Effect of Estimation Method on Incremental Fit Indexes for Covariance Structure Models

Hazuki M. Sugawara and Robert C. MacCallum

The Ohio State University

In a typical study involving covariance structure modeling, fit of a model or a set of alternative models is evaluated using several indicators of fit under one estimation method, usually maximum likelihood. This study examined the stability across estimation methods of incremental and nonincremental fit measures that use the information about the fit of the most restricted (null) model as a reference point in assessing the fit of a more substantive model to the data. A set of alternative models for a large empirical dataset was analyzed by asymptotically distribution-free, generalized least squares, maximum likelihood, and ordinary least squares estimation methods. Four incremental and four nonincremental fit indexes were compared. Incremental indexes were quite unstable across estimation methods—maximum likelihood and ordinary least squares solutions indicated better fit of a given model than asymptotically distribution-free and generalized least squares solutions. The cause of this phenomenon is explained and illustrated, and implications and recommendations for practice are discussed. *Index terms: covariance structure models, goodness of fit, incremental fit index, maximum likelihood estimation, parameter estimation, structural equation models.*

Covariance structure modeling (CSM; Bielby & Hauser, 1977; Bollen, 1988; Duncan, 1975; Goldberger & Duncan, 1973; Jöreskog, 1974, 1977) is a method of investigating theoretical relationships among a set of constructs or latent variables (LVs) and observable or measured variables (MVs) that serve as indicators of the LVs. Virtually all applications of CSM involve two primary objectives. The first is the estimation of the parameters of the model, where parameters may represent, for example, the linear effects of variables on other variables. The second is the assessment of the goodness of fit of the hypothesized model(s) to the observed data.

This study examined whether fit indexes perform differently depending on the type of estimation method used. Specific fit indexes may yield quite different values when a model is fit to a given dataset using different estimation methods. Tanaka (1987) observed this in a study focusing on the issue of sample size. La Du & Tanaka (1989) conducted a direct study of this issue and found that the goodness-of-fit index (GFI; Jöreskog & Sörbom, 1981) performed much more consistently across estimation methods than did the normed fit index (NFI; Bentler & Bonett, 1980). If this phenomenon generalizes beyond these specific indexes, it suggests that a researcher's evaluation of a model could depend on which fit index and estimation method were used. The present paper extends these earlier results in two ways.

First, the results of Tanaka (1987) and La Du & Tanaka (1989) are discussed. There are two distinct categories of fit indexes. In incremental indexes, the fit of a null or baseline model serves as a reference point for calculation of the index; nonincremental indexes do not use information regarding the fit of a null or baseline model. These categories have been described by Tanaka & Huba (1989), among others, but have not been associated in general terms with the phenomenon under study here.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 17, No. 4, December 1993, pp. 365-377

0146-6216/93/040365-13$1.90

Second, La Du & Tanaka's (1989) work is extended by providing an explanation and demonstration revealing how and why the incremental and nonincremental indexes behave differently across estimation methods.

Estimation Methods and Fit Indexes

Given a sample covariance matrix (S) and a hypothesized model, the parameters of the model can be estimated by a number of different procedures. The estimation procedures yield the covariance matrix $\hat{\Sigma} = \Sigma(\hat{\theta})$ implied by the model and written as a function of estimates of the parameters ($\hat{\theta}$) so that $\hat{\Sigma}$ will be as close to S as possible. In other words, the parameters are estimated so that the discrepancy between the implied covariance matrix $\Sigma(\hat{\theta})$ and S is minimal. This is achieved by minimizing some discrepancy function $F[\mathbf{S}, \Sigma(\hat{\theta})]$ that is a twice continuously differentiable, real valued nonnegative function of the positive definite (or positive semi-definite, for the generalized least squares estimation method discussed below) matrices $\Sigma(\hat{\theta})$ and S (see Bollen, 1989, chap. 4 for details). A general form of the discrepancy function is defined as

$$F = (\mathbf{s} - \boldsymbol{\sigma})' \mathbf{W}^{-1} (\mathbf{s} - \boldsymbol{\sigma}) \ . \tag{1}$$

Let p be the number of dependent MVs and q be the number of independent MVs. Then the terms s and $\boldsymbol{\sigma}$ are column vectors containing the $[(p + q)(p + q + 1)]/2$ distinct elements of the corresponding matrices S and $\Sigma(\theta)$ defined above (i.e., $\mathbf{s}' = [s_{11}, s_{21}, s_{22}, s_{31}, s_{32}, s_{33}, \ldots, s_{kk}]$ and $\boldsymbol{\sigma}' = [\sigma_{11}, \sigma_{21}, \sigma_{22}, \sigma_{31}, \sigma_{32}, \sigma_{33}, \ldots, \sigma_{kk}]$). Thus the entries in $(\mathbf{s} - \boldsymbol{\sigma})$ are residual variances and covariances representing the difference between observed sample values and values reconstructed from the model. W is a weight matrix, and serves to weight the residuals in $(\mathbf{s} - \boldsymbol{\sigma})$.

