STRUCTURAL EQUATION MODELING, *9*(2), 233–255

Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance

Gordon W. Cheung
Department of Management
The Chinese University of Hong Kong

Roger B. Rensvold
Department of Management
City University of Hong Kong

Measurement invariance is usually tested using Multigroup Confirmatory Factor Analysis, which examines the change in the goodness-of-fit index (GFI) when cross-group constraints are imposed on a measurement model. Although many studies have examined the properties of GFI as indicators of overall model fit for single-group data, there have been none to date that examine how GFIs change when between-group constraints are added to a measurement model. The lack of a consensus about what constitutes significant GFI differences places limits on measurement invariance testing.

We examine 20 GFIs based on the minimum fit function. A simulation under the two-group situation was used to examine changes in the GFIs (ΔGFIs) when invariance constraints were added. Based on the results, we recommend using Δcomparative fit index, ΔGamma hat, and ΔMcDonald's Noncentrality Index to evaluate measurement invariance. These three ΔGFIs are independent of both model complexity and sample size, and are not correlated with the overall fit measures. We propose critical values of these ΔGFIs that indicate measurement invariance.

Social science researchers are increasingly concerned with testing for measurement invariance; that is, determining if items used in survey-type instruments mean the same things to members of different groups. Measurement invariance is critically important when comparing groups. If measurement invariance cannot be established, then the finding of a between-group difference cannot be unambigu-

Requests for reprints should be sent to Gordon W. Cheung, Department of Management, The Chinese University of Hong Kong, Shatin, Hong Kong. E-mail: GORDONC@cuhk.edu.hk

ously interpreted. One does not know if it is due to a true attitudinal difference, or to different psychometric responses to the scale items. This is of particular concern in cross-cultural research when the cultures speak different languages, and researchers use translated versions of a survey instrument (Janssens, Brett, & Smith, 1995; Reise, Widaman, & Pugh, 1993; Riordan & Vandenberg, 1994; Steenkamp & Baumgartner, 1998). Other examples include groups having different levels of academic achievement (Byrne, Shavelson, & Muthén, 1989), working in different industries (Drasgow & Kanfer, 1985), of different genders (Byrne, 1994), and in experimental versus control groups (Pentz & Chou, 1994).

Measurement invariance is a general term that can be applied to various components of measurement models. Little (1997) identified two types of invariance. Category 1 invariance has to do with the psychometric properties of the measurement scales, and includes configural invariance (e.g., Buss & Royce, 1975; Irvine, 1969; Suzuki & Rancer, 1994), metric invariance (Horn & McArdle, 1992; also called weak factorial invariance by Meredith, 1993), measurement error invariance (Mullen, 1995; Singh, 1995), and scalar invariance (Meredith, 1993; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). Category 2 invariance has to do with between-group differences in latent means, variances, and covariances. Generally speaking, Category 1 invariance is a prerequisite for the interpretation of Category 2 differences, whereas Category 2 differences are usually the data having substantive research interest.

Structural equation modeling (SEM) is widely used in the social sciences. The suitability of a single-group measurement model is usually assessed using an SEM procedure known as confirmatory factor analysis (CFA). A model is considered suitable if the covariance structure implied by the model is similar to the covariance structure of the sample data, as indicated by an acceptable value of goodness-of-fit index (GFI).

The most commonly used GFI of SEM is the χ^2 statistic, defined as

$$\chi^2 .: (N \Lambda 1)\hat{F}_{\min} \tag{1}$$

where N is the sample size, and $\hat{F}_{\min}$ is the minimum value of the empirical fit function, estimated using an iterative procedure under the assumption that the data have multivariate normal distribution. A nonsignificant value of χ^2 indicates failure to reject the null hypothesis that the hypothesized covariance matrix is identical to the observed covariance matrix, which is usually accepted as evidence of adequate fit. A problem arises because of the statistic's functional dependence on N. For large sample sizes, the χ^2 statistic provides a highly sensitive statistical test, but not a practical test, of model fit. Owing to this and other considerations, many GFIs have been proposed as alternatives to χ^2. Some in common use include the comparative fit index (CFI; Bentler, 1990), Tucker–Lewis Index (TLI; Tucker & Lewis, 1973), Normed Fit Index (NNFI; Bentler & Bonett, 1980), and root mean squared

error of approximation (RMSEA; Steiger, 1989). Due to the fact that most of the "practical" GFIs do not have known sampling distributions, researchers have proposed many criterion values indicative of satisfactory model fit; examples include .90 or above for TLI and CFI. It is common practice to use multiple GFIs when evaluating and reporting overall model fit.

An extension of CFA, Multigroup Confirmatory Factor Analysis (MGCFA), tests the invariance of estimated parameters of two nested models across groups. The degree of invariance is most frequently assessed by the Likelihood Ratio Test (differences in χ^2 between two nested models), although researchers have demonstrated that differences in χ^2 are also dependent on sample size (Brannick, 1995; Kelloway, 1995). Reliance on the Likelihood Ratio Test is probably due to the lack of sampling distributions for GFI differences. In contrast to the CFA test for overall fit, there are no generally accepted criteria in MGCFA for determining if changes in the "practical" GFIs are meaningful when measurement invariance constraints are added. For example, there is no standard against which a researcher can compare changes in CFI when measurement invariance constraints are added, in order to determine if the constrained model fits the data less well than the less-constrained model (Vandenberg & Lance, 2000). Hence, the objective of this article is to assess the effects of sampling error and model characteristics on MGCFA outcomes; that is, differences in GFIs (ΔGFIs) obtained when an unconstrained model is compared with one having measurement invariance constraints, under the null hypothesis of invariance. We propose critical values of ΔGFIs that are independent from model characteristics, basing our proposals on sampling distributions of ΔGFIs obtained using simulations.

MEASUREMENT INVARIANCE

The series of tests that constitutes invariance testing through MGCFA are covered at length elsewhere (Bollen, 1989b; Byrne et al., 1989; Cheung & Rensvold, 2000; Drasgow & Kanfer, 1985; Jöreskog & Sörbom, 1993; Little, 1997; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). For the reader's convenience we briefly review them. There are eight invariance hypotheses that are frequently examined, as summarized in Table 1. The sequence of the invariance tests in Table 1 is only one of the many possible sequences that have been proposed, based on the substantive research questions at hand. The five hypotheses at the top of the table relate to measurement level invariance, whereas the three in the lower portion relate to construct level invariance (Little, 1997; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000).

