Journal of Applied Psychology
1989, Vol. 74, No. 4, 625–635

0021-9010/89/$00.75

Influence of Sample Size, Estimation Method, and Model Specification on Goodness-of-Fit Assessments in Structural Equation Models

Terence J. La Du and J. S. Tanaka
New York University

The problem of assessing fit of structural equation models is reviewed, and two sampling studies are reported that examine the effects of sample size, estimation method, and model misspecification on fit indices. In the first study, the behavior of indices in a known-population confirmatory factor analysis model is considered. In the second study, the same problem in an empirical data set is examined by looking at antecedents and consequences of work motivation. The findings across the two studies suggest that, (a) as might be expected, sample size is an important determinant in assessing model fit; (b) estimator-specific, as opposed to estimator-general, fit indices provide more accurate indications of model fit; and (c) the studied fit indices are differentially sensitive to model misspecification. Some recommendations for the use of structural equation model fit indices are given.

Structural equation models have become increasingly important methodological tools for studying linear relations in psychological data. Recent introductions to the area of structural equation modeling are provided in, for example, Anderson and Gerbing (1988), Bartholomew (1987), Connell and Tanaka (1987), Loehlin (1987), McDonald (1985), and Tanaka, Panter, Winborne, and Huba (in press). Specific applications have considered a wide variety of substantive psychological problems, including models of attitude–behavior relations (e.g., Bagozzi, 1981), stability of dysphoric affect (e.g., Tanaka & Huba, 1987), functional relations between job characteristics, satisfaction, and performance (Hogan & Martell, 1987) and the role of job satisfaction and organizational commitment in job turnover models (Williams & Hazer, 1986). As increasing numbers of applications of structural equation models appear in the psychological literature, issues surrounding the substantive robustness of obtained solutions to issues such as small sample size, data nonnormality, and model specification have been addressed (e.g., Bentler, 1983; Bentler & Bonett, 1980; Browne, 1982, 1984; Huba & Harlow, 1986, 1987; La Du, 1986; MacCallum, 1986; Mulaik, James, Van Alstine, Bennett, Lind, & Stilwell, 1989; Tanaka, 1987; Tanaka & Huba, 1985, 1987, 1988).

In particular, researchers interested in using these methods have become increasingly concerned with the issue of how to best ascertain model fit. A number of criteria have been suggested for evaluating the fit of structural equation models. The most well-known and well-accepted method of assessing fit in structural equation models is the chi-square goodness-of-fit statistic obtained under maximum likelihood (ML) and generalized least squares (GLS) methods of estimation (e.g., Jöreskog & Sörbom, 1985). In an *accept/support* strategy of model testing such as the one used in these methods, the null hypothesis being evaluated is that "the model fits the data." Hence, an investigator is looking to accept (or, more strictly speaking, to fail to reject) the null hypothesis or to look for models whose chi-square statistic values are small relative to their degrees of freedom.

Many criticisms have been raised regarding the validity of using this chi-square test to evaluate model fit. An early one noted by Bentler and Bonett (1980) is the fact that the chi-square is dependent on sample size. Thus, the larger the sample, the more difficult it is to fail to reject the null hypothesis. It is also known that the chi-square test is sensitive to data nonnormality (e.g., Bentler, 1983; Browne, 1982, 1984; Huba & Harlow, 1986, 1987; Tanaka & Huba, 1987). Although some research has been conducted looking at the seriousness of this problem (e.g., Mooijaart & Satorra, 1987; Satorra & Bentler, 1987; Shapiro & Browne, 1987), available data suggest that investigators should be appropriately skeptical about model fit when observed data do not follow "textbook" multivariate normal distributions.

Given these problems with the chi-square statistic, various alternative strategies for evaluating models have also been forwarded. Among the simplest of these is the ratio of the chi-square statistic for an obtained model to its degrees of freedom. Because the expected value for this ratio is 1.0, values close to

The research reported here derives, in part, from Terence J. La Du's doctoral dissertation, conducted under the supervision of J. S. Tanaka.

We wish to thank Jacob Cohen, Raymond Katzell, David Rindskopf, and Loriann Roberson for their participation on the doctoral examination committee as well as for their thoughtful comments on previous versions of this article. Several anonymous reviewers also provided helpful comments about this work.

Terence J. La Du is now affiliated with Bell Communications Research, Livingston, NJ.

Correspondence concerning this article should be addressed to Terence J. La Du, Bellcore, 290 West Mt. Pleasant Avenue, LCC-1B-132, Livingston, New Jersey 07039, or to J. S. Tanaka, Department of Psychology, New York University, 6 Washington Place, 7th Floor, New York, New York 10003.

1 are indicative of well-fitting models. The problem with using such a criterion in applications is that deviations from 1.0 are not well defined. For example, Marsh and Hocevar (1985) suggested that values of this ratio between 2.0 and 5.0 are indicative of acceptable models. Kaplan (1988) claimed that this ratio may be fairly insensitive to model misspecification.

Other approaches to understanding model fit have used other fit indices for structural equation models. Bentler and Bonett (1980), adapting earlier work by Tucker and Lewis (1973), proposed normed (Δ, hereafter referred to as NFI) and nonnormed (ρ) fit indices for covariance structure models. Both fit indices indicated the degree of fit associated with the tested model, with values close to 1.0 indicative of well-fitting models. Bentler and Bonett (1980) considered, in part, the issue of sample size and statistical power, correctly claiming that a measure of model fit, free of sample size considerations, was necessary to evaluate models appropriately, analogous to Cohen's (1977) notion of "effect sizes" in other linear statistical models. In addition, the Bentler–Bonett NFI was proposed as being nonestimator-specific in that it could be applied to any covariance structure estimator (e.g., of the estimators then well established: ML, GLS, and unweighted least squares [ULS]).

The interest sparked by the Bentler–Bonett article on addressing issues of fit in covariance structure models led to a number of alternative fit indices. For example, Jöreskog and Sörbom (1981, 1985) presented their own goodness-of-fit index (GFI) for ML and ULS estimation. James, Mulaik, and Brett (1982) introduced the parsimonious fit index, making a degrees-of-freedom correction to NFI. Cudeck and Browne (1983; Browne & Cudeck, in press) discussed Akaike and Schwarz information criteria to evaluate the fit of latent variable models under ML estimation for the purposes of cross-validation. Bentler (1983), in the context of asymptotically distribution-free estimation of latent variable models (cf., Browne, 1982, 1984), proposed a general fit index for such estimators. Tanaka and Huba (1985) showed that the index proposed in Bentler (1983) was the general case of the Jöreskog–Sörbom fit indices under ML and ULS estimation. The Tanaka–Huba result was subsequently extended to show that their fit index defined a general multivariate coefficient of determination for covariance structure models under arbitrary GLS estimation, including the case of asymptotically distribution-free estimators (Tanaka & Huba, 1988). Marsh, Balla, and McDonald (1988) provided a recent review of different fit indices available for covariance structure models.

