion of convergence for two indicators per factor models, although loadings of .9 also resulted in the highest proportion of improper solutions (for two indicators per factor models). Further, mixed loadings for the three indicators per factor models were the only three (or four) indicators per factor models which were found to give improper solutions when a sample size of 150 or 300 was employed.

The occurrence of improper solutions will additionally be lessened when the factor intercorrelations are moderate (.5), rather than small (.3). (It should be remembered that the atypical occurrence of zero correlation between factors in confirmatory models was not considered in this study.) Improper solutions which correspond to the close to zero case (van Driel, 1978) are much more frequent when there are only two indicators per factor. As a general finding, two indicator per factor models were problematic for obtaining a convergent and proper solution.

Goodness-of-Fit Indices

Given the range of sample sizes and the factor models studied, mean differences of practical significance due to sample size, number of indicators and number of factors were found for GFI, AGFI and RMR. As sample size increases, relatively larger values of GFI and AGFI are needed to indicate acceptable model fit, whereas relatively smaller values of RMR are needed. Increasing the number of indicators per factor or the number of factors in a model has obverse effects upon these indices. In addition, high factor loadings result in lower values of RMR than do either moderate or mixed loadings.

Sample size, by far, had the greatest effect upon the variances of the goodness-of-fit indices, having an expected inverse relationship with them. Variance differences of practical significance were also due to the number of indicators and number of factors. As noted earlier, factor correlation was found to not have any mean or variance effects of practical significance for any of the goodness-of-fit indices.

The lack of mean effects due to sample size for PRCHI can be explained by the assumptions under which the data were generated. Statements about a positive relationship between sample size and PRCHI have referred to (somewhat implicitly) a residual matrix containing *residuals of a fixed size*, where these residuals may in part be due to trivial discrepancies between the specified model and the data. Given these residuals of a fixed size, it can readily be seen that there will be a positive relationship between sample size and PRCHI. By contrast, the data in this study were generated from a known multivariate normal population, and perfectly specified models were fit to them. Because of this, the generated sample covariances became more accurate population estimates as sample size increased, thereby enabling the LISREL program to more accurately estimate the underlying covariance structure. The magnitude of the residuals was thus inversely related to sample size, as evinced by the findings for RMR, and the related values of the fitting function (3) were smaller as sample size increased. The net result of this is that, when the assumptions underlying the likelihood ratio test were *fully met*, there were no effects of practical significance for PRCHI due to sample size.

Some clarification also is needed with respect to the findings for GFI (and AGFI), and Joreskog and Sorbom's (1981) statement that these goodness-of-fit indices have the advantage of being independent of sample size. There is some ambiguity perhaps in their

use of "independent of sample size." Whereas GFI (and hence AGFI) is independent of sample size in that sample size is not an explicit part of the equation which defines GFI, the *distribution* of GFI values is strongly affected by sample size.

Smaller sample sizes led to sampling distributions of the likelihood ratio test statistic that were found to be different from the corresponding theoretical chi-square distribution. The total number of indicators in the model, as evidenced by the main effects of number of factors and indicators per factor, and their interaction, produced similar results. These results for number of factors and indicators per factor, however, are strongly attenuated as sample size increases. As only main effects and two-way interactions were fit (due to considerations of the degrees of freedom to estimated parameters ratio), it may be the case that number of factors, indicators per factor and sample size interact, though this was not tested. Throughout, significant departures of the obtained sampling distributions from the corresponding theoretical distributions resulted in conservative hypothesis tests where the number of solutions that had associated probability values greater than .05 and .10 was larger than expected. The results with regard to sample size are consistent with Boomsma (1982), and the results for sample size by number of factors are consistent with Bearden, Sharma, and Teel (1982).

The relationships between the goodness-of-fit indices revealed that the information provided by AGFI adds little to the information provided by GFI, although AGFI appears to be less sensitive to the interaction of the number of factors with sample size. RMR is also highly related to GFI and AGFI. RHO is unrelated to all of the remaining goodness-of-fit indices, perhaps because of the extreme sampling variability of RHO.