For any given weight matrix, model, and sample covariance matrix, the parameters of the model can be estimated so as to minimize the value of the discrepancy function. Obviously, the selection of different weight matrices results in the definition of different discrepancy functions. Different estimation methods are defined by the selection of the weight matrix in Equation 1. In this study, four estimation methods were considered—maximum likelihood (ML), generalized least squares (GLS), asymptotically distribution-free (ADF), and ordinary least squares (OLS).

Weight Matrices and Discrepancy Functions

Under ML, GLS, and ADF, elements of the weight matrix are defined to be consistent estimates of the asymptotic variances and covariances of the entries in S. That is, the typical element $[W_{\Sigma}]_{ij,kl}$ of W is defined as an estimate of the asymptotic variance or covariance of sample variance or covariances s_{ij} and s_{kl}. As will become clear later, the effect of estimation methods on fit is tied closely to the nature of the weight matrices used by the methods.

ML estimation. The ML method is by far the most widely used approach for fitting covariance structure models to sample data. For ML estimation, W in Equation 1 has the typical element

$$[\mathbf{W}_{\Sigma}]_{ij,kl} = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk} \ . \tag{2}$$

The discrepancy function is defined as

$$F_{\mathrm{ML}} = \log|\Sigma(\theta)| + \mathrm{tr}[\mathbf{S}\Sigma^{-1}(\theta)] - \log|\mathbf{S}| - (p + q) = (\mathbf{s} - \boldsymbol{\sigma})' \mathbf{W}_{\Sigma}^{-1} (\mathbf{s} - \boldsymbol{\sigma}) \ . \tag{3}$$

S approaches Σ asymptotically (Bollen, 1989). Because Σ is not known in practice, elements in $\mathbf{W}_{\Sigma}$ are computed from the model estimates of Σ, given by $\hat{\Sigma} = \Sigma(\hat{\theta})$. To maximize F_{ML}, θ estimation yields parameter estimates that maximize the joint likelihood of obtaining the observed data from the

population described by those parameter estimates under the assumption of multivariate normality.

GLS estimation. Under GLS estimation, the residuals in $(\mathbf{s} - \boldsymbol{\sigma})$ are weighted in accordance with their sample variances and covariances. The weight matrix in Equation 1 for this estimation method has the typical element of

$$[\mathbf{W}_s]_{ij,kl} = s_{ik}s_{jl} + s_{il}s_{jk} \ . \tag{4}$$

A general form of the GLS discrepancy function is

$$F_{\text{GLS}} = \frac{1}{2}\,\text{tr}\left\{\{[\mathbf{S} - \boldsymbol{\Sigma}(\theta)]\mathbf{S}^{-1}\}^2\right\} = (\mathbf{s} - \boldsymbol{\sigma})'\,\mathbf{W}_s^{-1}(\mathbf{s} - \boldsymbol{\sigma}) \ . \tag{5}$$

Note that the form of the weight matrices used in ML and GLS is similar—the typical elements of those weight matrices are

$$[\mathbf{W}_\Sigma]_{ij,kl} = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk} \tag{6}$$

for ML, and

$$[\mathbf{W}_s]_{ij,kl} = s_{ik}s_{jl} + s_{il}s_{jk} \tag{7}$$

for GLS. The difference is that under ML the entries in $\mathbf{W}_\Sigma$ are computed from elements of $\boldsymbol{\Sigma}(\hat{\theta})$, whereas under GLS the entries in $\mathbf{W}_s$ are computed from elements of $\mathbf{S}$. Thus, if the model fits well, meaning that $\mathbf{S} \approx \boldsymbol{\Sigma}(\hat{\theta})$, the weight matrices under these two methods will be similar.

ADF estimation. Browne (1982, 1984) extended the notion of GLS to the ADF method that yields optimal parameter estimates under less restricted assumptions than that of multivariate normality. For ADF, the elements of $\mathbf{W}_{\text{ADF}}$ are a complex function of the second-order (variances and covariances) and fourth-order moments (kurtosis and multivariate kurtosis) of the MVs. The discrepancy function to be minimized is

$$F_{\text{ADF}} = (\mathbf{s} - \boldsymbol{\sigma})'\,\mathbf{W}_{\text{ADF}}^{-1}(\mathbf{s} - \boldsymbol{\sigma}) \ . \tag{8}$$

The specific function defining the elements of $\mathbf{W}_{\text{ADF}}$ was given by Browne (1982, 1984). These elements are computed from entries in $\mathbf{S}$, and $\mathbf{W}_{\text{ADF}}$ can provide an estimate of the asymptotic covariance matrix for the elements of $\mathbf{S}$ (i.e., the same matrix estimated by $\mathbf{W}_\Sigma$ and $\mathbf{W}_s$), without any prior distributional assumptions.

