5-1-2016

Factorial Invariance Testing under Different Levels of Partial Loading Invariance within a Multiple Group Confirmatory Factor Analysis Model

Brian F. French
Washington State University, frenchb@wsu.edu

Holmes Finch
Ball State University, whfinch@bsu.edu

Factorial Invariance Testing under Different Levels of Partial Loading Invariance within a Multiple Group Confirmatory Factor Analysis Model

Brian F. French
Washington State University
Pullman, WA

Holmes Finch
Ball State University
Muncie, IN

Scalar invariance in factor models is important for comparing latent means. Little work has focused on invariance testing for other model parameters under various conditions. This simulation study assesses how partial factorial invariance influences invariance testing for model parameters. Type I error inflation and parameter bias were observed.

Keywords: Latent variable modeling, invariance, factor analysis

Introduction

Confirmatory factor analysis (CFA) is generally considered the preferred factor analytic approach for assessing scale dimensionality when both theory and empirical evidence support a particular latent structure. CFA is a model-based approach to examining whether there is empirical support for a theoretical latent structure and if the factor structure is equivalent across groups. Questions related to model equivalency across groups falls under the category of measurement invariance (MI; Bollen, 1989; Byrne, Shavelson, & Muthén, 1989; Millsap, 2011). A lack of MI is present when an assessment is used to measure a psychological (e.g., motivation) or educational (e.g., mathematical) ability and that assessment produces different results (i.e., scores) for persons from different groups (e.g., boys vs. girls) when those persons are of equal status on the ability assessed (Bollen, 1989; Drasgow & Kanfer, 1985; Millsap, 2011). Stated another way, the measurement properties of the instrument in relation to the ability assessed are the

same across pre-identified groups. Methods for identifying a lack of measurement invariance are well-studied. However, the influence of partial invariance, explained below, is not well-documented and becomes difficult to show analytically with many variables (Millsap, 2011). Thus, the goal of the current simulation study is to examine the impact of partial invariance on the invariance testing of factor model parameters (i.e., testing for factorial invariance), including factor intercepts, error variances, factor variances, factor covariances, and factor means under a variety of conditions, including differences in sample size, number of factors, and number of indicators per factor.

Factorial Invariance

Within the measurement invariance area, we focus our emphasis on factorial invariance (FI) examined through the use of multigroup CFA. FI has received increased attention in the past few years with sections of books devoted to the topic (e.g., Millsap, 2011; Schriesheim & Neider, 2001) as well as software being automated (e.g., MPLUS 7.2) to make the assessment of invariance accessible to a wide audience. This emphasis is a direct reflection of the essential role that assessment scores play in society (e.g., high-stakes decisions) ranging from education (e.g., teacher evaluations; international student achievement comparisons) to business (e.g., job applicant decisions). In fact, there is an increasing body of evidence that suggests that many observed differences that are cross-cultural may be contaminated by artifacts of measurement or a lack of factorial invariance (Baumgartner & Steenkamp, 2001; Church et al., 2011; Javaras & Ripley, 2007; Poortinga, 1989). These differences may be related to content meaning issues, translation problems, or even response style differences, and in turn, can result in incorrect decisions regarding individuals from different groups as well as group comparisons (French & Finch, 2008b; Millsap & Kwok, 2004; Steinmetz, 2013).

The examination of the internal structure of the instrument for FI is an important step in providing psychometric evidence supporting the validity argument for score use (American Educational Research Association et al., 2014; Wu, Li, & Zumbo, 2007). FI refers to the situation in which the latent factor structure underlying a scale is equivalent across predefined groups (Meredith & Millsap, 1992; Millsap, 2011; Millsap & Kwok, 2004). FI can be further described in terms of the factor model parameters being equivalent across groups. The investigation of invariance is dependent on a specific variable of interest that separates the groups (e.g., biological sex). This latter definition implies that several

parameters (e.g., factor loadings, intercepts, error variances) are equal across groups. We define the different levels of invariance in the factor model below to make explicit how levels of FI relate to one another.