Hypothesis H_{form} postulates configural invariance; that is, participants belonging to different groups conceptualize the constructs in the same way (Riordan & Vandenberg, 1994). If configural invariance exists, then data collected from each

TABLE 1
Hypotheses of Measurement Invariance

Model	*Hypothesis*	*Hypothesis Test*[a]	*Hypothesis Name*	*Symbolic Statement*[b]	*Conceptual Meanings of Hypotheses*
1	H_{form}	Overall fit	Configural invariance	$\Lambda^{(1)}_{form} \therefore \Lambda^{(2)}_{form}$	Both groups associate the same subsets of items with the same constructs (the cognitive domains are the same).
2	H_{Λ}	2 – 1	Construct-level metric invariance	$\Lambda^{(1)} \therefore \Lambda^{(2)}$	Overall, the strength of the relationships between items and their underlying constructs are the same for both groups. (The constructs are manifested in the same way across groups.)
3	H_{λ}	3 – 1	Item-level metric invariance	$\lambda^{(1)}_{ij} \therefore \Lambda^{(2)}_{ij}$	The strength of the relationship between *each* item and its underlying construct is the same for both groups.
4	$H_{\Lambda,\Theta(\delta)}$	4 – 2	Residual variance invariance	$\Theta^{(1)}_{\delta} \therefore \Theta^{(2)}_{\delta}$	Items have the same internal consistency for both groups. Alternatively: For both groups, items have the same quality as measures of the underlying construct.
5	$H_{\Lambda,\nu}$	5 – 2	Intercept invariance	$\tau^{(1)}_{i} \therefore \tau^{(2)}_{i}$	The cross-cultural differences latent means indicated by the items are the same across *items*. Alternatively: All items indicate the same cross-cultural differences.
6	$H_{\Lambda,\Phi(jj)}$	6 – 2	Equivalence of construct variance	$\Phi^{(1)}_{jj} \therefore \Phi^{(2)}_{jj}$	The range of responses given to each item is the same across groups. Alternatively: the variability/range of diversity with respect to the constructs are the same across groups.
7	$H_{\Lambda,\Phi(jj')}$	7 – 2	Equivalence of construct covariance	$\Phi^{(1)}_{jj'} \therefore \Phi^{(2)}_{jj'}$	The relationships among constructs (e.g., covariances) are the same across groups.
8	$H_{\Lambda,\nu,\kappa}$	8 – 5	Equivalence of latent means	$\kappa^{(1)}_{j} \therefore \kappa^{(2)}_{j}$	The mean level of each construct is the same across groups.

[a]In each case, the statistic for testing the hypothesis is the difference between the fit of the constrained model and that of a less constrained model. For example, the test statistic for H_{Λ} is the fit difference between Model 2 (with construct-level constraints) and Model 1 (with no constraints). This is indicated in the table as "2 – 1."

[b]Parenthetical superscripts indicate groups. For brevity, only the two-group case is shown, but each hypothesis generalizes to K groups. The general statement of Hypothesis 1, for example, is $\Lambda^{(1)}_{form} \therefore \Lambda^{(2)}_{form} \therefore \Lambda^{(3)}_{form} \therefore \text{L} \therefore \Lambda^{(k)}_{form}$.

group decompose into the same number of factors, with the same items associated with each factor (Meredith, 1993). Configural invariance may fail when, for example, the concepts are so abstract such that participants' perceptions of the constructs depend on their cultural context (Tayeb, 1994), or when participants from different groups use different conceptual frames of reference and attach different meanings to constructs (Millsap & Everson, 1991; Millsap & Hartog, 1988; Riordan & Vandenberg, 1994). Configural invariance may also fail owing to a host of other reasons, including data collection problems, translation errors, and so forth.

Hypothesis H_{Λ} postulates metric invariance (i.e., that all factor loading parameters are equal across groups). Samples drawn from two populations may provide data that indicate conceptual agreement in terms of the type and number of underlying constructs, and the items associated with each construct. Despite this, the strengths of the relations between specific scale items and the underlying constructs may differ. The data may indicate disagreement concerning how the constructs are manifested. Metric invariance is important as a prerequisite for meaningful cross-group comparison (Bollen, 1989b).

Hypothesis H_{λ} posits the invariance of factor loadings associated with a particular item. Due to the fact that the metric invariance requirement is usually difficult to satisfy, some researchers (Byrne, et al., 1989; Marsh & Hocevar, 1985) propose relaxing it. They suggest that if the noninvariant items constitute only a small portion of the model, then cross-group comparisons can still be made because the noninvariant items will not affect the comparisons to any meaningful degree. Before settling on this course of action, however, one must first identify the noninvariant items. Hence, if the metric invariance hypothesis H_{Λ} is rejected, a series of H_{λ} hypotheses are often tested in an attempt to locate items responsible for overall noninvariance.

Hypothesis $H_{\Lambda,\Theta(\delta)}$ states that residual variance is invariant across groups. Residual variance is the portion of item variance not attributable to the variance of the associated latent variable. Therefore, testing for the equality of between-group residual variance determines if the scale items measure the latent constructs with the same degree of measurement error. Residual invariance may fail when participants belonging to one group, compared with those of another, are unfamiliar with a scale and its scoring formats, and therefore respond to it inconsistently (Mullen, 1995). In addition, differences in vocabulary, idioms, grammar, syntax, and the common experiences of different cultures may produce residual noninvariance (Malpass, 1977).

Hypothesis $H_{\Lambda,\nu}$ proposes that in addition to metric invariance (H_{Λ}), the vectors of item intercepts are also invariant. The item intercepts are the values of each item corresponding to the zero value of the underlying construct. Support for $H_{\Lambda,\nu}$ indicates the existence of strong factorial invariance (Meredith, 1993), also referred to as scalar equivalence (Mullen, 1995). Strong factorial invariance is a prerequisite

for the comparison of latent means, because it implies that the measurement scales have the same operational definition across groups (i.e., have the same intervals and the zero points). In the absence of strong factorial invariance, the comparison of latent means is ambiguous, because the effects of a between-group difference in the latent means are confounded with differences in the scale and origin of the latent variable. Under these circumstances, Byrne et al. (1989) proposed to compare latent means under partial intercept invariance, assuming that the noninvariant item will not affect the latent means comparison at a great extent.

CONSTRUCT LEVEL INVARIANCE

The remaining hypotheses in Table 1 are concerned with construct-level invariance; that is, with the means and variances of constructs, and their equivalence across groups. Hypothesis $H_{\Lambda,\Phi(jj)}$ states that the variances of constructs (i.e., latent variables) are invariant across groups. This hypothesis may be rejected, for example, when a construct representing an attitude having different levels of diversity across groups is assessed. Hypothesis $H_{\Lambda,\Phi(jj)}$ must be supported before researchers can compare the correlations of constructs across groups (Byrne, 1994; Jackson, Wall, Martin, & Davids, 1993; Marsh, 1993). Hypothesis $H_{\Lambda,\Phi(jj')}$ posits that the covariances among the constructs are invariant across groups, whereas Hypothesis $H_{\Lambda,\nu,\kappa}$ posits that the latent means are invariant across groups. These two hypotheses address substantive research questions, and are usually based on theoretical considerations.

FIT STATISTICS

When testing measurement invariance under MGCFA, a series of models are estimated, and invariance is tested by comparing the GFI statistics of particular models with a model having additional between-group constraints. For example, testing construct-level metric invariance involves comparing the fit of an unconstrained model, which places no restrictions on model parameters,[1] with a constrained model in which all factor loadings associated with a particular construct are constrained to be equal across groups. If the imposition of additional constraints results in a signignicantly lower value of the fit statistic, then the constraint is "wrong." The parameters constrained to be equal across groups should not be constrained, because they are noninvariant. The fit differences used to test various invariance hypotheses under MGCFA are shown in Table 1.