OLS estimation. The OLS discrepancy function minimizes one-half the sum of squares of elements in the residual matrix. For purposes of comparing the OLS discrepancy function to the function defined for the other estimation methods, F_{OLS} is defined in terms of residual correlations as (Browne, 1969)

$$F_{\text{OLS}} = \frac{1}{2}\,\text{tr}\{\mathbf{D}_s^{-1}[\mathbf{S} - \boldsymbol{\Sigma}(\theta)]\}^2 \ , \tag{9}$$

where $\mathbf{D}_s$ is a diagonal matrix whose diagonal entries correspond to the sample standard deviations of the MVs. Under the usual definition of the OLS estimation method, $\mathbf{W}_{\text{OLS}}$ is defined as an identity matrix. However, this is rescaled to the standardized form so that discrepancy function values of F_{OLS} are comparable to other discrepancy function values, such as F_{ML} or F_{GLS}, which are invariant under changes in scale of the MVs. Equation 9 can be rewritten as

$$F_{\text{OLS}} = (\mathbf{s} - \boldsymbol{\sigma})'\,\mathbf{D}^{-1}(\mathbf{s} - \boldsymbol{\sigma}) \ , \tag{10}$$

where $\mathbf{D}$ is a diagonal weight matrix. For an element s_{jk} of $\mathbf{S}$, which becomes an entry in the vector

s, the corresponding entry in $\mathbf{D}$ is $s_{jj}s_{kk}$. Equations 3, 5, 8, and 10 represent the four discrepancy functions. They have the same general form as Equation 1 but involve different weight matrices. Differences among these weight matrices will be important in understanding the behavior of some fit indexes under different estimation methods.

Fit Indexes

One way of accomplishing the second primary objective of CSM—model evaluation—is to carry out a significance test examining the null hypothesis $\Sigma = \Sigma(\theta)$; that is, that the population covariance matrix for the MVs is exactly accounted for by the model. This null hypothesis can be tested by the likelihood ratio χ^2 test. Under the null hypothesis, $(N - 1)F$ (where N is sample size) approaches a χ^2 distribution as N becomes large, with degrees of freedom (df) equal to $[(p + q)(p + q + 1)/2] - t$, where t is the number of distinct parameters in θ. This test is problematic in practice because models that fit well will almost always be rejected when N is large. Such concerns have led to the development of more than 30 alternative measures of fit (Marsh, Balla, & McDonald, 1988).

Incremental fit indexes. Incremental fit measures use the information about the fit of a highly restricted model, often called a null model, as a reference in assessing the fit of a more substantive model to the data. Typically, a null or baseline model proposes that MVs are uncorrelated in the population (Bentler & Bonett, 1980). That is, the null model represents the hypothesis that Σ is a diagonal matrix. A conventional cutoff value of many incremental fit measures is .90, and larger values represent better fit (e.g., Bentler & Bonett, 1980). The incremental fit indexes considered here were the NFI (Δ_1), Bollen's delta index (Δ_2), the Tucker-Lewis Index (ρ_1), and Bollen's rho index (ρ_2).

Bentler & Bonett (1980) defined Δ_1 as

$$\Delta_1 = \text{NFI} = \frac{F_b - F_m}{F_b} = \frac{\chi_b^2 - \chi_m^2}{\chi_b^2}. \tag{11}$$

F_b and F_m are values of the sample discrepancy function of the baseline or null model and the maintained or target model, respectively. The values of Δ_1 range from 0 to 1. When the best possible fit is obtained (i.e., $F_m = 0$), $\Delta_1 = 1$. In order to take the df of the model into account and minimize the influence of the sample size on the mean of the index, Bollen (1988) modified Δ_1 by including in the denominator a term representing the expected value of F for a theoretically correct model; it is termed Δ_2 and defined as

$$\Delta_2 = \frac{F_b - F_m}{F_b - \dfrac{df_m}{N - 1}} = \frac{\chi_b^2 - \chi_m^2}{\chi_b^2 - df_m}. \tag{12}$$

ρ_1 and ρ_2 are non-normed incremental fit indexes that adjust for the df. Bentler & Bonett (1980) extended an index that was originally developed by Tucker & Lewis (1973) and defined ρ_1 as

$$\rho_1 = \text{TLI} = \frac{\dfrac{F_b}{df_b} - \dfrac{F_m}{df_m}}{\dfrac{F_b}{df_b} - \dfrac{1}{N-1}} = \frac{\dfrac{\chi_b^2}{df_b} - \dfrac{\chi_m^2}{df_m}}{\dfrac{\chi_b^2}{df_b} - 1}. \tag{13}$$

A theoretically correct model has $E(\chi_m^2/df_m) = 1$ when the underlying assumptions for the χ^2

approximation are met for the maintained model. Therefore, the value of ρ_1 representing the expected fit in the sample for a model that is correct in the population is 1 although it can be larger than 1 for an overfitting model.