Given that FI refers to a set of assumptions regarding the invariance of various parameters associated with the factor structure of an instrument, it is important to understand each aspect of FI. Millsap (2011) provides an excellent discussion of the levels of FI from weak, pattern, or metric invariance, which refers to pattern matrix invariance (Horn & McArdle, 1992; Millsap, 2011; Widaman & Reise, 1997) to strong or scalar factorial invariance (SI) referring to factor model intercepts being equal across groups (e.g. Steenkamp & Baumgartner, 1998). The method for assessing the invariance assumptions is based upon multiple groups confirmatory factor analysis (MGCFA) which begins by ensuring the general or configural form (CI) of the factor model is present across groups, where only the number of latent variables present, and the correspondence of observed indicators to factors is the same for all groups in the population. The weak and strong forms of invariance place corresponding constraints on the model. The presence of FI implies that the latent variables are being measured in the same way for the population subgroups under consideration (Wicherts & Dolan, 2010). Said another way, "the question of factorial invariance concerns the extent to which the factor structure underlying the measured variables is the same across multiple populations" (Millsap, 2011, p. 73). This implies that scores on the observed manifestation of the latent variable (i.e. expected score on the scale) are the same for members of different groups who have the same level of the latent trait being measured (Wicherts & Dolan, 2010).

When scalar variance does not hold, it is difficult, if not impossible, to know the extent to which group differences on a mean scale score are due to group differences on the latent trait of interest, or due to group differences on the intercepts (Steinmetz, 2013). The dependency of the differences in scale score means on a lack of invariance exists in applied studies (e.g., French & Mantzicopoulos, 2007). In addition to the three forms of invariance mentioned, it is also possible to assess whether there is group invariance with respect to the unique indicator variance (δ). This strict factorial invariance (SFI; Millsap, 2011) occurs when the factor loadings, intercepts, and unique variances are invariant. In addition, SFI is necessary in order to attribute group differences in the mean and covariance structure of the observed indicators to corresponding differences at the latent variable level (Millsap, 2011).

Multiple Group Confirmatory Factor Analysis (MGCFA)

FI can be assessed using MGCFA, where the standard CFA model is expressed as (Bollen, 1989):

$$x_g = \tau_g + \Lambda_g \xi_g + \delta_g \tag{1}$$

It is possible to have indicators (x_g), intercepts (τ_g), loadings ($\mathbf{\Lambda}_g$), and unique variances (δ_g) that are specific to each group within the population. Likewise, the indicator covariance matrix and associated latent means can be expressed respectively as:

$$\mathbf{\Sigma}_g = \mathbf{\Lambda}_g \mathbf{\Psi}_g \mathbf{\Lambda}_g' + \mathbf{\Theta}_g, \quad \mu_g = \tau_g + \mathbf{\Lambda}_g \mathbf{K}_g \tag{2}$$

such that groups are allowed unique observed covariance matrices ($\mathbf{\Sigma}_g$), factor loadings ($\mathbf{\Lambda}_g$), factor covariance matrices ($\mathbf{\Psi}_g$), and unique error matrices ($\mathbf{\Theta}_g$). In addition, the observed mean for group g is also a function of the intercept for that group (τ_g), the loadings, and the factor mean ($\mathbf{K}_g$). This implies the factor model holds in each population (Millsap, 2011).

MGCFA can be used to test each level or constraint on the factor model to evaluate FI that was described previously, using a series of nested models. For example, to assess CI, a model is fit such that the number of factors is the same across groups, as are the indicators associated with each of these factors. The model specification allows intercepts, loadings, and unique variances to vary across groups. Good model fit, based on appropriate indices (e.g., chi-square, CFI; Hu & Bentler, 1999), would indicate the presence of CI, and is necessary before the investigation of other types of invariance and placing additional constraints on the model. Often model comparison is made using a difference in chi-square (χ^2) statistic values between the less and more restrictive values. The use of the χ^2 statistic is somewhat problematic as a measure of absolute model fit (Bollen, 1989). However, there is evidence that the $\chi^2_{\text{difference}}$ statistic is an accurate tool for comparing the fit of two nested models (e.g., French & Finch, 2006). Other fit indices have been suggested (e.g., change in CFI, RMSEA) to asses FI. Given the lack of clear guidelines on the accuracy of the amount of change needed to indicate differences, the chi-square statistic remains the focus of this study.

