STRUCTURAL EQUATION MODELING, *14*(4), 611–635

Power and Precision in Confirmatory Factor Analytic Tests of Measurement Invariance

Adam W. Meade
Department of Psychology, North Carolina State University

Daniel J. Bauer
Department of Psychology, University of North Carolina at Chapel Hill

This study investigates the effects of sample size, factor overdetermination, and communality on the precision of factor loading estimates and the power of the likelihood ratio test of factorial invariance in multigroup confirmatory factor analysis. Although sample sizes are typically thought to be the primary determinant of precision and power, the degree of factor overdetermination and the level of indicator communalities also play important roles. Based on these findings, no single rule of thumb regarding the ratio of sample size to number of indicators can ensure adequate power to detect a lack of measurement invariance.

Measurement invariance (MI) can be considered the degree to which measurements conducted under different conditions yield equivalent measures of the same attributes (Drasgow, 1984, 1987; Horn & McArdle, 1992). These different conditions include stability of measurement over time (Chan, 1998; Chan & Schmitt, 2000), across different populations (e.g., cultures [Riordan & Vandenberg, 1994], gender [Marsh, 1985, 1987], age groups [Marsh & Hocevar, 1985]), rater groups (Facteau & Craig, 2001), or over different mediums of measurement administration (Chan & Schmitt, 1997; Ployhart, Weekley, Holtz, & Kemp, 2003).

Under all these conditions, tests of MI are often conducted via confirmatory factor analytic (CFA) methods. These methods have evolved substantially during

Correspondence should be addressed to Adam W. Meade, Department of Psychology, Campus Box 7650, Raleigh, NC 27695-7650. E-mail: awmeade@ncsu.edu

the past 20 years and are widely used in a variety of situations (Vandenberg & Lance, 2000). Although these methods have evolved greatly during this time, MI tests are still somewhat poorly understood, with little prior work investigating factors that determine the power of these tests to detect a lack of invariance. The goal of this study is to expand what is currently known about the quality of single-group factor analytic solutions to the case of multigroup MI tests.

MI can be defined in terms of probabilities such that for MI to exist, the probability of observed responses conditioned on latent scores must be unaffected by group membership (Meredith & Millsap, 1992; Millsap, 1995). Commonly used CFA tests of MI involve simultaneously fitting a measurement model to two or more data samples. The multigroup CFA measurement model between p observed variables and m latent factors is linear in nature and given by the equation:

$$\boldsymbol{X}_g = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g \boldsymbol{\xi}_g + \boldsymbol{\delta}_g \tag{1}$$

where $\boldsymbol{X}$ is a $p \times 1$ vector of observed scores, $\boldsymbol{\tau}$ is a $p \times 1$ vector of intercepts, $\boldsymbol{\Lambda}$ is a $p \times m$ matrix of factor loadings, $\boldsymbol{\xi}$ is a $m \times 1$ vector of latent variable scores, $\boldsymbol{\delta}$ is a $p \times 1$ vector of unique factor scores, and g denotes that these parameters are group specific. The observed variable means and covariances are then given as:

$$\boldsymbol{\mu}_g = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g \boldsymbol{\kappa}_g, \tag{2}$$

and

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Phi}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Theta}_g, \tag{3}$$

where $\boldsymbol{\mu}_g$ is a $p \times 1$ vector of observed score means, $\boldsymbol{\Sigma}_g$ is a $p \times p$ matrix of observed score covariances, $\boldsymbol{\kappa}_g$ is a $m \times 1$ vector of factor means, $\boldsymbol{\Phi}_g$ is a $m \times m$ latent variance-covariance matrix, and $\boldsymbol{\Theta}_g$ is a $p \times p$ diagonal matrix of unique variances.

MI can therefore exist for multiple parts of the CFA model.[1] For instance, if $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2$ for Groups 1 and 2, metric invariance is said to exist (Horn & McArdle, 1992); if $\boldsymbol{\tau}_1 = \boldsymbol{\tau}_2$, scalar invariance is indicated (Meredith, 1993); and if $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_2$, uniqueness invariance exists. If all three types of invariance are found, strict factorial invariance is indicated (Meredith, 1993) such that differences in observed score means or covariances are a product of differences in latent means (sometimes called impact; Holland & Wainer, 1993) or latent covariances.

[1]Note that some researchers prefer the term mean and covariances structures (MACS) to describe MI tests of the full factor analytic model (including item intercepts and latent mean structures; Chan, 1998; Ployhart & Oswald, 2004). Our discussion of CFA models includes these MACS models.

CFA tests of MI involve fitting a series of nested sequential measurement models to two or more data samples. Typically, in the first model, both data sets (representing groups, time periods, etc.) are examined simultaneously, holding only the pattern of factor loadings invariant. In other words, the same items are forced to load onto the same factors, but factor loading estimates themselves are allowed to vary between samples. This baseline model of identical factor patterns provides a chi-square value that reflects model fit for item parameters estimated separately for each group.

Next, a test of factor loading invariance (metric invariance; Horn & McArdle, 1992) is conducted by examining a model identical to the baseline model except that the matrix of factor loadings (Λ_x) is constrained to be equal across groups (Meredith, 1993; Millsap, 1997). The difference between the baseline and more restricted model is expressed as a chi-square statistic with degrees of freedom equal to the number of constrained parameters (i.e., likelihood ratio test [LRT]). A significant LRT indicates that factor loadings differ (i.e., items do not relate to the factors in the same way) across groups. Subsequent to tests of factor loadings' equality, several additional model parameters can also be tested for MI; however, there is firm consensus that tests of factor loadings are the most important for establishing that MI conditions exist (Meade & Kroustalis, 2006), thus our focus on these tests.

Although the focus on MI testing thus far is understandable, far less attention is typically given to the precision of the factor loading estimates or the differences between them across groups. This is an unfortunate omission in many applications of MI, given the increased emphasis on the use of confidence intervals in other areas of psychological research (Wilkinson & the Task Force on Statistical Inference, 1999). The precision of estimation of the factor loading difference may be expressed using the formula for the standard error,

$$SE_{(\hat{\lambda}_2-\hat{\lambda}_1)} = \sqrt{Var(\hat{\lambda}_1) + Var(\hat{\lambda}_2) - 2Cov(\hat{\lambda}_1, \hat{\lambda}_2)} \tag{4}$$

where the numeric subscripts refer to Groups 1 and 2 and *Var* and *Cov* are the sampling variances and covariances of the estimates. In a multiple groups model, the term involving the sampling covariance is zero due to the independence of the groups, making it possible to calculate the standard error of the difference solely as a function of the standard errors of the loading estimates from the two groups as

$$SE_{(\hat{\lambda}_2-\hat{\lambda}_1)} = \sqrt{SE(\hat{\lambda}_1)^2 + SE(\hat{\lambda}_2)^2} \tag{5}$$

In tests of longitudinal invariance, this simplification would not be possible due to the lack of independence. Once the standard error has been calculated, a

confidence interval can be formed for the factor loading difference in the usual way, as

$$CI_{\hat{\lambda}_2-\hat{\lambda}_2} = |\hat{\lambda}_2 - \hat{\lambda}_1| \pm z_{crit} \times SE_{(\hat{\lambda}_2-\hat{\lambda}_1)} \quad (6)$$

where z_{crit} is a critical z value chosen to give a confidence interval of specified precision (e.g., selecting $z_{crit} = 1.96$ would provide a 95% confidence interval).