ρ_2 (Bollen, 1986) is defined as

$$\rho_2 = \frac{\frac{F_b}{df_b} - \frac{F_m}{df_m}}{\frac{F_b}{df_b}} = \frac{\frac{\chi_b^2}{df_b} - \frac{\chi_m^2}{df_m}}{\frac{\chi_b^2}{df_b}} . \tag{14}$$

ρ_2 compares the discrepancy per *df* for the most restricted model relative to the target model. The maximum value of ρ_2 is 1. Although a value less than 0 is rare, ρ_2 does not have a lower bound.

Nonincremental fit indexes. Nonincremental fit indexes do not incorporate numerical information about the fit of the baseline model in their calculation. In the present study, four nonincremental fit indexes were used.

The minimal population discrepancy function (F_0) and minimal sample discrepancy function (F_s) are bounded by 0 and take on the value of 0 if and only if $\Sigma = \hat{\Sigma}$ or $\mathbf{S} = \hat{\Sigma}$, respectively. Although F_0 is defined as

$$F_0 = \text{Min } F(\Sigma, \hat{\Sigma}) . \tag{15}$$

$\hat{F}_0$ is used to estimate F_0 because Σ cannot be obtained in practice (Browne & Mels, 1990):

$$\hat{F}_0 = \text{Max}\left\{F_s - \frac{df}{N-1}, 0\right\} , \tag{16}$$

where

$$F_s = \text{Min } F(\mathbf{S}, \hat{\Sigma}) . \tag{17}$$

Jöreskog & Sörbom (1981) devised a measure of fit called the GFI:

$$\text{GFI} = 1 - \frac{(\mathbf{s} - \hat{\sigma})' \mathbf{W}^{-1} (\mathbf{s} - \hat{\sigma})}{\mathbf{s}' \mathbf{W}^{-1} \mathbf{s}} . \tag{18}$$

The numerator of the GFI can be recognized as $F(\hat{\theta})$, or the minimum value of the discrepancy function, and the denominator is the discrepancy function evaluated without fitting any model. The upper boundary of GFI is 1 and negative values can occur theoretically, although that is unlikely. In this study, GFI was computed using equations developed by Maiti & Mukherjee (1990).

Steiger & Lind (1980) developed the root mean square error of approximation (RMSEA). It is defined as

$$\text{RMSEA} = \left(\frac{F_0}{df}\right)^{1/2} , \tag{19}$$

where F_0 is the minimal population discrepancy function value, which is replaced with $\hat{F}_0$ in practice. For a hierarchy of nested models fit to a given sample, the values of RMSEA across the various models would be monotonically related to the χ^2/df ratio. Values below .10 represent a reasonable fit, and values below .05 represent a very good fit (Steiger, 1989). Values below .01, representing an outstanding fit, are rarely attained.

Calculating Fit Indexes Under Different Estimation Methods

All eight fit indexes (the four incremental indexes—Δ_1, Δ_2, ρ_1, ρ_2—and the four nonincremental indexes—$\hat{F}_0$, F_s, GFI, and RMSEA) defined above may be computed from discrepancy function values obtained under each of the four estimation methods (ADF, GLS, ML, OLS). None of the indexes is defined in such a way as to be meaningful only under certain estimation methods. Each is defined only in terms of discrepancy function values that are sensitive to the difference between $\mathbf{S}$ and $\boldsymbol{\Sigma}(\hat{\theta})$. Under specified distributional conditions and estimation methods, it is possible to define distributional properties of some fit indexes (Browne & Cudeck, 1992). However, such conditions are not necessary for the calculation and interpretation of these fit measures under any estimation method.

Issues Investigated

Fit measures may behave differently depending on the type of estimation method used. For instance, Tanaka (1987) noticed that GFI values were larger under ML than GLS in his study of sample size and goodness of fit. In addition, La Du & Tanaka (1989) found that NFI or Δ_1 values differed substantially under GLS and ML, whereas GLS-based and ML-based GFI values were similar.

Such observations represent specific instances of a much more general phenomenon involving the fundamental difference between incremental and nonincremental fit indexes. To begin to develop an explanation for such a phenomenon, consider some observations involving the discrepancy functions defined earlier.

The weight matrix for ML, $\mathbf{W}_{\Sigma}^{-1}$, is computed from the reproduced covariance matrix, $\boldsymbol{\Sigma}(\hat{\theta})$. The weight matrices for ADF and GLS are computed from $\mathbf{S}$. All three weight matrices are estimates of the same unknown matrix; that is, the matrix of asymptotic variances and covariances of the elements of $\mathbf{S}$.

Consider the ideal situation in which a model fits the data well [i.e., $\mathbf{S} \approx \boldsymbol{\Sigma}(\hat{\theta})$], the sample size is large, and the assumption of multivariate normality is approximately satisfied. Under these conditions, the residuals in the vector $(\mathbf{s} - \boldsymbol{\sigma})$ would be small and the weight matrices defined for the ML, GLS, and ADF discrepancy functions would be very similar. As a result, the values of the discrepancy functions produced by these three estimation methods would be quite similar. However, for the same model and data the value of the OLS discrepancy function would most likely be quite different because the OLS weight matrix (Equation 10) is diagonal. This weight matrix would be similar to the weight matrices for the other discrepancy functions only when the MVs were approximately mutually uncorrelated, which would be highly unusual.