Partial Factorial Invariance

Although FI is desirable for educational and psychological scales to possess, in practice it may be a rare commodity (Millsap & Meredith, 2007). For example, it seems with various types of scales it is unusual for complete FI to exist (Church et al., 2011; French & Gotch, 2013; French & Mantzicopoulos, 2007). Group equality on some but not all factor parameters is known as partial factorial invariance (PI) and does exist on major instruments such as intelligence measures (Maller & French, 2004). In one of the first studies to describe PI, Byrne et al. (1989) explored how researchers identify specific factor parameters (e.g. loadings) that are not group invariant after an initial rejection of the complete invariance hypothesis. Using this sensitivity analysis approach, it is possible to identify and release specific model parameters that differ across groups, leading to a PI model (Millsap, 2011). If PI is indeed found, the next question for researchers is to determine whether these differences in measurement structure are meaningful in practice. While the question of whether or not invariance holds can be addressed in a more or less straightforward manner using the MGCFA methodology described above, the issue of what to do about PI is not so clearly addressed, nor is the impact of PI at one level on assessing invariance at another level well understood.

Goals of the Current Study

The use of MGCFA for testing FI and latent mean differences has experienced growth (Vandenberg & Lance, 2000), with a focus on appropriate practices for testing invariance with attention on accuracy (French & Finch, 2006; Meade & Lautenschlager, 2004; Yoon & Millsap, 2007). Much of this work has stemmed from recommendations for researchers and practitioners not to assume the universal accuracy of MGCFA across many data conditions. While this recommendation has been followed with the implementation of Monte Carlo studies (French & Finch, 2006; Meade & Lautenschlager, 2004; Yoon & Millsap, 2007), gaps remain in the research on several issues related to FI and MGCFA. In particular, MGCFA procedures for identification of a lack of FI require further examination to evaluate accuracy under various conditions where no solution may be fully known with analytic work. There remains uncertainty as to the influence of PI on the assessment of invariance of factor model parameters, including intercepts, error variances, factor variances, and factor means.

There is evidence in the latent variable modeling literature that of the presence of non-invariance for one model parameter can lead to inflated Type I error rates for detection of non-invariance for another model parameter. In the context of IRT,

for example, group differences in item difficulty parameters are associated with inflated Type I error rates for the detection of group differences on the item discrimination parameter (French & Finch, 2008b), even in the presence of no population differences in discrimination parameters across groups. Similarly, when item discrimination values differ across groups, the Type I error rate for detecting group differences in item difficulty values are inflated. These findings have been documented in simulation work, by reviewers of such work, and in applied analysis (Finch & French, 2008a). Given the close link between IRT and CFA models (e.g. McDonald, 1999), the same confound may be present for MGCFA analysis. That is, group difference in factor loadings could lead to inflation of the Type I error rate for testing group differences on factor intercepts, error variances, factor variances, factor covariances, and factor means. Such inflated Type I error rates may in turn be especially problematic for specification searches (Millsap, 2011) to identify true group differences on CFA model parameters. Ideally, an analytic solution to address this issue could be employed. However, as noted by Millsap and Kwok (2004) deriving such an analytic solution becomes exceedingly difficult for models that consist of more than one latent variable and a small number of indicators, leading to the necessity of simulation research.

Given that most real world applications of CFA involve models with multiple factors and multiple indicators, an analytic solution to investigate the impact of non-invariant factor loadings on testing invariance for other model parameters will likely be too limited in scope to be informative for most applications. Thus, we turn to simulations to observe whether the presence of non-invariant loading parameters in a CFA model impacts the testing of invariance for other model parameters as has been reported for a similar situation in the context of IRT (Finch & French, 2008a). Moreover, the examination of bias of parameter estimates in such situations is difficult to derive analytically, whereas through simulation we can determine how PI influences bias (Boomsma, 2013).

Thus, the goal of this study was to begin providing insight to the impact of PI on FI assessment by addressing two research questions:

1. What is the influence of partial factor loading invariance on Type I error and power rates for invariance testing of other model parameters beyond the factor loadings? Specifically, what is the influence of incorrectly modeling such partial invariance? We hypothesize that incorrectly modeling or ignoring factor loading differences across groups will result in inflated Type I error rates when testing the invariance of other model parameters.

