STRUCTURAL EQUATION MODELING, *14*(3), 464–504
Copyright © 2007, Lawrence Erlbaum Associates, Inc.

Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance

Fang Fang Chen
Department of Psychology, University of Delaware

Two Monte Carlo studies were conducted to examine the sensitivity of goodness of fit indexes to lack of measurement invariance at 3 commonly tested levels: factor loadings, intercepts, and residual variances. Standardized root mean square residual (SRMR) appears to be more sensitive to lack of invariance in factor loadings than in intercepts or residual variances. Comparative fit index (CFI) and root mean square error of approximation (RMSEA) appear to be equally sensitive to all 3 types of lack of invariance. The most intriguing finding is that changes in fit statistics are affected by the interaction between the pattern of invariance and the proportion of invariant items: when the pattern of lack of invariance is uniform, the relation is nonmonotonic, whereas when the pattern of lack of invariance is mixed, the relation is monotonic. Unequal sample sizes affect changes across all 3 levels of invariance: Changes are bigger when sample sizes are equal rather than when they are unequal. Cutoff points for testing invariance at different levels are recommended.

Measurement invariance is a prerequisite for comparing different cultural, ethnic, gender, age, or experimental versus control groups. When groups are compared based on instruments that do not measure the same constructs, inference problems occur. In other words, the conclusions drawn from a study may be biased or invalid if the measures that we rely on do not have the same meanings across different groups. Consequently, an effective health prevention program might be deemed harmful whereas a detrimental early education program might be considered beneficial. For example, in a pioneering study (Millsap & Kwok, 2004) on the impact of lack of measurement invariance on group comparisons, selection bias under varying conditions of lack of invariance was examined. The

Correspondence should be sent to Fang Fang Chen, University of Delaware, Department of Psychology, Wolf Hall, Newark, DE 19716. E-mail: xiyu@udel.edu

results indicate that small group differences in factor structure can have a significant impact on selection accuracy. Selection bias was defined by the relations among true positive, true negative, false positive, and false negative.

Although measurement invariance has been increasingly tested in applied research (e.g., Byrne & Campbell, 1999; Chen, in press; Chen, Sousa, & West, 2005; Hendriks et al., 2003; Kwan, Bond, & Singelis, 1997; Little, 1997; Rhee, Uleman, & Lee; 1996; Steenkamp & Baumgartner, 1998), an important question has yet to be addressed: What criteria should be used to evaluate measurement invariance? A double standard has been applied to evaluate model fit in testing measurement invariance. On the one hand, a number of goodness of fit indexes have been used to judge the absolute model fit in the analysis of covariance and mean structures, in addition to chi-square tests. This is because chi-square tests are sensitive to sample size and to violation of the normality assumption, and thus a trivial discrepancy may lead to the rejection of a model (Bollen, 1989; Tucker & Lewis, 1973). On the other hand, when testing nested models, such as measurement invariance tests, researchers have mainly relied on the chi-square difference tests, which have the same drawbacks as the absolute chi-square tests (Brannick, 1995). The purpose of this study is to systematically examine changes in commonly used fit indexes under varying conditions of lack of measurement invariance.

MEASUREMENT INVARIANCE

Tests of measurement invariance examine whether an instrument has the same psychometric properties across heterogeneous groups. Conceptually, they test whether the same construct has been measured in different groups.[1] To answer different research questions, measurement invariance should be tested at corresponding levels. Meredith (1993) and Widaman and Reise (1997) described procedures for testing a series of models to establish measurement invariance. The first level of measurement invariance is configural invariance (Horn, McArdle, & Mason, 1983) or form invariance. It requires that the same item must be associated with the same factor in each group; however, the factor loadings may differ across groups. This level of invariance indicates that similar, but not identical, latent constructs have been measured in the groups (Widaman & Reise, 1997). The second level of invariance is tested at the factor loading level. Factor loadings represent the strength of the linear relation between each factor and its associated items (Bollen, 1989; Jöreskog & Sörbom, 1999). When the

[1]To ensure that the same construct has been measured in different groups, measurement invariance is necessary but not sufficient. Other criteria for checking validity of the measure, such as convergent and discriminant validity, external validity, and so on, should still be applied.

loading of each item on the underlying factor is equal in two (or more) groups, it suggests that the underlying factor has the same unit or same interval. This level of invariance is required for comparing regression slopes or change scores in longitudinal studies. The third level of invariance is tested at the intercept level. Intercepts represent the origin of the scale. When this level of invariance is achieved, it indicates that scores from different groups have the same unit of measurement (factor loading) as well as the same origin (intercept). This level of invariance is required for comparing latent mean differences across groups (Widaman & Reise, 1997). The fourth form of invariance is tested at the residual invariance level. When this level of invariance holds, all group differences on the items are due only to group differences on the common factors. However, a variety of reasons can make it difficult to achieve residual invariance.

Measurement invariance can be tested at more advanced levels, such as variance and covariance invariance (Widaman & Reise, 1997). However, these more advanced levels of invariance represent very strict standards that are often difficult to fulfill in practice. Therefore, configural, factor loading, intercept, and residual invariance are the most commonly tested forms of invariance.

GOODNESS OF FIT INDEXES AND MEASUREMENT INVARIANCE TESTS

As noted earlier, a double standard has been used to evaluate absolute and relative model fit. In testing absolute model fit, goodness of fit indexes have been used, in addition to the chi-square test. However, in testing measurement invariance, which involves comparing a more restricted model and a baseline model, the chi-square difference statistic has been the primary criterion, even though it has the same problems as the absolute chi-square test. That is, the chi-square statistic is sensitive to sample size and violation of the normality assumption. The performance of goodness of fit indexes in tests of measurement invariance has become an important yet unanswered question.

Goodness of fit indexes can be classified into different types, and one of them is the distinction between absolute and incremental fit indexes. Absolute fit indexes assess the degree to which the model-implied covariance matrix matches the observed covariance matrix. They gauge "badness of fit," and therefore the smaller the number, the better the model fit. A value of 0 indicates an optimal fit, and increasing values indicate greater departure of the implied covariance matrix from the observed matrix. In contrast, incremental fit indexes assess the degree to which the tested model is superior to an alternative model in reproducing the observed covariance matrix. They evaluate goodness of fit, and therefore the larger the number, the better the model fit. Larger values indicate greater improvement of model fit over an alternative model, which often assumes that

the observed variables are independent of each other (for detailed discussions on fit indexes, see Hu & Bentler, 1998, 1999; Marsh, Balla, & Hau, 1996; Marsh, Balla, & McDonald, 1988; Tanaka, 1993). The following section covers three commonly used fit indexes.

1. Root mean square error of approximation (RMSEA): Absolute Fit Index

$$RMSEA = \sqrt{\frac{\chi_t^2 - df_t}{df_t(N-1)}}$$

   where χ^2 is the chi-square for the tested model, *df* is the degrees of freedom, and N is sample size. RMSEA is a type of absolute fit index. RMSEA is a measure of discrepancy between the observed covariance matrix and model-implied covariance matrix per degree of freedom (Browne & Cudeck, 1993; Steiger, 1990; Steiger & Lind, 1980). RMSEA has a built-in penalty for lack of parsimony, but RMSEA tends to overreject a true model when sample sizes are small (Hu & Bentler, 1998). In a multiple-group analysis, RMSEA is adjusted by using a correction parameter. Specifically, the adjusted RMSEA is obtained by dividing the overall population discrepancy function (which is a weighted average of the sample-based discrepancies) by the average number of degrees of freedom per sample, then taking the square root to obtain a root-mean-square measure (Steiger, 1998).[2]
2. Standardized root mean square residual (SRMR): Absolute Fit Index

$$SRMR = \sqrt{\frac{2\Sigma\Sigma[(s_{ij} - \sigma_{ij})/(s_{ii}s_{jj})]^2}{p(p+1)}}$$

   where s_{ij} is the observed covariance, σ_{ij} is the model-implied covariance, s_{ii} and s_{jj} are the observed standard deviations, and p is the number of observed variables. SRMR is a measure of the average of the standardized residuals between the observed and model-implied covariance matrices (Bentler, 1995). SRMR is relatively independent of sample size.
3. The Comparative Fit Index (CFI): Incremental Fit Index

$$CFI = 1 - \left\{ \frac{\chi_t^2 - df_t}{\chi_n^2 - df_n} \right\}$$

[2] Another feature of RMSEA is that it has a point as well as interval estimate. For simplicity and ease of presentation, only the point estimate is discussed in this article. However, the rules regarding the interval estimate of RMSEA apply to measurement invariance tests as well.

where χ_t^2 is the chi-square for the tested model, χ_n^2 is the chi-square for the null model, and df_t and df_n are the degrees of freedom for the tested model and null model, respectively. A typical null model is the one in which only the variances of the observed variables are estimated, but no covariances are permitted. CFI assesses the extent to which the tested model is superior to an alternative model in reproducing the observed covariance matrix (Bentler, 1990; McDonald & Marsh, 1990). CFI ranges between 0 and 1. CFI is relatively independent of sample size and performs well in small samples (Hu & Bentler, 1998).

In early work on examining the performance of fit indexes in comparing nested models, Tucker and Lewis (1973) compared factor solutions of often-analyzed data sets. Tucker and Lewis rejected models based on a difference in Tucker-Lewis Index (TLI; Tucker & Lewis, 1973) of .008, .018, .019, and .022 across four different samples, respectively. In selecting the final models, Tucker and Lewis also considered whether a selected model was substantially meaningful. Similarly, Little (1997) proposed three criteria for comparing nested models, in addition to relying on TLI: (a) the overall model fit is acceptable, (b) indexes of local misfit (e.g., specific modification indexes, fitted residuals) are uniformly and unsystematically distributed for the restricted parameters, and (c) the restricted model is substantively more meaningful and parsimonious than the unrestricted model. However, there is lack of empirical support for these proposed standards.

Recently, Cheung and Rensvold (2002) conducted an extensive simulation study to examine the properties of 20 goodness of fit indexes under various levels of measurement invariance where all models contained small errors of approximation. Cheung and Rensvold concluded that changes in CFI ($\geq -.01$), Gamma hat ($\geq -.001$), and McDonald's Noncentrality Index ($\geq -.02$) provided the best performance, because these indexes are independent of sample size and model complexity, and are not correlated with the overall fit measures in the baseline model.

Cheung and Rensvold's study is the first one to empirically examine the sensitivity of fit indexes to lack of measurement invariance. However, this study has only focused on sampling variation of changes in goodness of fit indexes. The performance of fit indexes under varying degrees of noninvariance at different levels has yet to be studied. Moreover, there are several issues regarding proposing uniform standards for testing measurement invariance at all levels. First, evidence suggests that goodness of fit indexes are differentially sensitive to parameter misspecification in variance and mean structures. For example, compared to other fit indexes, SRMR is most sensitive to factor covariance misspecification, whereas CFI and RMSEA are most sensitive to factor loading misspecification (Hu & Bentler, 1998). Second, the sensitivity of goodness of

fit indexes may be affected by the specific pattern of noninvariance. Research indicates that it is harder to detect lack of loading invariance when the pattern is uniform (e.g., one group has higher loadings on all items than the other group) than when the pattern is mixed (e.g., one group has higher loadings on some of the items but lower values on others). (Meade & Lautenschlager, 2004). Although noninvariance was evaluated by the chi-square difference tests in that study, it is expected that the same pattern may be found in goodness of fit indexes, as the chi-square statistic is the basis for constructing goodness of fit indexes. Third, unequal sample sizes, which are often encountered in practice, might affect changes in goodness of fit indexes as well. For example, Kaplan and George (1995) found that as sample size becomes increasingly disparate across groups, the power to detect factor mean differences decreases. It is possible that the same pattern may be found in goodness of fit indexes.

Two Monte Carlo studies were conducted to systematically investigate the sensitivity of goodness of fit indexes to varying conditions of noninvariance. Study 1 was meant to investigate sampling variability of three commonly used fit indexes (i.e., SRMR, CFI, and RMSEA) and two promising indexes identified by Cheung and Rensvold (2002) (i.e., Gamma hat [Steiger, 1989] and McDonald's [1989] Non-Centrality Index [Mc]) under various levels of invariance tests. In other words, the goal of Study 1 is to establish cutoff points by examining the 1st/95th and 5th/99th percentiles of the goodness of fit indexes when the Type I error of these indexes is known. That is, there was only sampling variability across groups when various levels of invariance tests were conducted and no lack of invariance was simulated.

Based on the results of Study 1, cutoff points were proposed for different levels of invariance. Study 2 was conducted to investigate changes in goodness of fit indexes under different degrees of noninvariance at three commonly tested invariance levels: factor loadings, intercepts, and residual variances, when degrees of invariance are beyond sample variation. Another aim of Study 2 was to examine Type I errors (i.e., rejecting invariance falsely) and Type II errors (i.e., failing to find violations of invariance) based on cutoff points of goodness of fit indexes proposed in Study 1.

SIMULATION STUDY 1: SAMPLING VARIABILITY OF GOODNESS OF FIT INDEXES

Study 1 was conducted to examine sampling variability of goodness of fit indexes under seven levels of invariance: factor loading, intercept, residual variance, factor variance, factor covariance, factor mean, and the variance/covariance and mean structure. The Multiple Group Monte Carlo feature in Version 3.01 of M*plus* (Muthén & Muthén, 1998) was used to generate data and Mplus's

maximum likelihood was used to estimate models. First, data were randomly generated from a multivariate normal distribution to correspond to the parameters of a target model. Second, a baseline invariance model was fit to the simulated data, and the baseline model was the configural, loading, or intercept invariance model, depending on the level of invariance tests. The population values for each parameter were used as initial start values in the baseline invariance model. Finally, a more restricted invariance model was fit to the corresponding simulated data. Goodness of fit indexes and chi-square difference statistics obtained from the more restricted model were compared to those from the baseline model.

Experimental Conditions

To maximize the external validity of the findings from this study, the experimental conditions were designed to reflect commonly encountered situations in substantive research. In addition, these conditions were chosen based on pilot studies so that the research questions could be examined efficiently. As Hu and Bentler (1998) suggested, it is important to select a number of variables that is not too small but remains practical when conducting a large simulation study. Data properties are presented in the Appendix.

Number of groups. Two groups were chosen for simplicity.

Number of indicators. The number of observed measures for each factor was 8 or 12.

Number of factors. Only a one-factor model was examined except when examining invariance of factor covariance, for which a two-factor model was tested. A pilot study indicates that the pattern of the results is generalizable to multiple-factor models.

Sample size. Sample size for each group was 150, 250, or 500.

Data Characteristics

Distribution. Data were generated from a multivariate normal distribution.

Replications. Five hundred replications were generated for each condition.

Convergence. All models converged within 1,000 iterations.

Results

Tables 1 and 2 present the chi-square difference test statistics and rejection rates, the means, standard deviations, and 1st and 5th percentiles of CFI, Gamma hat, and Mc, and 95th and 99th percentiles of RMSEA and SRMR for the 8-indicator and 12-indicator conditions, respectively. The results are discussed with respect to two types of tests: The first is the invariance test of factor loadings, intercepts, and residual variances, and the second is the invariance test of the factor variance, factor covariance, and factor mean, as the impact of lacking invariance on model fit could differ across the two types of tests. There are several general findings in the study.