Given the number of different options available to assess the fit of any given structural equation model, it is unclear which (if any) of the alternatives might provide the best index of model fit. In other words, how might researchers summarize results of different models such that sufficient information about their fit is communicated? Furthermore, in trying to examine the cross-validity of models across different samples, in studies conducted by different investigators, or using different estimators, it is unclear whether different fit indices are indicating the same degree of fit. For example, Tanaka (1987) presented an example in which an index of model fit (NFI) applied to the same data and model was affected by choice of estimator.

To address issues about determinations of model fit based on these fit indices, some investigators have undertaken sampling studies to examine fit index behavior in different experimental conditions (e.g., sample size, number of observed variables in the models, etc.). These are reviewed next.

Previous Simulation Work in Structural Equation Fit Indices

Much of the simulation work done in the domain of structural equation models has focused on issues such as (a) properties of estimators under standard and nonstandard conditions (e.g., Boomsma, 1982; Muthén & Kaplan, 1985; Tanaka, 1984); (b) properties of the chi-square test statistic under conditions of small sample size or data nonnormality, or both (e.g., Bearden, Sharma, & Teel, 1982; Geweke & Singleton, 1980; Sharma, Durvasulva, & Dillon, 1989; Tanaka, 1984); or (c) model misspecification (Kaplan, 1988; MacCallum, 1986). None of these studies are directly central to the examination of the nonparametric fit indices.

A study explicitly comparing fit indices was conducted by Anderson and Gerbing (1984). They compared the following five fit indices: (1) the goodness-of-fit index (GFI) of Jöreskog and Sörbom (1981; Tanaka & Huba, 1985); (2) the Jöreskog–Sörbom (1981) adjusted goodness-of-fit index (AGFI); (3) the root-mean-squared residual (RMR); (4) the probability of the chi-square statistic given the null hypothesis; and (5) the Tucker–Lewis/Bentler–Bonett nonnormed coefficient. They concluded that GFI, AGFI, and RMR were all affected by sample size, the number of observed variable indicators per factor, and the number of factors; and that the Tucker-Lewis/Bentler–Bonett coefficient was unrelated to all of the other measures of fit.

By using simulated data and two types of models (a two-factor and a four-factor, both with three indicators each), Bearden et al. (1982) found that both the chi-square statistic and NFI were affected by sample size and model complexity.

Luijben, Boomsma, and Molenaar (1987) reported results from a Monte Carlo study investigating the behavior of covariance structure fit using a step-up procedure, that is, an initial baseline model is obtained that estimates one fewer parameter than the known population model. In their 300 replications using this step-up procedure, the modification process detected the correct model only 18% of the time. However, the focus of their study was not directly on the nonparametric fit indices we are concerned with here.

Marsh et al. (1988) examined a number of fit indices and evaluated them on the criterion of stability over differing sample sizes. Sampling from four data sets of both simulated and real data, they fitted a three-factor model with three indicators each and allowed the factors to be correlated in the four data sets. They found that sample size did affect the value of most commonly used fit indices (e.g., chi-square, chi-square/*df*, RMR, GFI, etc.). They then transformed these indices in two ways. The first (Type I) transformed indices relative to the fit generated from a null model (a reproduced matrix containing variances in the diagonal and zeros in the off-diagonal). The second (Type II) transformed indices relative to both a null model index and an expected fit index value that included sample size considerations. They concluded that Type II indices should be used in practice, because they did not exhibit sample size effects.

A number of issues are left unresolved by these studies. In particular, studies have typically excluded consideration of either the range of available fit indices or estimation method. For example, the Anderson and Gerbing and Luijben et al. studies omitted one of the most widely used of the fit indices for latent variable models, Bentler and Bonett's NFI. When the Anderson and Gerbing and Bearden et al. studies were conducted, the range of available estimation strategies was limited, with ML methods predominating. The more recent Marsh et al. (1988) study used only ML in their work. Given the possible effect of the choice of estimation method on the fit indices (e.g., Tanaka, 1984; Tanaka & Huba, 1985), it may be the case that estimator-specific, rather than estimator-general, fit indices are appropriate for ascertaining model fit. Our studies were conducted to investigate the following three questions in assessing fit in latent variable models:

1. Does sample size interact with either estimation method or model misspecification in affecting the magnitude of goodness-of-fit (GOF) indices? The main effect of sample size on these indices is well known from previous research (Anderson & Gerbing, 1984; Bearden et al., 1982; Marsh et al., 1988).

2. To what extent does the estimation method used in the solution of structural equation models affect the values of GOF indices? Does sample size moderate the influence of estimation method on the size of GOF indices? This examines an implication of Tanaka and Huba's (1985) derivation of estimator-specific fit indices. Given the asymptotic equivalence of ML and normal theory GLS, for example, we might expect differences between fit indices across estimation methods to disappear with increasing sample size. No previous study has explicitly looked at the impact of estimation method on fit index values.

3. Does model specification affect the values of GOF indices? More specifically, what happens to a GOF index value if a parameter is estimated whose population value is zero, or if a parameter is fixed to zero whose population value is nonzero? Does sample size moderate the influence of model misspecification on the GOF index sizes? Here, we address the sorts of specification issues discussed by, for example, Farley and Reddy (1987), Luijben et al. (1987), and MacCallum (1986; Silvia & MacCallum, 1988).

General Study Designs

Two sampling studies were conducted to address these questions. The first study consisted of drawing samples from a known population model. The design is a 3 × 2 × 2 (sample size: *N*s = 35, 100, 300; estimation method: ML, GLS; model specification: correct-population, incorrect–trivial misspecification) factorial design with the latter two factors being within-replicates. Note that although a sample size of 35 may be considered as a priori "too small" for latent variable modeling (e.g., Anderson & Gerbing, 1984; Boomsma, 1982), support for an examination of fit index behavior at this sample size is warranted both in theory and in practice. For example, Geweke and Singleton (1980) demonstrated that the chi-square statistic was not adversely affected in sample sizes of 30. Furthermore, models have been fit in samples close to this size in practice (e.g., Bornstein & Benasich, 1986).