[1]Except for the referent associated with each construct, which is set equal to unity across groups.

Model fit differences are often determined using the likelihood-ratio (LR) test, also known as the chi-square difference test (Bollen, 1989b). The chi-square difference ($\Delta\chi^2$) is calculated as

$$\Delta\chi^2 \therefore \chi^2_c \Lambda \chi^2_{uc} \tag{2}$$

where χ^2_c and χ^2_{uc} are the values for the constrained model and the unconstrained (or less constrained) model, respectively. Significance is evaluated with Δdf degrees of freedom, where

$$\Delta df = df_c - df_{uc} \tag{3}$$

The LR test, like the usual chi-square test, is a null-hypothesis significance test for a difference between the two groups. If there is no difference in fit, that is, if $(\hat{F}_{\min})_c \therefore (\hat{F}_{\min})_{uc}$, then $\Delta\chi^2 = 0$. If the sample sizes are large, however, even a small difference between $(\hat{F}_{\min})_c$ and $(\hat{F}_{\min})_{uc}$ may result in a significant value of $\Delta\chi^2$, indicating that the null hypothesis of no difference should be rejected even when the difference is trivial (Brannick, 1995; Kelloway, 1995). The question then becomes one of statistical significance versus practical significance.

When assessing overall model fit using CFA, it is common for researchers to use other GFIs in preference to the chi-square statistic because the models rarely seem to fit by that criterion due to its well-known dependence on sample size. On the other hand, the same researchers tend to adopt the LR test when testing invariance hypotheses using MGCFA (e.g., Byrne et al., 1989; Reise et al., 1993; Steenkamp & Baumgartner, 1998). Brannick (1995) and Kelloway (1995) warned against this double standard, and suggested that researchers should consistently use a single standard for all GFI tests, whether applied to tests of overall model fit, or to differences in fit between constrained and unconstrained models.

An alternative to $\Delta\chi^2$ is ΔGFI, defined as

$$\Delta\text{GFI} = \text{GFI}_c - \text{GFI}_{uc} \tag{4}$$

where GFI_c and GFI_{uc} are the values of some selected GFI estimated with respect to the constrained and unconstrained model. As opposed to $\Delta\chi^2$, however, there is in general no statistically based criterion (controlling for sampling error) for determining if the invariance hypothesis ought to be rejected based on a particular value of ΔGFI. Although many simulation studies have examined GFIs as indicators of overall model fit for single-group data, there have been no studies that examine various ΔGFI as indicators of measurement invariance. One exception was Little (1997), who proposed four criteria for assessing the relative fit indexes of two nested models: (a) the overall fit should be acceptable, (b) the difference in Tucker–Lewis Index (TLI) should be less than or equal to 0.05, (c) indexes of local misfit are uniformly and unsystematically distributed with respect to the constrained parameters, and (d) the constrained model is substantively more meaning-

ful and parsimonious than the unconstrained model. However, the 0.05 criterion has neither strong theoretical nor empirical support, and it is not widely used. In short, researchers to date have enjoyed considerable latitude in deciding whether a particular value of ΔGFI indicates noninvariance (e.g., Drasgow & Kanfer, 1985).

We review the properties of 20 different ΔGFIs when invariance constraints are added to two-group measurement models when the null hypotheses of invariance are true. The GFIs can be classified into six categories, as follows:

1. GFIs based on the minimum sample discrepancy, including chi-square (χ^2) and normed chi-square (χ^2/df; Wheaton, Muthen, Alwin, & Summers, 1977).
2. GFIs based on the population discrepancy, including the noncentrality parameter (NCP; Steiger, Shapiro, & Browne, 1985) and RMSEA (Steiger, 1989).
3. Information-theoretic GFIs, including the Akaike's Information Criterion (Akaike, 1987), Browne and Cudeck's Criterion (1989), and the Expected Cross-Validation Index (Browne & Cudeck, 1993).
4. Incremental GFIs, including the Normed Fit Index (Bentler & Bonett, 1980), Relative Fit Index (RFI; Bollen, 1986), Incremental Fit Index (IFI; Bollen 1989a), TLI (Tucker & Lewis, 1973), CFI (Bentler, 1990), and Relative Noncentrality Index (RNI; McDonald & Marsh, 1990).
5. Parsimony adjusted GFIs, including the parsimony-adjusted NFI (James, Muliak, & Brett, 1982) and parsimonious CFI (Arbuckle & Wothke, 1999).
6. Absolute GFIs, including gamma hat (Steiger, 1989), a rescaled version of Akaike's Information Criterion (Cudeck & Browne, 1983), cross-validation index (Browne & Cudeck, 1989), McDonald's (1989) Non-Centrality Index, and critical *N* (Hoelter, 1983). Algebraic definitions and properties of these GFIs, plus rules of thumb for using them to evaluate overall model fit, can be found in Arbuckle and Wothke (1999) and Hu and Bentler (1998).

The effect of sampling error is a critical concern. Knowledge of the sampling distribution is a prerequisite to deciding whether a value of ΔGFI should be attributed to noninvariance or to sampling error. In addition to sampling error, it seems likely that sample size, the characteristics of a model (number of factors, etc.) and the particular GFI being used should also affect the ΔGFI distribution. We conducted a simulation to address these issues.

SIMULATION

We used a Monte Carlo procedure to generate ΔGFI for the 20 GFIs described previously. The sampling distributions of the ΔGFIs were examined under the null hypothesis of invariance. A total of 48 different models were generated by varying the parameters shown in Table 2: the number of Factors F (either 2 or 3), the factor

TABLE 2
Model Parameters for Simulation

Code	*Variable*	*Values*
F	Number of factors (2)	2 or 3
V	Factor variance (2)	.36 or .81
C	Correlations of factors (2)	.3 or .5
I	Number of items per factor (3)	3, 4, or 5
L	Factor loadings (2 for each value of I)	If I = 3: 1, 1, 1 1, 1.25, 1.5 If I = 4: 1, 1, 1, 1 1, 1.25, 1.25, 1.5 If I = 5: 1, 1, 1, 1, 1 1, 1, 1.25 1.5, 1.5
N	Sample size per group (2)	150 or 300

variance V (.36 or .81), the correlations between Factors C (.3 or .5), the number of items per Factor I (3, 4, or 5), and the factor loadings L (two patterns for each value of I). This range of model parameters, which follows Anderson and Gerbing's (1984) simulation, was an attempt to represent a range of models encountered in practice, whereas keeping the size of the simulation within manageable limits. Two different sample sizes *N* per group (150 or 300) were also examined. All model parameters were estimated with maximum likelihood (ML) method.

The simulation procedure is shown schematically in Figure 1. It was assumed that the latent factors would not explain all the variance in the items; therefore, the reliability of each factor was set to 0.80. Appropriate item residual variances were calculated using Fornell and Larcker's (1981) equation. These, together with the model parameters F, C, V, I, and L, were used to calculate a population covariance matrix Σ for each model with χ^2 equal to zero.