It is important also to consider the behavior of these discrepancy functions under the null model. Clearly, in most empirical studies the null model would fit very badly [i.e., $\mathbf{S} \neq \boldsymbol{\Sigma}(\hat{\theta})$], resulting in large residuals in the vector $(\mathbf{s} - \boldsymbol{\sigma})$. However, if the assumption of multivariate normality is approximately satisfied and if sample size is large, the weight matrices for the ADF and GLS discrepancy functions would still be quite similar, because both of these weight matrices would be estimates of the asymptotic covariance matrix of the elements of $\mathbf{S}$ and both are computed from the entries in $\mathbf{S}$ under distributional conditions sufficient for both methods. Thus, under these conditions, obtained values of F_{ADF} and F_{GLS} would be very similar for the null model, even though that model fits very poorly.

On the other hand, if sample size were small and/or the normality assumption were substantially violated, these two discrepancy function values would be less similar because the weight matrices for ADF and GLS would be different. Under these same conditions, $\mathbf{W}_{\Sigma}^{-1}$ is computed from the elements of $\boldsymbol{\Sigma}(\hat{\theta})$, which is very different from $\mathbf{S}$ under the null model, and the weight matrix for OLS is diagonal; thus, the values of F_{ML} and F_{OLS} would likely be quite different from each other and

from the values of the other discrepancy functions for the null model.

These observations are important because they have implications for the behavior of incremental versus nonincremental fit measures. Because incremental fit indexes are functions of the fit of the null model, and because the discrepancy function values for the null model may vary substantially across estimation methods, it is to be expected that the values of incremental fit indexes may vary similarly even for models that fit well. On the other hand, nonincremental fit measures should be less prone to such instability because their values are not sensitive to the fit of the null model. Thus, if these phenomena hold in practice, a given model may be evaluated quite differently when different fit measures and methods of estimation are used. This rationale would provide further evidence and explanation for the results observed in the study by La Du & Tanaka (1989).

The practical importance of this investigation stems from its implication that if indicators of fit behave differently depending on the type of discrepancy function minimized, then the conclusion of model evaluation based on the same set of fit indexes would be dependent on the estimation method used. This issue was investigated by analyzing a large empirical dataset with four different estimation methods (ADF, ML, GLS, and OLS) and comparing four incremental fit measures (Δ_1, Δ_2, ρ_1, ρ_2) and four nonincremental fit measures ($\hat{F}_0$, F_s, GFI, RMSEA) for a set of alternative models.

Method

Dataset

Although three large empirical datasets (MacCallum, Roznowski, & Necowitz, 1992; Mels & Knoorts, 1989; Verhoef & Roos, 1970) were examined, results for only one dataset (Dataset 1) will be discussed here in detail because of the consistency of the outcomes across the datasets (Sugawara, 1992). Part of the data from a project by the Human Sciences Research Council (Verhoef & Roos, 1970) analyzed by Cudeck & Browne (1983) was used. A battery of six ability tests (Elder, 1957) was administered to 2,677 high school students (approximately 14, 16, and 18 years of age) in 1965, 1967, and 1969. The six ability tests were Number Series, Pattern Completion, Classification of Word Pairs, Verbal Reasoning, Figure Analogies, and Word Analogies.

Procedure

A set of alternative models was constructed. They consisted of (1) a null model in which all MVs were uncorrelated (Null); (2) a model with a single general factor (G); (3) an orthogonal three-factor model with three factors representing the three uncorrelated occasions (OR3); (4) a six-factor orthogonal factor model with six factors representing six uncorrelated abilities (OR6); (5) and a six-factor oblique factor model with six factors representing six correlated abilities (OB6). The models first were fit to the data by the ADF, GLS, ML, and OLS estimation methods. Then the four incremental fit indexes and four nonincremental fit indexes were computed to evaluate the fit of the models.

Computer Program

The RAMONA computer program (Browne & Mels, 1990) was used for the data analysis. This program is based on the reticular action model (McArdle & McDonald, 1984) with a slight modification. When ADF, GLS, or ML is used, RAMONA provides the values of χ^2, $\hat{F}_0$, F_s, and RMSEA. Only F_s is provided when OLS is used. The desired incremental indexes and GFI (or all indexes except F_s in the case of OLS) can be obtained from the information provided by RAMONA by hand calculation.

Results

Table 1 includes values of the eight fit indexes for the five models under the four estimation methods.