Although hundreds of studies using CFA tests of MI have been conducted in applied research, the factors that influence the power and precision of MI tests are poorly understood. Indeed, Vandenberg (2002) called for increased research on MI analyses and the logic behind them, stating that, "A negative aspect of this fervor, however, is characterized by unquestioning faith on the part of some that the technique is correct or valid under all circumstances" (p. 140). In answer to this call, this article provides a review of what is known about the factors that influence the precision of factor loading estimates and the power to detect factor loading differences between groups. Aside from the obvious influence of sample size, we also consider item communality (the proportion of variance in the item accounted for by the latent factors) and factor overdetermination (the number of indicators per factor). To this extent, we build directly on the influential work of MacCallum, Widaman, Zhang, and Hong (1999) and their investigation of the influence of sample size, factor overdetermination, and item communality on the quality of exploratory factor analytic solutions. We then supplement our review with a set of new simulation studies designed to further illustrate the effects of these data properties on precision and power in multiple groups factor analysis.

Past Research on the Power of CFA MI Tests

Although there are many examples of applications of tests of MI in the extant literature (see Riordan, Richardson, Schaffer, & Vandenberg, 2001; Vandenberg & Lance, 2000, for reviews), little has been done to determine the factors that affect the power of these tests (Meade & Lautenschlager, 2004; Vandenberg, 2002). In one published study, Meade and Lautenschlager (2004) simulated data for both a 6- and 12-item scale, then manipulated the number of items showing loading differences between simulated samples (sometimes called differential functioning [DF] items), and the directionality of the simulated differences (uniformly lower factor loadings for Sample 2, or some higher and some lower). As expected, they found that a lack of MI was more readily detected when more items were simulated to differ. Importantly, they found that a mixed pattern of simulated difference directionality was associated with more frequent detection of a lack of MI than were uniformly lower simulated loadings in one group. As the authors noted, the latter effects could be due to higher item communalities in the mixed pattern simulation conditions as compared to the uniformly lower simulated conditions (as the residual variances were held constant across

conditions). In another study, Kaplan (1989) specified a two-group, two-factor model in which one item in one group loaded on both factors. Power was highest when the communalities of the items in the factors were high and when the cross-loading was large.

Past Research on the Precision of Factor Loadings

Little research has been conducted on the factors that influence the precision of factor loading in multigroup models. Because the precision of the factor loading difference is a function of the precision of the loading estimate in each of the two groups, however, studies evaluating the precision of factor loading estimates in single samples are equally germane. Along these lines, several authors have demonstrated the relation between data properties and the quality of factor analytic solutions. Most notably, MacCallum et al. (1999) showed that increased sample size, higher communality, and greater factor overdetermination led to increased precision of estimated factor loadings in an exploratory factor analysis. Recently, Hogarty, Hines, Kromrey, Ferron, and Mumford (2005) used a highly similar approach but extended the number of conditions analyzed and found largely the same conclusions as MacCallum et al. (1999). Similarly, Marsh, Hau, Balla, and Grayson (1998) simulated a large number of data conditions and found that the standard deviations of factor loading estimates decreased both with increasing sample size and the indicator-to-factor ratio. Further, this decrease in variability of factor loadings was more pronounced for items with high communality than those with low communality. Gerbing and Anderson (1987) also found that the standard errors of CFA estimates were smaller when both sample size and the ratio of items to factors were larger. Thus, the effects of sample size, communality, and overdetermination on the accuracy of estimated factor loadings in both exploratory and single-group confirmatory factor analyses are well documented. We therefore expect these effects to generalize to the multigroup CFA case and to lead to increased power of CFA tests of MI.

This Study

In this study, we were interested in further understanding how data properties affect both the accuracy of parameter recovery in a multigroup situation and the power of tests of equality of factor loadings. Clearly, precision will be affected primarily by sample size, whereas power will be affected by both sample size and the effect size in the population (the true difference in the factor loadings). However, the effects of other data properties, such as communality and factor overdetermination, are less obvious but may be of substantial importance.

To better elucidate the importance of these data properties, we simulated a number of conditions of DF items (i.e., a lack of MI) between a pair of groups.

In each condition, we held the population magnitude of differences in factor loadings constant, but varied other data properties to determine the effect of these properties on the precision of the factor loading differences and the power of the LRT for assessing metric invariance. To this extent, we expand the work of MacCallum et al. (1999) to the case of two-group CFA tests of MI. Given the previous findings of MacCallum et al. and others, we expect that factor loading differences will be more precisely estimated and the power of the LRT will be highest when there are more items per factor, high item communalities, and large sample sizes.

METHOD

In this study, we simulated data for a 20-item survey for two groups (Sample 1 and Sample 2). DF of items was simulated by changing the magnitude of the factor loadings for four Sample 1 items by .20 to create Sample 2 item factor loadings. This magnitude of change is theoretically reasonable and pilot analyses suggest that this magnitude of difference was conducive to illustrating the effects proposed. The design factors that were varied to produce the conditions were similar to those used by MacCallum et al. (1999) and include the following.

Sample Size

We examined three sample size conditions ($N = 100$, 200, and 400) per group of respondents.[2] These values correspond closely to those chosen by MacCallum et al. (1999), and initial pilot analyses suggested that these values tended to illustrate the effects of sample size without encountering ceiling or floor effects for the significance of the LRT found with higher or lower values. In all analyses, sample sizes were simulated to be equal across Samples 1 and 2.

Factor Overdetermination

Based on both MacCallum et al.'s (1999) earlier work and pilot analyses, we choose to simulate a low factor overdetermination condition in which 20 items represent six factors and a high factor overdetermination condition in which 20 items represent three factors. Factor structures were always identical for Samples 1 and 2 (i.e., either a three-factor or six-factor solution was used to generate both Sample 1 and 2 data).

[2]We also pilot tested sample sizes of 60 per group (as per MacCallum et al., 1999), but those small sample sizes resulted in many inadmissible solutions and poor parameter estimation.