The second study entailed sampling from an actual data base whose variables were not multivariate normally distributed. The models tested contained only observed variables (i.e., only path models were examined), with the number of observed variables being 10 rather than 6 (as in the first study). The impact of employee morale on self-reported performance, inclination to stay, and attendance behavior is the focus of the model, with interest in the antecedents of worker morale being included in the system. These data are more fully described in Katzell (1982). The full data base consists of 3,328 survey responses. The design for Study 2 is a 3 × 2 × 3 (sample size: *N*s = 70, 200, 600; estimation method: ML, GLS; model specification: correct-baseline, that is, based on all 3,328 responses; incorrect–trivial misspecification; incorrect–nontrivial misspecification) factorial design with repeated measurements on the latter two factors.

These two studies differ from the previous research in three aspects. First, sampling is from a known population model with a multivariate normal distribution as well as from a data base whose correct or true model is only approximately known and whose variables are not multivariate normal. Previous research efforts have generally used the former to the exclusion of potentially unique effects in less structured data. Second, these studies address the influence of estimation method on the behavior of GOF indices. Previous studies have used the ML estimator almost exclusively and it is of interest to determine if the GOF indices for two asymptotically equivalent estimators, ML and GLS, are comparable in finite samples. Finally, our use of the model specification factor allows us to examine the behavior of GOF indices in nested models. Although nested models have been described in applications of the analysis of covariance structures (e.g., Bentler & Bonett, 1980; James et al., 1982), the behavior of nested models' GOF indices has not been previously examined. Model nesting represents one strong method for comparing competing psychological theories in data.

We believe these two studies are analogous to the laboratory (Study 1) and field (Study 2) experimental inquiry of a research question. In Study 1, we examine how experimental factors in a controlled design affect the variability of four GOF indices with samples drawn from a population distribution with known properties. In Study 2, with a slight variation on the model specification factor, we examine the impact of the same experimental factors on GOF index behavior by using samples drawn from a nonnormal multivariate baseline data base whose population properties are only approximately known. We believe this makes these two studies somewhat unique in addressing the issues raised.

Fit Indices

Four GOF indices are studied. The first two GOF indices considered are the Jöreskog and Sörbom (1981) GFI_{ML} and the Tanaka and Huba (1985) GFI_{GLS}, as follows:

$$GFI_{ML} = 1 - tr(\hat{\Sigma}^{-1}\mathbf{S} - \mathbf{I})^2/tr(\hat{\Sigma}^{-1}\mathbf{S})^2 \quad (1)$$

and

$$GFI_{GLS} = 1 - tr(\mathbf{I} - \hat{\Sigma}\mathbf{S}^{-1})^2/p, \quad (2)$$

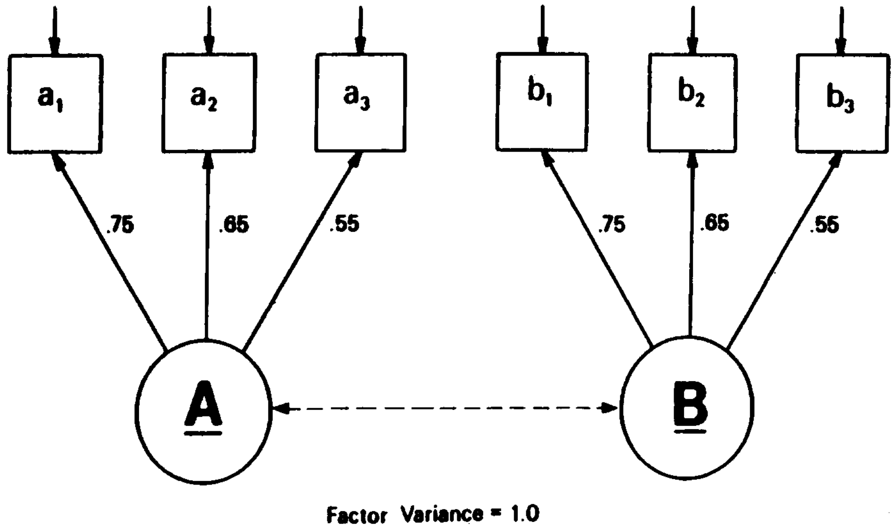

Figure 1. Correct two-factor model: Population covariance structure. (Dashed line indicates a trivial misspecified path.)

where S is the sample covariance matrix, $\hat{\Sigma}$ is the reproduced covariance matrix based on the specified model, and p is the rank of S (i.e., the total number of observed variables). As shown in Tanaka and Huba (1985), these represent special cases of a weighted coefficient of determination (Tanaka & Huba, 1988), originally presented in Bentler (1983). The normed fit index for both ML (NFI_{ML}) and GLS (NFI_{GLS}) are also included (Bentler & Bonett, 1980), as follows:

$$NFI = (F_k - F_l)/F_0, \tag{3}$$

where F_0 is the minimum of the fit function for any estimation method corresponding to the worst case model. Historically (e.g., Bearden et al., 1982; Marsh et al., 1988), this has been taken to be the case when observed variables are mutually uncorrelated (although, see Sobel & Bohrnstedt, 1985, for an objection to this particular "null model"). F_k is the fit function minimum for any substantive model, M_k, whereas F_l is the minimum of a substantive model, M_l, but one that is less restrictive than F_k. Because $F_0 \geq F_k \geq F_l$, the index is additive and lies between 0 and 1. For the results reported here, the NFIs measure improvement only over the *null model,* defined as one resulting in a $\hat{\Sigma}$ whose diagonals contain the variances of the measured variables while zeros occupy the off-diagonal positions. As a result, the NFIs are formulated as

$$NFI_{k0} = (F_0 - F_k)/F_0, \tag{4}$$

where F is the minimum of either ML or GLS minimization functions for the particular models.

Study One

Method

An orthogonal two-factor model with three indicators each in a simple cluster structure was used with factor loadings of .75, .65, and .55 (see Figure 1). Other population parameters (i.e., residual variances) were specified such that diagonal elements of the population covariance matrix, Σ, were equal to 1.0.

Three sample sizes of 35, 100, and 300 were used. The smallest, $N = 35$, was slightly over the minimum necessary number of observations to compute a positive definite covariance matrix with six observed variables, that is, $p(p + 1)/2 = 21$. The second sample size has been stated as a minimum for factor analytic models (Boomsma, 1982, p. 171). The third sample size chosen, $N = 300$, was above the previously published minimum of $N = 200$ recommended if the researcher wished to limit the risk of drawing improper conclusions (Boomsma, 1982, 1985).

Two estimation methods, ML and GLS, were examined. The behavior of GOF indices when estimation method has been varied has not been previously reported in the literature.