This variability of RHO was largely accounted for by a very small number of outliers whose values were either much smaller than zero or much larger than one. These outliers only occurred with small sample sizes, and with one exception, for models with two indicators per factor. For the remaining models and sample sizes, the values of RHO were much closer to their predicted values. An explanation for these outlying values can be seen in the relationships of the three components of RHO to one another when sample size is small (i.e., 50 or 75). For smaller sample sizes the LISREL program provides a much better fit of the specified model than expected, so M_q is much smaller than $E(M_q)$. The sign and degree of aberration in the corresponding value of RHO depends upon the sample values of covariances which determine M_0, and in particular, the difference between M_0 and $E(M_q)$. When $M_0 > E(M_q) > M_q$, aberrant positive values of RHO must occur, whereas when $E(M_q) > M_0 > M_q$, aberrant negative values must occur.

While this study was ambitious in scope, nonetheless, generalization of the findings of any Monte Carlo study is limited by its design. One limitation of the present study is that the number of indicators per factor was not varied within the same model, which is more typical in practice. However, where models "fall between" the models (and sample sizes) studied here, their related goodness-of-fit values will likely be between the corresponding tabled values. Unfortunately, appropriate rules for interpolation are presently not known.

Perfect model specification and multivariate normality in the population were assumed throughout this study. What the differential effects, if any, of model misspecification or the violation of multivariate normality would be upon the behavior of these goodness-of-fit indices is also not known. These limitations not withstanding, the models used in this study should give the researcher worthwile guidelines in the application of maximum likelihood confirmatory factor analysis.

REFERENCES

Bearden W. O., Sharma, S., & Teel, J. E. (1982). Sample size effects on chi-square and other statistics used in evaluating causal models. *Journal of Marketing Research, 19*, 425–430.

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

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606.

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

Burt, R. S. (1973). Confirmatory factor-analytic structures and the theory construction process. *Sociological Methods and Research, 2*, 131–187.

Finn, J. D. (1973). *A general model for multivariate analysis*. NY: Holt, Rinehart and Winston, Inc.

Games P. A., Keselman, H. J., & Clinch, J. J. (1979). Tests for homogeneity of variance in factorial designs. *Psychological Bulletin, 86*, 978–984.

Goodman, L. A. (1971). The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics, 13*, 33–61.

Gweke, J. F., & Singleton, K. J. (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. *Journal of the American Statistical Association, 75*, 133–137.

Hays, W. L. (1973). *Statistics for the social sciences (2nd ed.)*. NY: Holt, Rinehart and Winston, Inc.

Heise, D. R. (1975). *Causal analysis*. NY: John Wiley & Sons.

IMSL. (1980). *International Mathematical and Statistical Libraries*. Houston: IMSL, Inc.

Jackson, D. N., & Chan, D. W. (1980). Maximum-likelihood estimation in common factor analysis: A cautionary note. *Psychological Bulletin, 88*, 502–508.

Joreskog, K. G. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika, 31*, 165–178.

Joreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32*, 443–482.

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

Joreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika, 57*, 239–251.

Joreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109–133.

Joreskog, K. G. (1978). Structural analysis of covariance and correlation matric *Psychometrika, 43*, 443–477.

Joreskog, K. G., & Sorbom, D. (1978). *LISREL: Analysis of linear structural relationships by the method of maximum likelihood (Version IV)*. Chicago: National Educational Resources, Inc.

Joreskog, K. G., & Sorbom, D. (1981). *LISREL: Analysis of linear structural relationships by the method of maximum likelihood (Version V)*. Chicago: National Educational Resources, Inc.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method*. NY: American Elsevier Publishing Company, Inc.

Learmouth, G. P. and Lewis, P. A. W. (November 1973). *Statistical tests of some widely used and recently proposed uniform random number generators*. (Report No. NPS55LW73111A). Monterey, CA: Naval Postgraduate School.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. NY: McGraw-Hill Book Company.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1–10.

van Driel, O. P. (1978). On various causes of improper solutions of maximum likelihood factor analysis. *Psychometrika, 43*, 225–243.

Manuscript received 6/6/83
Final version received 2/3/84