Factor Correlation Equality

In all conditions, population latent factor variances were 1.0 for both groups. In Sample 1, the factor correlations were set to be, on average, around .3 (see Tables 2 and 3; cf. Cheung & Rensvold, 2002). We then manipulated factor correlation equality by creating Sample 2 data with either the same population correlation structure or correlations that were different from the Sample 1 data. For data with which the correlations differed between samples, the correlations of .25 in Tables 2 and 3 were set to .45 for Sample 2 data and correlations of .35 were set to .15.

Item Communalities

In the population model for the Sample 1 data, we chose to replicate MacCallum et al.'s (1999) wide item communalities conditions by assigning item communalities ranging from .2 to .7 within each factor, with factor loadings equaling the square root of the communality for these data.[3] Data were simulated to have equal average item communalities across conditions to the extent possible (see Tables 1 and 2). The Sample 1 population parameters were then modified to create Sample 2 population parameters. Item communalities were manipulated in two primary ways: by choice of which items exhibited DF (either low or high communality items), or the manner in which the items were modified (in uniform or mixed fashion) to create Sample 2 factor loadings. Both of these methods are discussed in more detail later. Note that the variances for the item uniqueness terms were set to give each indicator a population variance of 1.0.

High or low communality DF items. The choice of items for which to modify item factor loadings (thus creating DF items) was either low or high communality items (i.e., *which item* variable). For the low communality choice condition, the four items with the lowest Sample 1 factor loadings were modified to create the Sample 2 item factor loadings (by changing the Sample 1 loading by .2 to create the Sample 2 loading). For the high communality DF item condition, the four highest Sample 1 item factor loadings were modified to create the Sample 2 factor loadings. In all cases, factor loadings for DF items were modified by the same magnitude, .2. However, as item communality for these data is equal to the square of the factor loadings, a change of .2 for a higher magnitude factor loading constitutes a larger change in communality than does

[3]MacCallum et al.'s (1999) wide communality condition included communalities that ranged from .2 to .8. However, communalities of .8 were problematic for this study as some simulated difference conditions would have resulted in simulated communalities greater than 1.0. For our data, item communality is equal to the squared factor loading for that item.

TABLE 1
Factor Loadings and Correlations,
High Overdetermination Condition, Sample 1 Data

	F1	*F2*	*F3*
F1	1.0		
F2	.25	1.0	
F3	.30	.35	1.0
Item			
I1	.84[a]		
I2	.77		
I3	.71		
I4	.63		
I5	.63		
I6	.55		
I7	.45[b]		
I8		.84	
I9		.77[a]	
I10		.71	
I11		.63	
I12		.63	
I13		.55	
I14		.45[b]	
I15			.84[a]
I16			.77[a]
I17			.71
I18			.63
I19			.55[b]
I20			.45[b]

[a]Loadings were modified in the *which loading* high communality condition.

[b]Loadings were modified in the *which loading* low communality condition.

an equal change in a lower factor loading. See Tables 1 and 2 for the items that were modified under these conditions.

Mixed or uniform DF. For the *how modified* set of communality manipulations, Sample 1 item factor loadings were modified in either a uniform or mixed pattern. For the uniform pattern condition, a value of .2 was subtracted from Sample 1 factor loadings for all four DF items to create Sample 2 factor loadings. For the mixed pattern condition, a value of .2 was subtracted from two Sample 1 DF items and a value of .2 was added to two other Sample 1 DF items to create Sample 2 factor loadings. When factor loadings are uniformly lower in Sample 2, the overall communality of the set of items for that group

TABLE 2
Factor Loadings and Correlations, Low Overdetermination Condition, Sample 1 Data

	F1	*F2*	*F3*	*F4*	*F5*	*F6*
F1	1.0					
F2	.25	1.0				
F3	.30	.35	1.0			
F4	.25	.30	.35	1.0		
F5	.30	.25	.25	0.25	1.0	
F6	.35	.30	.30	0.35	.35	1.0
Item						
I1	.84[a]					
I2	.71					
I3	.63					
I4	.45[b]					
I5		.77				
I6		.71				
I7		.63				
I8		.55				
I9			.84			
I10			.55			
I11			.63			
I12				.77[a]		
I13				.45[b]		
I14				.77		
I15					.84[a]	
I16					.45[b]	
I17					.71	
I18						.77[a]
I19						.55[b]
I20						.71

[a]Loadings were modified in the *which loading* high communality condition.
[b]Loadings were modified in the *which loading* low communality condition.

is reduced compared to the Sample 1 data. Conversely, when factor loadings are changed in a mixed pattern for the second group, the overall communality remains approximately the same as that of the Sample 1 data.[4]

In total, our study constitutes a 3 (sample size) × 2 (factor overdetermination) × 2 (factor correlation invariance) × 2 (which items were DF) × 2 (how

[4]We realize that an analytically simpler manipulation of communality would have been to hold the items with differing factor loadings constant and to manipulate the residual variances for those items. However, we chose to manipulate which items differed and the directionality of the difference because, in our experience, these types of differences in the data are more often encountered in applied research.

items were DF) design, resulting in a total of 48 conditions. We chose a very small subset of possible data properties as our goal was not to delineate all possible data conditions. Rather, we sought to examine a small subset to determine whether, in general, the data properties manipulated in this study could affect the precision of the estimated factor loading differences and the power of the LRT of those differences.

For each cell of the design, 500 replications of data were simulated using the PRELIS 2.5 program accompanying LISREL 8.54 (Jöreskog & Sörbom, 1996). Sampling error was introduced by using different seed numbers for the each replication for Sample 1 and 2 data; however, the same seed numbers for Sample 2 data were used across conditions to keep the amount of sampling error uniform. All data were simulated to be multivariate normal and models were fit to the observed covariance matrices obtained for each replication.[5]

Analyses

A CFA baseline model was estimated in which the correct factor structure was specified for both Group 1 and Group 2. Next, a constrained model was estimated in which the entire factor loading matrix was constrained to be equal for the Group 1 and Group 2 data. Covariance matrices were analyzed and factor variances were standardized to achieve model identification for all conditions. Results from models with standardized latent variances are equal to those using referent indicators when latent variances are known to be invariant across groups. A probability value of .05 was used in computing LRTs; LISREL 8.54 (Jöreskog & Sörbom, 1996) was used for all analyses.