Table 1
Values of Incremental and Nonincremental Fit Indexes for Five Models Using Four Estimation Methods for Dataset 1

Fit Index and Estimation Method	Model				
	Null	OR6	G	OR3	OB6
Incremental					
Δ_1 Fit Index					
ADF	—	.38	.38	.09	.78
GLS	—	.34	.33	.14	.71
ML	—	.57	.79	.56	.93
OLS	—	.96	.96	.31	.99
Δ_2 Fit Index					
ADF	—	.40	.39	.09	.81
GLS	—	.36	.35	.15	.74
ML	—	.57	.80	.56	.93
OLS	—	.96	.96	.31	.99
ρ_1 Fit Index					
ADF	—	.31	.30	-.03	.76
GLS	—	.27	.26	.03	.66
ML	—	.51	.77	.50	.92
OLS	—	.96	.96	.22	.99
ρ_2 Fit Index					
ADF	—	.30	.29	-.03	.72
GLS	—	.26	.24	.03	.63
ML	—	.51	.77	.50	.91
OLS	—	.96	.96	.22	.99
Nonincremental					
$\hat{F}_0$ Fit Index					
ADF	1.26	.76	.77	1.15	.24
GLS	1.25	.81	.82	1.07	.33
ML	7.44	3.20	1.50	3.27	.49
OLS	21.49	.81	.76	17.8	.14
F_c Fit Index					
ADF	1.32	.82	.82	1.20	.28
GLS	1.31	.86	.87	1.12	.38
ML	7.49	3.25	1.54	3.32	.54
OLS	21.55	.86	.81	14.85	.19
GFI Fit Index					
ADF	.88	.92	.92	.89	1.00
GLS	.88	.92	.92	.89	1.00
ML	.55	.74	.86	.73	1.00
OLS	.30	.92	.93	.33	.99
RMSEA Fit Index					
ADF	.09	.08	.08	.09	.04
GLS	.09	.08	.08	.09	.05
ML	.22	.15	.10	.16	.06
OLS	.37	.08	.08	.36	.03

Incremental Fit Indexes

For a given model, values of the incremental fit indexes differed substantially depending on the estimation method used. When ADF, GLS, and ML are compared, the trend observed was systematic.

For OB6, the Δ and ρ fit indexes based on ML solutions had values higher than a conventional cutoff of .90, showing "good" fit; however, the values of ADF and GLS did not exceed the cutoff value. The highest values obtained were $\Delta_2 = .81$ and $\Delta_2 = .74$ for OB6 under ADF and GLS, respectively. Moreover, all incremental fit index values under the ADF and GLS estimation methods were substantially smaller than corresponding ML-based values.

Across estimation methods, the incremental fit measures suggested that OR3 had the worst fit and OB6 had the best fit. Although model selection in practice would also take into account factors such as parsimony or interpretability of solution, here the purpose was to evaluate similarities and differences in behavior of these fit measures specifically with respect to their relationships with the different estimation methods. The erratic behavior of the incremental fit indexes was particularly apparent under the OLS estimation method. For all models except OR3, the OLS-based incremental fit measures had values higher than .96, indicating very good fit. This was inconsistent with the results based on the other three estimation methods for which only the ML-based index values for OB6 exceeded .90. The incremental fit measures yielded values indicating poor fit for OR3 which was evaluated to have the worst fit among the four target models. For OR3, the ML-based values were higher than the OLS-based values; however, the trend was reversed for the other three models.

Nonincremental Fit Indexes

Results for the nonincremental indexes in Table 1 provide further information, as well as an explanation, for the behavior of the incremental indexes. Considering the results for the values of the sample discrepancy function, F_c, for the null model, ADF and GLS produced very similar values, whereas ML and OLS produced much larger values. The similarity of the ADF and GLS values was consistent with expectations and arose because of the large sample and the approximate normality of the variables. Under these conditions, the weight matrices for ADF and GLS became similar, resulting in similar discrepancy function values. The discrepant values produced by ML and OLS for the null model also were expected because the weight matrices defined by these functions were different from those defined by ADF and GLS, even when sample size was large and normality held.

These varied values of the discrepancy function for the null model accounted for the erratic behavior of the incremental fit indexes. When ML or OLS was used, the fit of the null model was indicated by much higher values of the respective discrepancy function, as compared to results obtained from the ADF or GLS method. The numerators of Equations 11–14 defining the incremental fit indexes compared function values for the null model and a substantive model. The wide variation in function values for the null model across estimation methods, as seen in Table 1, will naturally give rise to similar variation in values of incremental fit indexes, even for a model that fits the data quite well. More specifically, improvement in fit expressed by an incremental fit measure tends to be intensified under ML and OLS estimation due to the increased size of the null model discrepancy function values under these methods as compared to ADF and GLS. Thus, the incremental fit index values based on ML and OLS estimations were larger than corresponding values based on ADF or GLS.

Results in Table 1 also show more stable behavior for two of the nonincremental fit indexes. The GFI and RMSEA indexes were especially consistent for a good-fitting model, with consistency decreasing as model quality deteriorated. For OB6, these measures showed the model to be quite good under all estimation methods.

Regardless of the estimation method or fit index, the model selection process yielded a consistent result. That is, OB6 was the preferred model, and OR3 had the worst fit compared to the other models, excluding the null model. Overall however, the order of the models with respect to their fit was

consistent across the estimation methods although the size of the index values varied substantially.