First, when testing measurement invariance in loadings, intercepts, and residual variances, SRMR was more sensitive to random variation in factor loadings than in intercepts and residual variances, whereas CFI, RMSEA, Gamma hat, and Mc were equally sensitive to all three levels of random variation. For example, given a model with a sample size of 300 and 8 indicators per factor, for SRMR, the means were .021, .003, and .004, and the 95th percentiles were .0368, .0058, and .0102, for invariance tests of loadings, intercepts, and residual variances, respectively; for CFI, the means were .000, .000, and .000, and the 5th percentiles were −.0062, −.0053, and −.0058; for RMSEA, the means were .000, −.001, and −.001, and the 95th percentiles were .0206, .0163, and .0131; for Gamma hat, the means were .000, .000, and .000, and the 5th percentiles were −.0062, −.0063, and −.0064; for Mc, the means were .000, .000, and .000, and the 5th percentiles were −.0123, −.0126, and −.0127 for invariance tests of loadings, intercepts, and residual variances, respectively (also see Figure 1 for 5th/95th percentiles). It is worth noting that the mean value of SRMR decreased as sample size increased, whereas the mean value of other fit indexes remained zero. For example, for an 8-indicator model, SRMR varied from .021, .016, to .011 as sample size increased from 300, 500, to 1,000.

Second, when testing invariance in factor variance, covariance, and mean, SRMR is most sensitive to random variation in factor variance, moderately sensitive to random variation in covariance, and least sensitive to random variation in the latent mean, whereas CFI, RMSEA, Gamma hat, and Mc were equally sensitive to all three levels of random variation. For example, given a sample size of 300 and the 8-indicator case, for SRMR, the means were .018, .009, and .003, and the 95th percentiles were .0564, .0342, .0130, for invariance tests of factor variance, covariance, and the latent mean, respectively; for CFI, the means were .000, .000, and .000, and the 95th percentiles were −.0016, −.0019, and −.0020; for RMSEA, the means were .000, .000, and .000, and the 95th percentiles were .0057, .0042, and .0051; for Gamma hat, the means were .000, .000, and .000, and the 95th percentiles were −.0021, −.0022, and −.0025; for Mc, the means were .000, .000, and .000, and the 95th percentiles were −.0043,

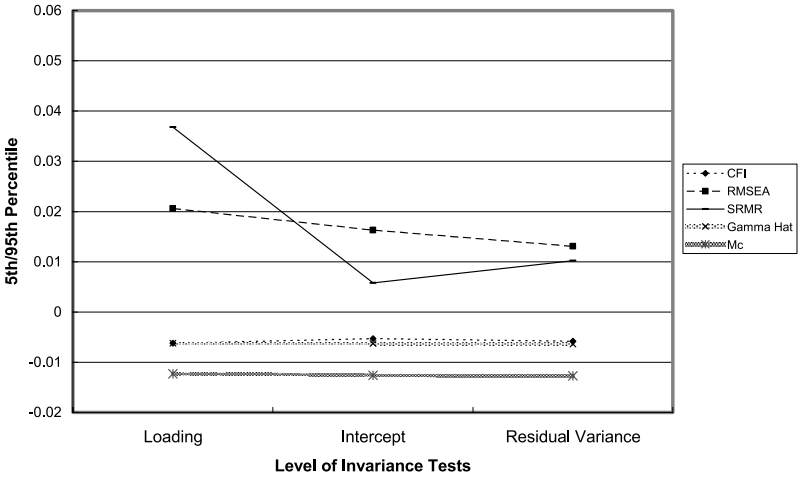

FIGURE 1 5th/95th percentile of fit indexes as a function of invariance tests.

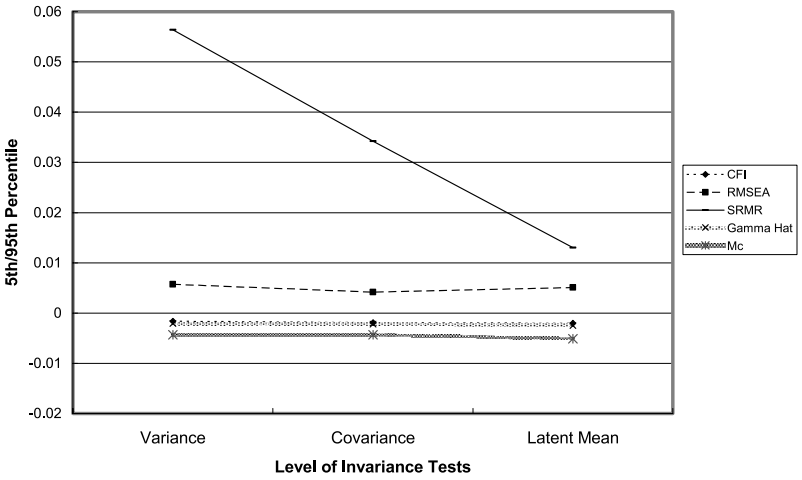

FIGURE 2 5th/95th percentile of fit indexes as a function of invariance tests.

−.0043, and −.0051, for invariance tests of factor variance, covariance, and the latent mean, respectively (see Figure 2). Note that the mean value of SRMR decreased as sample size increased, whereas the mean values of other fit indexes remained zero. For example, for an 8-indicator model, SRMR varied from .018, .013, to .009 as sample size increased from 300, 500, to 1,000.

TABLE 1
Sampling Variability of Goodness of Fit Indexes Under Different Levels of Invariance (8-Indicator)

	$\Delta\chi^2$			ΔCFI			$\Delta RMSEA$			$\Delta SRMR$			$\Delta Gamma\ Hat$			ΔMc		
N	*M*	*SD*	*Psig 5%*	*M*	*SD*	*5%/ 1%*	*M*	*SD*	*95%/ 99%*	*M*	*SD*	*95%/ 99%*	*M*	*SD*	*5%/ 1%*	*M*	*SD*	*5%/ 1%*
Loading Invariance (Baseline: Configural)																		
300	7.01	3.83	6.00	.000	.003	−.0062 −.0091	.000	.010	.0206 .0230	.021	.009	.0368 .0435	.000	.003	−.0062 −.0093	.000	.006	−.0123 −.0187
500	6.98	3.91	5.40	.000	.002	−.0030 −.0063	−.001	.007	.0139 .0213	.016	.007	.0295 .0354	.000	.002	−.0038 −.0060	.000	.004	−.0076 −.0120
1,000	7.00	3.70	5.40	.000	.001	−.0016 −.0026	.000	.005	.0072 .0135	.011	.005	.0196 .0236	.000	.001	−.0018 −.0027	.000	.002	−.0036 −.0054
Intercept Invariance (Baseline: Loading)																		
300	7.12	3.79	5.60	.000	.003	−.0053 −.0113	.000	.009	.0163 .0287	.003	.002	.0058 .0079	.000	.003	−.0063 −.0102	.000	.006	−.0126 −.0205
500	7.19	3.98	6.20	.000	.002	−.0038 −.0065	.000	.007	.0124 .0227	.002	.001	.0047 .0060	.000	.002	−.0037 −.0064	.000	.004	−.0074 −.0128
1,000	7.21	3.83	5.40	.000	.001	−.0015 −.0028	.000	.004	.0071 .0148	.002	.001	.0031 .0045	.000	.001	−.0019 −.0031	.000	.002	−.0038 −.0062
Residual Invariance (Baseline: Loading Intercept)																		
300	8.27	4.00	5.00	.000	.003	−.0058 −.0093	−.001	.008	.0131 .0265	.004	.004	.0102 .0120	.000	.003	−.0064 −.0101	.000	.007	−.0127 −.0202
500	8.12	4.05	4.20	.000	.002	−.0032 −.0069	.000	.006	.0118 .0198	.003	.003	.0081 .0103	.000	.002	−.0035 −.0069	.000	.004	−.0071 −.0138
1,000	8.26	3.99	5.40	.000	.001	−.0018 −.0027	.000	.004	.0084 .0149	.002	.002	.0055 .0066	.000	.001	−.0020 −.0030	.000	.002	−.0040 −.0060
Factor Variance Invariance (Baseline: Loading)																		
300	1.03	1.38	4.20	.000	.001	−.0016 −.0043	.000	.004	.0057 .0159	.018	.019	.0564 .0850	.000	.001	−.0021 −.0053	.000	.002	−.0043 −.0106
500	.95	1.33	5.00	.000	.001	−.0008 −.0023	.000	.002	.0034 .0109	.013	.014	.0441 .0592	.000	.001	−.0015 −.0027	.000	.001	−.0029 −.0054
1,000	.96	1.42	5.00	.000	.000	−.0006 −.0013	.000	.002	.0031 .0086	.009	.011	.0314 .0443	.000	.000	−.0007 −.0015	.000	.001	−.0014 −.0030
Factor Covariance Invariance (Without Mean Structure; Baseline: Loading)																		
300	.91	1.32	4.00	.000	.001	−.0019 −.0057	.000	.003	.0042 .0126	.009	.012	.0342 .0480	.000	.001	−.0022 −.0045	.000	.002	−.0043 −.0090
500	.91	1.26	4.40	.000	.001	−.0015 −.0032	.000	.003	.0039 .0076	.007	.009	.0260 .0352	.000	.001	−.0013 −.0024	.000	.001	−.0026 −.0047
1,000	1.11	1.56	6.20	.000	.000	−.0008 −.0021	.000	.002	.0042 .0108	.006	.007	.0196 .0321	.000	.000	−.0008 −.0018	.000	.001	−.0016 −.0036
Latent Mean Invariance (Baseline: Loading Intercept)																		
300	.98	1.43	5.40	.000	.001	−.0020 −.0050	.000	.003	.0051 .0157	.004	.004	.0130 .0227	.000	.001	−.0025 −.0056	.000	.002	−.0051 −.0111
500	.91	1.34	3.20	.000	.001	−.0006 −.0016	.000	.002	.0025 .0091	.002	.004	.0092 .0149	.000	.001	−.0011 −.0025	.000	.001	−.0023 −.0050
1,000	1.08	1.48	5.80	.000	.000	−.0005 −.0014	.000	.002	.0021 .0078	.002	.003	.0077 .0122	.000	.000	−.0009 −.0016	.000	.001	−.0018 −.0033
Variance/Covariance and Mean Structure Invariance (Baseline: Configural)																		
300	24.66	7.30	6.20	−.001	.005	−.0110 −.0179	−.001	.015	.0263 .0385	.043	.019	.0798 .1046	−.001	.006	−.0113 −.0169	−.001	.012	−.0226 −.0338
500	24.30	6.97	6.00	−.001	.003	−.0058 −.0091	−.001	.010	.0190 .0249	.033	.014	.0602 .0771	.000	.003	−.0063 −.0085	.000	.007	−.0127 −.0170
1,000	23.86	6.72	5.00	.000	.001	−.0027 −.0039	−.001	.007	.0134 .0195	.023	.009	.0389 .0479	.000	.002	−.0031 −.0049	.000	.003	−.0062 −.0097

Note. CFI = Comparative Fix Index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; Mc = McDonald's Non-Centrality Index; Δdf is 7 for loading and intercept invariance tests; Δdf is 8 for residual invariance test; Δdf is 1 for factor variance, factor covariance, and latent mean invariance tests; Δdf is 24 for variance/covariance and mean structure invariance test.

TABLE 2
Sampling Variability of Goodness of Fit Indexes Under Different Levels of Invariance (12-Indicator)

	$\Delta\chi^2$			ΔCFI			$\Delta RMSEA$			$\Delta SRMR$			$\Delta Gamma\ Hat$			ΔMc		
N	*M*	*SD*	*Psig 5%*	*M*	*SD*	*5%/ 1%*	*M*	*SD*	*95%/ 99%*	*M*	*SD*	*95%/ 99%*	*M*	*SD*	*5%/ 1%*	*M*	*SD*	*5%/ 1%*
Loading Invariance (Baseline: Configural)																		
300	11.07	4.85	5.20	.000	.002	−.0045 −.0073	.000	.006	.0118 .0176	.020	.007	.0322 .0383	.000	.003	−.0050 −.0076	.000	.008	−.0151 −.0225
500	11.00	4.79	5.00	.000	.001	−.0023 −.0044	−.001	.004	.0074 .0133	.015	.005	.0244 .0294	.000	.002	−.0028 −.0046	.000	.005	−.0084 −.0136
1,000	10.87	4.58	4.00	.000	.000	−.0009 −.0016	−.000	.003	.0047 .0100	.011	.004	.0170 .0208	.000	.001	−.0014 −.0023	.000	.002	−.0041 −.0068
Intercept Invariance (Baseline: Loading)																		
300	11.43	5.00	7.60	.000	.002	−.0043 −.0073	.000	.005	.0110 .0173	.002	.001	.0041 .0050	.000	.003	−.0054 −.0087	−.001	.008	−.0162 −.0261
500	11.29	4.75	5.20	.000	.001	−.0025 −.0046	.000	.004	.0060 .0155	.002	.001	.0028 .0035	.000	.002	−.0030 −.0048	.000	.005	−.0090 −.0141
1,000	11.02	4.65	5.20	.000	.001	−.0010 −.0022	.000	.003	.0048 .0093	.001	.000	.0020 .0026	.000	.001	−.0015 −.0025	.000	.002	−.0043 −.0076
Residual Invariance (Baseline: Loading & Intercept)																		
300	12.25	5.16	8.00	.000	.002	−.0050 −.0072	.000	.005	.0098 .0169	.002	.002	.0061 .0080	.000	.003	−.0059 −.0080	.000	.008	−.0174 −.0239
500	12.50	5.08	5.40	.000	.001	−.0026 −.0043	.000	.004	.0093 .0135	.002	.002	.0047 .0059	.000	.002	−.0031 −.0055	.000	.005	−.0095 −.0164
1,000	12.05	4.67	4.80	.000	.001	−.0011 −.0016	.000	.003	.0041 .0089	.001	.001	.0031 .0036	.000	.001	−.0015 −.0023	.000	.002	−.0045 −.0070
Factor Variance Invariance (Baseline: Loading)																		
300	1.03	1.42	4.80	.000	.001	−.0011 −.0026	.000	.002	.0030 .0074	.016	.018	.0529 .0826	.000	.001	−.0016 −.0035	.000	.002	−.0047 −.0107
500	1.02	1.53	6.00	.000	.000	−.0006 −.0014	.000	.002	.0016 .0067	.012	.015	.0435 .0624	.000	.001	−.0010 −.0019	.000	.002	−.0032 −.0058
1,000	1.03	1.50	6.20	.000	.000	−.0004 −.0011	.000	.001	.0014 .0039	.009	.010	.0322 .0466	.000	.000	−.0005 −.0012	.000	.001	−.0016 −.0036
Factor Covariance Invariance (Baseline: Loading)																		
300	1.02	1.40	4.40	.000	.001	−.0013 −.0030	.000	.002	.0026 .0087	.009	.011	.0303 .0453	.000	.001	−.0015 −.0033	.000	.002	−.0045 −.0100
500	1.02	1.51	6.00	.000	.000	−.0008 −.0024	.000	.002	.0028 .0084	.007	.009	.0267 .0403	.000	.001	−.0012 −.0022	.000	.002	−.0035 −.0066
1,000	1.05	1.42	5.60	.000	.000	−.0004 −.0008	.000	.001	.0013 .0028	.005	.006	.0172 .0255	.000	.000	−.0005 −.0010	.000	.001	−.0015 −.0031
Latent Mean Invariance (Baseline: Loading & Intercept)																		
300	.98	1.31	5.00	.000	.001	−.0012 −.0025	.000	.002	.0021 .0076	.003	.003	.0099 .0154	.000	.001	−.0016 −.0028	.000	.002	−.0048 −.0083
500	1.03	1.37	5.80	.000	.000	−.0008 −.0014	.000	.001	.0020 .0036	.002	.003	.0071 .0112	.000	.000	−.0011 −.0022	.000	.001	−.0032 −.0065
1,000	1.06	1.62	5.40	.000	.000	−.0003 −.0008	.000	.001	.0016 .0050	.001	.002	.0051 .0103	.000	.000	−.0006 −.0011	.000	.001	−.0017 −.0033
Variance/Covariance and Mean Structure Invariance (Baseline: Configural)																		
300	37.19	8.12	4.20	−.001	.003	−.0066 −.0113	−.001	.009	.0161 .0231	.039	.018	.0743 .1004	−.001	.004	−.0078 −.0136	−.002	.013	−.0232 −.0405
500	36.50	9.09	5.60	.000	.002	−.0043 −.0067	−.001	.007	.0114 .0209	.030	.015	.0591 .0876	.000	.003	−.0053 −.0081	.000	.009	−.0156 −.0244
1,000	36.53	8.33	5.00	.000	.001	−.0021 −.0037	.000	.004	.0080 .0112	.021	.009	.0425 .0535	.000	.001	−.0025 −.0041	.000	.004	−.0075 −.0123

Note. CFI = Comparative Fix Index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; Mc = McDonald's Non-Centrality Index; Δdf is 11 for loading and intercept invariance tests; Δdf is 12 for residual invariance test; Δdf is 1 for factor variance, factor covariance, and latent mean invariance tests; Δdf is 36 for variance/covariance and mean structure invariance test.