Outcome Variables

Two outcomes of the Monte Carlo study were of key interest: the power of the test of metric invariance and the precision of the estimated group differences in the factor loadings. As such, for each of the 500 replications for each condition, tests of metric invariance were conducted by analyzing the two models previously described. Because the two models are nested, the difference in chi-square values between the two models is itself distributed as a chi-square, with degrees of freedom equal to the number of newly constrained parameters. This then provides a test statistic (i.e., LRT) for assessing the null hypothesis of metric

[5]We did not explicitly model mean structure, as our focus was on metric invariance, but this is equivalent to fitting a model with saturated means (i.e., where the intercepts of all indicators are unconstrained across items and groups). Because loading equivalence is typically tested prior to intercept equivalence, this choice is reasonable given the goals of the study.

invariance (Horn & McArdle, 1992). Power was computed as the proportion of chi-square tests (correctly) rejected within each cell of the design.

The precision of the estimated difference in groups' factor loadings was assessed by examining the standard error (*SE*) of the factor loading difference (see Equation 4). Inadmissible solutions were omitted from all analyses, but this had little impact on the results.[6] A total of 23,963 of the 24,000 solutions were admissible and used in further analyses.

RESULTS

Results are presented in two parts. First, the effects of sample size, factor overdetermination, item communality, and factor correlation invariance on the precision of the estimated difference in factor loading (across groups) are presented. Next, the effects of the design factors on the significance of tests of metric invariance are presented. In neither case did the factor correlation invariance manipulation meaningfully affect the results, so we do not present results from the noninvariance factor correlation conditions except in aggregated analyses (i.e., analysis of variance [ANOVA] and logistic regression models).

Precision of Factor Loading Estimates

Results of the *SE* of the difference values by condition are presented in Table 3 and Figure 1 charts these same values. As can be seen in Figure 1, all other things held constant, larger sample sizes were always associated with smaller *SE* values than smaller sample sizes. Similarly, holding other variables constant, higher factor overdetermination (i.e., three factors) was associated with lower *SE* values than were lower factor overdetermination conditions. There was a small but significant two-way interaction between number of factors and the communality of items chosen as DF items (*which* variable). Specifically, *SE* values were considerably lower when high communality items were modified than when low communality items were modified when there were three factors (i.e., high overdetermination). However, for six-factor (low overdetermination) data, *SE* values were nearly identical for high and low communality DF items.

ANOVA results of these same data indicated that the study variables accounted for 96% of the variance in *SE* values, $F(47, 23915) = 13278.8$, $p < .0001$. As can be seen in Table 4, all four study variables had a main effect on *SE* values, although only two accounted for nontrivial ($\omega^2 \geq .01$) proportions

[6]Solutions were considered inadmissible if the program did not converge in 1,000 iterations or if estimated parameters were out of bounds (e.g., negative uniqueness terms, standardized factor loadings or factor correlations over 1.0, etc.).

TABLE 3
Standard Error of Difference in Groups' Factor Loading Estimates, by Study Conditions

	K = 3				*K* = 6			
	All Lower		*Mixed*		*All Lower*		*Mixed*	
N	*Lowest*	*Highest*	*Lowest*	*Highest*	*Lowest*	*Highest*	*Lowest*	*Highest*
100	0.145	0.133	0.141	0.122	0.152	0.157	0.149	0.149
200	0.103	0.094	0.100	0.087	0.108	0.111	0.106	0.104
400	0.072	0.066	0.071	0.061	0.077	0.078	0.075	0.074

Note. *K* = number of factors. *All lower* and *mixed* indicate the manner in which factor loadings were manipulated. *Lowest* and *highest* indicate which factor loadings were manipulated. Results are for conditions in which factor correlations were equal across groups. Results for factor correlation inequality were identical within rounding.

TABLE 4
Effects of Study Variables on Standard Error of Difference in Groups' Factor Loading Estimates

Source	*df*	*F*	ω^2
Sample size (*N*)	2	291585.00	0.90
No. of factors (*F*)	1	22432.60	0.03
Which items manipulated (*W*)	1	5470.89	0.01
How items manipulated (*H*)	1	3991.86	0.01
Factor correlation invariance (*I*)	1	34.16	0.00
*F***W*	1	6048.76	0.01
*N***F*	2	804.93	0.00
*W***H*	1	680.12	0.00
*N***F***W*	2	348.45	0.00
*N***W*	2	130.92	0.00
*N***H*	2	122.51	0.00
*N***W***H*	2	23.22	0.00
*F***I*	1	35.97	0.00
*F***H*	1	32.32	0.00
*F***W***I*	1	21.89	0.00
*W***I*	1	20.51	0.00
*F***W***H*	1	11.02	0.00
*N***F***H*	2	3.99	0.00

Note. All *F* values are significant beyond the .01 level. Interactions sorted by effect size. Nonsignificant interactions omitted.

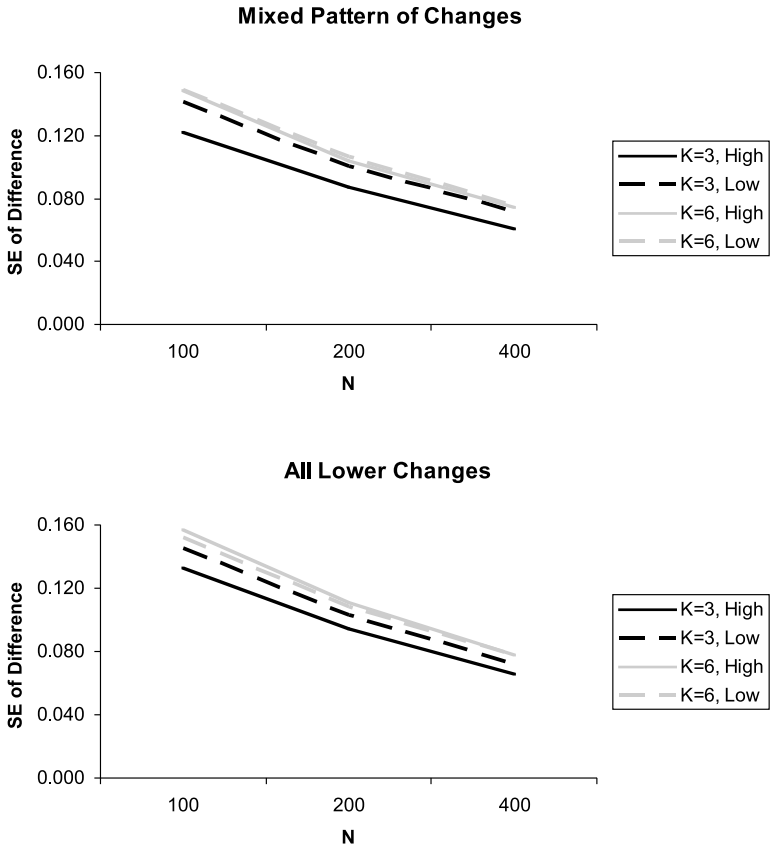

FIGURE 1 Standard error of difference in groups' factor loading estimates by study condition.

of variance in *SE*. As expected, sample size had the largest effect with factor overdetermination having a sizable effect as well; the communality manipulations had relatively small main effects, with a trivial effect of factor correlation invariance. In addition, eight two-way interactions were significant, as well as five three-way interactions. However, the effect sizes of the interactions were typically very small. The largest effect of these small interactions was that of factor overdetermination and the choice of DF items described earlier.