Model specification was the third experimental factor. As previously noted, in the population model, the two factors were orthogonal. The correctly specified population model (correct model) did not estimate the covariance between the two latent variables, whereas the model corresponding to the incorrect–trivial misspecification model allowed this parameter to be estimated.

Fifty replications were drawn for each sample size. Given the population factor loadings, the equations for each of six measured variables were derived and random normal deviates from the IMSL subroutine GGNML (International Mathematical and Statistical Libraries, IMSL, 1980) were used for the factor and residual scores. An initial random start seed was provided, with subsequent seeds generated by the algorithm for each following sample. After 25 samples were produced, a new seed was introduced for the 26th sample, with the GGNML algorithm generating the remaining 24 start seeds. Each sample size had its own unique set of start seeds. Values for the six measured variables were obtained for each sample size, with a total of 50 replications each. Factor loadings were given at starting values of .3 for each run. Estimated parameters included factor and unique loadings in the correctly specified model by using the modified specification of the Jöreskog–Sörbom (1985) LISREL model discussed in Rindskopf (1983). Unique variances, as well as factor variances, were treated as fixed parameters and set to 1.0.

In all, 600 solutions were required for this design. With the smallest sample size ($N = 35$), the LISREL program failed to converge in 250 iterations for six samples, and replacement samples were generated. For the two larger sample sizes, no convergence problems occurred. Summary statistics indicated that data were correctly generated as multivariate normal.

Table 1
Marginal Means for Goodness-of-Fit Indices: Study 1

Variable	GFI_{ML} *M*	GFI_{ML} *SD*	GFI_{GLS} *M*	GFI_{GLS} *SD*	NFI_{ML} *M*	NFI_{ML} *SD*	NFI_{GLS} *M*	NFI_{GLS} *SD*
Sample size								
N = 35	.880	.056	.882	.061	.729	.122	.915	.062
	.734–.973		.717–.974		.469–.941		.767–.982	
N = 100	.969	.012	.971	.011	.919	.035	.959	.019
	.932–.991		.935–.991		.802–.976		.902–.989	
N = 300	.989	.005	.989	.005	.967	.014	.980	.008
	.975–.998		.977–.998		.929–.993		.955–.995	
Estimation method								
ML	.962	.034	.930	.089	.902	.081	.940	.056
	.864–.998		.546–.998		.673–.993		.645–.995	
GLS	.930	.083	.964	.030	.842	.174	.964	.023
	.603–.998		.878–.998		.264–.993		.868–.996	
Model specification								
Correct	.942	.062	.942	.067	.861	.143	.947	.047
	.717–.998		.659–.998		.307–.993		.649–.995	
Incorrect	.950	.055	.952	.053	.883	.115	.956	.032
	.750–.998		.757–.998		.483–.994		.857–.996	
Overall	.946	.058	.947	.059	.872	.127	.952	.038
	.734–.998		.717–.998		.469–.993		.767–.995	

Note. Ranges are given below means and standard deviations. ML = maximum likelihood estimation; GLS = generalized least squares estimation; GFI_{ML} = goodness-of-fit index for ML estimation (Jöreskog & Sörbom, 1981); GFI_{GLS} = goodness-of-fit index for GLS estimation (Tanaka & Huba, 1985); NFI_{ML} = normed fit index using the ML fit function (Bentler & Bonett, 1980); NFI_{GLS} = normed fit index using the GLS fit function (Bentler & Bonett, 1980).

Results

By using SAS's PROC MATRIX routine (SAS, 1982), we computed the GOF indices using matrices produced by LISREL. These were cross-checked, when appropriate, against GOF values output by the LISREL program. To assess the overall effect of the experimental factors on goodness of fit, a multivariate analysis of variance (MANOVA) with one between-replicates variable (sample size) and three within-replicates variables (estimation method, model specification, and GOF type) was conducted on the Fisher's *z*-transformed GOF indices. We observed that the raw GOF indices were skewed with ceiling effects due to the upper bound of 1.0. Although this might not be the optimal transformation for these data, the Fisher's *z*-transformed indices did not have the skewed distributions observed in the raw indices. For example, the kurtosis and skew of the raw GFI_{ML} indices were 6.29 and −2.38, respectively, whereas the Fisher's *z*-transformed GFI_{ML} values were −0.52 and −0.25, respectively. Hence, the transformed data better met the distributional assumptions for inferential tests, although as might be expected, substantive results did not appreciably change as a function of this transformation.

As is the case in sampling studies, many effects are statistically significant. In both studies, we focus on the presentation of descriptive aspects of study results, noting the statistical significance of results where necessary. Statistical tests were conducted on transformed indices, whereas tables present the data in their original metric.

Table 1 confirms Anderson and Gerbing's finding that sample size affects the size of the fit index: as sample size increases, so does the value of the index. Table 1 findings also support the Tanaka and Huba (1985) proposal about fit index optimality being dependent on the estimation method. The average value of GFI_{ML} for a ML solution was .032 larger than for GLS solutions (.962 vs. .930), whereas, for GFI_{GLS}, a GLS solution is .034 larger than a ML solution (.964 vs. .930). A Wilcoxon matched-pairs signed ranks test indicated that, for the 300 comparisons of ML solutions versus GLS solutions on their respective GFI_{ML} values, all values of ML solutions were greater than their GLS solution values, $Z = 15.01$, $p \leq .001$. The same was true for GFI_{GLS}, where GLS solution values were always greater than ML solution values, $Z = 15.01$, $p \leq .001$. The NFI for each of the estimation methods was similarly affected: NFI_{ML} computed with ML solutions had a larger value than when it was computed with a GLS solution, with the reverse being true for NFI_{GLS}. There were relatively large Sample Size × Estimation Method interactions for NFI_{ML}, NFI_{GLS}, GFI_{ML}, and GFI_{GLS}, as Table 2 indicates. For small sample sizes, the discrepancy between estimator-appropriate versus estimator-inappropriate GOF index values was large, but as sample size increased, this difference decreased. For example, in looking at Table 2's GFI_{ML} column, the discrepancy when N = 35 was