2.   What influence does partial factor loading invariance have on estimation of other factor model parameters, including intercepts, error variances, factor variances, factor covariances, and factor means? We hypothesize that the estimates of other model parameters will be attenuated for the group with the larger values when groups' factor loading differences are ignored.
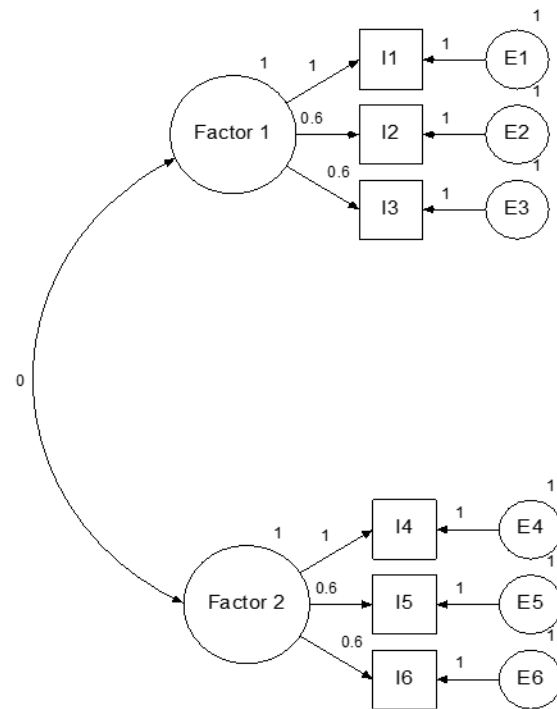
Figure 1. Example model used to simulate the data

Methods

To test our hypotheses, a Monte Carlo Simulation study (1000 replications per combination of conditions) was conducted. All simulations were conducted using Mplus 7.1 (Muthén & Muthén, 2013). An example of the model used to simulate the data appears in Figure 1. This example is the simplest model used, with 2 factors and 3 indicators per factor. Two groups were simulated across all conditions. The manipulated variables are described below. These conditions were completely crossed with one another, yielding a total of 240 different simulated conditions, or

design cells. In essence, each model parameter for which invariance was assessed can be viewed as representing a unique simulation study. The parameters that were tested for differences in the population were the factor intercepts, error variances, factor variances, factor covariances, and factor means. For each of these parameters group differences were simulated at varying levels, which are described below. In addition, when assessing group invariance for each of these parameters, factor loading values were allowed to vary at different levels. The experimental factors are described below. We also include sample Mplus code for the models in the appendix.

Experimental Factors Manipulated in the Simulation

Percent of factor loading PI To assess the influence of factor loading PI on assessment of other model parameters, loadings were simulated to differ between the groups. Specifically, in one condition loadings were simulated to be equal across groups, while in a second case, they were simulated to differ by 0.25, 0.50, and 1.00. The latter two conditions were included in order to assess the performance of the MGCFA model for invariance testing in more extreme cases of factor noninvariance. For each noninvariant condition, 34% of loadings lacked invariance. This allowed for the conditions where these differences could be ignored to examine what occurs in the case of incorrect modeling of PI or correct modeling of PI. Such conditions can occur with software with automatic testing routines with certain models (e.g., Mplus Analysis = Configural, Metric, Scalar with cross-loadings). Finally, in order to assess the performance of the MGCFA in extreme cases of noninvariance, 68% of the factor loadings were allowed to differ between groups.

Modeling of factor loading noninvariance The modeling of factor loading noninvariance was either correctly specified or incorrectly specified. Correct modeling meant that when the loadings were invariant, they were modeled to be so, whereas when they were not invariant they were correctly modeled to be noninvariant. Finally, incorrect modeling meant that when the factor loadings were simulated to be noninvariant, they were incorrectly modeled as being invariant. Again, this could be a result of Type II errors, automatic software routines, or direction of invariance testing.

Model parameter group differences For each model parameter tested for invariance, several conditions were used for group differences. The

intercept and error variances were simulated to be different across groups on the same indicator variables for which the factor loadings were simulated to be different. All differences were unidirectional (e.g., favoring one group). The intercept values and factor means differed by 0, .2, .5 or .8, representing no difference to a large difference (Cohen, 1988). These were standardized value differences. That is, intercept and latent mean differences are absolute differences whereby one group had a value of 0 for the intercept or factor mean, and the other group had a value of 0.0, 0.2, 0.5, or 0.8. The covariances between factors had standardized values (i.e., correlations) of either 0.0, 0.1, 0.3, or 0.5, representing differences. The factor variance for one group was set to 1 across conditions. The variance of the second group was then varied from 1 (factor variance invariance), 1.33, 1.66, and 2.00, in order to reflect different levels of factor variance noninvariance. 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 respective factor loading, as we set the loading values. This minimized potential confounding factors in examining the results. The differences simulated reflect a range of values that represent typical small to large differences. These were not tied to any content area as Cohen (1988) suggests but were broad strokes to capture situations that could be applied to many areas of work. In other words, the goal was to study a range of potential group parameter differences from none through moderate and large.