We also examined the mean estimated factor loading difference between Sample 1 and Sample 2 DF items. Although these estimates did vary slightly by condition, these differences were mostly trivial, centering around their population value of .20. Thus, as expected, the mean values of estimated differences in

TABLE 5
Standard Deviation of Difference Between Sample 1 and Sample 2 Factor Loading Estimates for Four DF Items

	K = 3				*K* = 6			
	All Lower		*Mixed*		*All Lower*		*Mixed*	
N	*Lowest*	*Highest*	*Lowest*	*Highest*	*Lowest*	*Highest*	*Lowest*	*Highest*
100	0.056	0.056	0.054	0.042	0.061	0.060	0.056	0.068
200	0.043	0.041	0.039	0.030	0.048	0.042	0.039	0.044
400	0.030	0.029	0.027	0.022	0.032	0.029	0.030	0.032

Note. *K* = number of factors. All lower and mixed indicate the manner in which factor loadings were manipulated. Lowest and highest indicate which factor loadings were manipulated. Results are for conditions in which factor correlations were equal across groups. Results for factor correlation inequality were identical within rounding.

item parameters were not biased across the 500 replications, even with sample sizes of 100. Additionally, we examined the standard deviation of this difference across the 500 replications in the study (see Table 5). Like the *SE* of this difference, these values indicate the accuracy with which the factor loadings were estimated for the four DF items in the two groups. As expected, analyses of these effects were highly similar to those of the *SE* values.

Power to Detect Factor Loading Differences

Results of the power of the LRT by study condition are presented in Table 6 and Figure 2. As with *SE*, holding other study variables constant, larger sample sizes were always associated with a larger number significant metric invariance tests. Large effects of sample size were expected as sample size directly affects the precision of estimated parameters and is incorporated into the χ^2-based LRT formula. Similarly, all other variables held constant, more metric invariance tests were significant when the pattern of changes were mixed rather than uniformly lower, although the magnitude of these differences varied greatly depending on the levels of other study variables. Finally, all other things being equal, more metric invariance tests were significant when factor overdetermination was high than low, although again the magnitude of this difference varied from very small to large by condition. Interpretations of other trends were complicated by interactions with other variables.

The result of the metric invariance LRT constitutes a significant-nonsignificant dichotomous dependent variable, necessitating the use of logistic regression to examine the effects of the study variables. The overall logistic regression model

TABLE 6
Percentage of Metric Invariance Tests Significant by Study Condition

	K = 3				*K* = 6			
	All Lower		*Mixed*		*All Lower*		*Mixed*	
N	*Lowest*	*Highest*	*Lowest*	*Highest*	*Lowest*	*Highest*	*Lowest*	*Highest*
100	7	12	10	40	8	8	9	11
200	34	53	43	95	30	22	29	36
400	90	99	96	100	81	79	88	91

Note. *K* = number of factors. All lower and mixed indicate the manner in which factor loadings were manipulated. Lowest and highest indicate which factor loadings were manipulated. Results are for conditions in which factor correlations were equal across groups. Results for factor correlation inequality were identical within rounding.

was significant (Wald = 6308.29, $p < .0001$). Although no directly comparable index of R^2 is available in logistic regression, Cox and Snell's have provided an approximation to R^2 (see Hair, Anderson, Tatham, & Black, 1998, for a review). The Cox and Snell R^2 value was .44, but the maximum value of this statistic is less than 1.0, thus we also report the Nagelkerke R^2 static, which has a maximum value of 1.0. The Nagelkerke R^2 statistic was .58 for this model; taken together, these indexes indicate that a moderately large proportion of variance in the significance of metric invariance test was due to the study conditions (with the remainder due to sampling variability). Wald significance statistics, standardized parameter estimates, odds ratios, and their associated confidence intervals for individual study variables can be found in Table 7.

There was a main effect for sample size, factor overdetermination, and the manner in which items differed (see Table 7). Conversely, there was no main effect for which items were chosen as DF items or the invariance of correlations among the latent factors. Additionally, there were several two-way interactions that were significant. For instance a large interaction was found between which items were chosen as DF and factor overdetermination. Examination of Table 6 and Figure 2 indicates that this interaction was due to a large effect of how communality was manipulated for three-factor data, particularly where $N = 200$. In the mixed high condition, items already high in communality were given even higher communalities, and at moderate sample sizes ($N = 200$), this boosted power when the factors were sufficiently overdetermined. Interestingly, although there was no main effect of which items were DF, this variable had a sizable effect as manifest through the interaction with factor overdetermination. Unlike *SE* analyses, the effect of the manner in which communality was manipulated seemed to vary greatly for three and six factors where $N = 200$. Namely, metric