Third, number of indicators did not have appreciable impact on changes in goodness of fit indexes. For example, given a model with a sample size of 300 and 12 indicators per factor, for SRMR, the means were .020 (vs. .021 in the 8-indicator case), .002 (vs. .003), and .002 (vs. .004), and the 95th percentiles were .0322 (vs. .0368), .0041 (vs. .0058), and .0061 (vs. .0102), for invariance tests of loadings, intercepts, and residual variances, respectively; for CFI, the means were .000, .000, and .000, exactly the same as in the 8-indicator cases, and the 95th percentiles were −.0045 (vs. −.0062), −.0043 (vs. −.0053), and −.0050 (vs. −.0058); for RMSEA, the means were .000 (vs. .000), .000 (vs. −.001), and .000 (vs. −.001), and the 95th percentiles were .0118 (vs. .0206), .0110 (vs. .0163), and .0098 (vs. .0131); for Gamma hat, the means were .000, .000, and .000, exactly the same as in the 8-indicator cases, and the 95th percentiles were −.0050 (vs. −.0062), −.0054 (vs. −.0063), and −.0059 (vs. −.0064); for Mc, the means were .000, .000, and .000, exactly the same as in the 8-indicator cases, and the 95th percentiles were −.0151 (vs. −.0123), −.0162 (vs. −.0126), and −.0174 (vs. −.0127), for invariance tests of loadings, intercepts, and residual variances, respectively.

Fourth, as expected, as sample size increased, there was less sampling variation in fit indexes, which was a reflection of less sampling variation in the data. For example, in testing factor loading invariance with an 8-indicator model, as sample size increased from 300 to 1,000, mean SRMR decreased from .021 to .011, the standard deviation decreased from .009 to .005, and the 95th percentile decreased from .0368 to .0196. A similar pattern was observed for CFI, RMSEA, Gamma hat, and Mc in standard deviations and percentiles, although not for the means, which were zeros across different sample sizes. These results suggest that it would be easier to commit Type I errors with small sample sizes, particularly when relying on SRMR, as both its mean value and standard deviation are larger in small samples.

When pooled, the results from 8-indicator and 12-indicator cases across sample sizes of 300 and 500,[3] are similar to those in Cheung and Rensvold's study (2002) except for Gamma hat. Specifically, in this study, the 99th percentiles for CFI were −.0068 (vs. −.0085 in Cheung & Rensvold's study), −.0074 (vs. −.0082), and −.0069 (vs. −.0094) for invariance tests of factor loadings, intercepts, and residuals, respectively. For RMSEA the 99th percentiles were .0188 (vs. .0126 in Cheung & Rensvold's study), .0211 (vs. .0091), and .0192 (vs. .0127). For Mc the 99th percentiles were −.0167 (vs. −.0160), −.0184 (vs. −.0060), and −.0186 (vs. −.0080). However, for Gamma hat the 99th percentiles were −.0069 (vs. −.0008 in Cheung & Rensvold's study), −.0075 (vs. −.0008), and −.0076 (vs. −.0009) for invariance tests of factor loadings, intercepts, and

[3]Cases with a sample size of 1,000 were not included as these were not examined in Cheung and Rensvold's (2002) study, and sample size affects the sampling variation of goodness of fit indexes.

residuals, respectively. It is unclear what factors might have contributed to the discrepancy between the two studies on Gamma hat.[4] It is not possible to compare SRMR as it was not examined in that study. Cheung and Rensvold (2002) manipulated the number of indicators for each factor (three, four, or five), number of factors (two or three), and total sample size (300 or 600).

Based on the results of Study 1, cutoff points are proposed for testing measurement invariance at factor loading, intercept, and residual variance levels, given that measurement invariance has been commonly tested at these three levels. The following proposed cutoff points are roughly based on the average value of 1st/95th or 5th/99th percentiles across different number of indicators and sample sizes. Results indicate that there is less sampling variation in changes of goodness of fit indexes when sample size is large than when it is small. It suggests that under a null hypothesis, it is easier to commit Type I errors when sample size is small. However, under an alternative hypothesis, as indicated by a pilot study conducted by the author, it is easier to commit Type II errors when sample size is large, as changes of goodness of fit indexes increase as sample size increases. A good cutoff value should minimize Type I and Type II errors simultaneously (Hu & Bentler, 1999). To compromise the differential impact of sample size on Type I and Type II errors, same cutoff points are proposed for different sample sizes. However, the reader is reminded that the following standards should be applied with caution because test of measurement invariance is a complex issue, and is affected by a number of factors, as discovered in study 2.

For CFI,[5] RMSEA, Gamma hat, and Mc, similar values are suggested for all three levels of invariance tests: a change of $\leq -.005$ or $-.010$ for CFI, a change of $\geq .010$ or .015 for RMSEA, a change of $\leq -.005$ or $-.008$ for Gamma hat, and a change of $\leq -.010$ or $-.015$ for Mc, as CFI, RMSEA, Gamma hat, and Mc are equally sensitive to these levels of invariance. However, for SRMR, different values are recommended for different levels of invariance tests: When testing loading invariance, a change of $\geq .025$ or .030 is proposed; when testing invariance at the intercept and residual variance levels, a change of $\geq .005$ or .010 is proposed as SRMR is more sensitive to loading invariance than to the other two levels of invariance. These proposed cutoff values are applied to the next study to examine the rejection rates under various degrees of invariance within each level and at various levels of invariance.

[4]In both Study 1, and as we will see in Study 2, the performance of Gamma hat is similar to that of CFI, which is consistent with the findings in Hu and Bentler's (1998) study.

[5]For CFI, the average value for 95th and 99th percentiles across sample sizes, indicators, and levels of invariance tests is .0031 and .0054, respectively. Given that these two values are very close, .01 is also suggested, in addition to .005. A similar rule was applied to Gamma hat.

STUDY 2: SENSITIVITY OF GOODNESS OF FIT INDEXES TO LACK OF MEASUREMENT INVARIANCE

Study 2 was conducted to examine changes in goodness of fit indexes under different degrees of noninvariance beyond sampling variation in factor loadings, intercepts, and residual variances, respectively. As discussed earlier, these three levels of invariance have been commonly tested. The second goal is to examine Type I errors (i.e., rejecting invariance falsely) and Type II errors (i.e., failing to find violations of invariance) based on cutoff points of goodness of fit indexes proposed in Study 1.

The data generation and estimation procedures were similar to those in Study 1. Four major factors that might affect the sensitivity of goodness of fit indexes to noninvariance were considered: factor complexity (i.e., number of indicators per factor), pattern of noninvariance (i.e., uniform vs. mixed), sample size, and ratio of sample size. After conducting a series of pilot studies, the following experimental conditions were chosen based on two criteria: (a) be able to efficiently test the factors that might affect the sensitivity of fit indexes to lack of invariance in the context of a large simulation study; and (b) to simulate situations that are commonly encountered in social science research, simultaneously. Data properties are presented in the Appendix.

Experimental Conditions

Number of groups. Two groups were chosen for simplicity.

Number of indicators. The number of observed measures for each factor was 8 or 12.

Number of factors. Only a one-factor model was examined, as a pilot study indicated that the pattern of the results is generalizable to multiple-factor models.

Proportion of invariance. The proportion of items that was invariant was 0%, 25%, 50%, 75%, or 100%. For example, in the 0% loading invariance condition, all factor loadings were different across the two groups except for the marker variable.[6] In the 100% condition, all parameters were set equal except for random variation, and this serves as a control condition.

[6]A model is identified when there is a unique solution for each of its parameters (Ullman, 2001). Two typical approaches have been used to identify measurement models: One is to fix one of the factor loadings to a value of 1 for each factor, and the other is to fix the variance of each factor to 1. In general, it is easier to interpret the findings when the first approach is used.

Pattern of invariance. Noninvariance was either uniform or mixed. For example, when lack of loading invariance was uniform, all loadings were set higher in the first group; when lack of loading invariance was mixed, about half of the loadings were set higher in the first group, whereas the other half were set higher in the second group.[7]

Ratio of Sample Size. Three sample size ratios were varied: 1:1, 2:1, or 4:1. Specifically, the sample sizes were 300 (150 vs. 150, 200 vs. 100, 240 vs. 60), 500 (250 vs. 250, 333 vs. 167, 400 vs. 100), or 1,000 (500 vs. 500, 666 vs. 334, 800 vs. 200).

Five hundred replications were generated for each condition. Any replication that resulted in a Heywood Case in either the baseline model or more restricted model was eliminated.[8] To maintain 500 proper replications for each condition, additional replications were generated to replace the improper solutions. All models converged within 1,000 iterations.

RESULTS

To determine factors that affected changes in goodness of fit indexes, a 2 (pattern of invariance: uniform and mixed) × 4 (proportion of invariance: 0%, 25%, 50%, and 75%) × 3 (ratio of sample size: 1 vs. 1, 2 vs. 1, and 4 vs. 1) × 3 (sample size: 300, 500, and 1,000) × 2 (number of indicators: 8 and 12) analysis of variance (ANOVA) on each fit index was conducted for testing invariance of factor loadings, intercepts, and residuals, respectively. The percentage of variance explained by each factor (η^2) is presented in Table 3. The most interesting finding is that when testing measurement invariance at the factor loading and intercept levels, a good proportion of variance in all fit indexes was explained by the interaction of pattern of invariance and proportion of invariance. However, there is no such interaction effect when invariance was tested at the residual level (the explained variance was less than 2%). Pattern of invariance affected all fit indexes except at the residual invariance level. Proportion of invariance and ratio of sample size affected all fit indexes at every level of invariance tests. In addition, number of indicators affected RMSEA and Mc at all levels of tests. Finally, CFI and Gamma hat were affected by similar factors across all levels of tests and the percentages of variance explained by these factors were also similar for CFI and Gamma hat.

---

[7]In the 8-indicator cases, three loadings were set higher in the first group, four loadings were set higher in the focal group, and one loading was set to 1 in both groups. A similar approach was followed for the 12-indicator cases.

[8]Improper solutions occurred in six conditions, where sample sizes were 240 versus 60. They involved only one or two replications in each case.

TABLE 3
Percentage of Variance Explained by Factors Affecting Goodness of Fit Indexes in Testing Invariance of Factor Loadings, Intercepts, and Residuals

Predictors	*Test Level*	$\Delta\chi^2$	ΔCFI	$\Delta RMSEA$	$\Delta SRMR$	$\Delta Gamma\ Hat$	ΔMc
Pattern of invariance	Loading	22.44	39.35	34.05	49.21	41.88	36.89
	Intercept	26.34	45.92	40.59	44.09	47.96	43.92
	Residual						
Proportion of invariance	Loading	6.50	16.87	7.33	2.94	11.72	10.25
	Intercept	7.55	13.27	8.38	11.85	13.16	11.73
	Residual	29.26	50.29	43.58	23.78	59.47	51.11
Ratio of sample size	Loading	4.31	6.43	7.44	7.43	8.07	7.04
	Intercept	3.57	6.18	5.71	3.50	6.38	5.72
	Residual	3.45	7.93	5.25	6.17	6.98	6.01
Sample size	Loading	24.34		2.29			
	Intercept	23.59			2.71		
	Residual	32.68		4.38	2.60		
Indicator	Loading	6.69		4.94			8.77
	Intercept	4.61		6.40	5.97		6.09
	Residual	9.97		2.29	6.67		12.77
Pattern of invariance × proportion of invariance	Loading	12.26	24.01	17.67	22.98	22.34	19.79
	Intercept	12.75	21.82	17.76	18.38	22.40	20.55
	Residual						
Pattern of invariance × sample size	Loading	5.49					
	Intercept	6.30					
	Residual						
Proportion of invariance × sample size	Loading						
	Intercept						
	Residual	7.13					
Proportion of invariance × indicator	Loading						
	Intercept						
	Residual						2.41

Note. Percentage that is less than 2% is not reported. CFI = Comparative Fix Index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; Mc = McDonald's Non-Centrality Index.

Tables 4, 5, and 6 present detailed results for testing lack of loading invariance, intercept invariance, and residual invariance, respectively. Given that the pattern of findings for conditions with sample size ratio of 2 versus 1 is similar to that for conditions with sample ratio of 4 versus 1, only conditions with sample ratio of 4 versus 1 were reported. Similarly, because the pattern of findings for 12-indicator cases is similar to that for 8-indicator cases, only conditions with a total sample size of 300 were reported for 12-indicator cases.[9] The results are discussed with respect to two outcome measures: changes in goodness of fit indexes and rejection rate based on the proposed cutoff points for goodness of fit indexes.

Lack of Loading Invariance

Changes in Goodness of Fit Indexes

As can be seen from Table 4, the sensitivity of goodness of fit indexes as well as the chi-square difference statistics to lack of loading invariance were affected by a number of factors, similar to those identified by ANOVA.

Interaction between the proportion of invariance and pattern of invariance. Changes in fit statistics were affected by the interaction between the degree of invariance and whether lack of invariance was uniform or mixed. When lack of loading invariance was uniform, the relation between the degree of invariance and changes in fit indexes as well as in the chi-square differences was nonmonotonic. Specifically, when 0% of the items were invariant, the changes were the smallest, whereas when 50% of the items were invariant, the changes were the largest. For example, given an 8-indicator model with sample sizes of 150 and 150, as the invariance proportion varied from 0%, 25%, 50%, to 75%, changes in CFI varied from −.015, −.028, −.035, to −.019; in RMSEA from .024, .041, .050, to .037; in SRMR from .036, .052, .071, to .063; in Gamma hat from −.012, −.021, −.027, to −.018; in Mc from −.024, −.041, −.055, to −.037; and similarly, in chi-square difference statistics from 21.35, 32.32, 40.86, to 29.62 (see Figure 3).

In contrast, when lack of loading invariance was mixed, the relation between the proportion of invariance and the changes in fit indexes and the chi-square differences was monotonic. Specifically, when 0% of the items were invariant, changes were the biggest, and when 75% of the items were invariant, changes were the smallest. For example, given an 8-indicator model with sample sizes of 150 and 150, as the invariance proportion varied from 0%, 25%, 50%, to 75%, changes in CFI varied from −.112, −.093, −.056 to −.024;

[9]The complete tables are available upon request to interested readers.

