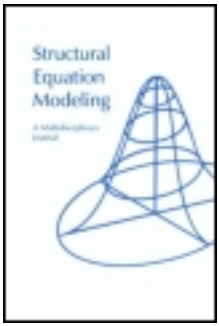

Structural Equation Modeling: A Multidisciplinary Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hsem20

Confirmatory Factor Analytic Procedures for the Determination of Measurement Invariance

Brian F. French & W. Holmes Finch

Available online: 19 Nov 2009

To cite this article: Brian F. French & W. Holmes Finch (2006): Confirmatory Factor Analytic Procedures for the Determination of Measurement Invariance, Structural Equation Modeling: A Multidisciplinary Journal, 13:3, 378-402

To link to this article: http://dx.doi.org/10.1207/s15328007sem1303_3

STRUCTURAL EQUATION MODELING, *13*(3), 378–402

Confirmatory Factor Analytic Procedures for the Determination of Measurement Invariance

Brian F. French
Purdue University

W. Holmes Finch
Ball State University

Confirmatory factor analytic (CFA) procedures can be used to provide evidence of measurement invariance. However, empirical evaluation has not focused on the accuracy of common CFA steps used to detect a lack of invariance across groups. This investigation examined procedures for detection of test structure differences across groups under several conditions through simulation. Specifically, sample size, number of factors, number of indicators per factor, and the distribution of the observed variables were manipulated, and 3 criteria for assessing measurement invariance were evaluated. Power and Type I error were examined to evaluate the accuracy of detecting a lack of invariance. Results suggest that the chi-square difference test adequately controls the Type I error rate in nearly all conditions, and provides relatively high power when used with maximum likelihood (ML) estimation and normally distributed observed variables. In addition, the power of the test to detect group differences for dichotomous observed variables with robust weighted least squares estimation was generally very low.

The measurement of underlying constructs, such as intellectual ability, psychological states (depression, anxiety, etc.), and attitudes, serves an important role in society, especially when test scores claiming to measure these abilities are used for high-stakes decisions in a variety of environments (e.g., educational, workplace). A sharp increase in the use of test scores for such decisions has been observed (Brennan, 2004), especially in relation to the No Child Left Behind Act of 2001

Correspondence should be addressed to Brian F. French, Department of Educational Studies, BRNG, Purdue University, West Lafayette, IN 47907–2098. E-mail: frenchb@purdue.edu

(NCLB; PL. 107–110). Thus, the statistical properties of tests must meet current validity standards (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999) to overcome both legal and technical challenges from the research community and the general public (Brennan, 2004).

The examination of measurement invariance (i.e., the extent to which items or subtests have equal meaning across groups of examinees) is one component used to gather score validity evidence and to evaluate construct-irrelevant variance (e.g., group membership). Note that issues of measurement invariance are not limited to cognitive tests, and use of the term *examinee* here refers to an individual who responds to any type of instrument. An examinee's score should not depend on construct-irrelevant variance. When decisions are made for individuals in the absence of measurement invariance, the decision maker risks committing a serious error (Bollen, 1989), as observed score differences can reflect (a) true group mean differences, and (b) differences in the relation between the construct and the observed score that is not equivalent across groups (Raju, Laffitte, & Byrne, 2002). Thus, to avoid undesirable social consequences, the measurement process must keep irrelevant variables from influencing scores (Messick, 1989) and employ methods to determine the extent to which scores are influenced by such variables (e.g., Standard 7.10, AERA et al., 1999). See Haladyna and Downing (2004) for a discussion of this topic.

Several degrees of measurement invariance have been defined (e.g., Little, 1997; Meredith, 1993; Millsap, 2005). For instance, Little (1997) illustrated two hierarchical levels of measurement invariance. The first level requires that the psychometric properties of an instrument be equivalent (i.e., configural, metric, measurement error, and scalar invariance; see Bollen, 1989; Horn & McArdle, 1992; Jöreskog, 1971; Meredith, 1993; Thurstone, 1947) before the second level, which includes group differences in latent means and covariances (e.g., Sörbom, 1971), is appropriately examined and results interpreted. A common method for examining these levels of measurement invariance is multisample confirmatory factor analysis (MCFA), which allows for testing an a priori theory of the test structure across groups (Alwin & Jackson, 1981; McGaw & Jöreskog, 1971) or across time (e.g., developmentally related questions; Mantzicopoulos, French, & Maller, 2004). Such an approach allows for the comparison of specific features of the factor model from one group to another. To the extent that these features are found to be equivalent across groups, the researcher can infer measurement invariance, and more specifically factorial invariance.

MCFA for invariance testing has experienced substantial growth recently (Vandenberg & Lance, 2000). Much of this development has focused on such issues as (a) goodness-of-fit indexes (Cheung & Rensvold, 2002), (b) analysis of ordinal data (Flora & Curran, 2004; Lubke & Muthén, 2004; Millsap & Yun-Tein, 2004), (c) appropriate practices for testing invariance (Vandenberg & Lance,

2000), (d) computer programs to make the process less time-intensive (e.g., Rensvold & Cheung, 1998), and (e) steps for testing invariance and latent mean structures (e.g., Bollen, 1989; Byrne, Shavelson, & Muthén, 1989). Recent recommendations urge practitioners not to assume that procedures are accurate under all circumstances and to continue to evaluate these procedures (Vandenberg, 2002). The latter recommendation is beginning to be followed with the implementation of simulation studies (e.g., Meade & Lautenschlager, 2004). However, continued evaluation of these methods is required as there are only a handful of studies that have empirically examined MCFA procedures for invariance testing. For instance, Meade and Lautenschlager (2004) pointed out that their study was the first to examine MCFA procedures through simulation. Furthermore, many questions remain about the procedures (e.g., accuracy, influence of partial invariance, influence of the selected reference indicator, etc.; Millsap, 2005; Vandenberg, 2002).

Evidence suggests that MCFA invariance tests appear to perform well under ideal conditions such as with large sample sizes and sufficient indicator variables (Meade & Lautenschlager, 2004). However, as Meade and Lautenschlager (2004) stated, some results were unexplainable and quite puzzling. For instance, in one condition in their study where 66% of factor loadings differed (with .25 of a difference for factor loadings) across groups, only factor variances were found to differ significantly. Additionally, the models and conditions generated were not complex (e.g., one factor, normal data, ML estimation). Such results suggest that evidence is needed to examine the performance of MCFA under less than ideal conditions, especially given that many models and data seen in practice are more complex.

One of these more complex circumstances is the analysis of ordinal variables (e.g., rating scale data) in MCFA. A common approach in practice is to use methods designed for continuous data when analyzing such variables. However, treatment of ordinal data as continuous in MCFA (a) violates multivariate normality, (b) may distort a factor structure across groups, and (c) could lead to inaccurate measurement invariance tests (Lubke & Muthén, 2004). A possible solution is the use of robust weighted least squares (RWLS) estimation. This methodology is based on work by Muthén, du Toit, and Spisic (1997), among others, which improved on the standard weighted least squares (WLS) approach to parameter estimation, which has been shown to work well in certain conditions for the CFA context using ordinal data.