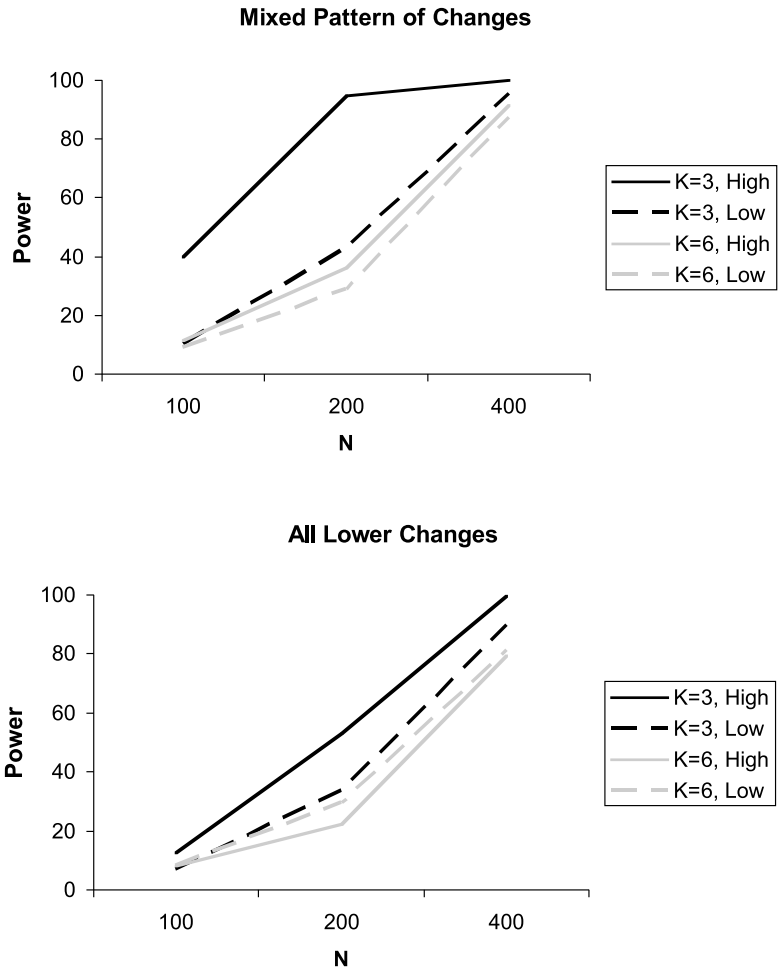

FIGURE 2 Percentage of metric invariance tests significant by study condition.

invariance tests were significant more often for mixed patterns of changes for three-factor data (particularly for $N = 200$), than for uniformly lower changes for three-factor data. Additionally, although the two-way interaction of sample size and the number of factors was significant, these results were contingent on which items were chosen as DF items, as indicated by the three way-interaction of these variables.

Finally, we assessed the relationship between *SE* of the difference in estimated factor loadings and power. As expected, the mean *SE* values were significantly

TABLE 7
Results of Logistic Regression of Metric Invariance Test on Study Variables

Parameter	*β*	*SE*	*Wald Statistic*	*Odds Ratio*	*Odds Ratio 95% Confidence Interval*	
Intercept	1.10	0.20	30.87**			
*N*100	−1.97	0.09	528.98**	0.02	0.01	0.03
*N*200	−1.18	0.05	479.44**	0.09	0.08	0.12
No. of factors (*F*)	0.32	0.07	18.01**	1.88	1.41	2.52
Which (*W*)	−0.07	0.07	0.92	0.87	0.65	1.16
How (*H*)	0.23	0.08	7.81**	1.59	1.15	2.21
Factor correlation invariance (*I*)	0.08	0.05	2.28	1.17	0.96	1.42
*N*200* *H*	0.48	0.20	5.85*	1.62	1.10	2.40
*N*100* *F*	0.78	0.21	13.38**	2.19	1.44	3.32
*N*200* *F*	0.41	0.16	6.35*	1.51	1.10	2.08
*W** *F*	−3.13	0.61	26.46**	0.04	0.01	0.14
*W** *H*	−0.55	0.26	4.61*	0.58	0.35	0.95
*N*100* *W** *F*	2.51	0.66	14.32**	12.27	3.35	44.94
*N*200* *W** *F*	1.97	0.63	9.82**	7.15	2.09	24.45

Note. $N = 100$ and $N = 200$ conditions were dummy-coded and used in these analyses. $df = 1$ for all analyses above. Nonsignificant interactions omitted. For the sample size factor, $N = 400$ was the reference category.

lower for replications producing significant metric invariance tests ($M = .085$, $SD = .022$) than nonsignificant metric invariance tests ($M = .125$, $SD = .025$), $F(1, 23961) = 17405.3$, $R^2 = .42$. Moreover, this trend held after controlling for all other study variables, indicating that even within a particular study condition, significant metric invariance tests tended occur more often when the groups' factor loadings were more precisely estimated within conditions. Differences in *SE* are due only to the influence of sampling variability and thus underscore the important relationship between precision and power.

Additional Analyses

One limitation of reliance on the chi-square-based test of metric invariance is that, like chi-square tests of overall model fit, the test is highly sensitive to sample size (Brannick, 1995; Kelloway, 1995; Meade & Lautenschlager, 2004). Thus, in large samples, power to detect even trivial differences in the properties of a measure between groups is extremely high, potentially leading to overidentification of a lack of invariance. Thus, Cheung and Rensvold (2002) examined the potential use of change in alternative fit indexes (AFIs) in MI investigations. As a result of their simulation work, they recommended that researchers

report the change in comparative fit index (CFI), Gamma-hat, and McDonald's Noncentrality Index (NCI) fit indexes. We report the difference in our baseline and constrained model fit for these indexes (and the root mean squared error of approximation due to its prevalence of use in overall determination of fit) in Table 8.

As can be seen in Table 8, as sample size increases, the change in each AFI increases and the standard deviation of these differences decreases, as would be expected. Additionally, higher communalities for the DF items were associated with greater differences in the AFI values between models, especially in the high overdetermination conditions (three factors). This was true both when the high communality items were chosen as the DF items and when a mixed pattern of DF was imposed. In the low overdetermination condition, the communalities of the DF items had less effect on the AFI differences. Factor correlation invariance seemed to have little or no effect. Overall, these results are similar to those obtained from the LRT. Of note is the fact that sample size continued to play a large role in the magnitude of the difference in AFI value, suggesting that to avoid the overidentification of trivial effects, direct assessments of effect size (loading differences) may often be a useful supplement to examining change in AFIs.

DISCUSSION

There are several important findings from this study. Perhaps the most important finding from this study is that the psychometric properties of the data, and not just sample sizes, affect the precision and power of MI tests. To our knowledge, this study is the first to illustrate these effects in multi-group CFAs. Our results indicated that under all conditions simulated, the precision of the estimated factor loading differences was uniformly high for sample sizes of 400, but varied somewhat by condition at sample sizes of 100 and 200. Conversely, the power of the LRT to detect these differences was uniformly low for sample sizes of 100, uniformly high for sample sizes of 400, and power differed greatly by condition for sample sizes of 200. Although sample size did have the strongest effect on both precision and power, by no means were the effects of other study variables trivial. For instance, for sample sizes of 200 per group, the percentage of metric invariance tests that were significant varied from 22% to 95%. Thus, as MacCallum et al. (1999) found with exploratory factor analysis, no one rule of thumb regarding the number of respondents per indicator is appropriate for all data. Instead, the sample sizes required to provide adequate power depend highly on the items' communality and factor overdetermination, and of course, the effect size of the DF of items.