Sample size   The total sample was simulated to be 300, 1000, and 2000 with equal group sizes. These values are designed to represent cases from the smallest samples generally seen in practice (French & Mantzicopoulos, 2007), to what would be considered a large sample in most social science applications and common conditions in simulation work (French & Finch, 2006; 2008b; Meade & Lautenschlager, 2004; Steinmetz, 2013) while maintaining adequate statistical power (Hancock & French, 2013). The smallest total sample size used here was 300, meaning that each group contained 150 individuals. Samples smaller than that were not used as it could lead to unstable parameter estimates, particularly for the more complex models (Kline, 2011), and confound the results.

Number of factors and indicators per factor   The number of indicators per factor simulated was 3 or 6 which were completely crossed with the number of factors of either 2 or 4. This range of values is designed to reflect both extensively measured and less extensively measured constructs, and was in accord with prior research in this area (e.g. Millsap & Kwok, 2004; Steinmetz, 2013) to

facilitate generalizability. In addition, the models used in this simulation study are similar to models published in actual practice (e.g. Bavarian et al., 2014; Hesse & Klingberg, 2014; Tam, 2014).

Percent of invariance in factor intercepts, error variances, factor variances, factor covariances, and factor means Three levels of the amount of non-invariance across groups were simulated. To assess Type I error (i.e., false identification of a lack of invariance) of the MGCFA methods employed, the case of complete invariance (i.e. no differences in model parameters across groups) was simulated. In addition, to assess power (i.e., correct identification of a lack of invariance) 34% of target model parameters lacked invariance.

Response variables Several outcome variables were examined including Type I error rates and power for the chi-square difference test χ^2_Δ at $\alpha = 0.05$, parameter estimation bias (sample parameter estimate – population parameter value), standard deviation of the parameter estimates, mean square error (MSE) for parameter estimates, and parameter coverage rates for 95% confidence intervals. To determine which of the manipulated variables or their interactions significantly impacted the Type I error and power rates, analysis of variance (ANOVA) was used (Paxton, Curran, Bollen, Kirby, & Chen, 2001), in addition to the effect size measure (η^2) to assist with identifying effects worth noting. The outcome variable for the ANOVA model was the total number of the 1000 replications for each combination of conditions in which the null hypothesis of invariance was rejected. With regard to criteria for acceptable performance, Type I error rates were considered acceptable if they were at the nominal 0.05 level, just as coverage rates for parameter estimates were acceptable when the actual coverage was at the nominal 0.95 value. For power, we considered the typical value of 0.80 to be the criterion.

Results

Model Parameter Invariance: Type I Error and Power when Factor Loadings Differed by 0.5 or 1.0

Following are the results for Type I error and power rates for assessing the invariance assumption for the various model parameters. As noted above, simulations were conducted for factor loading differences between groups of 0.25, 0.50, and 1.00. Results showed that the Type I error rates for assessing invariance

of other model parameters when factor loadings differed between the groups by 0.50 and 1.00 were highly inflated when the lack of loading invariance was not properly modeled. When group loading differences were not properly modeled and the loadings differed by 0.50 or 1.00, the Type I error rates for incorrectly identifying noninvariance for factor intercepts, error variances, factor variances, factor covariances, and factor means were at or above 0.80 in the 34% noninvariant case, and at or above 0.95 in the 68% noninvariant case. These inflated rates were present regardless of the other manipulated conditions, including sample size, number of indicators, and number of factors. On the other hand, when the lack of factor loading invariance was correctly modeled, the Type I error rates for assessing invariance of the other model parameters were between 0.045 and 0.058 for all model parameters, across sample size, number of factors, and number of indicators. Given the uniformly inflated Type I error rates for the incorrectly modeled factor loading noninvariance condition, power for this case was not investigated for any of the model parameters, as these rates cannot be interpreted with any confidence.