Table 2
Sample Size by Estimation Method Marginal Means: Study 1

Variable	GFI_{ML} *M*	GFI_{ML} *SD*	GFI_{GLS} *M*	GFI_{GLS} *SD*	NFI_{ML} *M*	NFI_{ML} *SD*	NFI_{GLS} *M*	NFI_{GLS} *SD*
N = 35								
ML	.922	.028	.835	.099	.809	.070	.885	.063
	.864–.975		.546–.972		.673–.946		.645–.980	
GLS	.837	.084	.928	.025	.649	.177	.946	.024
	.603–.971		.878–.976		.264–.935		.868–.984	
N = 100								
ML	.974	.010	.968	.014	.928	.029	.955	.022
	.946–.991		.923–.991		.834–.977		.887–.988	
GLS	.965	.015	.974	.009	.911	.041	.964	.017
	.919–.990		.947–.991		.770–.975		.917–.989	
N = 300								
ML	.989	.004	.989	.005	.969	.013	.980	.009
	.979–.998		.974–.998		.932–.993		.953–.995	
GLS	.988	.005	.990	.004	.966	.015	.981	.008
	.972–.998		.979–.998		.926–.993		.957–.996	

Note. Ranges are given below means and standard deviations. ML = maximum likelihood estimation; GLS = generalized least squares estimation; GFI_{ML} = goodness-of-fit index for ML estimation (Jöreskog & Sörbom, 1981); GFI_{GLS} = goodness-of-fit index for GLS estimation (Tanaka & Huba, 1985); NFI_{ML} = normed fit index using the ML fit function (Bentler & Bonett, 1980); NFI_{GLS} = normed fit index using the GLS fit function (Bentler & Bonett, 1980).

.922 − .837 = .085, whereas for *N* = 300, this becomes .989 − .988 = .001.

When the covariance between the two latent variables is estimated resulting in a misspecified model, a slight improvement in fit is indicated for all four GOF indices, supporting the effect of model specification on these indices. Higher order interactions involving model specification, although occasionally reaching statistical significance at the .05 level, are substantively uninteresting given the small effect sizes associated with these effects.

Discussion

The results of this study replicate the sample size findings of Anderson and Gerbing (1984), Bearden et al. (1982), and Marsh et al. (1988). Estimation method effects are present, as suggested by Tanaka and Huba (1985) for GFI_{ML} and GFI_{GLS}, as well as the two versions of the NFI. To repeat an earlier observation, it is noteworthy that, for the same model and data, the average GOF index values for solutions using different estimation methods might lead to different conclusions, for example, for *N* = 100, the correct model, and ML estimation method: NFI_{ML} = .922; for the same sample size and model, but GLS estimation method: NFI_{GLS} = .961. This would seem to suggest that the normed fit index does not differentiate well over some threshold of fit.

A comment should also be made about investigator sensitivity in model selection. As can be seen from Table 1, in all cases, fitting an extra parameter that is extraneous to the true model leads to increments in fit of between 1% and 2%. Depending on an investigator's sensitivity threshold regarding model evaluation, this additional increment may be interpreted as being substantively important. If the heuristics of statistical significance were to be used in making decisions about model fit, this increment should be more important to interpret at larger sample sizes, all other things being equal.

Study 2 was conducted to investigate the robustness of these findings in a naturalistic context by using an empirical data base with nonnormally distributed data.

Study 2

The questionnaire responses serving as the data base for Study 2 were collected as part of a larger project examining antecedents and consequences of work motivation. Integrating need theory, attitude theory, expectancy theory, goal theory, extrinsic–intrinsic theory, equity theory, reward/reinforcement theory, and role theory, as well as more comprehensive theories discussed in Katzell and Thompson (1987), Katzell (1982) defined constructs pertinent to employee motivation and productivity. Furthermore, he outlined a research program for collecting self-reports on these constructs, supervisory ratings of work groups, objective performance data, and an inventory of the personnel policies with respect to linkages to the work motivation constructs. The self-report data served as the baseline data base for Study 2.

Table 3 gives the variable labels, number of scale items, and the internal consistencies for each scale. Figure 2 graphically depicts the correct-baseline model. The model has two exogenous observed variables and eight endogenous observed variables. A line with a single arrowhead indicates a hypothesized, unidirectional relationship between two variables, and a bidi-

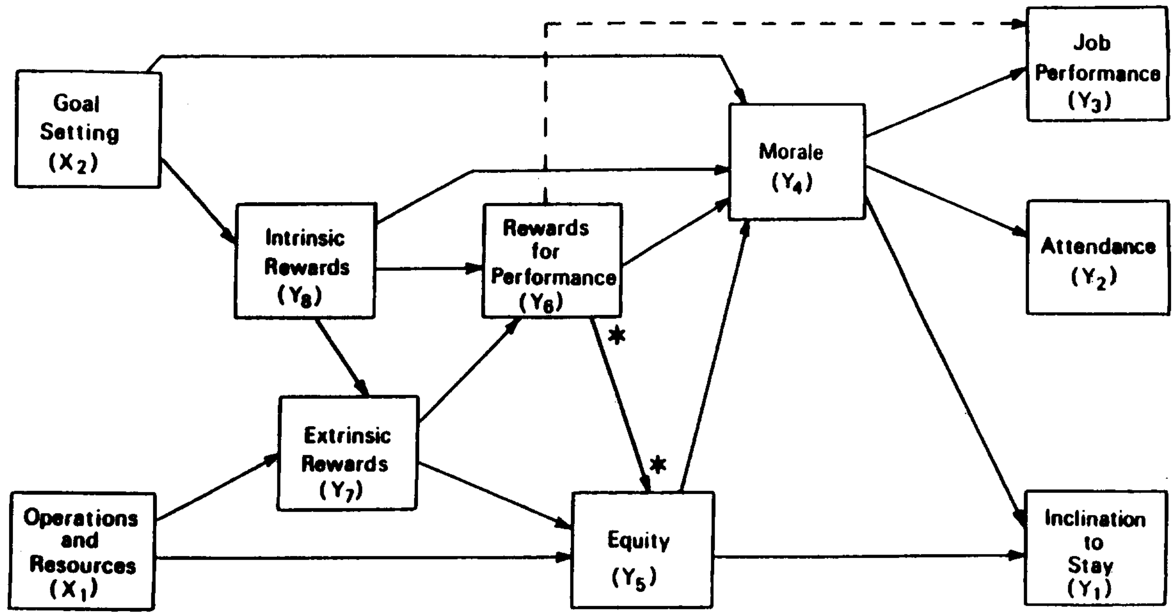

Figure 2. Path model of Study 2. (Solid lines are correct model; dashed line indicates a trivial misspecified path; starred solid line indicates a nontrivial misspecified path deletion.)

rectional arrow indicates a correlation. Sample multivariate kurtosis values indicate that these data are not multivariate normal.