TABLE 4
Changes in Fit Indexes and Rejection Rates Based on Changes in Fit Indexes as a Function of Proportion of Loading Invariance and Ratio of Sample Size

			ΔCFI			ΔRMSEA			ΔSRMR			ΔGamma Hat			ΔMc		
Inva %	$\Delta\chi^2$ ($\Delta df = 7$)	Psig % ($\alpha = .05$)	*M*	*SD*	$\leq -.005$ $\leq -.010$	*M*	*SD*	$\geq .010$ $\geq .015$	*M*	*SD*	$\geq .025$ $\geq .030$	*M*	*SD*	$\leq -.005$ $\leq -.008$	*M*	*SD*	$\leq -.010$ $\leq -.015$
8-Indicator/Total N = 300 (150 vs. 150)																	
0%	21.35	82.40[a]	−.015	.011	**80.20**[b] 64.20	.024	.017	**78.60**[c] 67.40	.036	.010	**85.00**[d] 70.60	−.012	.007	**86.00**[e] 68.80	−.024	.013	**86.00**[f] 72.00
25%	32.32	98.00	−.028	.014	**95.00** 90.60	.041	.018	**94.80** 91.60	.052	.012	**98.60** 96.40	−.021	.008	**98.60** 95.20	−.041	.017	**98.60** 96.00
50%	40.86	99.60	−.035	.014	**99.00** 96.00	.050	.018	**99.60** 98.60	.071	.015	**100** 99.80	−.027	.009	**99.80** 99.40	−.055	.019	**99.80** 99.60
75%	29.62	96.20	−.019	.011	**93.20** 81.20	.037	.018	**94.60** 90.80	.063	.016	**99.60** 98.60	−.018	.008	**97.40** 92.00	−.037	.017	**97.40** 92.80
100%	7.27	5.80	.000	.003	**5.00** 0.80	−.001	.009	**10.00** 7.00	.021	.009	**32.20** 15.60	.000	.003	**7.80** 2.00	.000	.006	**7.80** 2.40
8-Indicator/Total N = 300 (240 vs. 60)																	
0%	14.19	47.20	−.006	.006	**46.60** 22.20	.012	.014	**49.60** 37.20	.026	.009	**53.80** 33.00	−.006	.005	**53.80** 31.60	−.012	.010	**53.80** 35.00
25%	19.76	74.60	−.011	.009	**70.00** 48.80	.022	.017	**73.80** 63.40	.036	.012	**80.00** 64.40	−.010	.007	**79.80** 62.80	−.021	.013	**79.80** 65.00
50%	25.36	90.00	−.016	.010	**83.60** 71.40	.031	.019	**87.60** 79.40	.052	.016	**96.60** 93.20	−.015	.008	**93.00** 81.20	−.030	.016	**93.00** 82.80
75%	21.36	78.20	−.011	.009	**72.60** 53.40	.024	.017	**78.20** 67.40	.054	.021	**93.80** 88.40	−.012	.007	**81.80** 68.20	−.024	.015	**81.80** 72.00
8-Indicator/Total N = 500 (250 vs. 250)																	
0%	30.97	96.80	−.017	.009	**90.40** 76.60	.030	.014	**94.00** 86.20	.036	.008	**91.40** 74.20	−.012	.005	**92.20** 77.20	−.024	.010	**92.20** 81.00
25%	50.31	100	−.031	.010	**99.60** 98.60	.047	.014	**99.80** 99.20	.054	.010	**100** 99.80	−.021	.006	**100** 99.20	−.042	.013	**100** 99.20
50%	62.90	100	−.035	.011	**100** 99.20	.055	.014	**100** 99.80	.073	.013	**100** 100	−.027	.008	**100** 100	−.054	.015	**100** 100
75%	45.50	100	−.021	.008	**97.00** 89.60	.044	.014	**99.20** 97.60	.065	.013	**100** 99.80	−.019	.006	**99.00** 95.60	−.038	.013	**99.00** 96.00
100%	7.25	5.80	.000	.002	**2.80** 100	.000	.007	**6.40** 4.40	.016	.007	**11.60** 4.20	.000	.002	**2.40** .40	.000	.004	**2.40** 1.00
8-Indicator/Total N = 500 (400 vs. 100)																	
0%	17.99	68.40	−.005	.005	**47.00** 15.00	.015	.012	**59.80** 47.60	.024	.007	**44.80** 19.00	−.005	.004	**50.40** 20.40	−.011	.007	**50.40** 25.60
25%	27.49	93.40	−.011	.007	**77.40** 49.80	.027	.014	**88.00** 80.20	.034	.009	**83.40** 67.60	−.010	.005	**83.40** 62.80	−.020	.010	**83.40** 67.60
50%	37.50	99.60	−.016	.007	**95.40** 81.00	.036	.013	**98.40** 95.20	.052	.012	**99.60** 96.60	−.015	.006	**97.40** 89.60	−.030	.011	**97.40** 91.80
75%	30.19	96.20	−.011	.006	**85.00** 56.60	.030	.014	**92.80** 84.20	.054	.016	**97.60** 94.20	−.011	.005	**91.40** 72.80	−.023	.010	**91.40** 77.00
8-Indicator/Total N = 1,000 (500 vs. 500)																	
0%	55.83	100	−.018	.006	**99.60** 92.20	.036	.010	**99.80** 98.40	.038	.006	**98.20** 91.20	−.012	.003	**99.20** 89.80	−.024	.007	**99.20** 93.00
25%	91.84	100	−.031	.007	**100** 99.60	.051	.010	**100** 100	.056	.008	**100** 100	−.021	.005	**100** 99.80	−.042	.009	**100** 100
50%	115.77	100	−.035	.007	**100** 100	.059	.011	**100** 100	.076	.009	**100** 100	−.026	.005	**100** 100	−.053	.010	**100** 100
75%	83.47	100	−.022	.005	**100** 99.20	.048	.010	**100** 100	.068	.009	**100** 100	−.019	.004	**100** 99.60	−.038	.008	**100** 99.80
100%	6.93	5.40	.000	.001	**0** 0	.000	.005	**4.00** 2.00	.011	.005	**0.20** 0	.000	.001	**0** 0	.000	.002	**0** 0

(*continued*)

TABLE 4
(*Continued*)

Inva %	$\Delta\chi^2$ ($\Delta df = 7$)	Psig % ($\alpha = .05$)	ΔCFI M	ΔCFI SD	ΔCFI $\leq -.005$ / $\leq -.010$	ΔRMSEA M	ΔRMSEA SD	ΔRMSEA $\geq .010$ / $\geq .015$	ΔSRMR M	ΔSRMR SD	ΔSRMR $\geq .025$ / $\geq .030$	ΔGamma Hat M	ΔGamma Hat SD	ΔGamma Hat $\leq -.005$ / $\leq -.008$	ΔMc M	ΔMc SD	ΔMc $\leq -.010$ / $\leq -.015$
8-Indicator/Total N = 1,000 (800 vs. 200)																	
0%	29.50	96.40	−.006	.003	**59.60** 11.00	.021	.009	**88.00** 73.40	.024	.006	**42.00** 13.40	−.006	.002	**57.80** 15.20	−.011	.005	**57.80** 20.80
25%	47.86	100	−.011	.004	**96.00** 61.40	.032	.010	**99.60** 97.40	.036	.007	**94.20** 78.60	−.010	.003	**95.40** 71.20	−.020	.007	**95.40** 76.20
50%	67.76	100	−.017	.005	**99.80** 94.00	.041	.010	**100** 99.80	.054	.009	**100** 99.80	−.015	.004	**100** 97.80	−.030	.008	**100** 98.60
75%	53.92	100	−.012	.004	**96.80** 71.00	.035	.010	**99.60** 98.20	.056	.011	**99.60** 99.20	−.012	.003	**98.20** 84.80	−.023	.007	**98.20** 89.00
Mixed Invariance Pattern/8-Indicator/Total N = 300 (150 vs. 150)																	
0%	91.20	100	−.112	.027	**100** 100	.090	.021	**100** 100	.114	.018	**100** 100	−.065	.015	**100** 100	−.130	.030	**100** 100
25%	83.26	100	−.093	.023	**100** 100	.087	.020	**100** 100	.112	.017	**100** 100	−.060	.014	**100** 100	−.120	.028	**100** 100
50%	59.59	100	−.056	.017	**100** 99.60	.068	.018	**100** 100	.096	.017	**100** 100	−.042	.012	**100** 100	−.084	.023	**100** 100
75%	33.92	97.60	−.024	.011	**94.60** 87.80	.042	.017	**96.20** 94.20	.070	.016	**99.80** 99.60	−.022	.009	**98.80** 95.00	−.044	.017	**98.80** 95.60
100%	7.27	5.80	.000	.003	**5.00** 0.80	−.001	.009	**10.00** 7.00	.021	.009	**32.20** 15.60	.000	.003	**7.80** 2.00	.000	.006	**7.80** 2.40
Mixed Invariance Pattern/8-Indicator/Total N = 300 (240 vs. 60)																	
0%	61.77	100	−.074	.023	**100** 100	.069	.019	**100** 99.60	.088	.018	**100** 100	−.043	.013	**100** 100	−.087	.025	**100** 100
25%	53.36	99.80	−.055	.020	**99.60** 98.80	.061	.020	**99.80** 99.40	.081	.019	**100** 99.80	−.037	.013	**100** 99.80	−.074	.025	**100** 99.80
50%	39.12	99.80	−.033	.014	**98.00** 94.80	.049	.019	**98.60** 96.80	.075	.020	**100** 100	−.026	.010	**99.80** 98.40	−.052	.019	**99.80** 99.00
75%	23.62	83.60	−.014	.009	**80.20** 63.40	.028	.017	**82.60** 73.80	.057	.019	**97.80** 92.40	−.014	.007	**86.00** 75.20	−.027	.015	**86.00** 77.60
100%	7.22	5.20	.000	.002	**6.00** 0.40	−.001	.008	**8.60** 4.00	.021	.008	**30.00** 15.60	.000	.003	**7.40** 1.60	.000	.006	**7.40** 2.20
12-Indicator/Total N = 300 (150 vs. 150)																	
0%	29.20	85.00	−.012	.008	**77.20** 54.40	.017	.011	**69.80** 54.60	.030	.008	**73.80** 48.60	−.010	.005	**83.60** 60.80	−.030	.015	**91.20** 83.20
25%	53.38	99.80	−.028	.011	**98.80** 94.20	.034	.012	**98.60** 95.80	.053	.010	**99.80** 99.40	−.023	.007	**99.80** 99.20	−.068	.021	**100** 99.80
50%	64.51	100	−.033	.011	**99.80** 98.40	.039	.013	**99.60** 99.00	.070	.012	**100** 100	−.029	.008	**100** 99.80	−.085	.023	**100** 100
75%	45.91	99.00	−.018	.008	**94.00** 84.00	.029	.012	**94.40** 87.00	.061	.014	**99.80** 99.40	−.019	.007	**98.80** 95.40	−.056	.021	**99.40** 98.80
100%	10.98	5.40	.000	.002	**3.00** 0.20	−.001	.006	**5.60** 2.20	.020	.007	**21.00** 7.40	.000	.003	**3.80** .80	.000	.008	**10.00** 3.80
12-Indicator/Total N = 300 (240 vs. 60)																	
0%	20.00	47.60	−.004	.004	**39.80** 11.00	.008	.009	**34.80** 19.40	.022	.007	**33.20** 13.80	−.005	.004	**46.20** 19.00	−.015	.011	**61.20** 45.80
25%	31.73	86.80	−.011	.007	**79.60** 53.20	.017	.011	**72.80** 53.00	.035	.010	**83.20** 69.80	−.011	.006	**85.60** 69.60	−.033	.017	**91.60** 85.40
50%	41.06	96.40	−.016	.008	**92.20** 75.20	.024	.012	**87.00** 75.20	.052	.014	**97.80** 94.80	−.016	.007	**96.20** 88.80	−.048	.021	**98.80** 96.00
75%	32.34	89.80	−.010	.006	**79.80** 48.20	.018	.011	**74.80** 55.80	.052	.016	**97.40** 93.80	−.012	.006	**88.00** 70.80	−.034	.017	**95.20** 88.00
100%	11.50	5.80	.000	.002	**3.40** .20	−.001	.006	**4.60** 1.40	.021	.007	**25.80** 10.40	.000	.003	**5.40** 1.20	.000	.008	12.60 5.40

Note. CFI = Comparative Fix Index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; Mc = McDonald's Non-Centrality Index. The rejection rates are based on two criteria. The bolded numbers correspond to the rejection rates when the first criterion is used.

[a]Rejection rate based on chi-square difference statistic at α level of .05. [b]Rejection rate based on change in CFI. [c]Rejection rate based on change in RMSEA. [d]Rejection rate based on change in SRMR. [e]Rejection rate based on change in Gamma Hat. [f]Rejection rate based on change in Mc.

TABLE 5
Changes in Fit Indexes and Rejection Rates Based on Changes in Fit Indexes as a Function of Proportion of Intercept Invariance and Ratio of Sample Size

			ΔCFI			ΔRMSEA			ΔSRMR			ΔGamma Hat			ΔMc		
Inva %	$\Delta\chi^2$ ($\Delta df = 7$)	Psig % ($\alpha = .05$)	*M*	*SD*	$\leq -.005$ / $\leq -.010$	*M*	*SD*	$\geq .010$ / $\geq .015$	*M*	*SD*	$\geq .005$ / $\geq .010$	*M*	*SD*	$\leq -.005$ / $\leq -.008$	*M*	*SD*	$\leq -.010$ / $\leq -.015$
8-Indicator/Total N = 300 (150 vs. 150)																	
0%	26.85	92.60[a]	−.015	.009	**87.60**[b] 69.40	.030	.016	**88.80**[c] 81.80	.012	.004	**97.00**[d] 61.60	−.016	.008	**94.60**[e] 86.40	−.032	.015	**94.60**[f] 88.80
25%	38.43	99.40	−.026	.011	**98.60** 94.40	.045	.016	**98.60** 96.80	.016	.005	**99.80** 88.60	−.025	.009	**99.60** 98.60	−.050	.018	**99.60** 99.20
50%	47.57	100	−.034	.013	**99.80** 97.20	.053	.016	**100** 99.80	.018	.005	**100** 97.80	−.033	.010	**100** 100	−.065	.020	**100** 100
75%	36.89	99.40	−.025	.011	**97.20** 93.20	.042	.016	**98.60** 96.40	.014	.004	**99.60** 81.20	−.024	.009	**99.60** 98.60	−.048	.018	**99.60** 98.80
100%	7.43	6.40	−.001	.003	**6.80** 1.40	.000	.008	**9.80** 6.20	.003	.002	**10.00** 0.20	.000	.003	**8.20** 2.60	.000	.006	**8.20** 3.00
8-Indicator/Total N = 300 (240 vs. 60)																	
0%	19.95	76.20	−.010	.007	**71.60** 44.60	.020	.015	**72.20** 60.20	.009	.004	**85.40** 35.20	−.011	.006	**80.80** 62.80	−.021	.013	**80.80** 67.20
25%	26.28	91.80	−.015	.009	**86.80** 69.00	.029	.017	**86.00** 80.40	.012	.005	**95.60** 60.80	−.016	.008	**93.20** 85.00	−.031	.015	**93.20** 86.00
50%	32.78	98.40	−.021	.010	**94.40** 86.00	.038	.017	**95.40** 90.20	.013	.005	**97.80** 77.40	−.021	.009	**98.60** 94.80	−.042	.017	**98.60** 95.80
75%	26.54	91.60	−.016	.009	**88.60** 71.60	.030	.016	**88.60** 81.80	.011	.005	**94.00** 55.80	−.016	.008	**93.60** 83.80	−.032	.015	**93.60** 86.20
8-Indicator/Total N = 500 (250 vs. 250)																	
0%	39.12	99.80	−.016	.007	**95.20** 77.20	.035	.012	**98.80** 95.60	.012	.003	**99.80** 74.00	−.016	.006	**99.20** 93.20	−.032	.011	**99.20** 93.80
25%	58.16	100	−.026	.009	**99.80** 98.00	.048	.013	**100** 100	.017	.004	**100** 96.40	−.025	.007	**100** 99.80	−.050	.015	**100** 100
50%	74.42	100	−.035	.009	**100** 100	.058	.014	**100** 100	.020	.004	**100** 99.20	−.033	.008	**100** 100	−.065	.015	**100** 100
75%	56.16	100	−.025	.008	**99.80** 98.20	.047	.013	**100** 99.60	.016	.004	**100** 93.00	−.024	.007	**100** 99.80	−.048	.014	**100** 100
100%	7.33	7.40	.000	.002	**1.80** 0	.000	.007	**8.20** 3.60	.002	.001	**3.80** 0	.000	.002	**2.00** .40	.000	.004	**2.00** .40
8-Indicator/Total N = 500 (400 vs. 100)																	
0%	27.14	94.40	−.009	.005	**78.60** 42.00	.024	.012	**88.20** 75.80	.009	.003	**90.00** 34.80	−.010	.004	**87.80** 65.80	−.020	.008	**87.80** 70.60
25%	39.89	99.80	−.016	.006	**96.60** 81.60	.036	.013	**99.20** 94.60	.013	.004	**99.60** 73.20	−.016	.005	**99.40** 94.80	−.032	.011	**99.40** 96.60
50%	49.47	99.80	−.021	.008	**97.60** 91.20	.041	.013	**99.00** 98.00	.014	.004	**99.60** 86.80	−.021	.007	**99.80** 98.80	−.041	.013	**99.80** 99.00
75%	38.48	98.60	−.016	.007	**93.80** 79.00	.034	.013	**97.20** 93.40	.012	.004	**97.60** 66.40	−.015	.006	**97.80** 93.00	−.031	.011	**97.80** 94.00
8-Indicator/Total N = 1,000 (500 vs. 500)																	
0%	71.03	100	−.016	.005	**99.80** 92.20	.040	.009	**100** 99.60	.014	.003	**100** 92.40	−.016	.004	**100** 98.80	−.032	.008	**100** 99.40
25%	108.34	100	−.027	.006	**100** 100	.052	.010	**100** 100	.020	.004	**100** 100	−.025	.005	**100** 100	−.049	.009	**100** 100
50%	140.92	100	−.036	.007	**100** 100	.061	.011	**100** 100	.023	.004	**100** 99.80	−.032	.006	**100** 100	−.065	.011	**100** 100
75%	104.67	100	−.026	.006	**100** 100	.052	.009	**100** 100	.018	.003	**100** 99.60	−.024	.005	**100** 100	−.048	.009	**100** 100
100%	7.03	6.00	.000	.001	**0** 0	.000	.004	**3.60** 0.80	.002	.001	**0.40** 0	.000	.001	**0** 0	.000	.002	**0** 0