Instead of running the simulation using models with perfect fit, we used models containing approximation errors (Cudeck & Henly, 1991). A residual Matrix E consisting of random terms was generated, and the random off-diagonal elements were added to Σ to produce a population covariance Matrix S. This matrix was transformed using Cudeck and Browne's (1992) procedure into a population covariance matrix, having a model χ^2 equal to the number of degrees of freedom with sample size equals to 300. The choice of $\chi^2 = df$ in the population produced samples with unconstrained fit statistics representative of actual cases; for example, the mean values of CFI, TLI, and RMSEA for samples generated during this study were .97, .96 and .057, respectively. These were less than the perfect values (e.g., CFI = 1.0), which are seldom encountered in practice. On the other hand, they were high enough to justify invariance testing. If a test of overall fit produced a value of CFI less than .90, for example, it is unlikely that the model would receive further consideration.

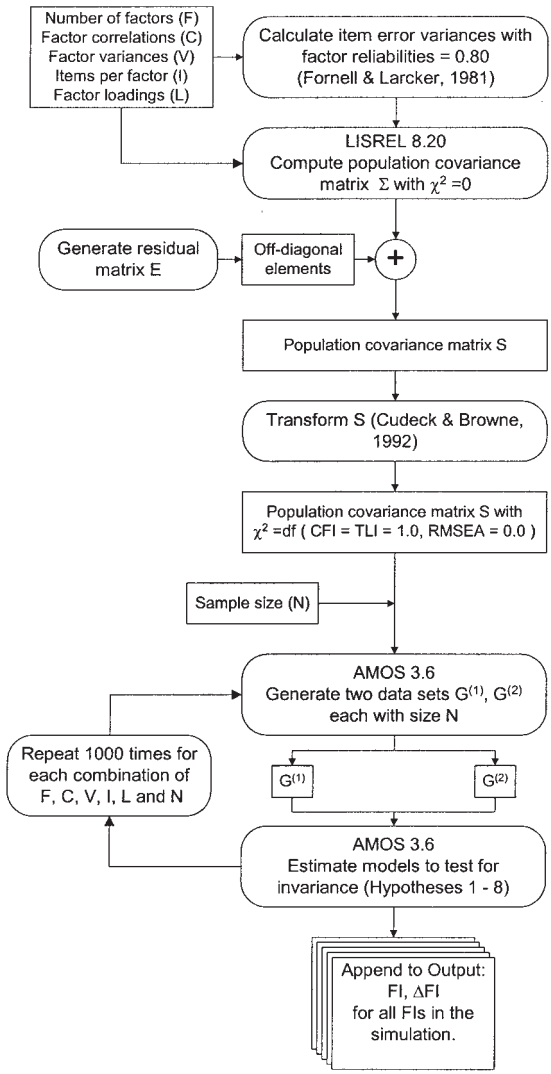

FIGURE 1 Simulation flowchart.

The Cudeck and Browne (1992) procedure warrants a brief digression. For a hypothesized Model K and a specified value of the fit function $F_{\min}$, there exists a corresponding covariance matrix $\Sigma_k(F_{\min})$ and a corresponding covariance matrix of measurement errors, $E(F_{\min})$, such that

$$\Sigma(F_{\min}) \therefore \hat{\Sigma}_k \vartheta E(F_{\min}) \tag{5}$$

For any value of a scalar multiplier κ, there exists a specific degree of model fit.

$$\Sigma(F_{\min}) \therefore \hat{\Sigma}_k \vartheta\ \kappa E(F_{\min}) \tag{6}$$

If $\kappa = 0$, then $\Sigma(F_{\min}) \therefore \hat{\Sigma}_k$; that is, the fit is perfect. If κ is greater than zero, then the fit is less than perfect. In the context of a particular model, there exists a one-to-one correspondence between the value of κ and model fit, as represented by the value of some GFI. Cudeck and Browne (1992) demonstrated that finding κ for any value of an GFI is straightforward when the generalized least squares function is used to estimate the model. If the ML function is used, as in this study, then an iterative method for finding κ is required.

Amos 3.6 was used to generate two multivariate-normal samples of size *N*, representing samples from two different groups that were completely invariant except for the effects of sampling error. These two samples were tested for the eight types of invariance represented by Hypotheses 1 through 8, and the relevant statistics were appended to an output file. The portion of the process beginning with the generation of the two data sets was repeated 1,000 times for each combination of F, C, V, I, L, and N. The same population covariance matrices were used for models with sample sizes of 150 and 300.

SIMULATION RESULTS

The quality of the simulation was assessed by examining the distribution of the $\Delta\chi^2$ statistics. As expected, these closely followed the distributions of χ^2 having Δdf degrees of freedom (Steiger, Shapiro, & Browne, 1985).

Configural Invariance (Hypothesis H_{form})

The effects of model parameters on the GFIs were assessed by performing a separate six-way analysis of variance (ANOVA) for each index. The factor loadings, factor variances, and factor correlations had no effect on any of the GFIs for the configural invariance (H_{form}) model. The percentages of total variance explained by number of items (I), number of factors (F), sample size (*N*), and Item × Factor interaction (I × F) are shown in Table 3. The number of items, number of factors, and Item × Factor interactions had significant effects on the values of the GFIs, showing that they were affected by the complexity of the model. The result for χ^2 is an artifact of the design of the simulation that set χ^2 equals to the degrees of freedom in the population. This procedure generates consistent theoretical *p* values of the χ^2 statistics, which fall between .45 to .48 across various levels of model complexity considered in this study.

TABLE 3
Factors Affecting Goodness-of-Fit Indexes (GFIs) in Testing Configural Invariance

	Percent Variance Due to:			
GFIs	*I*	*F*	*I* × *F*	*N*
GFIs based on the minimum sample discrepancy				
Chi-square (χ^2)	45.6	38.9	8.0	3.4
Normed chi-square (χ^2/df)				33.4
GFIs based on the population discrepancy				
Noncentrality parameter[a]	32.9	27.9	5.8	12.9
Root mean square error of approximation				
Information-theoretic GFIs				
Akaike's information criterion	45.7	44.3	6.4	
Browne and Cudeck criterion	44.7	43.3	6.3	
Expected value of the cross-validation index	35.7	34.8	5.0	18.1
Incremental GFIs				
Normed fit index	40.9	26.1		10.7
Relative fit index	32.1	23.4		15.1
Incremental fit index	34.2	21.9		
Tucker-Lewis index	24.8	18.1		
Comparative fit index	34.0	21.9		
Relative noncentrality index	34.0	21.9		
Parsimony adjusted GFIs				
Parsimonious normed fit index	67.7	21.8	6.1	
Parsimonious comparative fit index	70.0	25.6		
Other absolute GFIs				
Gamma hat	11.7	6.3		
Rescaled Akaike's information criterion	35.8	34.8	5.0	18.0
Cross-validation index[b]	27.0	23.6	10.5	17.9
McDonald's (1989) noncentrality index	39.1	33.4	6.3	
Critical *N*	10.0	7.0		35.0

Note. *I* = number of items (manifest variables); *F* = number of factors (latent variables); *N* = sample size.