Besides the correct-baseline model, two additional model specifications were used in this study. In Figure 2, the dotted line from rewards for performance to self-reported level of performance is a path that, in the baseline, resulted in a trivial improvement in fit. The RMR difference between these models for a ML solution on the total data base (N = 3,328) was zero. For GLS solutions, this difference was .003. The model including this path will be referred to as the incorrect–trivial model reflecting a trivial model misspecification. The asterisked path from rewards for performance to equity is fixed to zero in the third model specification, the incorrect–nontrivial misspecification model. Note that this corresponds to the Luijben et al. (1987) suggestion of conducting sampling studies when a model has one less parameter than the true (i.e., baseline) model. The RMR difference for a ML solution for this model is .147, whereas the GLS RMR difference is .310, demonstrating the empirical effects of misspecification for these data. Theoretically, this path is crucial because perceptions of equity in the work situation are based in part on whether rewards are tied to performance (Mowday, 1983). Thus, this incorrect–nontrivial model represents both an empirical and a theoretical model misspecification. Our study investigates the behavior of the GOF indices under these trivial and nontrivial misspecifications.

Method

The research design is similar to that of Study 1, with one slight modification in the model specification factor. The design is a 3 × 2 × 3 (sample size: 70, 200, 600; estimation method: ML, GLS; model specification: correct, incorrect–trivial, incorrect–nontrivial), with repeated measures on the latter two.

The 3,328 questionnaire responses were collected from nine different organizations participating in the full research project outlined by Katzell (1982). All variables had less than 5% missing data. When missing data occurred, the variable's mean calculated across all responses was substituted. Table 3 contains the baseline covariance matrix, means, and standard deviations. To randomly sample with replacement from the data base, a Fortran program utilizing IMSL's GGSRS routine was written (IMSL, 1980). One hundred fifty samples were drawn, 50 for each of the three sample sizes, with a new seed introduced after the first 25 samples were drawn. In all, 900 solutions were required. Using the baseline ML estimates for start values, all solutions converged in less than 250 iterations for all sample sizes.

Results

By using SAS's PROC MATRIX program (SAS, 1982), we computed the GOF indices from the matrix materials produced by LISREL and again cross-checked against LISREL output when appropriate. In assessing the overall effect of the experimental factors on goodness of fit, a MANOVA with a single between-replicates variable (sample size) and three within-replicates variables (model specification, estimation method, and GOF index type) was conducted on the Fisher's z-transformed GOF indices. All reported inferential statistics used results from transformed indices, whereas all tables providing descriptive statistics, such as means, present the raw, untransformed GOF indices. Again, the interpretative focus is on descriptive results, supplemented with statistical findings when necessary.

All GOF indices were affected by sample size such that as sample size increased, so did the value of the GOF index. Calculated effect sizes showed that the GFI indices and NFI_{ML} were about equally affected with eta squares ranging from .709 to

Table 3
Population Means, Standard Deviations, and Covariance Matrix: Study 2

Name	Scale	*M*	*SD*	No. items	α	1	2	3	4	5	6	7	8	9	10
1. Inclination to stay	Y1	4.542	1.720	4	.84	—									
2. Attendance	Y2	6.064	1.073	3	.59	0.401	—								
3. Job performance	Y3	5.517	0.873	9	.66	0.214	0.230	—							
4. Morale	Y4	5.252	1.164	9	.81	1.435	0.337	0.274	—						
5. Equity	Y5	4.357	1.297	7	.76	1.259	0.257	0.135	0.866	—					
6. Rewards for performance	Y6	3.916	1.308	8	.86	0.990	0.222	0.130	0.748	0.920	—				
7. Extrinsic rewards	Y7	4.108	1.199	7	.76	1.060	0.220	0.199	0.789	1.072	0.867	—			
8. Intrinsic rewards	Y8	4.334	1.218	7	.82	0.958	0.189	0.281	0.791	0.724	0.767	0.858	—		
9. Operations & resources	X1	5.102	1.163	11	.85	0.795	0.121	0.126	0.543	0.726	0.444	0.622	0.410	—	
10. Goal setting	X2	4.410	1.050	6	.66	0.859	0.208	0.258	0.674	0.738	0.637	0.630	0.635	0.571	—

.740. The surprise was that NFI_{GLS} was only slightly influenced by sample size in comparison with the other GOF indices. It is apparent from Table 4 that, with the average level of this index across sample sizes (.944–.961) and the relatively small standard deviations, sample size had little effect on the value of NFI_{GLS}.

As is evident from the within-replicate variables of Table 4, the contrast of the correct model with the incorrect–trivial model resulted in no improvement in fit as expected. The model specification contrast of the incorrect–nontrivial model versus the average of the correct and incorrect–trivial models showed the expected effect. As to method effects, both the GFI and NFI indices were strongly affected by the estimation method used.

By averaging GFI_{ML} and GFI_{GLS} values across model specifications, we conducted a Wilcoxon matched-pairs signed ranks test. For the 150 comparisons of a GLS solution, GFI_{GLS} was greater than GFI_{ML} in every case, $Z = 10.64$, $p < .0001$. The reverse was true for ML solutions, where GFI_{ML} was greater than GFI_{GLS} in each case. The same results were obtained for NFI_{ML} and NFI_{GLS}.

The interaction effects were as follows: For the sample size by estimation method effect, only GFI_{GLS} ($\eta^2 = .07$), NFI_{GLS} ($\eta^2 = .059$), and NFI_{ML} ($\eta^2 = .059$) were of substantive interest. Table 5 presents marginal means for this interaction. In each instance, as sample size increased, the difference between the ML and GLS estimated models decreased. Remaining higher order interactions were not substantively interesting.

The results of this study replicate findings that sample size strongly influences the size of GOF indices' values (e.g., Anderson & Gerbing, 1984, Bearden et al., 1982, Marsh et al., 1988, and Study 1). Method effects predicted by Tanaka and Huba (1985) and found in Study 1 were also obtained in this study.

Overall Discussion

Our initial motivation in conducting these studies was to examine the impact of sample size, estimation method, and model misspecification on some commonly used goodness-of-fit indices in structural equation models. We predicted that sample size would have an impact on these fit indices as had been previously shown by Anderson and Gerbing (1984), Bearden et al. (1982), and Marsh et al. (1988). Previous work had not investigated the effects of estimation method or model misspecification. With regard to the former, results presented in Jöreskog and Sörbom (1981) and Tanaka and Huba (1985, 1988) suggested that estimator-specific fit indices might be more appropriate in finite samples than were the estimator-general fit indices of Bentler and Bonett (1980). The question of model misspecification was open: Would clear substantive changes in model specification cause corresponding increments or decrements in GOF indices? The two studies presented here (a) affirm our initial hypothesis about sample size, (b) confirm our speculation that estimator-specific rather than estimator-general fit indices are more appropriate in finite samples, and (c) demonstrate that substantive model changes are detected and trivial model changes are ignored when investigating issues of model specification.