(*continued*)

TABLE 5
(*Continued*)

Inva %	$\Delta\chi^2$ ($\Delta df = 7$)	Psig % ($\alpha = .05$)	ΔCFI			ΔRMSEA			ΔSRMR			ΔGamma Hat			ΔMc		
			M	*SD*	$\leq -.005$ $\leq -.010$	*M*	*SD*	$\geq .010$ $\geq .015$	*M*	*SD*	$\geq .005$ $\geq .010$	*M*	*SD*	$\leq -.005$ $\leq -.008$	*M*	*SD*	$\leq -.010$ $\leq -.015$
8-Indicator/Total N = 1,000 (800 vs. 200)																	
0%	47.22	100	−.010	.004	**93.20** 47.20	.029	.009	**98.60** 95.80	.010	.003	**98.00** 52.80	−.010	.003	**97.20** 73.60	−.020	.006	**97.20** 79.60
25%	72.31	100	−.017	.005	**99.80** 94.80	.040	.009	**100** 99.80	.014	.003	**100** 90.40	−.016	.004	**100** 99.60	−.032	.008	**100** 99.60
50%	92.31	100	−.022	.005	**100** 99.40	.048	.009	**100** 100	.017	.003	**100** 99.20	−.021	.004	**100** 100	−.042	.009	**100** 100
75%	70.70	100	−.016	.005	**100** 93.80	.040	.009	**100** 100	.014	.003	**100** 88.80	−.016	.004	**100** 98.80	−.031	.008	**100** 99.00
Mixed Invariance Pattern/8-Indicator/Total N = 300 (150 vs. 150)																	
0%	140.40	100	−.118	.022	**100** 100	.113	.019	**100** 100	.044	.007	**100** 100	−.100	.015	**100** 100	−.199	.030	**100** 100
25%	122.87	100	−.102	.022	**100** 100	.102	.018	**100** 100	.039	.007	**100** 100	−.088	.015	**100** 100	−.175	.030	**100** 100
50%	83.53	100	−.067	.018	**100** 100	.079	.018	**100** 100	.029	.006	**100** 100	−.060	.013	**100** 100	−.119	.026	**100** 100
75%	44.42	96.40	−.031	.013	**99.20** 96.80	.050	.016	**99.60** 99.20	.017	.005	**99.80** 93.00	−.030	.010	**99.80** 99.40	−.060	.020	**99.80** 99.60
100%	7.43	6.40	−.001	.003	**6.80** 1.40	.000	.008	**9.80** 6.20	.003	.002	**10.00** 0.20	.000	.003	**8.20** 2.60	.000	.006	**8.20** 3.00
Mixed Invariance Pattern/8-Indicator/Total N = 300 (240 vs. 60)																	
0%	88.19	100	−.071	.016	**100** 100	.080	.018	**100** 100	.034	.008	**100** 100	−.063	.012	**100** 100	−.126	.025	**100** 100
25%	78.55	100	−.062	.016	**100** 100	.076	.017	**100** 100	.031	.008	**100** 100	−.056	.012	**100** 100	−.112	.024	**100** 100
50%	54.81	99.60	−.041	.013	**99.80** 99.20	.058	.016	**100.00** 99.80	.023	.007	**100.00** 98.80	−.038	.010	**100** 100	−.076	.020	**100** 100
75%	30.99	75.40	−.019	.010	**94.00** 80.80	.036	.016	**95.80** 92.20	.014	.006	**98.00** 73.60	−.020	.008	**98.20** 94.80	−.039	.016	**98.20** 95.00
12-Indicator/Total N = 300 (150 vs. 150)																	
0%	31.46	88.80	−.009	.009	**73.20** 40.60	.017	.016	**73.80** 54.80	.007	.004	**80.00** 13.00	−.011	.006	**87.40** 68.20	−.033	.017	**95.20** 86.80
25%	55.72	100	−.022	.008	**98.40** 92.80	.033	.012	**98.20** 95.00	.012	.003	**99.40** 66.20	−.024	.007	**100** 99.40	−.071	.022	**100** 100
50%	71.37	100	−.030	.009	**99.80** 99.20	.040	.012	**99.80** 99.40	.013	.003	**100** 86.60	−.032	.008	**100** 100	−.095	.024	**100** 100
75%	56.15	100	−.022	.008	**98.00** 93.00	.033	.011	**98.20** 94.20	.010	.003	**98.60** 51.80	−.024	.007	**100** 99.40	−.072	.021	**100** 100
100%	11.42	5.60	.000	.002	**3.00** 0.20	.000	.005	**4.80** 1.60	.002	.001	**0.40** 0	.000	.003	**5.00** 1.00	.000	.008	13.60 5.00
12-Indicator/Total N = 300 (240 vs. 60)																	
0%	24.36	69.00	−.006	.005	**54.40** 18.00	.011	.009	**49.00** 31.60	.006	.002	**55.40** 4.60	−.007	.005	**67.60** 40.00	−.022	.014	**79.40** 67.40
25%	39.60	97.60	−.014	.007	**89.20** 66.20	.021	.011	**87.20** 68.80	.008	.003	**90.60** 26.80	−.015	.006	**97.20** 89.80	−.046	.018	**98.80** 97.20
50%	49.01	99.60	−.019	.008	**97.20** 86.40	.028	.011	**96.20** 88.80	.010	.003	**96.20** 43.80	−.021	.007	**99.60** 97.80	−.061	.020	**99.80** 99.80
75%	40.16	98.00	−.014	.007	**90.00** 71.40	.022	.011	**87.60** 71.20	.008	.003	**88.40** 22.40	−.016	.006	**97.80** 92.20	−.047	.018	**99.00** 97.80

Note. CFI = Comparative Fix Index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; Mc = McDonald's Non-Centrality Index. The rejection rates are based on two criteria. The bolded numbers correspond to the rejection rates when the first criterion is used.

[a]Rejection rate based on chi-square difference statistic at α level of .05. [b]Rejection rate based on change in CFI. [c]Rejection rate based on change in RMSEA. [d]Rejection rate based on change in SRMR. [e]Rejection rate based on change in Gamma Hat. [f]Rejection rate based on change in Mc.

TABLE 6
Changes in Fit Indexes and Rejection Rates Based on Changes in Fit Indexes as a Function of Proportion of Residual Invariance and Ratio of Sample Size

			ΔCFI			ΔRMSEA			ΔSRMR			ΔGamma Hat			ΔMc		
Inva %	$\Delta\chi^2$ (Δdf = 8)	Psig % (α = .05)	*M*	*SD*	≤ −.005 ≤ −.010	*M*	*SD*	≥ .010 ≥ .015	*M*	*SD*	≥ .005 ≥ .010	*M*	*SD*	≤ −.005 ≤ −.008	*M*	*SD*	≤ −.010 ≤ −.015
8-Indicator/Total N = 300 (150 vs. 150)																	
0%	50.39	100	−.029	.011	**98.80** 97.40	.050	.016	**99.60** 99.20	.014	.004	**98.00** 82.20	−.034	.011	**100** 99.80	−.068	.022	**100** 99.80
25%	39.99	99.20	−.022	.010	**95.20** 88.80	.041	.016	**97.20** 95.80	.013	.006	**92.80** 65.80	−.026	.009	**99.60** 98.60	−.052	.019	**99.60** 98.60
50%	27.90	92.40	−.014	.009	**82.00** 62.60	.028	.016	**86.60** 77.60	.009	.005	**79.20** 44.80	−.016	.008	**95.00** 85.60	−.033	.016	**95.00** 87.60
75%	18.07	58.40	−.007	.007	**53.60** 27.00	.015	.015	**55.60** 44.40	.007	.005	**65.80** 29.00	−.008	.006	**68.00** 49.60	−.017	.013	**68.00** 52.00
100%	8.12	4.40	.000	.003	**6.00** 1.00	−.001	.007	**8.60** 4.20	.004	.004	**31.00** 5.20	.000	.003	**7.80** 2.20	.000	.007	**7.80** 2.80
8-Indicator/Total N = 300 (240 vs. 60)																	
0%	37.23	99.20	−.018	.009	**94.60** 82.20	.038	.016	**97.00** 93.40	.008	.004	**75.80** 30.20	−.024	.010	**99.80** 97.80	−.047	.019	**99.80** 98.00
25%	29.22	93.20	−.013	.008	**83.60** 63.40	.027	.015	**87.40** 78.20	.007	.005	**66.60** 24.40	−.017	.008	**95.60** 87.60	−.034	.017	**95.60** 89.40
50%	22.17	76.80	−.009	.007	**68.20** 39.60	.020	.015	**69.00** 57.40	.006	.005	**59.00** 20.20	−.012	.007	**84.60** 64.80	−.023	.014	**84.60** 68.60
75%	14.19	35.60	−.004	.005	**35.20** 15.00	.009	.013	**39.40** 27.40	.005	.005	**47.60** 14.80	−.005	.006	**44.40** 24.80	−.010	.012	**44.40** 28.00
8-Indicator/Total N = 500 (250 vs. 250)																	
0%	78.68	100	−.030	.008	**100** 99.60	.056	.012	**100** 100	.016	.004	**99.80** 92.80	−.034	.008	**100** 100	−.068	.016	**100** 100
25%	59.02	100	−.022	.007	**99.40** 96.00	.045	.012	**100** 99.40	.013	.004	**97.80** 72.80	−.025	.007	**100** 99.60	−.050	.014	**100** 100
50%	40.93	100	−.015	.006	**94.60** 74.60	.033	.012	**98.00** 92.60	.010	.005	**85.80** 53.20	−.016	.006	**99.40** 93.40	−.032	.011	**99.40** 94.60
75%	23.35	81.60	−.006	.005	**56.20** 20.40	.018	.012	**68.40** 57.60	.006	.004	**61.00** 20.60	−.008	.004	**70.80** 42.40	−.015	.009	**70.80** 48.20
100%	7.95	6.60	.000	.002	**2.20** 0	.000	.006	**6.20** 3.20	.003	.003	**19.20** 1.00	.000	.002	**2.60** .20	.000	.004	**2.60** .20
8-Indicator/Total N = 500 (400 vs. 100)																	
0%	58.28	100	−.019	.007	**99.60** 93.60	.044	.013	**99.80** 98.80	.009	.003	**89.60** 38.60	−.024	.007	**100** 100	−.049	.014	**100** 100
25%	43.33	99.40	−.014	.006	**95.60** 73.20	.034	.012	**98.20** 94.20	.008	.004	**75.80** 23.20	−.017	.006	**99.00** 95.00	−.035	.012	**99.00** 96.20
50%	31.16	94.40	−.009	.006	**75.80** 42.60	.025	.013	**86.60** 78.00	.007	.004	**63.40** 19.40	−.011	.006	**89.80** 70.40	−.023	.011	**89.80** 75.20
75%	18.46	60.40	−.004	.004	**34.40** 7.80	.012	.011	**52.20** 37.60	.005	.004	**45.60** 8.40	−.005	.004	**47.80** 23.60	−.010	.008	**47.80** 28.00
8-Indicator/Total N = 1,000 (500 vs. 500)																	
0%	151.43	100	−.032	.006	**100** 100	.060	.010	**100** 100	.018	.003	**100** 99.80	−.035	.006	**100** 100	−.069	.011	**100** 100
25%	111.45	100	−.024	.005	**100** 100	.049	.009	**100** 100	.015	.004	**100** 92.20	−.025	.005	**100** 100	−.050	.010	**100** 100
50%	73.30	100	−.015	.004	**99.40** 90.00	.037	.009	**100** 99.80	.011	.004	**95.00** 64.60	−.016	.004	**99.60** 98.40	−.032	.008	**99.60** 99.00
75%	40.32	99.80	−.007	.003	**75.20** 19.80	.023	.008	**95.00** 81.60	.007	.003	**75.60** 23.00	−.008	.003	**85.80** 45.20	−.016	.006	**85.80** 54.29
100%	8.28	6.40	.000	.001	**0** 0	.000	.004	**3.00** 1.20	.002	.002	**6.20** 0	.000	.001	**0** 0	.000	.002	**0** 0

(*continued*)

TABLE 6
(*Continued*)