Additional Datasets

In order to examine the effect of estimation methods further, two other large empirical datasets were examined (see Sugawara, 1992, for a detailed description of the analysis). Dataset 2 consisted of responses to self-report measures from 3,148 hospital employees (MacCallum et al., 1992). The models examined for this study included six LVs and 15 MVs, whereas the original model evaluated by MacCallum et al. (1992) included seven LVs and 21 MVs. Dataset 3 consisted of 213 nurses' responses to a job satisfaction questionnaire (Browne & Mels, 1990).

Table 2 summarizes F_c values for the null model under ADF, GLS, ML, and OLS for the three datasets. The deviation from normality is reflected by the degree to which the relative kurtosis value deviated from 1. Results in Table 2 provide further evidence for the phenomena described above. Specifically, for data in which N is large and normality is approximately satisfied (Datasets 1 and 2) the ADF and GLS methods produced similar values of the discrepancy function for the null model, whereas the ML and OLS methods produced substantially different values. For these datasets, incremental fit indexes behaved erratically, even for good-fitting models, across estimation methods. As a result of the smaller sample in the third dataset, some divergence was seen between the ADF and GLS discrepancy function values for the null model, indicating that values of incremental fit measures produced for good-fitting models under these methods would be moderately different.

Table 2
Sample Discrepancy Function Values for the Null Model Under ADF, GLS, ML, and OLS From Three Datasets

Estimation Method, N, and Kurtosis	Dataset		
	1	2	3
ADF	1.32	1.78	1.84
GLS	1.31	1.67	1.08
ML	7.49	5.83	3.07
OLS	21.55	7.64	3.95
N	2,677	3,148	213
Kurtosis	1.12	1.08	1.16

Discussion

The most important finding of this study is that for a given model, values of incremental fit indexes varied substantially across estimation methods. Considering the results presented in Table 1, this phenomenon held across all models from poor to good fitting models. The incremental fit index values based on ML estimates were always larger than corresponding GLS-based or ADF-based incremental index values. The OLS-solution based values were the largest, except for OR3 for which the ML-solution based values were the largest. Index values above the cutoff of .90 were observed only for OB6 under ML and for all models except OR3 under OLS.

On the other hand, variation in values of nonincremental indexes across the estimation methods was more dependent on the quality of the model. For instance, the values of GFI and RMSEA for OB6 were consistent under all estimation methods. As the quality of the model became poorer, nonincremental indexes varied more in comparing different estimation methods. This tendency was particularly obvious for the null model under the ML and OLS methods, which consistently yielded less

desirable values for $\hat{F}_0$, F_t, and GFI as compared to the ADF or the GLS methods. The index values for the OLS solution based on the null model were the largest. It follows that the same model is suggested to have a better fit by incremental fit indexes using a ML or OLS solution because of the degree of the fit of the null model. In other words, because the null model had a worse fit when ML or OLS was used, even minor improvement in the fit of a non-null model is magnified although the degree of fit of the non-null models is expressed as a consistent, if not identical, index value. ADF and GLS estimation methods yielded similar values when the sample size was large and the data were normally distributed.

The second observed pattern is related to the issue of model selection. Based on ADF, GLS, or ML, OB6 was recommended regardless of the type of fit measure or estimation method used. However, this was not the case when the incremental indexes under OLS were used.

These findings confirm and extend those reported by Tanaka (1987) and La Du & Tanaka (1989). Nonincremental fit indexes tend to behave much more consistently across estimation methods than do incremental fit indexes, especially for good models. Furthermore, an explanation for why this phenomenon occurs was provided, showing that it arises from the fact that different estimation methods yield very different discrepancy function values for a null model, due to the differences in the definition of the weight matrix used in the various discrepancy functions. As a result, incremental fit measures, which use the discrepancy function value for the null model in their calculation, tend to behave erratically across estimation methods, even for a model that is quite consistent with the observed data.

Thus, a researcher's choice of estimation method and fit indexes may substantially influence the evaluation of each model's fit to the observed data. That is, depending on the fit measures and the type of estimation method used, the conclusion regarding the degree of model fit tends to vary. For example, the researcher should be aware that a model's fit may be enhanced when the judgment is based on incremental indexes under the ML or OLS estimation method.

It has been recognized in the literature that the behavior of some fit measures may be influenced by other factors, including sample size (e.g., Bollen, 1990; Marsh et al., 1988) and violations of distributional assumptions (e.g., Boomsma, 1983). Although a strong and systematic effect of estimation methods on the behavior of incremental fit indexes was demonstrated here, it was not systematically investigated how this effect might vary as a function of various characteristics of data. However, this does not represent a major limitation of the study. Results in Table 2 clearly indicate that the critical phenomenon (i.e., the variability in discrepancy function values observed for the null model) can occur in both large and small samples. Furthermore, the explanation for the cause of this phenomenon, in terms of the behavior of weight matrices used in the various discrepancy functions, should be valid even when distributional assumptions are violated. These arguments could be tested using simulation.