Whereas WLS has been found to perform poorly when sample sizes are relatively small, RWLS does not appear to have as many problems (Flora & Curran, 2004; Jöreskog & Sörbom, 1996), and is thus preferred in most situations in which categorical indicators are used in CFA. The RWLS approach to model parameter estimation does not require the inversion of the weight matrix used in the standard WLS approach, which in turn leads to greater stability even when samples are as small as 100 (Flora & Curran) with dichotomous and five-category responses. In addition, Muthén et al. (1997) introduced a mean and variance adjusted chi-square

goodness-of-fit test, and an accompanying difference test that can be conducted using the M*plus* software package (Muthén & Muthén, 2004). Evaluation of MCFA invariance procedures with RWLS has not been conducted to the authors' knowledge and requires examination as this estimation method is anticipated to increase in use given recent results and suggestions (e.g., Flora & Curran, 2004; Millsap, 2005).

Beyond the assessment of MCFA procedures under more complex conditions is the issue of selecting the most appropriate goodness-of-fit index (GFI) to detect a lack of invariance. Measurement invariance testing involves the comparison of increasingly more restricted models by sequentially constraining various matrices to be equal across groups. A significant decline in fit between models indicates differences between groups in the constrained matrix. The decline in fit is indicated by a significant change in the likelihood ratio test (i.e., $\chi^2_{difference}$) between models. The dependence on the $\chi^2_{difference}$ test is most likely due to alternative GFIs not having sampling distributions (Cheung & Rensvold, 2002). However, differences in other GFIs have been suggested (Cheung & Rensvold, 2002; Rensvold & Cheung, 1998; Vandenberg & Lance, 2000), due to inflation of the χ^2 goodness-of-fit test with large samples, which in turn, they argue, may influence the $\chi^2_{difference}$ test (Brannick, 1995; Kelloway, 1995). As Brannick (1995) and Cheung and Rensvold (2002) pointed out, using several indexes to judge model fit yet only employing the $\chi^2_{difference}$ test to determine differences in nested models creates a double standard. However, chi-square may not be inflated with large samples when the model is properly specified. See Bollen (1989, 1990) for a detailed explanation of the influences of sample size on measures of model fit.

Evaluation of differences in many fit indexes may be superior to the $\chi^2_{difference}$ test, as these indexes are not influenced by sample size (Cheung & Rensvold, 2002). Cheung and Rensvold examined 20 GFIs and recommended 3 (change in comparative fit index [CFI], gamma hat, and McDonald's noncentrality index) for invariance testing, as these were relatively robust to small errors of approximation. Of the indexes recommended, change in the CFI is noteworthy as it is (a) the only one of the three available in most structural equation modeling programs (e.g., M*plus,* LISREL), (b) commonly reported in the literature, and (c) the recommended index to report showing, tentatively at least, promise in identifying noninvariance in the data (Cheung & Rensvold, 2002; Rensvold & Cheung, 1998). However, the CFI's performance has not been evaluated (a) under various conditions (e.g., nonnormal data), (b) in combination with the $\chi^2_{difference}$ test, and (c) in more than one empirical evaluation to the knowledge of the authors. Furthermore, the need for evaluation of this index for the detection of a lack of invariance, particularly in regard to power, has been suggested (Chen, Sousa, & West, 2005). Thus, given this combination of promise in detecting group differences, wide availability in standard statistical software, and a lack of examination in a variety of conditions, the performance of the CFI difference is one focus of this study.

MCFA procedures for detecting a lack of measurement invariance require further examination to evaluate their accuracy. For instance, uncertainty remains as to (a) which GFI results in the most accurate detection of a lack of invariance, (b) how a nonnormal distribution of items and estimation procedures influences detection, and (c) how methods function with complex models. This investigation aims to provide evidence for methodological issues in invariance testing with the goal of informing practice. Specifically, the purpose of the article is to evaluate the recovery of simulated test structure differences (i.e., detection of a lack of invariance) across conditions and classification criteria through the examination of power and Type I error rates.

METHOD

Simulated data were employed to control factors that can influence detection of test structure differences. Replications (N = 1,000) were conducted for each combination of conditions to ensure stable results. Data were simulated with known model differences across two groups under varying conditions including sample size, normality of the observed variables, number of factors and indicators per factor, and criteria for determining a lack of invariance. Simulations were completed in M*plus* 3.11 (Muthén & Muthén, 2004).

Number of Factors and Indicators

Data were simulated from both two- and four-factor models, with interfactor correlations set at .50 to represent moderately correlated factors. Correlations were not varied to minimize unwanted confounds. There were two conditions of the number of indicators per factor (three and six), resulting in three test lengths (two-factor model, 6 and 12 indicators; four-factor model, 12 and 24 indicators). Model parameters were selected to reflect real data, following procedures used by Cheung and Rensvold (2002) with the intent of accurately reflecting models seen in practice, yet keeping the simulations manageable.

Sample Size

The necessary sample size to obtain adequate power in factor analysis varies depending on the data conditions (e.g., level of communalities; see MacCallum, Widaman, Zhang, & Hong, 1999). Additionally, groups often are not of equal size, especially with invariance studies involving low-incidence populations (e.g., deaf examinees; Maller & French, 2004). Therefore, two individual group sample sizes were employed—150 and 500—resulting in three sample size combinations: 150/150, 150/500, and 500/500, which will be referred to as N = 300, 650, and 1,000,

respectively. These sample sizes are consistent with previous simulation research examining measurement invariance issues (e.g., Cheung & Rensvold, 2002; Lubke & Muthén, 2004; Meade & Lautenschlager, 2004) and represent a range of the number of participants available to practitioners working in a variety of settings where the number of available participants might vary a great deal (e.g., educational, industrial/organizational, psychological).

Percentage of Invariance

Three levels of the amount of factor loading differences across groups were simulated. To assess Type I error (i.e., false identification of a lack of invariance) of the MCFA methods employed, the case of complete invariance (i.e., no differences in loadings across groups) was simulated. In addition, to assess power (i.e., correct identification of a lack of invariance) 17% (i.e., low contamination condition) and 33% (i.e., high contamination condition) of the factor loadings differed across groups. These percentages were selected (a) to reflect what may be found in actual test data, and (b) for practical reasons (i.e., resulted in a whole number of differing loadings).