A second notable finding of this study was the finding that the power of MI tests and the precision with which factor loadings are estimated are largely

TABLE 8
Difference in Alternative Fit Indexes Between Baseline and Constrained Models, by Condition

No. of Factors	*Which Items DF*	*How DF Simulated*	*Factor Correlation Invariance*	*RMSEA*			*CFI*		*McDonald's NCI*		*Gamma-Hat*	
				N	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
3	Lowest	All lower	Equal	100	−0.0002	0.0050	0.0011	0.0046	0.0028	0.0152	0.0003	0.0016
				200	0.0024	0.0042	0.0025	0.0030	0.0103	0.0097	0.0011	0.0010
				400	0.0054	0.0042	0.0035	0.0024	0.0148	0.0061	0.0015	0.0006
			Not equal	100	−0.0003	0.0049	0.0012	0.0047	0.0032	0.0154	0.0004	0.0016
				200	0.0025	0.0042	0.0026	0.0030	0.0107	0.0098	0.0011	0.0010
				400	0.0055	0.0043	0.0035	0.0024	0.0149	0.0062	0.0015	0.0006
		Mixed	Equal	100	0.0002	0.0051	0.0016	0.0045	0.0054	0.0159	0.0006	0.0017
				200	0.0031	0.0042	0.0030	0.0029	0.0130	0.0099	0.0013	0.0010
				400	0.0064	0.0046	0.0040	0.0025	0.0175	0.0065	0.0018	0.0007
			Not equal	100	0.0003	0.0051	0.0017	0.0045	0.0058	0.0156	0.0006	0.0017
				200	0.0032	0.0043	0.0031	0.0029	0.0133	0.0099	0.0014	0.0010
				400	0.0066	0.0046	0.0041	0.0025	0.0177	0.0065	0.0018	0.0007
	Highest	All lower	Equal	100	0.0005	0.0053	0.0027	0.0054	0.0082	0.0160	0.0009	0.0017
				200	0.0036	0.0045	0.0042	0.0036	0.0160	0.0103	0.0016	0.0011
				400	0.0077	0.0047	0.0055	0.0031	0.0204	0.0070	0.0021	0.0007
			Not equal	100	0.0007	0.0054	0.0028	0.0054	0.0086	0.0159	0.0009	0.0017
				200	0.0037	0.0046	0.0043	0.0036	0.0163	0.0105	0.0017	0.0011
				400	0.0078	0.0047	0.0056	0.0031	0.0206	0.0071	0.0021	0.0007
		Mixed	Equal	100	0.0039	0.0068	0.0066	0.0060	0.0236	0.0200	0.0025	0.0021
				200	0.0086	0.0066	0.0077	0.0043	0.0321	0.0124	0.0033	0.0013
				400	0.0133	0.0049	0.0089	0.0031	0.0359	0.0082	0.0037	0.0008
			Not equal	100	0.0040	0.0069	0.0067	0.0060	0.0239	0.0198	0.0026	0.0021
				200	0.0086	0.0066	0.0077	0.0044	0.0323	0.0124	0.0033	0.0013
				400	0.0133	0.0048	0.0089	0.0032	0.0360	0.0083	0.0037	0.0009

(*continued*)

TABLE 8
(*Continued*)

No. of Factors	*Which Items DF*	*How DF Simulated*	*Factor Correlation Invariance*	*RMSEA*			*CFI*		*McDonald's NCI*		*Gamma-Hat*	
				N	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
6	Lowest	All lower	Equal	100	−0.0003	0.0049	0.0015	0.0065	0.0034	0.0154	0.0004	0.0016
				200	0.0024	0.0043	0.0034	0.0040	0.0092	0.0093	0.0009	0.0010
				400	0.0051	0.0042	0.0043	0.0031	0.0127	0.0060	0.0013	0.0006
			Not equal	100	0.0002	0.0055	0.0023	0.0065	0.0054	0.0159	0.0006	0.0017
				200	0.0028	0.0045	0.0037	0.0041	0.0101	0.0096	0.0010	0.0010
				400	0.0054	0.0042	0.0044	0.0031	0.0132	0.0061	0.0013	0.0006
		Mixed	Equal	100	−0.0002	0.0048	0.0019	0.0062	0.0039	0.0154	0.0004	0.0016
				200	0.0024	0.0040	0.0033	0.0036	0.0099	0.0087	0.0010	0.0009
				400	0.0057	0.0044	0.0044	0.0030	0.0138	0.0059	0.0014	0.0006
			Not equal	100	0.0001	0.0048	0.0025	0.0061	0.0057	0.0156	0.0006	0.0017
				200	0.0026	0.0041	0.0035	0.0036	0.0106	0.0089	0.0011	0.0009
				400	0.0059	0.0043	0.0044	0.0029	0.0141	0.0060	0.0014	0.0006
	Highest	All lower	Equal	100	−0.0005	0.0050	0.0011	0.0072	0.0020	0.0157	0.0002	0.0017
				200	0.0018	0.0041	0.0030	0.0040	0.0077	0.0089	0.0008	0.0009
				400	0.0047	0.0041	0.0041	0.0031	0.0117	0.0058	0.0012	0.0006
			Not equal	100	−0.0003	0.0049	0.0018	0.0073	0.0038	0.0161	0.0004	0.0017
				200	0.0021	0.0042	0.0033	0.0041	0.0087	0.0092	0.0009	0.0009
				400	0.0051	0.0042	0.0043	0.0032	0.0125	0.0061	0.0013	0.0006
		Mixed	Equal	100	0.0003	0.0052	0.0021	0.0065	0.0052	0.0165	0.0006	0.0018
				200	0.0029	0.0043	0.0037	0.0039	0.0112	0.0096	0.0011	0.0010
				400	0.0063	0.0045	0.0047	0.0030	0.0148	0.0058	0.0015	0.0006
			Not equal	100	0.0007	0.0052	0.0030	0.0067	0.0076	0.0170	0.0008	0.0018
				200	0.0034	0.0046	0.0042	0.0041	0.0126	0.0101	0.0013	0.0010
				400	0.0068	0.0046	0.0050	0.0030	0.0159	0.0061	0.0016	0.0006

Note. DF = differential functioning; RMSEA = root mean squared error of approximation; CFI = comparative fit index; NCI = XXXXXX.

affected by the same variables. Our analyses for the effects of study variables on the precision of estimated of factor loading differences largely mirrored those of the results of the metric invariance tests. Moreover, significant metric invariance tests had more precise estimates of the differences in factor loadings for DF items than did nonsignificant metric invariance tests, even when controlling for all study variables. These results suggest that the results of metric invariance tests are closely related to accurate estimation of DF item parameters. The precision of the estimates of factor loadings is, in turn, dependent on sample size, overidentification of the factors, and item communalities.

Although several of the interactions between study variables were significant, only one of these was a disordinal (or crossing) interaction, making the interpretation of the results simpler. Holding other study variables constant, larger sample sizes and a mixed pattern of changes in factor loadings were always associated with both a larger number of significant metric invariance tests and more accurate DF item factor loading estimates than were smaller sample sizes and uniformly lower factor loading changes. These results parallel those from Meade and Lautenschlager (2004). However, as evidenced by the interactions of the study variables, the magnitude of these differences varied greatly by the amount of factor overidentification and which items were chosen as DF items.