Inva %	$\Delta\chi^2$ ($\Delta df = 8$)	*Psig* % ($\alpha = .05$)	Δ*CFI*			Δ*RMSEA*			Δ*SRMR*			Δ*Gamma Hat*			Δ*Mc*		
			M	*SD*	$\leq -.005$ / $\leq -.010$	*M*	*SD*	$\geq .010$ / $\geq .015$	*M*	*SD*	$\geq .005$ / $\geq .010$	*M*	*SD*	$\leq -.005$ / $\leq -.008$	*M*	*SD*	$\leq -.010$ / $\leq -.015$
8-Indicator/Total N = 1,000 (800 vs. 200)																	
0%	105.76	100	−.019	.005	**100** 98.40	.048	.009	**100** 100	.011	.003	**99.40** 62.80	−.024	.005	**100** 100	−.048	.010	**100** 100
25%	81.12	100	−.015	.004	**99.00** 88.00	.039	.009	**100** 100	.010	.003	**93.40** 44.40	−.018	.005	**100** 99.00	−.036	.009	**100** 99.40
50%	54.68	100	−.010	.004	**91.80** 46.00	.030	.009	**98.80** 93.40	.007	.003	**77.60** 21.20	−.012	.004	**98.40** 82.20	−.023	.008	**98.40** 85.20
75%	30.05	93.20	−.005	.003	**44.00** 4.00	.017	.009	**77.00** 59.00	.005	.003	**47.00** 5.40	−.005	.003	**57.00** 17.00	−.011	.005	**57.00** 20.60
Mixed Invariance Pattern/8-Indicator/Total N = 300 (150 vs. 150)																	
0%	45.36	100	−.026	.011	**97.80** 93.40	.047	.016	**99.60** 98.40	.016	.008	**92.00** 77.40	−.030	.010	**100** 99.60	−.060	.020	**100** 99.60
25%	35.83	97.80	−.020	.010	**94.80** 84.80	.037	.016	**95.40** 92.00	.013	.007	**88.20** 65.40	−.023	.009	**99.20** 9.640	−.045	.018	**99.20** 96.60
50%	26.17	89.40	−.013	.008	**84.00** 61.60	.026	.015	**83.60** 74.40	.010	.006	**78.80** 50.00	−.015	.007	**93.00** 83.40	−.030	.014	**93.00** 84.60
75%	16.90	52.20	−.006	.006	**49.00** 24.00	.013	.014	**51.40** 40.60	.007	.005	**62.00** 24.20	−.007	.006	**61.80** 41.00	−.015	.012	**61.80** 43.80
100%	8.12	4.40	.000	.003	**6.00** 1.00	−.001	.007	**8.60** 4.20	.004	.004	**31.00** 5.20	.000	.003	**7.80** 2.20	.000	.007	**7.80** 2.80
Mixed Invariance Pattern/8-Indicator/Total N = 300 (240 vs. 60)																	
0%	32.77	96.40	−.016	.009	**90.00** 76.20	.034	.016	**93.00** 87.80	.012	.008	**80.60** 60.60	−.020	.009	**97.00** 94.60	−.040	.017	**97.00** 95.20
25%	26.18	86.40	−.012	.008	**80.20** 57.40	.026	.016	**83.20** 72.80	.010	.008	**76.20** 48.00	−.015	.008	**91.80** 79.80	−.030	.016	**91.80** 81.20
50%	20.41	71.00	−.008	.007	**62.80** 34.40	.018	.014	**67.00** 53.00	.009	.007	**74.00** 40.00	−.010	.007	**76.40** 60.20	−.020	.013	**76.40** 64.20
75%	13.80	36.80	−.004	.005	**35.20** 11.20	.008	.012	**35.80** 23.80	.006	.006	**54.60** 21.80	−.005	.005	**44.80** 23.80	−.010	.010	**44.80** 26.80
12-Indicator/Total N = 300 (150 vs. 150)																	
0%	79.99	100	−.029	.008	**100** 99.20	.042	.012	**99.80** 99.60	.011	.003	**97.80** 63.60	−.036	.009	**100** 100	−.106	.026	**100** 100
25%	62.37	100	−.021	.008	**98.00** 92.80	.034	.011	**99.00** 96.40	.009	.003	**90.20** 37.00	−.027	.008	**100** 99.80	−.080	.023	**100** 100
50%	43.31	98.00	−.013	.007	**88.40** 69.00	.023	.011	**88.20** 76.20	.007	.003	**70.80** 16.80	−.017	.006	**98.00** 91.80	−.050	.019	**99.20** 98.00
75%	27.64	76.00	−.007	.005	**59.00** 24.60	.013	.010	**56.60** 37.60	.005	.003	**42.40** 5.60	−.009	.005	**76.00** 53.00	−.025	.014	**86.60** 75.60
100%	11.88	5.40	.000	.002	**2.20** 0	−.001	.005	**2.60** 1.60	.002	.002	**11.80** 0	.000	.003	**5.00** .20	.000	.008	**11.00** 4.80
12-Indicator/Total N = 300 (240 vs. 60)																	
0%	57.89	100	−.017	.007	**96.80** 87.40	.030	.012	**98.40** 91.80	.006	.003	**57.60** 6.20	−.025	.008	**100** 99.80	−.073	.024	**100** 100
25%	45.61	97.40	−.013	.007	**89.80** 67.60	.023	.011	**89.20** 75.80	.005	.003	**47.60** 5.60	−.018	.007	**97.20** 93.20	−.054	.021	**98.80** 97.00
50%	34.65	89.00	−.009	.006	**74.20** 40.20	.016	.011	**68.80** 52.20	.004	.003	**36.00** 4.60	−.012	.006	**88.80** 71.40	−.036	.019	**95.00** 88.60
75%	22.80	53.20	−.004	.004	**38.40** 10.60	.008	.009	**36.00** 19.20	.004	.003	**29.20** 3.00	−.006	.005	**53.00** 32.00	−.018	.014	**66.20** 52.40

Note. CFI = Comparative Fix Index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; Mc = McDonald's Non-Centrality Index.

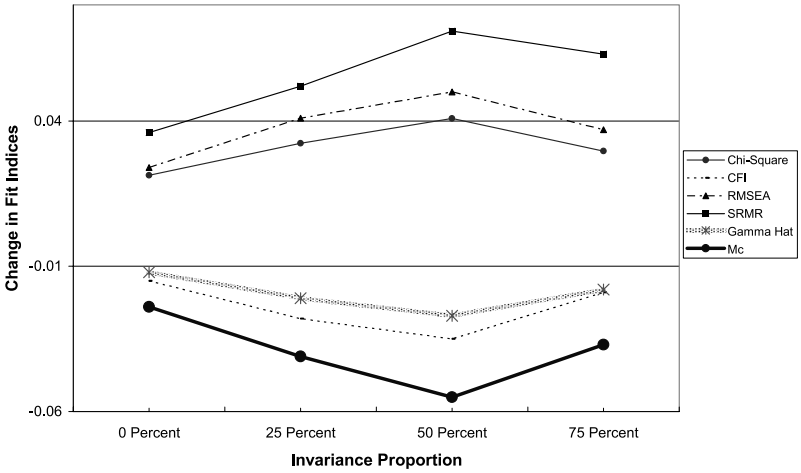

FIGURE 3 When lack of loading invariance is uniform: Changes in fit indexes as a function of invariance proportion. *Note*. To make the scales comparable, changes in the chi-square statistic were scaled down 1,000 times.

in RMSEA from .090, .087, .068, to .042; in SRMR from .114, .112, .096, to .070; in Gamma hat from −.065, −.060, −.042 to −.022; in Mc from −.130, −.120, −.084 to −.044; and in chi-square difference statistics from 91.20, 83.26, 59.59, to 33.92 (see Figure 4).

Pattern of invariance. Changes in goodness of fit indexes and the chi-square difference statistics were also affected by the pattern of invariance. The changes were bigger when lack of invariance was mixed than when it was uniform. For example, given an 8-indicator model with sample sizes of 150 and 150, as the invariance proportion varied from 0%, 25%, 50%, to 75%, changes in CFI varied from −.015 (vs. −.112), −.028 (vs. −.093), −.035 (vs. −.056), to −.019 (vs. −.024) when the pattern was uniform (vs. when the pattern was mixed). The same pattern was observed for RMSEA, SRMR, Gamma hat, Mc, and for the chi-square difference statistic.

Ratio of sample sizes. Changes were bigger when sample sizes were equal than when sample sizes were unequal. For example, given an 8-indicator model with a uniform pattern of invariance, as the invariance proportion varied from 0%, 25%, 50%, to 75%, changes in CFI varied from −.015 (vs. −.006), −.028 (vs. −.011), −.035 (vs. −.016), to −.019 (vs. −.011) when sample sizes were 150 and 150 (vs. when sample sizes were 240 and 60). The same pattern was

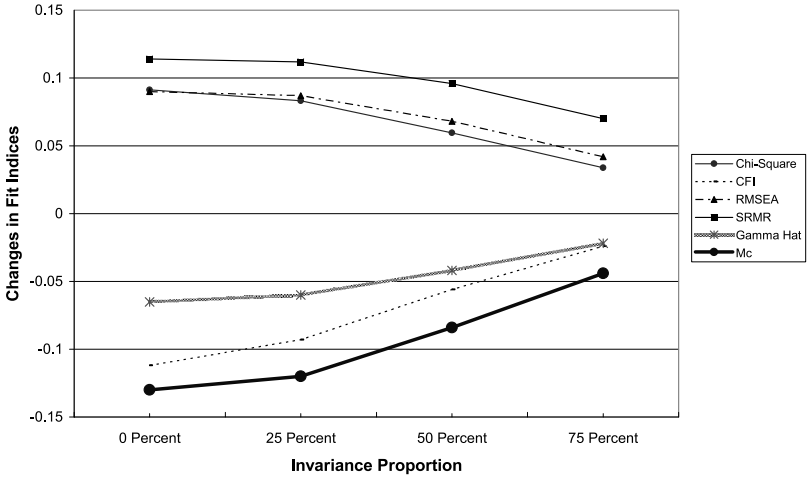

FIGURE 4 When lack of loading invariance is mixed: Changes in fit indexes as a function of invariance proportion.

observed for RMSEA, SRMR, Gamma hat, Mc, and the chi-square difference statistic as well. The same pattern was also found when invariance was mixed.

Sample size. Sample size did not have any appreciable impact on changes in CFI, SRMR, Gamma hat, and Mc, although for RMSEA, as sample size increased, changes were slightly increased. For example, given an 8-indicator model with a uniform pattern of invariance, and ratio of sample sizes of 1 versus 1, as the invariance proportion varied from 0%, 25%, 50%, to 75%, changes in CFI varied from −.015 (vs. −.017 vs. −.018), −.028 (vs. −.031 vs. −.031), −.035 (vs. −.035 vs. −.035), to −.019 (vs. −.021 vs. −.022) for a total sample size of 300 (vs. 500 vs. 1,000); changes in SRMR varied from .036 (vs. .036 vs. .038), .052 (vs. .054 vs. .056), .071 (vs. .073 vs. .076), to 0.63 (vs. .065 vs. .068); changes in Gamma hat varied from −.012 (vs. −.012 vs. −.012), −.021 (vs. −.021 vs. −.021), −.027 (vs. −.027 vs. −.026), to −.018 (vs. −.019 vs. −.019); and in Mc varied from −.024 (vs. −.024 vs. −.024), −.041 (vs. −.042 vs. −.042), −.055 (vs. −.054 vs. −.053), to −.037 (vs. −.038 vs. −.038). However, changes in RMSEA varied from .024 (vs. .030 vs. .036), .041 (vs. .047 vs. .051), .050 (vs. .055 vs. .059), to .037 (vs. .044 vs. .048) for a total sample size of 300 (vs. 500 vs. 1,000). As expected, as sample size increased, the standard deviation of all three fit indexes decreased.

Number of indicators. Number of indicators did not have an appreciable impact on changes in CFI, SRMR, and Gamma hat, although for RMSEA, the

changes were bigger in the 8-indicator models than in the 12-indicator models, and for Mc, the pattern was the opposite. For example, given a model with a uniform pattern of invariance and sample sizes of 150 and 150, as the invariance proportion varied from 0%, 25%, 50%, to 75%, the changes in CFI varied from −.015 (vs. −.012), −.028 (vs. −.028), −.035 (vs. −.033), to −.019 (vs. −.018) in 8-indicator (vs. 12-indicator) cases; in SRMR from .036 (vs. .030), .052 (vs. .053), .071 (vs. .070), to .063 (vs. .061); and in Gamma hat from −.012 (vs. −.010), −.021 (vs. −.023), −.027 (vs. −.029), to −.018 (vs. −.019). However, changes in RMSEA varied from .024 (vs. .017), .041 (vs. .034), .050 (vs. .039), to .037 (vs. .029); and in Mc changes varied from −.024 (vs. −.030), −.041 (vs. −.068), −.055 (vs. −.085), to −.037 (vs. −.056) in 8-indicator (vs. 12-indicator) cases. The standard deviations of all three fit indexes were slightly higher in the 8-indicator cases than in the 12-indicator cases except for Mc.

Rejection Rate

The same factors that affected changes in fit statistics also affected rejection rates.

Interaction between the proportion of invariance and pattern of invariance. When noninvariance was uniform, the relation between the proportion of invariance and the rejection rates was nonmonotomic. Specifically, the rejection rates were the smallest when the proportion of invariance was 0%, whereas the rejection rates were the highest when the proportion of invariance was 50%. In contrast, when noninvariance was mixed, the relation between the proportion of invariance and rejection rates was monotonic.

Pattern of invariance. Rejection rates based on fit indexes and the chi-square difference statistics were higher when lack of invariance was mixed than when it was uniform.

Ratio of sample sizes. Rejection rates based on fit indexes and the chi-square difference statistics were higher when sample sizes were equal than when sample sizes were unequal.

Sample size. As sample size increased, rejection rates based on fit indexes also tended to increase, although to a lesser degree compared to those based on chi-square difference tests. This finding appears to be inconsistent with the result that mean changes in CFI, SRMR, Gamma hat, and Mc were relatively independent of sample size and changes in RMSEA were slightly dependent on sample size. A possible explanation is that, as sample size increased, the

standard deviations of all fit indexes decreased, which made it easier to reject a model with a larger sample size.

Number of indicators. Number of indicators did not have any appreciable impact on rejection rates.

Type I and Type II errors. When relying on RMSEA, Gamma hat, and Mc, there was a small chance of committing Type I errors in small samples, but the chance was much higher when relying on SRMR due to the fact that the mean value of SRMR decreases as sample size increases. Type II errors tended to occur when the pattern of noninvariance was uniform and the degree of noninvariance was the highest, sample sizes were small, or sample sizes were unequal.

Summary

The sensitivity of goodness of fit indexes and of the chi-square difference statistics to lack of loading invariance was affected by the interaction between the proportion of invariance and pattern of invariance, pattern of invariance, and ratio of sample size. When noninvariance was uniform, the relation between the proportion of invariance and changes in fit statistics is nonmonotonic. Specifically, changes in fit indexes and the chi-square difference statistics were smallest when the proportion of invariance was 0%, whereas changes were largest when the proportion of invariance was 50%. In contrast, when noninvariance was mixed, the relation between the proportion of invariance and changes in fit statistics was monotonic. Changes were larger when lack of invariance was mixed than when it was uniform. Changes were also larger when sample sizes were equal than when they were unequal.

The same pattern was found in rejection rates. In addition, rejection rates based on fit statistics increased as sample size increased even though the mean changes in CFI, SRMR, Gamma hat, and Mc were relatively independent of sample size and changes in RMSEA were slightly dependent on sample size. This is because as sample size increased, the standard deviations of all fit indexes also decreased, which made it easier to reject a model with a larger sample size.

Lack of Intercept Invariance

Changes in Goodness of Fit Indexes

As can be seen from Table 5, the sensitivity of goodness of fit indexes to lack of intercept invariance was affected by similar factors that affected lack of loading invariance.

Interaction of proportion of invariance and pattern of invariance. Changes in goodness of fit indexes were affected by the relation between the degree of invariance and pattern of invariance. When lack of intercept invariance was uniform, the relation between the degree of noninvariance and changes in fit indexes as well as in the chi-square differences was not monotonic. Specifically, when 0% of the items were invariant, the changes were the smallest, whereas when 50% of the items were invariant, the changes were the largest. For example, given an 8-indicator model with sample sizes of 150 and 150, as the invariance proportion varied from 0%, 25%, 50%, to 75%, changes in CFI varied from −.015, −.026, −.034, to −.025; in RMSEA from .030, .045, .053, to .042; in SRMR from .012, .016, .018, to .014; in Gamma hat from −.016, −.025, −.033, to −.024; in Mc from −.032, −.050, −.065, to −.048; and in chi-square difference statistics from 26.85, 38.43, 47.57, to 36.89 (see Figure 5).