Model Simulation

Data were generated, using the M*plus* 3.11 (Muthén & Muthén, 2004) Monte Carlo command, to represent data at the subtest level (i.e., normal distributions) and the item level (i.e., dichotomous data) to assess performance under two conditions commonly found in practice. Data with known covariance matrices, factor loadings, and interfactor correlations were generated. Group 1 data represented the initial model, whereas Group 2 differed on specified factor loadings. All other aspects of the factor model were constant between the groups, including factor variances and covariances. Factor loadings (i.e., lambdas) were set at .60 and factor variances (i.e., Phi) were constrained to 1.0, as is common in previous simulation work and for identification purposes (Hu & Bentler, 1998; Meade & Lautenschlager, 2004; Millsap, 2005). Note that if variances are constrained to 1.0 when not actually equal across groups, contaminated invariance results can occur (Cheung & Rensvold, 1999). However, this was not a concern here as variances were simulated to be equal across groups. The error terms (i.e., theta-deltas) for the observed variables can be comprised of two components: specific variance and error variance. In the simulated models, the assumption was made that there was no specific variance (i.e., all error variance) and the value of the theta-deltas was 1.0 minus the square of the factor loadings. This minimized potential confounding factors in examining the results and was not central to the questions of interest in this study.

For the dichotomous data, thresholds were not manipulated but rather were held constant at .25 and kept invariant to meet identification conditions for dichotomous variables (Millsap, 2005). The difference in the factor loadings for the noninvariant variables for conditions was .25 (i.e., loadings = .85) and is consistent with previous simulation work with CFA and invariance (Meade & Lautenschlager, 2004). All differences were unidirectional (i.e., favored one group). Note that a .25 difference with the dichotomous data approximately corresponds to a moderate difference (.60; Swaminathan & Rogers, 1990) in the item response theory (IRT) *a*-parameter (i.e., discrimination parameter) in a differential item functioning framework. See Lord and Novick (1968), Muthén and Lehman (1985), and Thissen, Steinberg, and Wainer (1993) for a discussion on the relation between the IRT and CFA parameters. ML estimation and RWLS estimation were used for the normal and dichotomous data, respectively.

Invariance Testing Criteria and Analysis

Data were analyzed using M*plus* 3.11 (Muthén & Muthén, 2004). Two models were evaluated for a lack of measurement invariance. The first model (Model 1) was the baseline model for obtaining the first chi-square and CFI values for comparison with more constrained models. The baseline models were properly specified and results are based on such models. Thus, as misspecification was not a condition in the simulation, results cannot be generalized to such situations. For Model 1 all parameters for Groups 1 and 2 were estimated simultaneously and were free to vary across groups. Model 2 constrained the matrix of factor loadings across groups to be equal. The resulting chi-square and CFI values from Model 2 were then used with the Model 1 values to create the difference tests to evaluate invariance. The final step in the analysis was to evaluate single indicator invariance detection. In practice, this step would follow an initial determination of a lack of invariance in the lambda matrix. This involves the investigation, indicator by indicator, of factor loadings to determine where exactly this lack of invariance occurs. In this study, the decision was made to compare indicator loadings one by one after an initial finding of invariance (see Millsap & Tien, 2004, for a discussion of evaluation strategies).

Three tests of invariance were conducted, one for each invariance classification criterion. The criteria were the (a) chi-square difference test ($\chi^2_{difference}$), (b) CFI difference ($CFI_{difference}$), and (c) the combination of the two indexes (i.e., $\chi^2_{difference}$ and $CFI_{difference}$). A lack of invariance is indicated with a statistically significant $\chi^2_{difference}$ (i.e., $p < .05$) and a $CFI_{difference}$ less than –.01 (Cheung & Rensvold, 2002). An alpha level of .01 for the χ^2 also was evaluated. In the combined criteria condition, models were considered different across groups when both criteria were met. The analysis was constrained to testing metric invariance (equality of factor loadings), which is the most critical concern regarding construct validity (Keith,

1997), because factor loadings indicate the relation between the variable and factor. Additionally, the CFI may be relatively insensitive to certain aspects of invariance testing, mainly mean structures (Chen et al., 2005). Type I error and power were evaluated to determine accuracy of the criteria for detection of a lack of measurement invariance.

RESULTS

Normally Distributed Data

Type I error rates. Table 1 displays the Type I error rates for each statistic across levels of the manipulated factors. The $\chi^2_{difference}$ criterion appears to maintain the nominal Type I error rate at both the .05 and .01 levels across all factors. In contrast, the $CFI_{difference}$ has an inflated Type I error rate in several of the conditions, most notably with two factors, 300 participants, or three indicators per factor. The combined criterion appeared to control Type I error rates better than the single criterion. A more detailed demonstration of the Type I error rate by combinations of factors appears in Table 2.

Table 2 illustrates that the Type I error rate of the $CFI_{difference}$ tends to vary a great amount. Indeed, in the two-factor case when the sample size is 300 and the number of indicators is three per factor, the actual Type I error rate is 10 times the nominal level. Alternatively, in the majority of cases, when the sample size is 650 or 1,000 the Type I error rate is just above or below .01. In contrast, it appears that the chi-square statistics are both consistently near the nominal values. The only real elevation in these rates (and this elevation is not very great) occurs for the small sample size when there are three indicators per factor. The combination cri-

TABLE 1
Type I Error Rates for Chi-Square and CFI by Number of Factors, Number of Participants, and Number of Indicators per Factor

Variable	$\chi^2_{difference.05}$	$\chi^2_{difference.01}$	$CFI_{difference}$	$\chi^2_{difference.05}$ & $CFI_{difference}$
Factors				
2	0.055	0.011	0.045	0.007
4	0.054	0.012	0.020	0.007
Participants				
300	0.059	0.011	0.076	0.009
650	0.055	0.013	0.016	0.008
1,000	0.049	0.010	0.005	0.004
Indicators per factor				
3	0.056	0.011	0.044	0.008
6	0.053	0.013	0.021	0.006

TABLE 2
Type I Error by Number of Factors (F), Participants (N), and Number of Indicators per Factor (I)

F	*N*	*I*	$\chi^2_{difference.05}$	$\chi^2_{difference.01}$	$CFI_{difference}$	$\chi^2_{difference.05}$ & $CFI_{difference}$
2	300	3	0.070	0.010	0.130	0.009
		6	0.051	0.007	0.070	0.005
	650	3	0.049	0.012	0.037	0.010
		6	0.055	0.015	0.014	0.013
	1,000	3	0.048	0.008	0.016	0.005
		6	0.054	0.013	0.002	0.002
4	300	3	0.060	0.011	0.067	0.008
		6	0.056	0.017	0.036	0.015
	650	3	0.063	0.014	0.013	0.008
		6	0.053	0.012	0.001	0.001
	1,000	3	0.045	0.008	0.0	0.0
		6	0.049	0.011	0.0	0.0

terion, on average, resulted in lower Type I errors than the $\chi^2_{difference}$ and the $CFI_{difference}$ criteria when used alone.