One consistent yet somewhat surprising finding in this study was the disordinal interaction between factor overidentification and the choice of which items were DF items. For the case of high levels of overidentification (i.e., three factors and 20 items), when DF items had high communality, factor loading differences were more accurately estimated and metric invariance tests were more likely to be significant than when DF items had low communality. However, when factor overidentification was low (i.e., six factors and 20 items), factor loading differences were estimated with equal precision and metric invariance tests were more likely to be significant when DF items had low communalities than when DF items had high communalities (but only when DF was uniform in nature). In this case, it might be that there are so few high communality items to determine the factors that reducing the communality of even one of these items leads to less factor reliability and hence lower power and precision in general. Modifying the low communality items, in contrast, would not necessarily compromise the determination of the entire factor. Across all conditions, factor correlation invariance (or lack thereof) had little effect on either precision or power rates.

Summary and Recommendations

The most significant finding from this study as it relates to applied researchers is that sample size alone does not determine the power of CFA MI tests. Factor overdetermination and communality also play large roles in the efficacy of these tests to detect a lack of MI. One peculiarity of MI research is that typically a

lack of MI is seen as undesirable. In other words, most researchers treat MI as a statistical hurdle that must be passed before progressing to more interesting substantive questions (Vandenberg & Lance, 2000). As a result, it typically is in researchers' best interest to create conditions under which the null hypothesis is not rejected and MI is found. As this study suggests, these conditions include low sample sizes, few indicators per factor, and relative low communality in the items. Clearly, this is an undesirable psychometric situation.

As seen in this study, favorable psychometric properties can lead to a greater chance of detecting a lack of MI. Thus, applied researchers who deal with large sample sizes and well-developed psychometric instruments can reference this study as an example of how these factors work against them in MI tests. To offset this manner of thinking, we recommend that researchers discontinue viewing MI as an either–or proposition. Instead, if a lack of metric invariance is found, researchers may wish to calculate effect sizes and confidence intervals for the factor loading differences. A significant decrement in fit could be due to large differences in estimated factor loadings for a relatively small number of items or small differences in estimated factor loadings for a large number of items. Inspection of the estimated factor loadings and the confidence intervals around these estimates can clarify the nature of the lack of fit, allowing researchers the option of allowing for partial invariance (Byrne, Shavelson, & Muthén, 1989), or removal of items. Moreover, if these intervals are small and close to (but exclude) zero, then one might still be justified in pursuing more substantive research questions, particularly if latent rather than observed score means are compared. There might also be occasions, particularly when sample sizes are extremely large, when even small differences in factor loading estimates lead to significant metric invariance test results. Presumably, though, this will be reflected in narrow confidence intervals around very small effect sizes. Thus, inspection of the magnitude of the factor loading difference and the corresponding confidence interval could aid researchers in making a case that the chi-square-based metric invariance tests are detecting differences that are largely inconsequential for the assessment of their research hypotheses. On the other hand, if the intervals fall farther from zero, the loading differences may be of greater practical importance. One way to evaluate this importance was suggested by Millsap and Kwok (2004), who evaluated the impact of a lack of MI on selection decision accuracy. With large sample sizes and favorable psychometric properties, it could well be the case that a statistical lack of MI makes little practical difference in usage of the instrument. Conversely, in small samples, wide confidence intervals may indicate that, although there are no significantly different factor loadings, the precision of the estimated difference is too low to make firm conclusions about the presence or absence of MI.

Regardless of whether factor loading differences are a nuisance or of substantive interest, applied researchers should take care to ensure that they have

adequate power to detect meaningful departures from MI. As shown here, it may be possible to compensate for a relatively small sample by collecting data on more factor indicators, ideally with high reliability. As such, sample size limitations or the difficulty and expense of recruiting additional participants should not preclude sound psychometric research.

Limitations and Future Research

Even though this was a large simulation study, like all simulation studies, this study was limited in its scope. There are an infinite number of factor models and ways in which data can vary between two groups and we considered only a very small subset of these possibilities. As a result, most researchers would be unlikely to be able to reference this study as a justification of adequate power to detect a lack of MI with their own data. However, our goal was only to provide an initial exploration of the general data and model properties that affect the precision and power to detect factor loading differences. Additionally, we only considered tests of metric invariance. Although these tests are typically considered to be the most important MI tests (presuming the same general factor structure holds in both samples), the effects of data properties on other MI tests will be important to consider in follow-up research.

REFERENCES

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

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

Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal mean and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM). *Organizational Research Methods, 1,* 421–483.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143–159.

Chan, D., & Schmitt, N. (2000). Interindividual differences in intraindividual changes in proactivity during organizational entry: A latent growth modeling approach to understanding newcomer adaptation. *Journal of Applied Psychology, 85,* 190–210.

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

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95,* 134–135.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19–29.

Facteau, J. D., & Craig, S. B. (2001). Are performance appraisal ratings from different rating sources comparable? *Journal of Applied Psychology, 86,* 215–227.

Gerbing, D. W., & Anderson, J. C. (1987). Improper solutions in the analysis of covariance structures: Their interpretability and a comparison of alternate respecifications. *Psychometrika, 52,* 99–111.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.

Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement, 65,* 202–226.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

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

Kaplan, D. (1989). Power of the likelihood ratio test in multiple group confirmatory factor analysis under partial measurement invariance. *Educational and Psychological Measurement, 49,* 579–586.

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

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

Marsh, H. W. (1985). The structure of masculinity/femininity: An application of confirmatory factor analysis to higher-order factor structures and factorial invariance. *Multivariate Behavioral Research, 20,* 427–449.

Marsh, H. W. (1987). The factorial invariance of responses by males and females to a multidimensional self-concept instrument: Substantive and methodological issues. *Multivariate Behavioral Research, 22,* 457–480.

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181–220.

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

Meade, A. W., & Kroustalis, C. M. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods, 9,* 369–403.

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.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57,* 289–311.

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research, 30,* 577–605.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2,* 248–260.

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

Ployhart, R. E., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods, 7,* 27–65.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata and situational judgment tests comparable? *Personnel Psychology, 56,* 733–752.

Riordan, C. M., Richardson, H. A., Schaffer, B. S., & Vandenberg, R. J. (2001). Alpha, beta, and gamma change: A review of past research with recommendations for new directions. In C. A. Schriesheim & L. L. Neider (Eds.), *Equivalence of measurement* (pp. 51–97). Greenwich, CT: Information Age.

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

Vandenberg, R. J. (2002). Toward a further understanding of an 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.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.