In contrast, when lack of intercept invariance was mixed, the relation between proportion of noninvariance and changes in fit indexes as well as in the chi-square differences was monotonic. Specifically, when 0% of the items were invariant, the changes were the largest, and when 75% of the items were invariant, the changes were the smallest. For example, given an 8-indicator model with sample sizes of 150 and 150, as the invariance proportion varied from 0%, 25%, 50%, to 75%, the change in CFI varied from −.118, −.102, −.067 to −.031; in RMSEA from .113, .102, .079, to .050; in SRMR from .044, .039, .029, to .017; in Gamma hat from −.100, −.088, −.060, to −.030; in Mc from −.199, −.175,

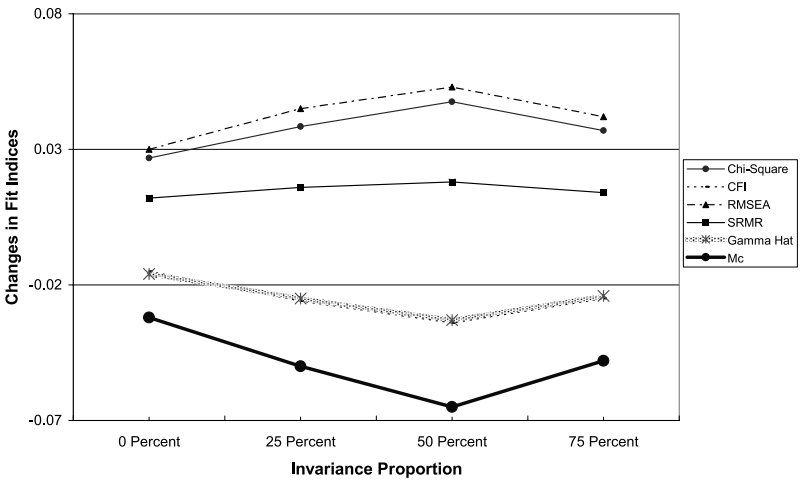

FIGURE 5 When lack of intercept invariance is uniform: Changes in fit indexes as a function of invariance proportion.

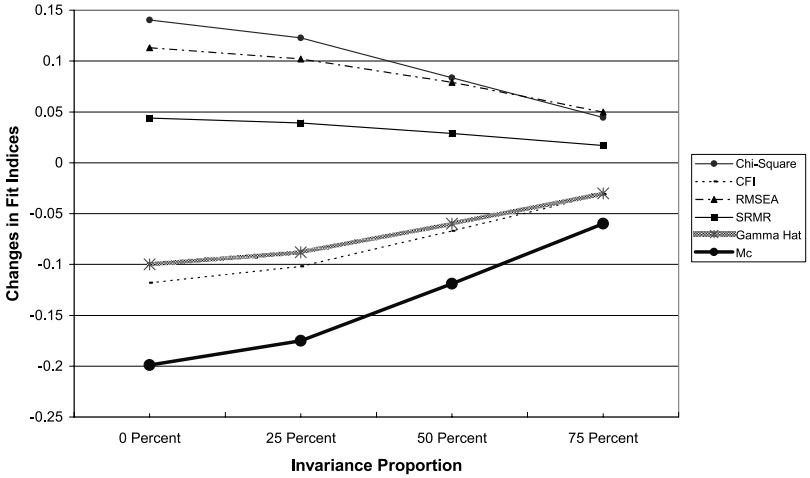

FIGURE 6 When lack of intercept invariance is mixed: Changes in fit indexes as a function of invariance proportion.

−.119, to −.060; and in chi-square difference statistics from 140.40, 122.87, 83.53, to 44.42 (see Figure 6).

Pattern of invariance. Changes in goodness of fit indexes and in the chi-square difference statistics were also affected by the pattern of invariance. The changes were larger when invariance was mixed than when it was uniform. For example, given an 8-indicator model with sample sizes of 150 and 150, as the invariance proportion varied from 0%, 25%, 50%, to 75%, changes in CFI varied from −.015, −.026, −.034, to −.025 when the pattern was uniform, whereas the change in CFI varied from −.118, −.102, −.067, to −.031 when the pattern was mixed. The same pattern was observed for RMSEA, SRMR, Gamma hat, Mc, and for the chi-square difference statistic as well.

Ratio of sample sizes. The sensitivity of fit indexes and chi-square difference statistics were affected by the ratio of sample size. Changes were larger when sample sizes were equal than when sample sizes were unequal. For example, given an 8-indicator model with a uniform pattern of invariance, as the invariance proportion varied from 0%, 25%, 50%, to 75%, the change in CFI varied from −.015, −.026, −.034, to −.025 when sample sizes were 150 and 150, whereas the change in CFI varied from −.010, −.015, −.021, to −.016 when sample sizes were 240 and 60. The same pattern was observed for RMSEA, SRMR, Gamma hat, Mc, and for the chi-square difference statistic as well. The same pattern was also found when invariance was mixed.

Sample size. Sample size did not have any appreciable impact on changes in CFI, Gamma hat, and Mc, although for RMSEA and SRMR, as sample size increased, changes were slightly increased. For example, given an 8-indicator model with a uniform pattern of invariance and sample sizes of 150 and 150, as the invariance proportion varied from 0%, 25%, 50%, to 75%, for a total sample size of 300 (vs. 500 vs. 1,000), changes in CFI varied from −.015 (vs. −.016 vs. −.016), −.026 (vs. −.026 vs. −.027), −.034 (vs. −.036 vs. −.036), to −.025 (vs. −.025 vs. −.026); changes in Gamma hat varied from −.016 (vs. −.016 vs. −.016), −.025 (vs. −.025 vs. −.025), −.033 (vs. −.033 vs. −.032), to −.024 (vs. −.024 vs. −.024); changes in Mc varied from −.032 (vs. −.032 vs. −.032), −.050 (vs. −.050 vs. −.049), −.065 (vs. −.065 vs. −.065), to −.048 (vs. −.048 vs. −.048). However, changes in RMSEA varied from .030 (vs. .035 vs. .040), .045 (vs. .048 vs. .052), .053 (vs. .058 vs. .061), to .042 (vs. .047 vs. .052) for a total sample size of 300 (vs. 500 vs. 1,000); changes in SRMR varied from .012 (vs. .012 vs. .014), .016 (vs. .017 vs. .020), .018 (vs. .020 vs. .023), to .014 (vs. .016 vs. .018). As expected, as sample size increased, the standard deviation of all fit indexes decreased.

Number of indicators. The changes were slightly larger in the 8-indicator model than in the 12-indicator model, particularly for RMSEA and Mc. For example, given a model with a uniform pattern of invariance and sample sizes of 150 and 150, as the invariance proportion varied from 0%, 25%, 50%, to 75%, the changes in CFI varied from −.015 (vs. −.009), −.026 (vs. −.022), −.034 (vs. −.030), to −.025 (vs. −.022) in the 8-indicator (vs. 12-indicator) cases; in SRMR from .012 (vs. .007), .016 (vs. .012), .018 (vs. .013), to .014 (vs. .010); and in Gamma hat from −.016 (vs. −.011), −.025 (vs. −.024), −.033 (vs. −.032), to −.024 (vs. −.024). However, in RMSEA the changes varied from .030 (vs. .017), .045 (vs. .033), .053 (vs. .040), to .042 (vs. .033); and in Mc they varied from −.032 (vs. −.033), −.050 (vs. −.071), −.065 (vs. −.095), to −.048 (vs. −.072). The standard deviations of all three fit indexes were slightly higher in the 8-indicator cases than in the 12-indicator cases.

Rejection Rate

The same factors that affected changes in fit statistics also affected rejection rates.

The interaction between the proportion of invariance and pattern of invariance. When noninvariance was uniform, the relation between the proportion of invariance and rejection rates is not monotonic. Specifically, the rejection rates were the smallest when the proportion of invariance was 0%, whereas the rejection rates were the highest when the proportion of invariance was 50%. In

contrast, when noninvariance was mixed, the relation between the proportion of invariance and rejection rates was monotonic.

Pattern of invariance. Rejection rates based on fit indexes and the chi-square difference statistics were higher when invariance was mixed than when it was uniform.

Ratio of sample sizes. Rejection rates based on fit indexes and the chi-square difference statistics were higher when sample sizes were equal than when sample sizes were unequal.

Sample size. As sample size increased, rejection rates based on fit indexes also tended to increase, although to a lesser degree compared to those based on chi-square difference tests.

Number of indicators. Number of indicators did not have any appreciable impact on rejection rates.

Type I and Type II errors. When relying on RMSEA, SRMR, Gamma hat, and Mc, there was a small chance of committing Type I errors in small samples. Type II errors tended to occur when the pattern of noninvariance was uniform, sample sizes were small, and sample sizes were unequal.

Summary

The sensitivity of goodness of fit indexes and of the chi-square difference statistics to lack of intercept invariance were affected by the interaction between the proportion of invariance and pattern of invariance, main effect of pattern of invariance, and ratio of sample sizes. When noninvariance was uniform, changes in fit indexes and the chi-square difference statistics were smallest when the proportion of invariance was 0%, whereas changes were largest when the proportion of invariance was 50%. In contrast, when noninvariance was mixed, the relation between proportion of invariance and changes in fit statistics was monotonic. Changes were larger when invariance was mixed than when it was uniform. Changes were also larger when sample sizes were equal than when they were unequal.

The same pattern was found in rejection rates. In addition, rejection rates based on fit statistics increased as sample size increased even though the mean changes in CFI, SRMR, Gamma hat, and Mc were relatively independent of sample size and changes in RMSEA were slightly dependent on sample size. This is because as sample size increased, the standard deviations of all fit indexes also decreased, which made it easier to reject a model with a larger sample size.

Lack of Invariance in Residual Variances

Changes in Goodness of Fit Indexes

As can be seen from Table 6, the sensitivity of goodness of fit indexes to lack of residual variance invariance was affected by a number of factors.

Proportion of invariance. Different from the findings in lack of loading or intercept invariance, the relation between the degree of noninvariance and changes in goodness of fit indexes as well as in the chi-square differences was monotonic regardless of whether noninvariance was uniform or mixed. For example, given an 8-indicator model with uniform invariance and sample sizes of 150 and 150, as the invariance proportion varied from 0%, 25%, 50%, to 75%, changes in CFI varied from −.029, −.022, −.014, to −.007; in RMSEA from .050, .041, .028, to .015; in SRMR from .014, .013, .009, to .007; in Gamma hat from −.034, −.026, −.016, to −.008; in Mc from −.068, −.052, −.033, to −.017; and in chi-square difference statistics from 50.39, 39.99, 27.90, to 18.07 (see Figure 7). Similar results were found when lack of residual invariance was mixed. For example, given an 8-indicator model with sample sizes of 150 and 150, as the invariance proportion varied from 0%, 25%, 50%, to 75%, changes in CFI varied from −.026, −.020, −.013, to −.006; in RMSEA from .047, .037, .026, to .013; in SRMR from .016, .013, .010, to .007; in Gamma hat from −.030, −.023, −.015, to −.007; in Mc from −.060, −.045, −.030, to −.015;

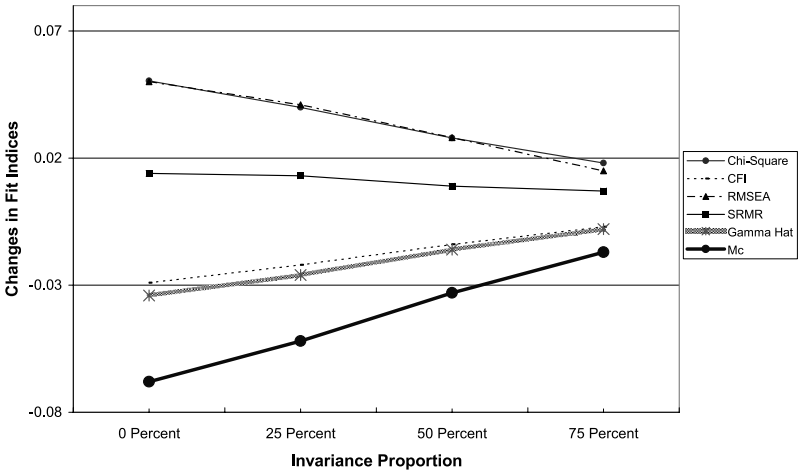

FIGURE 7 When lack of residual invariance is uniform: Changes in fit indexes as a function of invariance proportion.

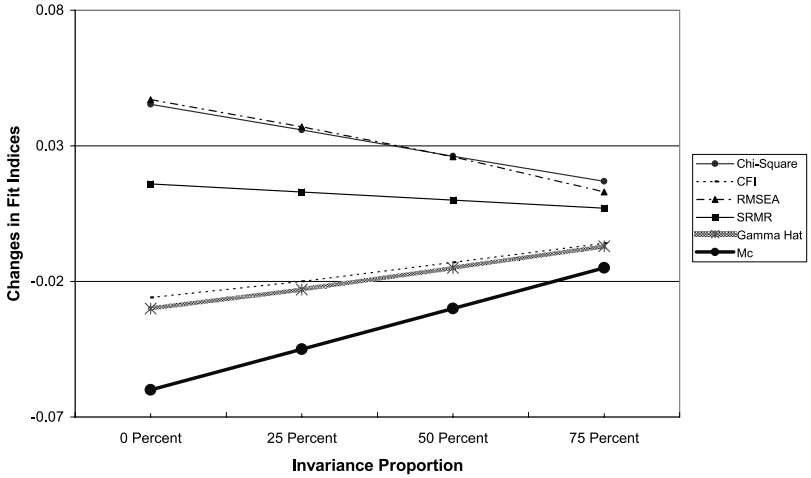

FIGURE 8 When lack of residual invariance is mixed: Changes in fit indexes as a function of invariance proportion.

and in chi-square difference statistics from 45.36, 35.83, 26.17, to 16.90 (see Figure 8).

Ratio of sample sizes. The sensitivity of fit indexes and the chi-square difference statistics was affected by the ratio of sample sizes. Changes were larger when sample sizes were equal than when sample sizes were unequal. For example, given an 8-indicator model with a uniform pattern of invariance, as the invariance proportion varied from 0%, 25%, 50%, to 75%, changes in CFI varied from −.029, −.022, −.014, to −.007 when sample sizes were 150 versus 150, whereas changes in CFI varied from −.018, −.013, −.009, to −.004 when sample sizes were 240 versus 60. The same pattern was observed for RMSEA, SRMR, Gamma hat, Mc, and the chi-square difference statistic. The same pattern was also found when noninvariance was mixed.

Sample size. Sample size did not have any appreciable impact on changes in CFI, SRMR, Gamma hat, and Mc, although for RMSEA, as sample size increased, changes were slightly increased. For example, given an 8-indicator model with a uniform pattern of invariance and sample sizes of 150 and 150, as the invariance proportion varied from 0%, 25%, 50%, to 75%, for a total sample size of 300 (vs. 500 vs. 1,000), changes in CFI varied from −.029 (vs. −.030 vs. −.032), −.022 (vs. −.022 vs. −.024), −.014 (vs. −.015 vs. −.015), to −.007 (vs. −.006 vs. −.007); changes in SRMR varied from .014 (vs. .016 vs. .018), .013 (vs. .013 vs. .015), .009 (vs. .010 vs. .011), to .007 (vs. .006 vs. .007); in

Gamma hat from −.034 (vs. −.034 vs. −.035), −.026 (vs. −.025 vs. −.025), −.016 (vs. −.016 vs. −.016), to −.008 (vs. −.008 vs. −.008); and in Mc from −.068 (vs. −.068 vs. −.069), −.052 (vs. −.050 vs. −.050), −.033 (vs. −.032 vs. −.032), to −.017 (vs. −.015 vs. −.016); however, changes in RMSEA varied from .050 (vs. .056 vs. .060), .041 (vs. .045 vs. .049), .028 (vs. .033 vs. .037), to .015 (vs. .018 vs. .023) for a total sample size of 300 (vs. 500 vs. 1,000). As expected, as sample size increased, the standard deviation of all three fit indexes decreased.