To examine the effects of the manipulated factors on the error rates, a variance components analysis was conducted treating the error rates for each criterion as the dependent variable and the manipulated factors as the independent variables. This analysis allows for an examination of the proportion of variance in the dependent variable accounted for by the manipulated factors. Percentages above 10% are reported. Results indicate that for both $\chi^2_{difference}$ statistics the interaction of number of factors, indicators per factor, and sample size accounted for more than 65% of the variation in Type I error rates, and approximately 20% was accounted for by the total sample size main effect for both test statistics. In the case of the $CFI_{difference}$, approximately 91% of variance in the Type I error rate was due to the sample size. For the combination of $CFI_{difference}$ and $\chi^2_{difference}$, the interaction of the number of indicators per factor by number of factors by sample size (73%) and the number of indicators per factor (12%) accounted for 85% of the variation in Type I error.

Power rates. Power by the levels of each factor appear in Table 3. As expected, given the $\chi^2_{difference.05}$ higher nominal Type I error rate, the $\chi^2_{difference.05}$ criterion had higher power compared to the other criteria. For both chi-square difference statistics, higher power is associated with a greater number of (a) factors, (b) participants, (c) indicators per factor, and (d) contaminated loadings. In contrast, the power of the $CFI_{difference}$ test was higher with fewer factors and a smaller sam-

TABLE 3
Power for Chi-Square and CFI by Number of Factors, Number of Participants, Number of Indicators per Factor, and Level of Contamination

Variable	$\chi^2_{difference\ .05}$	$\chi^2_{difference\ .01}$	$CFI_{difference}$	$\chi^2_{difference\ .05}$ & $CFI_{difference}$
Factors				
2	0.628	0.453	0.487	0.380
4	0.748	0.597	0.374	0.333
Participants				
300	0.570	0.387	0.555	0.368
650	0.612	0.425	0.275	0.247
1,000	0.882	0.763	0.461	0.454
Indicators per factor				
3	0.586	0.387	0.386	0.281
6	0.790	0.662	0.475	0.432
Contamination				
High	0.781	0.635	0.560	0.481
Low	0.595	0.415	0.301	0.232

ple size. In fact, the $CFI_{difference}$ power was greatest in the high contamination condition and smallest sample size condition. The use of both the $CFI_{difference}$ and the $\chi^2_{difference}$ results in lower power than for either individually. Indeed, at best, the power of the combined criterion is constrained to be no higher than the power of the least powerful individual criterion. Recall that the $CFI_{difference}$ had a higher Type I error rate in the same conditions in which it exhibited higher power, suggesting that the power results may not be very useful under certain conditions.

The levels of power by the four manipulated variables appear in Table 4. Again, when interpreting these results, it should be noted that the Type I error rate for the $CFI_{difference}$ was somewhat elevated in a number of conditions. That is, power might appear to be occasionally adequate, but at the expense of the loss of control of Type I error. Nonetheless, it was within appropriate bounds frequently enough that an examination of these more detailed results for power is warranted.

The variance components analysis for both $\chi^2_{difference}$ tests (results for the $\chi^2_{difference.01}$ appear later, as $\chi^2_{difference.05}$ results were nearly identical) shows that approximately 80% of the variation in power can be accounted for by five of the manipulated factors (or combinations of them), including the four-way Number of Factors × Indicators Per Factor × Sample Size × Contamination interaction (21%), the three-way Number of Factors × Indicators Per Factor × Contamination interaction (12%), the two-way Sample Size × Level of Contamination interaction (11%), and the main effects of sample size (15%) and indicators per factor (21%). On the other hand, 78% of the variance in the power of the $CFI_{difference}$ is accounted for by the Number of Factors × Indicators Per Factor × Sample Size × Contamination in-

TABLE 4
Power by Number of Factors (F), Participants (N), Number of Indicators per Factor (I), and Level of Contamination (C)

F	*N*	*I*	*C*	$\chi^2_{difference.05}$	$\chi^2_{difference.01}$	$CFI_{difference}$	$\chi^2_{difference.05}$ & $CFI_{difference}$
2	300	3	High	0.605	0.360	0.658	0.349
			Low	0.342	0.146	0.416	0.142
		6	High	0.787	0.590	0.830	0.589
			Low	0.293	0.124	0.295	0.108
	650	3	High	0.513	0.293	0.363	0.254
			Low	0.291	0.121	0.214	0.102
		6	High	0.955	0.863	0.752	0.750
			Low	0.459	0.237	0.146	0.145
	1,000	3	High	0.866	0.706	0.633	0.616
			Low	0.573	0.325	0.293	0.257
		6	High	1.000	1.000	0.983	0.983
			Low	0.852	0.674	0.263	0.263
4	300	3	High	0.530	0.280	0.475	0.249
			Low	0.295	0.115	0.268	0.092
		6	High	0.998	0.990	0.974	0.969
			Low	0.710	0.487	0.527	0.446
	650	3	High	0.765	0.528	0.337	0.336
			Low	0.442	0.234	0.146	0.142
		6	High	0.907	0.780	0.217	0.217
			Low	0.561	0.342	0.027	0.027
	1,000	3	High	0.992	0.951	0.618	0.618
			Low	0.815	0.589	0.215	0.215
		6	High	1.000	0.994	0.569	0.569
			Low	0.956	0.868	0.112	0.112

teraction (21%), the Number of Factors × Indicators Per Factor × Contamination interaction (25%), the Sample Size × Contamination interaction (17%), and the main effect of contamination (14%). For the combination of $CFI_{difference}$ and the $\chi^2_{difference}$, six terms accounted for 95% of the variance. These include the four-way Sample Size × Number of Factors × Indicators Per Factor × Contamination interaction (18%), the three-way Number of Factors × Indicators Per Factor × Contamination interaction (22%), the three-way Number of Factors × Indicators Per Factor × Sample Size interaction (13%), the two-way Sample Size × Contamination interaction (18%), and the main effects of contamination (12%) and indicators per factor (12%).

In general, for all three criteria, power is greater when the level of contamination is high, sample size is large, and there are more indicators per factor. This latter result is not universal, however. For instance, the power of the $CFI_{difference}$ declines in the low contamination condition when the number of indicators per factor

increases from three to six, except when there are four factors and 300 participants. In addition, power for $CFI_{difference}$ is not generally higher with four factors. Alternatively, for both chi-square difference tests, power is greater with four compared to two factors, except with $N = 300$ and three indicators per factor.