When the factor loadings were correctly modeled, power rates in the 0.50 and 1.00 factor loading difference cases were very similar to power rates in the 0.25 factor loading difference condition. Therefore, in order to save space, we report only the power values for the 0.25 factor loading difference condition in the following section of the paper. In addition, given that the Type I error rates were extremely inflated when the factor loadings were simulated to differ between groups by 0.50 and 1.00, and that they were essentially identical to those obtained when the loadings differed by 0.25 and this lack of invariance was correctly modeled, it was felt that reporting results for the two larger noninvariant conditions would be redundant. Therefore, the results that appear below reflect only the cases where the loadings were truly invariant, or where they differed between the groups by 0.25.

Intercept Invariance: Type I Error and Power

The ANOVA for the Type I error rate reveal no significant interactions or main effects ($p = 0.05$). The Type I error rate, parameter bias for the indicators on which intercepts were simulated to differ, parameter standard deviation (SD), mean squared error (MSE), and coverage rates appear in Table 1. Across conditions, the Type I error rate for χ^2_Δ was at the nominal 0.05 rate. Parameter estimates were somewhat negatively biased, and coverage rates were close to the nominal 0.95 both when the intercepts were constrained to be equal and when they were not. In addition, the SD and MSE of the parameter estimates were very comparable under

both unconstrained and constrained conditions. Notably given the goals of this study, the Type I error rate was not influenced by factor loading PI, nor by whether that PI was correctly or incorrectly modeled.

The ANOVA identified sample size (N) as being significantly related to power rates ($F_{2,88} = 8.4$, $p < 0.001$, $\eta^2 = 0.16$), as well as degree of intercept difference ($F_{2,88} = 41.8$, $p < 0.001$, $\eta^2 = 0.59$). Power rates increased from 0.69 for a total sample size of 300, to nearly 1.00 for samples of 1000 and 2000. Table 1 includes the power, bias, SD, MSE, and coverage rates for the unconstrained and constrained modeling conditions, by the degree of intercept difference.

Table 1. Intercept invariance testing Type I error rate and power, parameter bias, parameter estimate standard deviation, MSE, and coverage rates for intercept across simulated conditions: Unconstrained parameters/Constrained parameters

Difference	Type I error	Bias	SD	MSE	Coverage
0.0	0.051	-0.01 / -0.01	0.06 / 0.05	0.005 / 0.004	0.93 / 0.93
	Power				
0.2	0.780	0.03 / -0.08	0.06 / 0.05	0.010 / 0.020	0.93 / 0.92
0.5	0.990	0.05 / -0.15	0.06 / 0.05	0.030 / 0.050	0.92 / 0.66
0.8	1.000	0.05 / -0.25	0.06 / 0.06	0.030 / 0.110	0.93 / 0.54

Power for detecting intercept differences increased concomitantly with increases in the population difference between the groups' intercept values, which would be expected. In addition, for the group with the larger intercept when the intercept estimates were simulated to differ, the parameter estimate displayed greater bias than when no group differences were simulated (Table 1). For the unconstrained condition, there was a positive bias in the intercept estimate for the group with the larger intercept, while for the constrained condition there was negative bias that increased with greater group differences in the population intercept value. This result was expected for the constrained condition because the groups' intercepts were forced to be equal, thus driving down the value for the group with the larger intercept. In addition, whereas the SD of the estimates was comparable for both conditions across the size of intercept difference, the MSE increased and coverage decreased with greater such differences in the constrained condition but remained largely unchanged in the unconstrained case. Additionally, as was true with the Type I error, the loading PI condition, along with how it was modeled, had no impact on the assessment of intercept differences.