[a] $I \times N$ (3.0%) also significantly affects NCP. [b] $I \times F \times N$ (5.2%), *I* x *N* (8.6%), and $F \times N$ (6.8%) also significantly affect cross-validation index.

Although most GFIs attempt to accommodate the effects of model complexity in one way or another, the results show that except for RMSEA, all the GFIs fail to fully do so. The findings are consistent with previous suggestions that the same cutoff values should not be applied to evaluate models having different levels of complexity (Anderson & Gerbing, 1984; Bearden, Sharma, & Teel, 1982; Boomsma, 1982; Gerbing & Anderson, 1993; LaDu & Tanaka, 1989; Marsh, Balla, & McDonald, 1988). Table 4 lists the means, first (or 99th) percentiles, and standard deviation of the GFIs for the test of configural invariance for each combination of I and F. None of the factors had a significant effect on RMSEA. The mean

TABLE 4
Summary Statistics of Fit Indexes for Testing Configural Invariance

	CFI			*Gamma Hat*			*Mc*		
	M	*SD*	*1%*	*M*	*SD*	*1%*	*M*	*SD*	*1%*
Overall	.9728	.0180	.9196	.9980	.0009	.9957	.9385	.0474	.8077
I*F									
3, 2	.9899	.0085	.9613	.9987	.0011	.9956	.9865	.0110	.9547
3, 3	.9793	.0109	.9449	.9982	.0008	.9958	.9598	.0182	.9106
4, 2	.9824	.0094	.9549	.9982	.0009	.9956	.9681	.0164	.9229
4, 3	.9672	.0123	.9306	.9977	.0007	.9958	.9153	.0257	.8467
5, 2	.9715	.0116	.9389	.9979	.0008	.9956	.9432	.0216	.8856
5, 3	.9468	.0156	.9017	.9974	.0006	.9957	.8583	.0317	.7733

	IFI			*RNI*		
	M	*SD*	*1%*	*M*	*SD*	*1%*
Overall	.9733	.0178	.9212	.9729	.0181	.9196
I*F						
3, 2	.9903	.0087	.9625	.9902	.0089	.9613
3, 3	.9797	.0107	.9461	.9793	.0110	.9449
4, 2	.9826	.0094	.9556	.9824	.0095	.9549
4, 3	.9676	.0121	.9320	.9672	.0123	.9305
5, 2	.9719	.0114	.9400	.9715	.0116	.9389
5, 3	.9476	.0152	.9042	.9468	.0156	.9017

	χ^2			χ^2/df			*NCP*		
	M	*SD*	*99%*	*M*	*SD*	*99%*	*M*	*SD*	*99%*
Overall	131.95	97.09	389.08	1.77	.418	2.90	57.61	50.01	215.12
I*F									
3, 2	28.21	9.79	55.75	1.76	.612	3.48	12.21	9.79	39.75
3, 3	84.73	19.18	133.60	1.77	.400	2.78	36.73	19.18	85.60
4, 2	67.14	16.59	109.70	1.77	.437	2.89	29.14	16.59	71.70
4, 3	181.00	33.17	258.21	1.77	.325	2.53	79.00	33.17	156.21
5, 2	120.32	24.69	180.40	1.77	.363	2.65	52.32	24.69	112.40
5, 3	310.29	50.74	419.13	1.78	.292	2.41	136.29	50.74	245.13

	RMSEA			*AIC*			*BCC*		
	M	*SD*	*99%*	*M*	*SD*	*99%*	*M*	*SD*	*99%*
Overall	.0568	.0157	.0926	219.95	121.66	521.08	225.59	124.40	528.54
I*F									
3, 2	.0533	.0256	.1075	80.21	9.79	107.75	82.11	9.54	109.05
3, 3	.0567	.0148	.0883	168.73	19.18	217.60	173.20	18.27	220.51
4, 2	.0561	.0171	.0918	135.14	16.59	177.70	138.38	15.98	179.84
4, 3	.0582	.0097	.0806	289.00	33.17	366.21	296.62	31.24	371.11
5, 2	.0574	.0123	.0844	204.32	24.69	264.40	209.27	23.57	267.61
5, 3	.0589	.0073	.0768	442.29	50.74	551.13	453.96	47.37	558.59

(*continued*)

TABLE 4 (*Continued*)

	ECVI			*NFI*			*RFI*		
	M	*SD*	*99%*	*M*	*SD*	*1%*	*M*	*SD*	*1%*
Overall	.538	.330	1.451	.938	.0358	.828	.916	.0409	.792
I*F									
3, 2	.198	.062	.321	.977	.0118	.936	.956	.0222	.880
3, 3	.414	.119	.630	.951	.0174	.896	.927	.0261	.844
4, 2	.332	.096	.516	.958	.0134	.921	.939	.0197	.883
4, 3	.706	.191	1.037	.925	.0223	.857	.903	.0289	.815
5, 2	.500	.139	.753	.934	.0180	.886	.912	.0238	.849
5, 3	1.077	.278	1.542	.883	.0303	.804	.858	.0366	.763
	TLI			*PNFI*			*PCFI*		
	M	*SD*	*1%*	*M*	*SD*	*1%*	*M*	*SD*	*1%*
Overall	.9633	.0219	.9012	.6595	.0726	.5096	.6855	.0843	.5201
I*F									
3, 2	.9816	.0166	.9275	.5208	.0063	.4991	.5280	.0045	.5127
3, 3	.9690	.0165	.9174	.6343	.0116	.5973	.6529	.0073	.6299
4, 2	.9741	.0140	.9136	.6503	.0091	.6248	.6666	.0064	.6480
4, 3	.9575	.0160	.9101	.7147	.0172	.6622	.7473	.0095	.7191
5, 2	.9623	.0154	.9191	.7056	.0136	.6694	.7340	.0087	.7094
5, 3	.9358	.0188	.8814	.7313	.0251	.6662	.7844	.0129	.7471
	CAIC			*CK*			*Critical N*		
	M	*SD*	*99%*	*M*	*SD*	*99%*	*M*	*SD*	*99%*
Overall	.5365	.3290	1.4459	.7796	.7883	3.2803	348.95	122.49	779.35
I*F									
3, 2	.1978	.0616	.3195	.2158	.0722	.3488	452.96	188.99	1075.12
3, 3	.4133	.1185	.6282	.4873	.1656	.7512	347.26	96.64	601.84
4, 2	.3309	.0959	.5139	.3755	.1233	.5873	362.27	106.62	659.02
4, 3	.7042	.1896	1.0336	.9359	.3533	1.4329	310.45	75.72	482.35
5, 2	.4985	.1379	.7507	.5974	.2023	.9166	328.18	85.16	540.03
5, 3	1.0741	.2766	1.5368	2.0654	1.1121	3.3712	292.56	67.30	426.23

Note. CFI = comparative fit index; Mc = McDonald's Noncentrality Index; IFI = Incremental Fit Index; RNI = Relative Noncentrality Index; NCP = noncentrality parameter; RMSEA = root mean squared error of approximation; AIC = Akaike's Information Criterion; BCC = Browne and Cudeck Criterion; ECVI = Expected Cross-Validation Index; NFI = Normed Fit Index; RFI = Relative Fit Index; TLI = Tucker–Lewis Index; PNFI = parsimony-adjusted NFI; PCFI = parsimonious CFI; CAIC = rescaled Akaike's Information Criterion; CK = cross-validation index.