In terms of comparative power among the three criteria, as might be expected when alpha was higher, the $\chi^2_{difference.05}$ had higher power than the other two criteria. When comparing the $\chi^2_{difference.01}$ and the $CFI_{difference}$, it appears that $CFI_{difference}$ has higher power when the total sample size is 300 (corresponded to when its Type I error rate was highly inflated), but that in most, but not all, other cases, the chi-square statistic has equivalent or higher power. This advantage for the $\chi^2_{difference.01}$ is more marked in the four-factor case, particularly with larger samples and more indicators per factor. Under select conditions, the power of the two chi-square statistics is very similar, suggesting that when the proportion of loadings exhibiting group differences is sufficiently large, the alpha used may be immaterial in terms of power. Specifically, when there are 1,000 participants, four factors, and six indicators per factor, the two criteria have essentially equivalent power. A similar result is evident when there are four factors, six indicators per factor, a sample size of 300, and a high level of contamination, or when there are 1,000 participants, four factors, three indicators per factor, and a high level of contamination. One final comparative difference in power to note is that for the smallest sample size condition, coupled with two factors, the power of the $CFI_{difference}$ is actually higher than that of $\chi^2_{difference.05}$. However, it is important to note that in general, with 300 participants, the Type I error rate of the $CFI_{difference}$ is above the .05 level of this chi-square criterion. Therefore, these results for $CFI_{difference}$ must be interpreted with caution, as the higher power comes with the price of an artificially elevated Type I error rate.

Power for testing a single loading. In addition to identifying a lack of invariance among a set of factor loadings, power for detecting group differences on an individual loading was examined. This evaluation in practice would follow an initial determination of a lack of invariance in the lambda matrix, and involves testing each factor loading separately to determine where exactly this lack of invariance occurs. Therefore, contamination conditions remain as they were in the analysis described previously, although only one of the contaminated loadings is tested in this second stage of analysis. That is, power represents the detection of a lack of invariance when only one loading simulated to be different was constrained across groups while the other loadings were freely estimated. The same criteria were used to examine a single indicator as were used with testing the entire matrix. Table 5 displays the power for detecting lack of invariance for an individual loading by the manipulated factors.

The power of the $\chi^2_{difference.05}$ to identify an individual loading is uniformly higher compared to the other criteria, which might be expected given its nominally

TABLE 5
Power for Detecting Group Difference of a Single Indicator Loading

Variable	$\chi^2_{difference.05}$	$\chi^2_{difference.01}$	$CFI_{difference}$	$\chi^2_{difference.05}$ & $CFI_{difference}$
Factors				
2	0.382	0.221	0.153	0.127
4	0.323	0.235	0.012	0.011
Participants				
300	0.269	0.230	0.131	0.094
650	0.276	0.125	0.043	0.041
1,000	0.513	0.329	0.073	0.073
Indicators per factor				
3	0.412	0.252	0.160	0.134
6	0.293	0.204	0.005	0.005
Contamination				
High	0.403	0.216	0.069	0.056
Low	0.302	0.240	0.096	0.082

(but only sometimes) larger alpha value. The $\chi^2_{difference.01}$ had higher power than the $CFI_{difference}$ in all cases. Finally, for all three criteria, the ability to detect the difference for a single loading is lower than the ability to find differences for an entire set of loadings.

For all criteria, the power to detect group differences for an individual loading is somewhat lower for four factors as opposed to two for the $\chi^2_{difference.05}$ and the $CFI_{difference}$, but not for the $\chi^2_{difference.01}$. In addition, power is lower when there are more indicators per factor, regardless of the level of contamination. Power for the $\chi^2_{difference.05}$ is lower for the low contamination condition compared to high contamination, whereas the other two criteria had slightly higher power in the low contamination condition. Finally, it appears that power for all three criteria is highest when the sample size is 1,000, but some differences exist for the other two sample size conditions. For instance, for the $\chi^2_{difference.05}$ power is not greatly different for sample sizes of 300 and 650. However, both $CFI_{difference}$ and $\chi^2_{difference.01}$ actually had a decline in power from 300 to 650, before increasing with 1,000.

Table 6 provides more detailed results for each criterion by the combination of the manipulated variables. Perhaps most striking in Table 6 is the number of cases in which the $CFI_{difference}$ (and by extension the combination criteria) does not detect the group difference for a single indicator. This result is particularly prevalent when there are four factors. Indeed, despite the fact that the $CFI_{difference}$ has comparable power to the $\chi^2_{difference.01}$ for some conditions when detecting overall group differences, the criteria rarely performs as well for a single indicator. A second interesting result is that power for all criteria is much lower when detecting differences for an individual indicator versus the set of indicators. For both chi-square

TABLE 6
Power for Detecting Single Item Difference by Number of Factors (F), Participants (N), Number of Indicators per Factor (I), and Level of Contamination (C)

F	*N*	*I*	*C*	$\chi^2_{difference.05}$	$\chi^2_{difference.01}$	$CFI_{difference}$	$\chi^2_{difference.05}$ & $CFI_{difference}$
2	300	3	High	0.290	0.119	0.239	0.099
			Low	0.798	0.575	0.661	0.506
		6	High	0.208	0.073	0.040	0.040
			Low	0.062	0.012	0.005	0.005
	650	3	High	0.405	0.196	0.147	0.143
			Low	0.460	0.234	0.174	0.163
		6	High	0.208	0.079	0.001	0.001
			Low	0.066	0.015	0.0	0.0
	1,000	3	High	0.715	0.498	0.271	0.271
			Low	0.775	0.567	0.293	0.293
		6	High	0.519	0.265	0.001	0.001
			Low	0.077	0.013	0.0	0.0
4	300	3	High	0.274	0.115	0.081	0.079
			Low	0.054	0.017	0.011	0.011
		6	High	0.242	0.091	0.004	0.004
			Low	0.222	0.084	0.04	0.004
	650	3	High	0.398	0.206	0.022	0.022
			Low	0.048	0.011	0.0	0.0
		6	High	0.283	0.102	0.0	0.0
			Low	0.340	0.155	0.0	0.0
	1,000	3	High	0.678	0.470	0.017	0.017
			Low	0.054	0.013	0.0	0.0
		6	High	0.616	0.375	0.0	0.0
			Low	0.668	0.428	0.0	0.0

tests, there is greater power in the low contamination case for the two-factor models when there are three indicators per factor. On the other hand, when there are four factors, power is typically higher when the indicators exhibit greater contamination. However, with the two larger sample size conditions with six indicators per factor this does not occur. These results were somewhat unexpected given the assumption that greater contamination would be associated with higher power, regardless of the levels of the other manipulated variables. The outcome suggests that to detect a group difference in factor loadings when more loadings differ may actually be more difficult.

The variance components analysis indicated that for $\chi^2_{difference.05}$, four terms accounted for 89% of the observed variation in power, including the three-way Contamination × Indicators Per Factor × Number of Factors interaction (53%), the two-way Indicators Per Factor × Number of Factors interaction (12%), contamina-

tion (14%), and sample size (10%). Results for the $\chi^2_{difference.01}$ also are presented as they differed from the $\chi^2_{difference.05}$. For $\chi^2_{difference.01}$, three terms accounted for 83% of the variance in power: the four-way Sample Size × Contamination × Indicators Per Factor × Number of Factors interaction (19%), the three-way Contamination × Indicators Per Factor × Number of Factors interaction (47%), and the two-way Sample Size × Contamination interaction (17%). For the $CFI_{difference}$, 75% of the variance in power was accounted for by the Sample Size × Contamination × Indicators Per Factor × Number of Factors interaction (29%) and the Indicators Per Factor × Number of Factors interaction (46%). Finally, 82% of the variance in power for the combination criterion of the $CFI_{difference}$ and the $\chi^2_{difference}$ was accounted for by the Sample Size × Contamination × Indicators Per Factor × Number of Factors interaction (32%), the Indicators Per Factor × Number of Factors interaction (39%), and the number of indicators per factor (11%).