Table 2. Error variance invariance testing Type I error rate, parameter bias, parameter estimate standard deviation, MSE, and coverage rates for error variance by partial loading invariance and modeling conditions, and by number of factors and indicators per factor: Unconstrained parameters/Constrained parameters

Loading invariance	Type I error	Bias	SD	MSE	Coverage
Full	0.05	0.002 / 0.002	0.01 / 0.01	0.001 / 0.001	0.94 / 0.94
Partial correct	0.05	-0.002 / 0.002	0.01 / 0.01	0.001 / 0.001	0.96 / 0.94
Partial incorrect	0.44	-0.010 / -0.020	0.01 / 0.01	0.010 / 0.010	0.62 / 0.57

Correct modeling of partial loading invariance or Full loading invariance					
Factors / indicators per factor	Type I error	Bias	SD	MSE	Coverage
2 / 3	0.05	0.002 / 0.002	0.01 / 0.01	0.0020 / 0.0010	0.94 / 0.95
2 / 6	0.05	0.002 / 0.002	0.01 / 0.01	0.0010 / 0.0010	0.95 / 0.95
4 / 3	0.05	-0.010 / 0.002	0.01 / 0.01	0.0050 / 0.0010	0.78 / 0.94
4 / 6	0.05	0.002 / 0.002	0.01 / 0.01	0.0003 / 0.0003	0.94 / 0.94

Incorrect modeling of partial loading invariance					
Factors / indicators per factor	Type I error	Bias	SD	MSE	Coverage
2 / 3	0.34	0.005 / -0.001	0.01 / 0.01	0.003 / 0.002	0.93 / 0.94
2 / 6	0.18	0.000 / -0.002	0.01 / 0.01	0.001 / 0.001	0.95 / 0.92
4 / 3	0.98	-0.040 / -0.060	0.01 / 0.01	0.020 / 0.040	0.08 / 0.02
4 / 6	0.27	-0.010 / -0.010	0.01 / 0.01	0.002 / 0.002	0.53 / 0.41

Error Variances: Type I Error and Power

The ANOVA identified the interaction between factor loading difference and modeling of that difference ($F_{1,88} = 16.8$, $p < 0.001$, $\eta^2 = 0.51$), and the interaction of the number of factors by number of indicators per factor ($F_{1,88} = 6.2$, $p = 0.024$, $\eta^2 = 0.28$) as significantly related to the Type I error rate for detecting differences in group error variances. Table 2 contains Type I error rates, parameter bias, SD, MSE, and coverage rates by loading PI and modeling conditions. These results show that when the loadings are fully invariant, or PI with the invariance being correctly modeled, the Type I error rates are at the nominal 0.05 level. However, when the loadings are PI but modeled as fully invariant, the Type I error rate for testing error variance was inflated to 0.44. A further examination of the results in Table 2 reveals that under the fully invariant or partial correct conditions, parameter bias, SD, MSE are all relatively low, and the coverage rates are at the nominal level for both the unconstrained and constrained models. However, parameter bias was 5 times larger for the unconstrained model and 10 times larger for the constrained model in the partial incorrect condition, while the coverage rates for both modeling conditions was well below the nominal 0.95 level.

In addition, these results in Table 2 are divided into those in which full factor loading held, or loading PI was correctly modeled, and those in which loading PI was not correctly modeled. The Type I error rate was found to be at the nominal (0.05) level when loadings were fully invariant or PI was correctly modeled. However, when loadings were PI but not correctly modeled the error rates were inflated, with greater inflation occurring for 3 indicators per factor. In addition, the greatest inflation occurred when there were 4 factors each with 3 indicators. A condition for which parameter estimate bias was also greatest, and coverage was lowest.

Error Variances: Power

The ANOVA indicated that none of the manipulated factors were significantly related to the power to detect error noninvariance. Across conditions, power for detecting error noninvariance was extremely high (0.99). There was a much larger negative bias for the constrained parameter model than in the unconstrained case (-0.060 vs -0.003). In addition, the MSE was more than 10 times larger for the constrained model, and displayed coverage rates of just 0.10, well below the nominal 0.95 level. For the unconstrained model, the parameter coverage rate was also below the nominal level, at 0.83, but much higher than for the constrained model. Of particular interest in this study, power rates were not significantly influenced by PI of the factor loadings, unlike in the Type I error case. This result would appear to be in large part due to the extremely high power for detecting error noninvariance across conditions.

Factor Covariance: Type I Error and Power

With respect to the Type I error rate when testing the invariance hypothesis of factor covariances, the ANOVA results showed that the interaction of factor loading difference and the modeling of that difference ($F_{2,35} = 8.8$, $p = 0.001$, $\eta^2 = 0.35$) was statistically significant. Table 3 includes the outcome variables of interest by the factor loading difference and modeling of the difference, and by the number of factors. The Type I error rate showed some inflation when there was loading PI that was not properly modeled. Accompanying this Type I error inflation was greater negative bias of parameter estimates, particularly for the constrained parameter model, which in turn was associated with inflation of the MSE, again particularly in the case of incorrect modeling of the factor loading PI. Finally, coverage rates were lower in the PI incorrect condition for both the constrained and unconstrained model, with coverage higher for the unconstrained model.