Number of indicators. Number of indicators did not have any appreciable impact on fit indexes except for Mc, for which the changes were smaller in the 8-indicator model than in the 12-indicator model. For example, given a model with a uniform pattern of invariance and sample sizes of 150 and 150, as the invariance proportion varied from 0%, 25%, 50%, to 75%, the changes in CFI varied from −.029 (vs. −.029), −.022 (vs. −.021), −.014 (vs. −.013), to −.007 (vs. −.007) in 8-indicator (vs. 12-indicator) cases; in RMSEA they varied from .050 (vs. .042), .041 (vs. .034), .028 (vs. .023), to .015 (vs. .013); in SRMR from .014 (vs. .011), .013 (vs. .009), .009 (vs. .007), to .007 (vs. .005); and in Gamma hat from −.034 (vs. −.036), −.026 (vs. −.027), −.016 (vs. −.017), to −.008 (vs. −.009). However, the changes in Mc varied from −.068 (vs. −.106), −.052 (vs. −.080), −.033 (vs. −.050), to −.017 (vs. −.025). The standard deviations of all three fit indexes were slightly higher in the 8-indicator cases than in the 12-indicator cases.

Rejection Rate

The same factors that affected changes in fit statistics also affected rejection rates.

Proportion of invariance. The relation between proportion of invariance and rejection rate was monotonic regardless of whether the pattern of invariance was uniform or mixed. Specifically, the rejection rates were the smallest when the proportion of invariance was 0%, whereas the rejection rates were highest when the proportion of invariance was 75%.

Ratio of sample sizes. Rejection rates based on fit indexes and the chi-square difference statistics were higher when sample sizes were equal than when sample sizes were unequal.

Sample size. As sample size increased, rejection rates based on fit indexes also tended to increase, although to a lesser degree compared to those based on chi-square difference tests.

Number of indicators. Number of indicators did not have any appreciable impact on rejection rates.

Type I and Type II errors. When relying on RMSEA, Gamma hat, and Mc, there was a small chance of committing Type I errors in small samples, but the chance was much higher when relying on SRMR.

Summary

The sensitivity of goodness of fit indexes and of the chi-square difference statistics to lack of residual variance invariance was affected by the proportion of invariance and ratio of sample sizes. Different from the results found in testing lack of loading or intercept invariance, changes in fit indexes and the chi-square difference statistics were smallest when the proportion of invariance was the lowest, and changes were the largest when the proportion of invariance was the highest regardless of the invariance pattern. Also different from the findings of lack of loading or intercept invariance, the magnitude of changes was similar regardless of the pattern of lack of invariance. Changes were larger when sample sizes were equal than when they were unequal.

The same pattern was found in rejection rates. In addition, rejection rates based on fit statistics increased as sample size increased even though the mean changes in CFI, SRMR, Gamma hat, and Mc were relatively independent of sample size and changes in RMSEA were only slightly dependent on sample size.

DISCUSSION

Measurement invariance has been increasingly tested when comparing different groups (Vandenberg & Lance, 2000), as it is important to ensure that groups are compared based on instruments that measure the same constructs. However, an important question has yet to be answered: What statistical criteria should be used to evaluate measurement invariance? Two Monte Carlo studies were conducted to fill in this gap. Study 1 examined random variations of three commonly used goodness of fit indexes (i.e., CFI, RMSEA, and SRMR) and two promising indexes (i.e., Gamma hat and Mc), under three commonly tested invariance conditions. Results indicate that SRMR appears to be more sensitive to noninvariance in loadings than in intercepts or residual variances. Cheung and Rensvold (2002) did not examine SRMR, which prevented comparison of the results from the two studies. However, the performance of CFI, RMSEA, and Mc were consistent across the two studies. That is, these three indexes appear to be equally sensitive to all three types of invariance tests. In addition, Gamma hat is also equally sensitive to all three types of invariance tests. Therefore, different cutoff points were proposed for SRMR across different levels of invariance tests,

whereas the same cutoff points were recommended for CFI, RMSEA, Gamma hat, and Mc across different levels of invariance tests.

Study 2 tested the sensitivity of fit indexes under varying degrees of invariance. The most intriguing finding is that when testing invariance at the factor loading or intercept level, changes in fit indexes were affected by the interaction between the pattern of invariance and degree of invariance. When noninvariance was uniform, the relation between the degree of invariance and changes in fit statistics was nonmonotonic: When the degree of noninvariance was the highest, changes were the smallest, whereas when the degree of noninvariance was only moderate, changes were the largest. In contrast, when noninvariance was mixed, the relation between the degree of invariance and changes in fit statistics was monotonic: When the degree of noninvariance was the highest, changes were the largest, and when the degree of noninvariance was the lowest, changes were the smallest. The pattern of invariance also affected changes in fit statistics: Changes were larger when the pattern of invariance was mixed than when it was uniform.

The picture was quite different when testing invariance at the residual variance level. The relation between changes in fit statistics and degrees of invariance was monotonic regardless of the pattern of invariance. Also different from the findings in testing lack of loading invariance or intercept invariance, the pattern of invariance did not affect the magnitude of changes.

Unequal sample sizes affected changes across all three levels of invariance: Changes were larger when sample sizes were equal than when they were unequal. Unequal sample size is a common issue when different cultural or ethnic groups are involved. Given that changes in fit statistics were reduced when sample sizes were unequal, invariance tests are more likely to fail to detect noninvariance.

The same factors affecting the changes in fit indexes also affected rejection rates based on fit indexes. In addition, sample size had an undesirable effect on rejection rates based on fit indexes. That is, as sample size increased, rejection rates based on fit indexes tended to increase, although to a lesser degree compared to rejection rates based on the chi-square difference tests. The average values in ΔCFI, ΔSRMR, ΔGamma hat, and ΔMc were relatively independent of sample size, which is consistent with previous findings that CFI (Bentler, 1990), SRMR (Hu & Bentler, 1998), Gamma hat (Hu & Bentler, 1998), and Mc (McDonald, 1989) were less sensitive to sample size. However, standard errors of all fit indexes decreased as sample size increased, which led to higher rejection rates. Hu and Bentler (1999)[10] also reported rejection rates across different sample sizes based on fit indexes, although that study focused on performance

[10]Marsh, Hau, and Wen (2004) issued warnings on overgeneralizing Hu and Bentler's (1999) findings, and stressed the importance of interpreting their results with caution. However, Marsh et al. also stated that all goodness of fit indexes are more appropriate in comparing nested models than in establishing the absolute fit of a model.

of fit indexes under model misspecification in single group cases, rather than changes in fit indexes in model comparisons. Their results did not show a clear pattern regarding the relation of rejection rates to sample sizes. One possible reason might be that the rejection rates were pooled together across seven different distributional assumptions, which might have obscured the patterns.

These studies uncovered the great complexity of testing measurement invariance and have important implications in substantive research. When groups are compared based on measures that lack measurement invariance, test statistics such as means and regression coefficients can be biased or invalid. Consequently, we may miss true group differences that have been masked by the artifacts of measurements or we may discover pseudo group differences that are in fact due to these artifacts. For example, a series of simulation studies have been conducted to examine the impact of lack of factor loading invariance on regression slope and mean comparisons (Chen, 2007). The results indicate that when Group 1 has higher loadings in the predictor, the regression slope is underestimated in Group 1 but overestimated in Group 2, and consequently, the regression slope is found higher in Group 2 than in Group 1, when in fact there is no group difference. The opposite pattern is found when Group 1 has higher loadings in the criterion. Spurious gruop mean differences were also found in that study.

It is expected that the degree of bias in group comparisons should directly correspond to the degree of noninvariance in an instrument, which in turn is expected to correspond to changes in fit statistics when testing measurement invariance. In other words, sensitivity of measurement invariance tests should directly reflect the degree of invariance and potential bias in group comparisons. However, the story is complicated when examining the relation among testing invariance at the factor loading or intercept level, degree of noninvariance, and degree of bias in group comparisons. On the one hand, the relation between the degree of invariance and changes in fit statistics was nonmonotonic; on the other hand, the relation between the degree of invariance and bias in group comparisons was monotonic (Chen, 2007). As a result, when the degree of noninvariance is the largest, invariance tests may indicate that a noninvariant instrument is invariant. Consequently, invalid group comparisons may be made, as the invariance tests have failed to detect noninvariance due to the fact that changes in fit indexes are the smallest when the degree of noninvariance is the largest. In contrast, when the degree of noninvariance is minimal, invariance tests may indicate that a scale is noninvariant. Consequently, valid group comparisons may be avoided, as the invariance tests have falsely alarmed that there is noninvariance due to a mixed pattern of invariance or large sample sizes. Similarly, there is a lack of corresponding relation among sensitivity of invariance tests, the degree of noninvariance, and bias in group comparisons when lack of loading or intercept invariance was mixed. On the one hand, invariance tests are more likely to indicate there is noninvariance when noninvariance was

mixed than when it was uniform; on the other hand, bias was minimized when noninvariance was mixed (Chen, 2007).

Recommendations

Given the complex relations between the sensitivity of measurement tests invariance, degrees of noninvariance, and potential bias in group comparisons, it is difficult to propose statistical standards for testing measurement invariance. However, testing invariance is the first step in group comparisons, and some guidelines derived from these two studies may still be useful to substantive researchers. Cutoff points based on the three routinely used fit indexes (i.e., CFI, RMSEA, and SRMR) are recommended for evaluating invariance at the three commonly tested levels. Although ΔGamma performed as well as did ΔCFI, it is highly correlated with ΔCFI ($r = .97$, .98, and .95 at the loading, intercept, and residual invariance level, respectively).[11] Given that CFI is more widely used than Gamma hat, CFI was chosen over Gamma hat. ΔMc did not outperform ΔRMSEA or ΔSRMR, and therefore was not recommended. When sample size is small (total $N \leq 300$), sample sizes are unequal, and the pattern of noninvariance is uniform, the following cutoff criteria are suggested: for testing loading invariance, a change of $\leq -.005$ in CFI, supplemented by a change of $\geq .010$ in RMSEA or a change of $\geq .025$ in SRMR would indicate noninvariance; for testing intercept or residual invariance, a change of $\geq -.005$ in CFI, supplemented by a change of $\geq .010$ in RMSEA or a change of $\geq .005$ in SRMR would indicate noninvariance. In other words, similar values are suggested for CFI and RMSEA across all three levels of invariance tests, but different values are proposed for SRMR, as SRMR is more sensitive to noninvariance in loadings than to noninvariance in intercepts or residual variances. When sample size is adequate (total $N > 300$) and sample sizes are equal across the groups, particularly when lack of invariance is mixed, more stringent criteria are suggested. For testing loading invariance, a change of $\geq -.010$ in CFI, supplemented by a change of $\geq .015$ in RMSEA or a change of $\geq .030$ in SRMR would indicate noninvariance; for testing intercept or residual invariance, a change of $\geq -.010$ in CFI, supplemented by a change of $\geq .015$ in RMSEA or a change of $\geq .010$ in SRMR would indicate noninvariance. Among the three indexes, CFI was chosen as the main criterion because RMSEA and SRMR tend to overreject an invariant model when sample size is small, particularly when using SRMR for testing loading or residual variance invariance. In addition, changes in RMSEA are more likely to be affected by sample size and model complexity.

[11] Gamma hat also performed similarly as did CFI in Hu and Bentler's (1998) study, and the correlation between the two indexes was .985.

However, these criteria should be used with caution, because testing measurement invariance is a very complex issue. As uncovered in this investigation, a number of factors can affect the magnitude of changes in fit statistics, such as pattern of noninvariance, sample size, ratio of sample size, and model complexity. Furthermore, as in any simulation studies, the findings are limited to the conditions investigated in the studies reported here. To conclude, as Bollen and Long (1983) pointed out, "The test statistics and fit indices are very beneficial, but they are no replacement for sound judgment and substantive expertise" (p. 8). Similar views have been expressed by many other researchers (e.g., Browne & Cudeck, 1993; Byrne, 2001; Marsh et al., 1988; Steiger & Lind, 1980).

ACKNOWLEDGMENTS

I would like to thank Kristopher Preacher and Donna Coffman for their thoughtful comments on an earlier version of this article. I also express appreciation to Stephen West and Roger Millsap for their insights on measurement invariance. I am grateful to the Quantitative Forum in the Psychology Department at the University of North Carolina at Chapel Hill for fruitful discussions at the early stages of this work.

REFERENCES

Bentler, P. M. (1990). Comparative fit indices in structural equation models. *Psychological Bulletin, 107,* 238–246.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software, Inc.

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

Bollen, K. A., & Long, J. S. (Eds.). (1983). *Testing structural equation models.* Newbury Park, CA: Sage.

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

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

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications and programming.* Mahwah, NJ: Lawrence Erlbaum Associates.

Byrne, B. M., & Campbell, T. L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: A look beneath the surface. *Journal of Cross-Cultural Psychology, 30,* 555–574.

Chen, F. F. (2007). What happens if we compare chopsticks with forks? The Impact of making inappropriate comparisons in cross-cultural research. Manuscript under review.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling, 12,* 471–492.

Chen, F. F., & West, S. G. (in press). Measuring individualism and collectivism: The importance of considering different components, reference groups, and measurement invariance. *Journal of Research in Personality.*

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255.

Hendriks, A. J., Perugini, M., Angleitner, A., Ostendorf, F., Johnson, J. A., De-Fruyt, F., et al. (2003). The Five-Factor Personality Inventory: Cross-cultural generalizability across 13 countries. *European Journal of Personality, 17,* 347–373.

Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist, 4,* 179–188.

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

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8: User's reference guide* (2nd ed.). Chicago: Scientific Software International.

Kaplan, D., & George, R. (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling, 2,* 101–118.

Kwan, V. S. Y., Bond, M. H., & Singelis, T. M. (1997). Pancultural explanations for life satisfaction: Adding relationship harmony to self-esteem. *Journal of Personality and Social Psychology, 73,* 1038–1051.

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

Marsh, H. W., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical properties. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 315–353). Mahwah, NJ: Lawrence Erlbaum Associates.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: Effects of sample size. *Psychological Bulletin, 103,* 391–411.

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu & Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320–341.

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.

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling, 11,* 60–72.

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

Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection of multiple populations. *Psychological Methods, 9,* 93–115.

Muthén, L. K., & Muthén, B. O. (1998). *M*plus *user's guide.* Los Angeles: Muthén & Muthén.

Rhee, E., Uleman, J. S., & Lee, H. K. (1996). Variations in collectivism and individualism by ingroup and culture: Confirmatory factor analysis. *Journal of Personality and Social Psychology, 71,* 1037–1054.

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: A supplementary module for SYSTAT and SYGRAPH.* Evanston, IL: SYSTAT.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173–180.

Steiger, J. H. (1998). A note on multiple sample extensions of the RMSEA fit index. *Structural Equation Modeling, 5,* 411–419.

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

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

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

Ullman, J. B. (2001). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using Multivariate Statistics* (pp. 653–771). Boston: Allyn and Bacon.

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, 2,* 4–69.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

APPENDIX
Model Parameters in Simulation Study 1 and Study 2

	Study 1			*Study 2*		
		Nonmarker Indicators			*Nonmarker Indicators*	
	Marker Indicator	*Group 1*	*Group 2*	*Marker Indicator*	*Group 1*	*Group 2*
Loadings	1	.9	.9	1	.9	.5
Intercepts	0	1	1	0	1	.6
Residuals		.6	.6		.4	.6
Variance		.8	.8		.8	.8
Latent mean		5	5		5	5

Note. For Study 1, the values were the same for all nonmarker indicators in both Group 1 and Group 2. For Study 2, the values were the same for all nonmarker indicators in Group 1, but in Group 2, the number of indicators that takes a different value depends on the proportion of invariance.