Dichotomous Data

Type I error rates. Based on the results for the ML estimation with normally distributed data, described earlier, we decided not to use the $CFI_{difference}$ or the combination criteria for determining a lack of invariance with the RWLS estimation. We should note that a few conditions where power was highest for $CFI_{difference}$ in the normal case were examined for power with dichotomous data using RWLS estimation. Specifically, the power of $CFI_{difference}$ with four factors, three indicators per factor, 1,000 participants, and high contamination was .0457, whereas power with four factors, six indicators per factor, 300 participants, and low contamination was 0. Finally, the power of $CFI_{difference}$ with four factors, six indicators per factor, 1,000 participants, and high contamination was .1537. As expected, in all three cases power was much lower for dichotomous data than in the normal case, supporting our reasoning for not including these two criteria in the remaining analyses. Thus, in the remaining results, only the $\chi^2_{difference.05}$ is used. The Type I error rate for the RWLS by the number of factors, participants, and items per factor appears in Table 7.

The Type I error rate for the chi-square difference test with the RWLS estimation method is near the nominal rate regardless of the number of factors present in the model or the number of items per factor. However, the number of participants in the sample does seem to have a substantial impact on the error rate. With a total sample size of 1,000, the Type I error rate is somewhat inflated above the nominal .05 level, whereas with 300 participants, the error rate is below the nominal level. This result stands in contrast to the normal case, where the sample size had a rather small influence on the Type I error rate. A more detailed display of the Type I error rate by the manipulated variables appears in Table 8.

Perhaps the most revealing finding in Table 8 is that the impact of sample size on the Type I error rate is influenced, to some degree, by the number of items per

TABLE 7
Type I Error Rates for Chi-Square for Robust Weighted Least Squares Difference Test by Number of Factors, Number of Participants, and Number of Items per Factor

Variable	$\chi^2_{difference.05}$
Factors	
2	0.052
4	0.054
Participants	
300	0.028
650	0.041
1,000	0.090
Items per factor	
3	0.066
6	0.040

TABLE 8
Type I Error Rate for Detecting Lack of Model Invariance for Robust Weighted Least Squares Estimation by Number of Factors (F), Participants (N), Number of Items per Factor (I), and Level of Contamination (C)

F	*N*	*I*	$\chi^2_{difference.05}$	*Nonconvergence Rate*
2	300	3	0.031	0.477
		6	0.027	0.441
	650	3	0.049	0.091
		6	0.022	0.180
	1,000	3	0.111	0.010
		6	0.073	0.015
4	300	3	0.037	0.310
		6	0.016	0.568
	650	3	0.053	0.089
		6	0.041	0.226
	1,000	3	0.114	0.010
		6	0.060	0.026

factor. Variance components analysis for the Type I error rate indicates that approximately 92% of the variance is due to the sample size. When there are fewer items (three), the error rate at all three sample sizes is somewhat higher than when there are more items, although this effect is most dramatic when sample size was largest. Indeed, when there are four factors and six items per factor, the error rate with 1,000 participants is only slightly above the nominal rate of .05. Additionally,

with 300 participants and six items per factor, regardless of the number of factors, the Type I error rate is noticeably lower than the nominal rate. In short, the impact of the number of participants on the Type I error rate is determined in part by the size of the model being estimated, as expressed by the number of items per factor.

Power rates. Power by the number of factors, sample size, number of items per factor, and level of contamination appears in Table 9. Certainly, the most obvious finding is the very low power across all conditions. The highest power (i.e., .313), which occurs with the largest sample size, is well below the lowest power found in the normal case (Table 4). This outcome is especially interesting given that the amount of group difference simulated in the dichotomous case is comparable to that with the normal variables. Of importance to remember was that the Type I error rate is inflated for the largest sample size. The power of the chi-square test by all combinations of the manipulated variables appears in Table 10.

As is evident in Table 10, the power of the chi-square difference test with RWLS estimation is very low. The highest power in this context occurs with the largest sample size, four factors, three items per factor, and a high level of contamination (e.g., power = .479). In fact, only for this combination of conditions and a similar combination, with two factors rather than four, was power greater than .40. These results are a sign of minimal or no real detection capability. Predictably, based on Table 9, the lowest power values observed in Table 10 occur with the smallest sample size. These values are sometimes comparable to the Type I error rates under similar conditions. Additionally, the results of the variance compo-

TABLE 9
Power for Detecting Group Difference With Robust Weighted Least Squares Estimation

Variable	$\chi^2_{difference.05}$
Factors	
2	0.163
4	0.152
Participants	
300	0.051
650	0.108
1,000	0.313
Items per factor	
3	0.188
6	0.127
Contamination	
High	0.193
Low	0.122

TABLE 10
Power of the Chi-square With Robust Weighted Least Squares Estimation by Number of Factors (F),Participants (N), Number of Items per Factor (I), and Level of Contamination (C)

F	*N*	*I*	*C*	$\chi^2_{difference.05}$	*Nonconvergence Rate*
2	300	3	High	0.088	0.197
			Low	0.055	0.197
		6	High	0.068	0.370
			Low	0.042	0.329
	650	3	High	0.185	0.094
			Low	0.098	0.013
		6	High	0.102	0.144
			Low	0.064	0.196
	1,000	3	High	0.405	0.012
			Low	0.289	0.010
		6	High	0.360	0.027
			Low	0.213	0.020
4	300	3	High	0.054	0.244
			Low	0.047	0.362
		6	High	0.037	0.667
			Low	0.026	0.601
	650	3	High	0.164	0.193
			Low	0.114	0.138
		6	High	0.081	0.397
			Low	0.058	0.347
	1,000	3	High	0.479	0.025
			Low	0.290	0.020
		6	High	0.302	0.085
			Low	0.168	0.064

nents analysis show that 75% of the variance in power is accounted for by the sample size.

Given these very low power values observed with RWLS estimation, coupled with the result for the normal case (Tables 5 and 6) that suggests that the power for detecting group differences for a single variable is lower than for detecting a lack of invariance across all variables, no further analyses were conducted with RWLS. The assumption is reasonable that the power for detecting differences for individual items will be very low in most cases when RWLS estimation is used with dichotomous data.