RMSEA for all models was .0401, with a standard deviation of 0.011 and a 99th percentile value of .0655. These values are consistent with Browne and Cudeck's (1993) 0.05 criterion for RMSEA.

Other Invariance Tests (Hypotheses H_{Λ} Through $H_{\Lambda,\nu,\kappa}$)

As noted previously, the test statistics for hypotheses H_{Λ} through $H_{\Lambda,\nu}$ were the ΔGFIs. These were calculated for 1,000 simulated tests for each hypothesis, for each GFI. Correlations were calculated between the overall GFI (obtained from the test of hypothesis H_{form}: configural invariance) and the ΔGFI. One desirable property of a ΔGFI is that there should be a nonsignificant correlation between overall fit (GFI) and incremental fit (ΔGFI), owing to the independence between the configural invariance test and other invariance tests. Except for NCP, IFI, CFI, RNI, Gamma hat, McDonald's NCI, and critical *N*, the value of each ΔGFI was correlated with the overall GFI. For example, the values of TLI obtained with testing for configural invariance were signficantly correlated with values of ΔTLI obtained when testing for metric invariance, residual invariance, and equivalence of construct covariance (–.385, –.426, and –.408). These results indicate that the tests of these three invariance hypotheses were dependent on configural invariance. The correlations between RMSEA and ΔRMSEA varied from –.108 (equivalence of construct variance) to –.257 (metric invariance). On the other hand, the correlations between CFI and ΔCFI varied from .000 (equivalence of construct variance) to –.021 (equivalence of construct covariance).

A second desirable property of a ΔGFI is that it should not be affected by model complexity. The effects of the model parameters on ΔNCP, ΔIFI, ΔCFI, ΔRNI, ΔGamma hat, ΔMcDonald's NCI, and Δcritical *N* were tested using a six-way ANOVA. Except for ΔNCP and Δcritical *N*, which were affected by sample size, none of the model parameters explained more than 5% of the variance in these statistics.

A third desirable property of a ΔGFI is that it should not be redundant with other ΔGFIs. Correlations among the overall fit statistics of IFI, CFI, RNI, Gamma hat, and McDonald's NCI were calculated. Correlations among IFI, CFI, and RNI were at the .99 level, and correlations among ΔIFI, ΔCFI and ΔRNI for Hypotheses 2 through 8 were all above .98. This indicates that reporting all six of these statistics would be redundant. Due to the fact that CFI is the most frequently used index of this group, we recommend reporting only CFI and ΔCFI, and not IFI, ΔIFI, RNI, or ΔRNI. Gamma hat, ΔGamma hat, McDonald's NCI, and ΔMcDonald's NCI are not redundant with CFI and ΔCFI; therefore, we recommend that these also be reported when the results of measurement invariance tests are reported.

The means, standard deviations, and first (or 99th) percentiles of the ΔGFIs are listed in Table 5. The mean values in Table 5 show that the expected value for ΔCFI, ΔGamma hat, and ΔMcDonald's NCI are all zero. The percentile values

TABLE 5
Summary Statistics of Test Statistics for Hypotheses 2 Through 8

	ΔCFI			ΔGamma Hat			ΔMc		
	M	*SD*	*1%*	*M*	*SD*	*1%*	*M*	*SD*	*1%*
H2	–.0001	.0025	–.0085	–.0000	.0002	–.0008	–.0001	.0049	–.0160
H3	.0000	.0010	–.0039	–.0000	.0001	–.0004	.0000	.0018	–.0071
H4	–.0001	.0029	–.0094	–.0000	.0003	–.0009	–.0001	.0055	–.0180
H5	.0000	.0024	–.0082	–.0000	.0002	–.0008	–.0000	.0048	–.0160
H6	.0000	.0015	–.0056	–.0000	.0001	–.0006	–.0000	.0028	–.0100
H7	.0000	.0013	–.0048	–.0000	.0001	–.0005	–.0000	.0025	–.0094
H8	.0000	.0015	–.0055	–.0000	.0001	–.0005	–.0000	.0028	–.0100

	ΔIFI			ΔRNI		
	M	*SD*	*1%*	*M*	*SD*	*1%*
H2	–.0002	.0025	–.0086	–.0001	.0025	–.0085
H3	–.0000	.0010	–.0039	.0000	.0010	–.0040
H4	–.0002	.0029	–.0097	–.0001	.0029	–.0095
H5	–.0000	.0015	–.0056	–.0000	.0015	–.0056
H6	–.0000	.0013	–.0048	–.0000	.0013	–.0048
H7	–.0001	.0025	–.0084	–.0000	.0025	–.0083
H8	–.0000	.0015	–.0056	–.0000	.0015	–.0056

	$\Delta\chi^2$			$\Delta\chi^2/df$			ΔNCP		
	M	*SD*	*99%*	*M*	*SD*	*99%*	*M*	*SD*	*99%*
H2	7.6158	4.7728	22.3030	–.088	.1053	.1856	.1158	3.9876	12.605
H3	0.9901	1.4059	6.6320	–.017	.0456	.1378	–.0099	1.4059	5.632
H4	10.1180	5.3916	26.0820	–.089	.1013	.1758	.1180	4.5319	13.757
H5	2.5282	2.3153	10.5190	–.028	.0544	.1457	.0282	2.2612	7.846
H6	2.0178	2.2571	10.1330	–.020	.0410	.1241	.0178	2.0169	7.438
H7	7.5301	4.6801	21.7879	–.068	.0826	.1655	.0301	3.9248	12.209
H8	2.5101	2.2995	10.4890	–.022	.0453	.1261	.0101	2.2437	7.794

	ΔRMSEA			ΔAIC			ΔBCC		
	M	*SD*	*99%*	*M*	*SD*	*99%*	*M*	*SD*	*99%*
H2	–.0034	.0050	.0126	–7.3842	4.7105	4.9360	–8.3574	5.0638	4.0818
H3	–.0007	.0024	.0074	–1.0099	1.4059	4.6320	–1.1283	1.4071	4.5253
H4	–.0037	.0053	.0127	–9.8820	5.3547	3.7919	–11.1629	5.7887	2.8599
H5	–.0011	.0031	.0091	–2.4718	2.3163	5.3710	–2.7795	2.3486	5.0888
H6	–.0008	.0023	.0072	–1.9822	2.2452	5.0940	–2.2422	2.3250	4.8233
H7	–.0028	.0044	.0117	–7.4699	4.6980	4.5940	–8.4431	5.0634	3.9208
H8	–.0038	.0052	.0127	–2.4899	2.2980	5.3009	–2.7976	2.3290	5.0018

(*continued*)

TABLE 5 (*Continued*)