Another motivation for conducting these studies was to ex-

Table 4
Marginal Means for Goodness-of-Fit Indices: Study 2

	GFI_{ML}		GFI_{GLS}		NFI_{ML}		NFI_{GLS}	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Sample size								
$N = 70$	.782	.053	.776	.067	.711	.083	.944	.019
	.608–.868		.557–.881		.475–.849		.883–.977	
$N = 200$	.877	.036	.878	.030	.844	.048	.954	.013
	.750–.933		.795–.930		.648–.911		.917–.978	
$N = 600$	.917	.013	.913	.015	.893	.019	.961	.008
	.871–.942		.859–.938		.814–.930		.929–.957	
Estimation method								
ML	.914	.030	.800	.120	.884	.041	.936	.025
	.820–.957		.279–.926		.761–.943		.823–.972	
GLS	.803	.108	.912	.026	.748	.154	.970	.008
	.350–.928		.832–.950		.096–.916		.938–.987	
Model specification								
Correct	.878	.062	.873	.067	.845	.083	.958	.015
	.645–.956		.566–.950		.574–.946		.890–.984	
Incorrect–trivial	.878	.063	.876	.065	.847	.084	.959	.015
	.622–.956		.574–.949		.544–.946		.891–.984	
Incorrect–nontrivial	.820	.086	.818	.095	.756	.132	.941	.019
	.557–.915		.320–.915		.397–.898		.860–.968	
Overall	.859	.068	.856	.072	.816	.096	.953	.016
	.608–.942		.557–.938		.475–.930		.883–.978	

Note. Ranges are given below means and standard deviations. ML = maximum likelihood estimation; GLS = generalized least squares estimation; GFI_{ML} = goodness-of-fit index for ML estimation (Jöreskog & Sörbom, 1981); GFI_{GLS} = goodness-of-fit index for GLS estimation (Tanaka & Huba, 1985); NFI_{ML} = normed fit index using the ML fit function (Bentler & Bonett, 1980); NFI_{GLS} = normed fit index using the GLS fit function (Bentler & Bonett, 1980).

Table 5
Sample Size by Estimation Method Marginal Means: Study 2

	GFI_{ML}		GFI_{GLS}		NFI_{ML}		NFI_{GLS}	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
$N = 70$								
ML	.878	.021	.669	.116	.838	.034	.918	.030
	.820–.924		.279–.838		.761–.900		.823–.967	
GLS	.686	.092	.883	.021	.584	.143	.970	.010
	.350–.834		.832–.923		.096–.804		.938–.987	
$N = 200$								
ML	.922	.013	.838	.047	.894	.018	.939	.019
	.894–.952		.703–.914		.849–.931		.884–.972	
GLS	.832	.062	.918	.013	.794	.081	.968	.007
	.606–.913		.887–.946		.448–.892		.947–.983	
$N = 600$								
ML	.942	.006	.891	.021	.919	.010	.951	.011
	.929–.957		.812–.926		.886–.943		.901–.968	
GLS	.892	.020	.935	.009	.867	.027	.970	.005
	.812–.928		.906–.950		.743–.916		.950–.978	

Note. Ranges are given below means and standard deviations. ML = maximum likelihood estimation; GLS = generalized least squares estimation; GFI_{ML} = goodness-of-fit index for ML estimation (Jöreskog & Sörbom, 1981); GFI_{GLS} = goodness-of-fit index for GLS estimation (Tanaka & Huba, 1985); NFI_{ML} = normed fit index using the ML fit function (Bentler & Bonett, 1980); NFI_{GLS} = normed fit index using the GLS fit function (Bentler & Bonett, 1980).

amine the behavior of the Bentler–Bonett normed fit index, NFI, across different estimation methods. Although this index is perhaps the most popular to use in applications of structural equation models in psychology, previous comprehensive studies of fit indices by Anderson and Gerbing (1984) and Marsh et al. (1988) did not study NFI across different estimation methods. The findings of these two studies suggest that NFI is not a good summary of model fit, particularly when different methods of estimation are being compared. This supports in a systematic way the claim made in Tanaka (1987) and tested in La Du (1986). It appears that comparing NFIs across different methods of estimating the same model in the same data could lead to different substantive conclusions. This is explicit when one looks at Table 5, in which, for $N = 200$, the average of NFI_{ML} for ML estimation is .894 and NFI_{GLS} for GLS is .968 (an average difference of .074), whereas, for the same sample size but with the GFIs, GFI_{ML} is .922 and GFI_{GLS} is .918 (a difference of .003). Note, however, that these comments are strictly true only for the version of NFI computed for a null model in which measured variables are hypothesized to be mutually uncorrelated.

As is the case in any sampling investigation, the findings presented here may lack generalizability beyond the conditions examined in our two studies. However, the external validity of the conclusions in these studies were enhanced by the sampling experiments of Study 2, which used real data in a model of substantive interest. Study 2 thus can be thought to simulate conditions observed in actual observed data to a greater extent than do the studies of Anderson and Gerbing (1984), Bearden et al. (1982), or our Study 1, because they are based on real data. Furthermore, in Study 2 we investigated the behavior of GOF indices under conditions of data nonnormality; we suggest that model fit can be adequately described even when nonoptimal estimation strategies are used, as might be expected given the results of Tanaka and Huba (1988).

Although there would appear to be no best index of fit in examining latent variable models, the findings obtained in these studies allow some conclusions to be drawn. First, it is clear that the NFI of Bentler and Bonett should be used with caution, particularly when making comparisons across different methods of estimation or different-sized samples. Second, when trying to compare findings across different studies, sample size and estimation method must be considered. This focus on sample size can be obscured by using the chi-square goodness-of-fit statistics in these models, because degrees-of-freedom calculations are not explicitly based on sample size as in an ANOVA or multiple regression. Finally, for these estimation methods, it does appear that the estimator-specific fit indices studied are more sensitive in picking up important model misspecifications when they occur and in ignoring trivial model misspecifications, although in some sense this is dependent on differential investigator sensitivity to the detection of small changes in these fit indices. However, the NFI under GLS estimation appears to be particularly insensitive to important model misspecification. The fit indices examined here and, in particular, those given in the LISREL program and in Tanaka and Huba (1985) provide important adjunctive information in ascertaining the fit of latent variable models. These empirical findings, along with the Tanaka and Huba (1988) theoretical results, support the use of Jöreskog–Sörbom/Tanaka–Huba GFI indices in evaluating structural equation models.