An additional issue pertinent to understanding the performance of the RWLS difference testing is the rate of nonconvergence. Although this was a nonissue with the ML estimation, the rates of nonconvergence for RWLS under certain conditions was very high. Examination of Tables 8 and 10 suggests that the problem is

particularly acute when the sample size is small and the model is more complex (larger numbers of items and factors). Indeed, the only factor that appears to guarantee low rates of nonconvergence is a large sample. This lack of convergence is especially problematic when there were 300 participants, where the rates were above 0.6 in some cases, and never below 0.2. Based on this result, it appears that researchers who want to use the RWLS difference testing with smaller sample sizes may run into serious problems in obtaining proper convergence and stable estimates. Note that all Type I error rate and power results are based on a full set of 1,000 replications. In cases where nonconvergence occurred, replacement samples were generated until 1,000 simulations were run for all conditions. Nonconvergence rates were based on the first 1,000 replications run.

To better understand the relation between the performance of the $\chi^2_{difference.05}$ and the ability of the RWLS estimation procedure to converge, Pearson correlations between convergence rates and both the Type I error and power were computed. In both cases, the correlation was strong and negative (–.765 for Type I error and –.702 for power). These results may indicate that the low power for detecting a lack of invariance associated with using RWLS estimation is due in part to difficulties for the algorithm in correctly estimating parameters. Thus, even when convergence is achieved, the resulting estimates, in some cases, may not be stable. This instability may overshadow any differences when the constrained and unconstrained models are compared. Clearly, regardless of the cause, testing for a lack of invariance suffers from low power with the use of RWLS for parameter estimation.

DISCUSSION

The results described suggest that when the indicator variables are normally distributed and ML estimation is used, the $\chi^2_{difference.05}$ test of invariance offers researchers generally good control of the Type I error rate, along with relatively high power in most situations. Furthermore, the $\chi^2_{difference.01}$ also adequately controls the Type I error rate, although, predictably, has lower power compared to the $\chi^2_{difference.05}$. In contrast, the $CFI_{difference}$ statistic behaved somewhat more erratically than the $\chi^2_{difference.01}$ test. In cases where power comparisons between the $CFI_{difference}$ and the $\chi^2_{difference.01}$ were possible due to noninflated Type I error rates, the latter had higher power than the former. Using both criteria in conjunction yielded very low power, and did not appear to be a worthwhile alternative to either in isolation. As mentioned previously, power with a combination of criteria will be constrained to the least powerful individual criterion, at best.

In terms of the factors that influence the Type I error rates of the criteria examined, most important for the $CFI_{difference}$ test appears to be sample size, where larger samples were associated with lower Type I error rates. Indeed, for samples of 1,000, the rate is actually below .01. On the other hand, with samples of 300, the

Type I error rate can be up to 10 times more than this rate. The $CFI_{difference}$ test also appears to have subnominal Type I error rates in certain cases when there are six variables per factor. In contrast, none of the manipulated variables had a marked impact on the error rates of either of the chi-square criteria. In sum, this finding suggests that in some sense, this test is more stable than the $CFI_{difference}$, and that indeed the chi-square tests are generally a dependable tool in terms of maintaining the nominal Type I error rate in factorial invariance studies.

With respect to power, the primary variables influencing the chi-square criteria are sample size and the number of indicators per factor. In general, the more participants included in the analysis, the greater the power for the chi-square difference test, regardless of the nominal alpha. In addition, irrespective of the sample size, there is greater power for detecting a lack of invariance with more indicators per factor. The number of factors has some impact on the power of both the chi-square difference test as well as the $CFI_{difference}$, although the nature of the effect is very different. For both chi-square criteria, power was higher for models with four factors than those with two, and conversely, the power of the $CFI_{difference}$ was higher in the two-factor than the four-factor case. The number of indicators per factor also has a noticeable impact on the power of $CFI_{difference}$, with more indicators associated with higher power, as was seen with the chi-square statistics.

When differences in models have been detected between two groups, a researcher will naturally be interested in isolating the specific items or subtests that differ across the groups. For this to be evaluated, the data analyst will need to conduct model invariance tests for individual indicators by allowing each to be constrained as others are free to vary across groups. The results of this simulation study suggest that for any of the statistics used here, power is much reduced when testing invariance for a single indicator versus that of the entire set. The power of the $CFI_{difference}$ test is particularly low when testing for invariance of a single indicator, although even the chi-square statistics suffer reductions in their ability to detect group differences. The only instance when the $CFI_{difference}$ is more powerful than the $\chi^2_{difference.01}$ is when the sample size is 300 and the model is simple (i.e., three indicators per factor and two factors). However, note that this combination of conditions is associated with an inflated Type I error rate in the testing of overall invariance so that interpretation of power in this context is not meaningful. In contrast, often with four factors the $CFI_{difference}$ power for testing a single indicator was actually zero.

The chi-square difference criteria also suffer from a diminution in power in the single indicator case, although not as severe as seen with the $CFI_{difference}$ test. Additionally, some of the effects of the manipulated variables when testing all of the indicators are reversed when examining an individual indicator. For example, when testing a single indicator, models of greater complexity (more factors and more indicators) were actually associated with lower power, in contrast to the situation when testing all indicators. This result seems logical, in that with more indicators

and more factors that presumably do not differ by group, the single indicator that differs might be hidden in the crowd. This is in contrast to recent suggestions that it may be easier to locate invariance when only a few variables lack invariance (Millsap, 2005). On the other hand, as with testing the entire set of indicators, larger sample sizes are associated with greater power for the chi-square difference tests.

An interesting result present for all of the criteria examined is that power in the low contamination condition was higher compared to the high contamination for two factors and three indicators per factor with samples of size 300 or 650 and for four factors with six indicators per factor. In other words, power for detecting group differences for a single item is greater when the number of items that actually differ between the groups is smaller, in accord with Millsap (2005), if the problem itself is somewhat simpler (fewer factors and fewer participants). This outcome would suggest that when a researcher is interested in isolating individual indicator differences (or lack thereof), one must be aware of not only the individual target indicator, but also the potential of group differences for other indicators in the set.

As noted earlier, recent work has found that for dichotomous items, RWLS appears to be preferable in terms of estimating item parameters in the CFA context (Flora & Curran, 2004). Therefore, this approach was selected for testing model invariance with dichotomous items. The results presented here suggest that the Type I error rate generally tended to be at or below the nominal .05 level, except when the sample size was 1,000 and the number of items per factor was small. Of greater concern in this context than Type I error control is the high rate of nonconvergence in certain instances. With samples of 1,000, this problem occurred very infrequently; however, when there were 300 participants, nonconvergence was nearly as common (and in one case more common) than convergence. Clearly, this result suggests that for certain applications, researchers may have difficulty applying RWLS because the method may not yield estimates due to small sample sizes.

The power of the RWLS chi-square test for testing invariance is very low in comparison with that for the normal data using ML estimation. In fact, in a number of cases, power is nearly indistinguishable from the Type I error rate. Specifically, it is lowest for the smallest sample size, as was found with ML estimation. However, even when the sample is 1,000, power never reaches .50, and is most often below .40. As with the ML estimation, power is somewhat lower for situations in which there are six indicators per factor, and for the lower level of contamination. Again, however, even in the best situations, power for the RWLS remains very low relative to the normally distributed variables with ML estimation.