	ΔECVI			*ΔNFI*			*ΔRFI*		
	M	*SD*	*99%*	*M*	*SD*	*1%*	*M*	*SD*	*1%*
H2	–.019	.0139	.0123	–.0041	.0029	–.014	.0034	.0036	–.0073
H3	–.0025	.0038	.0122	–.0006	.0010	–.0047	.0006	.0015	–.0049
H4	–.025	.0164	.0093	–.0055	.0035	–.017	.0036	.0036	–.0070
H5	–.0062	.0065	.0141	–.0014	.0016	–.0075	.0010	.0019	–.0055
H6	–.005	.0062	.0135	–.0011	.0014	–.0063	.0008	.0016	–.0048
H7	–.019	.0139	.0113	–.0041	.0028	–.014	.0028	.0031	–.0065
H8	–.0062	.0065	.0139	–.0014	.0016	–.0075	.0008	.0017	–.0049
	ΔTLI			*ΔPNFI*			*ΔPCFI*		
	M	*SD*	*1%*	*M*	*SD*	*1%*	*M*	*SD*	*1%*
H2	.0036	.0039	–.0078	.0815	.0268	.0398	.0875	.0256	.0502
H3	.0006	.0016	–.0052	.0137	.0092	.0029	.0145	.0093	.0036
H4	.0038	.0038	–.0073	.1114	.0427	.0486	.1206	.0417	.0630
H5	.0011	.0020	–.0059	.0309	.0172	.0093	.0331	.0173	.0117
H6	.0008	.0017	–.0051	.0212	.0104	.0080	.0229	.0107	.0093
H7	.0029	.0033	–.0069	.0808	.0266	.0390	.0875	.0256	.0502
H8	.0009	.0018	–.0052	.0307	.0171	.0090	.0331	.0173	.0116
	ΔCAIC			*ΔCK*			*ΔCN*		
	M	*SD*	*99%*	*M*	*SD*	*99%*	*M*	*SD*	*99%*
H2	–.018	.0139	.0123	–.062	.0973	.0037	10.329	27.503	75.533
H3	–.0025	.0038	.0121	–.0068	.0087	.0090	1.869	13.445	28.289
H4	–.025	.0163	.0093	–.080	.1211	.0006	11.232	28.428	81.915
H5	–.0062	.0065	.014	–.018	.0246	.0083	3.282	16.642	40.205
H6	–.0050	.0062	.0134	–.016	.0252	.0084	2.530	11.702	24.038
H7	–.019	.0139	.0113	–.062	.0975	.0033	8.642	23.843	64.015
H8	–.0062	.0064	.0139	–.018	.0246	.0078	2.789	14.583	34.613

Note. CFI = comparative fit index; Mc = McDonald's Noncentrality Index; IFI = Incremental Fit Index; RNI = Relative Noncentrality Index; NCP = noncentrality parameter; RMSEA = root mean squared error of approximation; AIC = Akaike's Information Criterion; BCC = Browne and Cudeck Criterion; ECVI = Expected Cross-Validation Index; NFI = Normed Fit Index; RFI = Relative Fit Index; TLI = Tucker–Lewis Index; PNFI = parsimony-adjusted NFI; PCFI = parsimonious CFI; CAIC = rescaled Akaike's Information Criterion; CK = cross-validation index. H2 = metric invariance (weak factorial invariance); H3 = partial metric invariance (partial measurement invariance); H4 = metric invariance + invariance of residual variance; H5 = strong factorial invariance (metric invariance + scalar invariance); H6 = metric invariance + invariance of construct variance; H7 = metric invariance + invariance of construct covariance; H8 = strong factorial invariance + invariance of latent means.

shown in Table 5 are the critical values for rejecting the null hypothesis of equivalence, with an alpha of 0.01 and assuming multivariate normal distributions.

DISCUSSION OF SIMULATION RESULTS

Although the formulas for many GFIs (e.g., CFI and TLI) involve terms that adjust for degrees of freedom, this study shows that number of items per factor and number of factors in the model affect most of the GFIs (except for RMSEA). When overall fit is examined, models with more items and more factors can be expected to yield smaller values of these GFIs. This is due to the omission of small, theoretically insignificant factor loadings and correlated error terms in the model (Hall, Snell, & Foust, 1999; Hu & Bentler, 1998). In exploratory factor analysis, these terms are usually ignored; in CFA, and other SEM, they are hypothesized to be zero. This assumption has a negative impact on overall fit. This should serve as a warning to researchers who judge model fit in accordance with some generally accepted criterion (e.g., CFI = .90) while ignoring the effects of model complexity. RMSEA was not affected by any of the model parameters examined in this study, but its standard error was affected. Models with fewer items and factors were associated with larger standard errors in RMSEA.

Unlike the LR test in which $\Delta\chi^2$ is always greater than or equal to zero, many difference indexes (e.g., ΔRFI and ΔTLI) can assume both positive and negative values. This is because the underlying GFIs are functions of the number of degrees of freedom in the model. If the null hypothesis of invariance is true, and there is no sampling error, then decreasing the number of degrees of freedom can produce a value of ΔGFI greater than zero. If a ΔGFI is less than zero, then its value does not represent a change from a baseline value of zero, but rather a change from its hypothetical positive value. Hence, a slight reduction in the difference indexes may indicate a substantial change in the minimum value of the fit function.

Many ΔGFIs are superior to $\Delta\chi^2$ as tests of invariance because they are not affected by sample size. In many cases, however, ΔGFI is correlated with the GFI of the overall model. This implies that less accurately specified models produce larger values of difference statistics when measurement invariance constraints are added. As shown previously, the only difference statistics not having this undesirable characteristic are ΔCFI, ΔGamma hat, ΔMcDonald's NCI, ΔNCP, ΔIFI, ΔRNI, and Δcritical *N*.

The aforementioned results show that ΔCFI, ΔGamma hat, and ΔMcDonald's NCI are robust statistics for testing the between-group invariance of CFA models. These results are unexpected because many simulation studies have demonstrated that GFIs are affected by model complexity when evaluating overall model fit.

Although the standard errors and critical values differ for the different invariance models, the between-model variations are so small that a general crite-

rion for all hypotheses can be proposed. A value of ΔCFI smaller than or equal to –0.01 indicates that the null hypothesis of invariance should not be rejected. For ΔGamma hat and ΔMcDonald's NCI, the critical values are –.001 and –.02, respectively.

Limitations of the Simulation

One limitation of this study is that only GFIs based on the minimum value of the fit function are examined. Future studies are required to review other GFIs, such as the Jöreskog–Sörbom GFI (Jöreskog & Sörbom, 1993), adjusted GFI (Jöreskog & Sörbom, 1993), and root mean squared residual (Jöreskog & Sörbom, 1993).

Although 768,000 CFA models were estimated in the course of this study, generalizability is limited by the fact that only ML estimation was used and only 96 combinations of model parameters and sample sizes were examined. Further studies are needed to assess the effects of a wider range of parameters and other estimation methods. In addition, this study stipulated that the data distributions were multivariate normal. Deviations from multivariate normality may affect the results.

Another limitation of this study is that we only examined the Type I error when testing for measurement invariance. Although the Type II error or power may not be a major concern in most CFA work because of the large sample sizes, future efforts are needed to examine the power of the proposed criteria for determining measurement invariance. However, before the power can be examined, we must first meaningfully define the effect size of deviations from measurement invariance. Otherwise, using any arbitrary non-invariant model to examine the power will not be meaningful.