References

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness of fit indices for maximum likelihood confirmatory factor analyses. *Psychometrika, 49,* 155–173.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411–423.

Bagozzi, R. P. (1981). Attitudes, intentions, and behavior: A test of some key hypotheses. *Journal of Personality and Social Psychology, 41,* 607–627.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis.* London, England: Griffin.

Bearden, W. O., Sharma, S., & Teel, J. (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. *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. Jöreskog & H. Wold (Eds.), *Systems under indirect observation* (Part 1, pp. 149–173). Amsterdam, The Netherlands: North-Holland.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika, 50,* 229–242.

Bornstein, M. H., & Benasich, A. A. (1986). Infant habituation: Assessments of individual differences and short-term reliability at five months. *Child Development, 57,* 87–99.

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.), *Topics 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., & Cudeck, R. (in press). Single sample cross validation indices for covariance structures. *Multivariate Behavioral Research.*

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

Connell, J. P., & Tanaka, J. S. (Eds.). (1987). Structural equation models in developmental psychology [Special section]. *Child Development, 58,* 1–175.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18,* 147–167.

Farley, J. U., & Reddy, S. K. (1987). A factorial evaluation of effects of model specification and error on parameter estimation in a structural equation model. *Multivariate Behavioral Research, 22,* 71–90.

Geweke, 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.

Hogan, E. A., & Martell, D. A. (1987). A confirmatory structural equations analysis of the job characteristics model. *Organizational Behavior and Human Decision Processes, 39,* 242–263.

Huba, G. J., & Harlow, L. L. (1986). Robust estimation for causal models: A comparison of methods in some developmental datasets.

In P. B. Baltes, D. M. Featherman, & R. M. Lerner (Eds.), *Life-span developmental psychology* (Vol. 6, pp. 69–111). Hillsdale, NJ: Erlbaum.

Huba, G. J., & Harlow, L. L. (1987). Robust structural equation models: Implications for developmental psychology. *Child Development, 58,* 147–166.

International Mathematical and Statistical Libraries. (1980). *International mathematical and statistical libraries.* Houston, TX: Author.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data.* Beverly Hills, CA: Sage.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by the method of maximum likelihood.* Mooresville, IN: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1985). *LISREL VI: Analysis of linear structural relationships by the method of maximum likelihood.* Mooresville, IN: Scientific Software.

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

Katzell, R. A. (1982, August). *Measuring employee motivation: A systems approach.* Division 14 invited address at the 90th Annual Convention of the American Psychological Association, Washington, DC.

Katzell, R. A., & Thompson, D. E. (1987). *An integrative model of work attitudes and motivation.* Unpublished manuscript, Department of Psychology, New York University, New York.

La Du, T. J. (1986). *Assessing the goodness of fit of linear structural equation models: The influence of sample size, estimation method, and model specification.* Unpublished doctoral dissertation, Department of Psychology, New York University, New York.

Loehlin, J. C. (1987). *Latent variable models: An introduction to factor, path, and structural analysis.* Hillsdale, NJ: Erlbaum.

Luijben, T., Boomsma, A., & Molenaar, I. W. (1987). *Modification of factor analysis models in covariance structure analysis: A Monte Carlo study* (Tech. Rep. HB-87-851-EX). Groningen, The Netherlands: Heymans Bulletins Psychologische Instituten Rijksuniversiteit.

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

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–410.

Marsh, H. W., & Hocevar, D. (1985). Applications of confirmatory factor analysis to the study of self-concept: First- and higher order factor models and their invariance across groups. *Psychological Bulletin, 97,* 562–582.

McDonald, R. P. (1985). *Factor analysis and related methods.* Hillsdale, NJ: Erlbaum.

Mooijaart, A., & Satorra, A. (1987). *Robustness of ML and ADF methods in the analysis of covariance structures: Heteroscedastic models.* Paper presented at the meeting of the Psychometric Society, Montréal, Quebec, Canada.

Mowday, R. T. (1983). Equity theory predictions of behavior in organizations. In R. M. Steers & L. W. Porter (Eds.), *Motivation and work behavior* (pp. 91–112). New York: McGraw-Hill.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105,* 430–445.

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

Rindskopf, D. (1983). Parameterizing inequality constraints on unique variances in linear structural models. *Psychometrika, 48,* 73–83.

SAS. (1982). *SAS user's guide: Statistics.* Cary, NC: Author.

Satorra, A., & Bentler, P. M. (1987). *Adjustments to the chi-square goodness of fit statistic in covariance structure analysis: Comparative performance of a specific scaling correction.* Paper presented at the meeting of the Psychometric Society, Montréal, Quebec, Canada.

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

Sharma, S., Durvasula, S., & Dillon, W. R. (1989). Some results on the behavior of alternate covariance structure estimation procedures in the presence of non-normal data. *Journal of Marketing Research, 26,* 214–221.

Silvia, E. S. M., & MacCallum, R. C. (1988). Some factors affecting the success of specification searches in covariance structure modeling. *Multivariate Behavioral Research, 23,* 297–326.

Sobel, M. E., & Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. In N. B. Tuma (Ed.), *Sociological methodology 1985* (pp. 152–178). San Francisco: Jossey-Bass.

Tanaka, J. S. (1984). Some results on the estimation of covariance structure models (Doctoral dissertation, University of California, Los Angeles, 1984). *Dissertation Abstracts International, 45,* 924B.

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.

Tanaka, J. S., & Huba, G. J. (1985). A fit index for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology, 38,* 197–201.

Tanaka, J. S., & Huba, G. J. (1987). Assessing the stability of depression in college students. *Multivariate Behavioral Research, 22,* 5–19.

Tanaka, J. S., & Huba, G. J. (1988). *A general coefficient of determination for covariance structure models under arbitrary GLS estimation.* Manuscript submitted for publication.

Tanaka, J. S., Panter, A. T., Winborne, W. C., & Huba, G. J. (in press). Theory testing in personality and social psychology with latent variable models: A primer in 20 questions. In C. Hendrick & M. S. Clark (Eds.), *Review of personality and social psychology* (Vol. 11). Newbury Park, CA: Sage.

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

Williams, L. J., & Hazer, J. T. (1986). Antecedents and consequences of satisfaction and commitment in turnover models: A reanalysis using latent variable structural equation methods. *Journal of Applied Psychology, 71,* 219–231.

Received March 7, 1988
Revision received December 5, 1988
Accepted December 6, 1988 ■