CONCLUSIONS

The results of this study present several implications for the practitioner who plans to conduct factorial invariance analyses using either the ML or RWLS approaches

for normally distributed or dichotomous variables, respectively. First, when it is appropriate to use ML, the chi-square difference test maintains the nominal Type I error rate at both .05 and .01, across a variety of sample size and model complexity conditions. In addition, the chi-square difference test provides comparable or better power than the $CFI_{difference}$ test, whereas the latter does have some inflation of the Type I error rate for small sample sizes. In short, the chi-square difference test appears to be a solid candidate for testing model invariance in the ML case, and may be more appropriate in many cases than the $CFI_{difference}$ test.

Second, although the chi-square test for overall lack of invariance appears to offer satisfactory power in many instances, the same cannot be said of the tests for individual indicators. In most applied situations, a researcher will need to follow up a significant overall test for a lack of invariance with tests of the individual indicators to isolate where differences occur. However, the results presented in this study seem to suggest that the power for such tests might often be fairly low. Perhaps most surprising is that even in some cases where N is 1,000, the power is below .20. The incidence of low power is exacerbated by a more complex model structure (more variables and more factors). This result could potentially present the practitioner with the conundrum of having sufficient power to detect an overall difference between groups but not having the power to find such differences for individual indicators. Such an outcome could lead to the unsatisfying conclusion that there is a difference in factor loadings between the groups somewhere, but it is not possible to say where.

A third major implication of this study is that power appears to be a major problem for the chi-square difference test when the RWLS method of estimation is employed. Indeed, the possibility exists that power would be nearly as low as the Type I error rate with small sample sizes. Certainly, in practice, having such a low probability of detecting group differences in terms of factor loadings hampers the ability of a researcher to make substantive statements about invariance. Based on the results presented, it would seem that only in the largest sample size condition is it realistic for researchers to expect to find differences between groups when they should find them. For many applied researchers, this could suggest limited utility in applying the chi-square difference test when RWLS estimation is used because obtaining such large samples is often not feasible, especially with low-incidence populations. A separate but related problem to that of power in the RWLS case is the fairly high rate of nonconvergence when sample sizes are small. Although this problem did not appear to be severe with 1,000 participants, it was increasingly likely to occur when only 300 participants were available. Again, for many studies, obtaining as many as 1,000 participants is not possible, making the application of the preferred method of estimation, RWLS, in group invariance testing less feasible.

The purpose of this investigation is to provide evidence of MCFA performance for measurement invariance testing under a variety of practical and applied condi-

tions. However, although several thousand samples were examined, simulation of exhaustive conditions is not practically possible. For instance, an important aspect of this study to recall is that properly specified baseline models were simulated. Results cannot be assumed to apply to situations when baseline models do not meet this condition. That is, results of this study are dependent on properly specified baseline models. Additionally, the sample size of 650 deserves further exploration as it may represent effects of uneven sample sizes and not just a moderate total sample size. Therefore, further simulation work is encouraged to continue to examine MCFA analyses under various additional conditions (e.g., referent indicator, longitudinal data, etc.) as there are several problems that remain to be solved in invariance testing (Millsap, 2005). That said, this research should inform practice for those using model invariance analyses to better understand phenomena in their disciplines. The results described here should (a) allow practitioners to make informed decisions about the use of GFIs to determine measurement invariance, (b) inform practice by highlighting the strengths and limitations of MCFA given certain conditions (e.g., complex models, nonnormality), and (c) stimulate new research surrounding the implementation of the MCFA. For example, it is clear that a measure of effect size would be very helpful in allowing power results to be more meaningfully examined (Cheung & Rensvold, 2002; Millsap, 2005).

Continued examination of methods, such as MCFA, for gathering score validity evidence is crucial in many fields. Practitioners rely on methodologies such as these to compare instruments across groups, thus ensuring that constructs, and ultimately test scores, have equal meaning and interpretation for a variety of individuals. Knowledge of when such methods have difficulties in comparing groups is essential so that conclusions about group differences (or lack of differences) can be interpreted accurately. Results presented here would suggest that when ML is used for normally distributed indicators, the MCFA approach using chi-square difference tests will often yield reasonable outcomes regarding group invariance overall, but may have difficulties in pinpointing differences for individual factor loadings. Furthermore, when RWLS is used with dichotomous items, care must be taken in the interpretation of findings suggesting that group invariance holds, particularly for smaller sample sizes, in light of the low power results and nonconvergence rates.

REFERENCES

Alwin, D. F., & Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. In D. Jackson & E. Borgatta (Eds.), *Factor analysis and measurement in sociological research: A multi-dimensional perspective* (pp. 249–279). Beverly Hills, CA: Sage.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

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

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

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

Brennan, R. L. (2004). *Revolutions and evolutions in current educational testing* (CASMA Research Rep. No. 4). Iowa City: University of Iowa.

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

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

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25,* 1–27.

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

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466–491.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23,* 17–27.

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

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 1,* 424–451.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 57,* 409–426.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago: Scientific Software.

Keith, T. Z. (1997). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 373–403). New York: Guilford.

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

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 31,* 53–76.

Lord, F., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11,* 514–534.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84–99.

Maller, S. J., & French, B. F. (2004). Factor invariance of the UNIT across deaf and standardization samples. *Educational and Psychological Measurement, 64,* 647–660.

Mantzicopoulos, P. Y., French, B. F., & Maller, S. J. (2004). Factor structure of the pictorial scale of perceived competence and social acceptance with two pre-elementary samples. *Child Development, 75,* 1214–1228.

McGaw, B., & Jöreskog, K. G. (1971). Factorial invariance of ability measures in groups differing in intelligence and socio-economic status. *British Journal of Mathematical and Statistical Psychology, 24,* 154–168.

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.

Messick, S. (1989). Meaning and values in test validation. The science and ethics of assessment. *Educational Researcher, 18,* 5–11.

Millsap, R. E. (2005). Four unresolved problems in studies of factorial invariance. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 153–172). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Millsap, R. E., & Tein, J.-Y. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39,* 479–515.

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* Unpublished manuscript, UCLA.

Muthén, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10,* 133–142.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide, version 3*. Los Angeles: Author.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87,* 517–529.

Rensvold, R. B., & Cheung, G. W. (1998). Testing measurement models for factorial invariance: A systematic approach. *Educational and Psychological Measurement, 58,* 1017–1034.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361–370.

Sörbom, D. (1974). A general model for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology, 27,* 229–239.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Thurstone, L. L. (1947). *Multiple factor analysis: A development and expansion of the vectors of the mind.* Chicago: University of Chicago Press.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5,* 139–158.

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