Finally, this simulation is limited to measurement models with two groups. Suitability of the recommended GFIs for testing across three or more groups is an interesting topic for future study.

CONCLUSION

The purpose of this article is to assist the decision-making processes of researchers engaged in examining measurement invariance. The various tests for between-group equivalence are laid out in a hierarchical sequence that researchers should find useful.

This study contributes to the literature in two ways. First, it invites the attention of researchers to different forms of measurement invariance that can be detected using structural equation modeling. MGCFA provides an elegant way to examine a large number of issues via a single procedure, rather than many separate procedures. This should facilitate empirical tests of many conceptual models having

cross-group differences. In addition, SEM provides a direct measure of how much a measurement model is improved or degraded by various between-group constraints, which offers a clear advantage over other techniques currently in use.

MGCFA is an extremely powerful procedure for detecting a range of cross-group differences, particularly when criterion values of ΔCFI, ΔGamma hat, and ΔMcDonald's NCI criteria are utilized. These cross-group differences, rather than being viewed as impediments to research, should be considered valid objects of research. Metric invariance, for example, need not be seen merely as an obstacle that must be surmounted before the equality of latent means can be assessed; rather, it should seen as a source of potentially interesting and valuable information about how different groups view the world. The same comment can be made with respect to any one of the measurement invariance failures considered.

Second, this article provides practical guidelines for determining when measurement invariance should be rejected. Currently, measurement invariance is most commonly examined with the LR test, which is based on the chi-square statistic. However, it makes no sense to argue against the usefulness of the chi-square and rely on various GFIs to evaluate the overall model fit and then argue for the usefulness of the chi-square instead of various GFIs to test for measurement invariance. We proposed critical values of ΔCFI, ΔGamma hat, and ΔMcDonald's NCI, that are independent of model parameters and sample size for testing measurement invariance and more generally, two nested models.

ACKNOWLEDGMENTS

Preparation of this article was supported in part by a Competitive Earmarked Research Grant funded by the Research Grants Council of the Hong Kong University Grants Committee (Project Code No. CUHK 4032/00H) to Gordon W. Cheung.

We thank the editor and three anonymous reviewers for their invaluable help and input.

An earlier version of this article was presented at the 1999 Annual Meeting of the Academy of Management, Chicago, Illinois.

REFERENCES

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52,* 317–332.

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 analysis. *Psychometrika, 49,* 155–173.

Arbuckle, J. L., & Wothke, W. (1999). *Amos 4.0 user's guide.* Chicago: SmallWaters.

Bearden, W. O., Sharma, S., & Teel, J. R. (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. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246.

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

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

Bollen, K. A. (1989a). A new incremental fiIt index for general structural equation models. *Sociological Methods and Research, 17,* 303–316.

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

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: Causality, structure, prediction* (Part 1, pp. 149–173). Amsterdam: North-Holland.

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior, 16,* 201–213.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24,* 445–455.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fiIt. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equations models* (pp. 136–162). Newbury Park, CA: Sage.

Buss, A. R., & Royce, J. R. (1975). Detecting cross-cultural commonalties and differences: Intergroup factor analysis. *Psychological Bulletin*, *82,* 128–136.

Byrne, B. M. (1994). Testing for the factorial validity, replication, and invariance of a measurement instrument: A paradigmatic application based on the Maslach Burnout Inventory. *Multivariate Behavioral Research, 29,* 289–311.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equation modeling. *Journal of Cross-Cultural Psychology, 31,* 187–212.

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

Cudeck, R., & Browne, M. W. (1992). Constructing a covariance matrix that yields a specified minimizer and a specified minimum discrepancy function value. *Psychometrika, 57,* 357–369.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin, 109,* 512–519.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70,* 662–680.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18,* 39–50.

Gerbing, D. W., & Anderson, J. C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 40–65). Newbury Park, CA: Sage.

Hall, R. J., Snell, A. F., & Foust, M. S. (1999). Item parcelling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. *Organizational Research Methods, 2,* 233–256.

Hoelter, J. W. (1983). The analysis of covariance structures: Goodness-of-fit indices. *Sociological Methods and Research, 11,* 325–344.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18,* 117–144.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424–453.

Irvine, S. H. (1969). Contributions of ability and attainment testing in Africa to a general theory of intellect. *Journal of Biosocial Science, 1,* 91–102.

Jackson, P., Wall, T., Martin, R., & Davids, K. (1993). New measures of job control, cognitive demand, and production responsibility. *Journal of Applied Psychology, 78,* 753–762.

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

Janssens, M., Brett, J. M., & Smith, F. J. (1995). Confirmatory cross-cultural research: Testing the viability of a corporation-wide safety policy. *Academy of Management Journal, 38,* 364–382.

Jöreskog, K., & Sörbom, D. (1993). *LISREL 8: User's reference guide.* Chicago: Scientific Software International.

Kelloway, E. K. (1995). Structural equation modelling in perspective. *Journal of Organizational Behavior, 16,* 215–224.

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

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32,* 53–76.

Malpass, R. S. (1977). Theory and method in cross-cultural psychology. *American Psychologist, 32,* 1069–1079.

Marsh, H. W. (1993). The multidimensional structure of academic self-concept: Invariance over gender and age. *American Educational Research Journal, 30,* 841–860.

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). Application 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. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6,* 97–103.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin, 107,* 247–255.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543.

Millsap, R. E., & Everson, H. (1991). Confirmatory measurement model comparisons using latent means. *Multivariate Behavioral Research, 26,* 479–497.

Millsap, R. E., & Hartog, S. B. (1988). Alpha, beta, and gamma changes in evaluation research: A structural equation approach. *Journal of Applied Psychology, 73,* 574–584.

Mullen, M. (1995). Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies, 3,* 573–596.

Pentz, M. A., & Chou, C. (1994). Measurement invariance in longitudinal clinical research assuming change from development and intervention. *Journal of Consulting and Clinical Psychology, 62,* 450–462.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552–566.

Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management, 20,* 643–671.

Singh, J. (1995). Measurement issues in cross-national research. *Journal of International Business Studies, 26,* 597–619.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross- national consumer research. *Journal of Consumer Research, 25,* 78–90.

Steiger, J. H. (1989). *EzPATH: Causal modeling.* Evanston, IL: SYSTAT.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50,* 253–263.

Suzuki, S., & Rancer, A. S. (1994). Argumentativeness and verbal aggressiveness: Testing for conceptual and measurement equivalence across cultures. *Communication Monographs, 6,* 256–279.

Tayeb, M. (1994). Organizations and national culture: Methodology considered. *Organization Studies, 15,* 429–446.

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

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–69.

Wheaton, B., Muthén, B., Alwin, D. F., & Summers, G. F. (1977). Assessing reliability and stability in panel models. In D. R. Heise (Ed.), *Sociological methodology* (pp. 84–136). San Francisco: Jossey-Bass.

Windle, M., Iwawaki, S., & Lerner, R. M. (1988). Cross-cultural comparability of temperament among Japanese and American preschool children. *International Journal of Psychology, 23,* 547–567.