These findings are important if the typical selection of estimation methods and fit measures is considered. ML estimation is the default method in CSM computer programs and is used routinely in empirical applications. To illustrate the selection of fit measures, 42 applications of CSM published in the *Journal of Applied Psychology* from February 1986 to August 1991 used the following fit indexes with the indicated frequencies: root mean squared residual (28), GFI (26), NFI (20), ρ_1 (14), adjusted GFI (14), parsimonious fit index (10), χ^2/df ratio (5), relative NFI (2), ECVI (1), parsimonious GFI (1), relative fit index 2 (1), relative parsimonious ratio (1), and ρ_1 (1). Researchers apparently tend to use indexes provided by LISREL (e.g., adjusted GFI, χ^2, GFI, root mean squared residual) and to calculate additional incremental indexes (e.g., NFI, ρ_1). The present results clearly indicate that the use of these latter indexes may be problematic due to their lack of stability across estimation methods.

The use of nonincremental fit measures instead of incremental fit measures is recommended when the ML estimation method is the only method used. Based on these findings, RMSEA would be one of the most appropriate practical choices among the nonincremental fit measures because it behaves consistently across estimation methods for good models and also has an interpretable scale associated with it for determining the degree of fit.

References

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

Bielby, W. T., & Hauser, R. M. (1977). Structural equation models. *Annual Review of Sociology, 3,* 137–163.

Bollen, K. A. (1986). Sample size and Bentler and Bonett's non-normed fit index. *Psychometrika, 51,* 375–377.

Bollen, K. A. (1988). A new incremental fit index for general structural equation models. *Sociological Methods and Research, 17,* 303–316.

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

Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin, 107,* 256–259.

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and nonnormality.* Amsterdam, The Netherlands: Sociometric Research Foundation.

Browne, M. W. (1969). Fitting the factor analysis model. *Psychometrika, 34,* 375–394.

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. (1992). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills CA: Sage Publications.

Browne, M. W., & Mels, G. (1990). *RAMONA PC: User's guide* [Computer program manual]. Department of Statistics, University of South Africa.

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

Duncan, O. D. (1975). *Introduction to structural equation models.* New York: Academic Press.

Elder, C. M. (1957). The statistical procedure adopted in the construction of the New South African Group Test. *Journal for Social Research, 8,* 1–12.

Goldberger, A. S., & Duncan, O. D. (1973). *Structural equation models in the social sciences.* New York: Seminar Press.

Jöreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In R. C. Atkinson, D. H. Krantz, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. II) (pp. 1–56). San Francisco: Freeman.

Jöreskog, K. G. (1977). Structural equation models in the social sciences: Specification, estimation, and testing. In P. R. Krishnaiah (Ed.), *Applications of statistics* (pp. 265–287). Amsterdam, The Netherlands: North-Holland.

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

La Du, T. J., & Tanaka, J. S. (1989). Influence of sample size, estimation method, and model specification on goodness-of-fit assessments in structural equation models. *Journal of Applied Psychology, 74,* 625–635.

MacCallum, R., Roznowski, M., & Necowitz, L. B. (1992). Model modification in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111,* 490–504.

Maiti, S. S., & Mukherjee, B. N. (1990). A note on distribution properties of the Jöreskog-Sörbom fit indexes. *Psychometrika, 55,* 721–726.

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.

McArdle, J., & McDonald, R. P. (1984). Some algebraic properties of the reticular action model for moment structures. *British Journal of Mathematical and Statistical Psychology, 37,* 234–254.

Mels, G., & Knoorts, A. S. (1989). Causal models for various job aspects. *SAIPA Journal of Public Administration, 24,* 144–156.

Steiger, J. H. (1989). *EzPATH: A supplementary module for SYSTAT and SYSGRAPH* [Computer program]. Evanston IL: SYSTAT, Inc.

Steiger, J. H., & Lind, J. (1980, May). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric

Society, Iowa City IA.
Sugawara, H. M. (1992). *An empirical study of fit indices for covariance structure models.* Unpublished master's thesis, The Ohio State University, Columbus OH.
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. (1989). A general coefficient of determination for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology, 42,* 233–239.
Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.
Verhoef, W., & Roos, W. L. (1970). *The aim and experimental design of project talent survey* (Report No. MT 1). Pretoria, South Africa: Human Sciences Research Council.

Acknowledgments

The authors thank Michael W. Browne for his helpful suggestions and insights provided throughout this project. The authors also express appreciation to Jeffrey Tanaka, who served as a reviewer and whose comments were most helpful in our efforts to improve this paper. This paper is dedicated to his memory.

Author's Address

Send requests for reprints or further information to Hazuki Sugawara, Nihon Keizai Shimbun, Inc., 6-6-5 Ginza (9F) Chuo-ku, Tokyo 104, Japan.