524

The ANOVA for power in testing covariance parameter invariance identified sample size as the only statistically significant variable ($F_{2,35} = 5.7$, $p = 0.004$, $\eta^2 = 0.12$). Power rates, bias, SD, MSE, and coverage rates appear in Table 4. Power increased from 0.61 for $N = 300$ to 0.88 for $N = 2000$. Parameter bias, MSE, and coverage rates were largely unaffected by sample size, though SD for both the constrained and unconstrained models was lower for the two larger sample sizes than for $N = 300$. In addition, bias was lower for the unconstrained model as compared to the constrained model, and MSE and coverage rates were higher. Power for detecting noninvariant factor covariances was not found to be influenced by loading PI or how it was modeled.

ANOVA did not identify any significant effects for the Type I error rate in the detection of factor mean differences between groups. In this case, bias, SD, MSE, and coverage rates appear only for the unconstrained model because in the constrained case, both groups' means were set equal to 0. The Type I error rate for testing group mean invariance was at the nominal 0.05 rate, with low bias for the mean that was allowed to vary, and a coverage rate at the nominal 0.95 level. Of particular interest was the fact that the loading PI condition and the way that this was modeled, were not significantly related to the Type I error rate when testing for factor mean differences between groups.

Table 3. Factor covariance invariance testing Type I error rate, parameter bias, parameter estimate standard deviation, MSE, and coverage rates for testing covariance invariance by partial loading invariance and modeling conditions: Unconstrained parameters/Constrained parameters

Loading invariance	Type I error	Bias	SD	MSE	Coverage
Full	0.05	-0.0001 / -0.0900	0.07 / 0.05	0.12 / 0.19	0.96 / 0.82
Partial correct	0.05	0.0002 / -0.1000	0.07 / 0.05	0.19 / 0.25	0.93 / 0.80
Partial incorrect	0.07	-0.0003 / -0.2100	0.07 / 0.07	0.19 / 0.37	0.91 / 0.71

Table 4. Factor covariance invariance testing power, parameter bias, parameter estimate standard deviation, MSE, and coverage rates for intercept by sample size (N): Unconstrained parameters/Constrained parameters

N	Power	Bias	SD	MSE	Coverage
150 / 150	0.61	-0.002 / -0.130	0.10 / 0.09	0.20 / 0.24	0.94 / 0.83
500 / 500	0.76	-0.004 / -0.140	0.06 / 0.05	0.19 / 0.24	0.94 / 0.80
1000 / 1000	0.88	0.007 / -0.140	0.04 / 0.04	0.19 / 0.24	0.95 / 0.82

Factor means: Type I error and Power

Table 5. Factor mean invariance testing power, parameter bias, parameter estimate standard deviation, MSE, and coverage rates for intercept by sample size (*N*): Unconstrained parameters

N	Power	Bias	SD	MSE	Coverage
150 / 150	0.78	-0.015	0.14	0.35	0.93
500 / 500	0.95	-0.013	0.07	0.33	0.94
1000 / 1000	0.99	-0.013	0.05	0.30	0.93
Difference	**Power**	**Bias**	**SD**	**MSE**	**Coverage**
0.2	0.73	-0.006	0.09	0.05	0.96
0.5	0.99	-0.014	0.09	0.27	0.94
0.8	1.00	-0.038	0.09	0.71	0.90

The ANOVA results showed that the sample size ($F_{2,91} = 22.3$, $p < 0.001$, $\eta^2 = 0.33$) and the degree of factor mean difference ($F_{2,91} = 46.4$, $p < 0.001$, $\eta^2 = 0.50$) were significantly related to power rates. Table 5 includes power, bias, SD, MSE, and coverage rates by sample size, and group factor mean difference, for the unconstrained model only. Power increased concomitantly with sample size, as would be expected. Furthermore, power was above 0.75 in the worst case, and at 0.95 or above for samples of 1000 or more. Parameter bias appears to not have been influenced by sample size, though the SD and MSE declined somewhat with increasing sample sizes. Coverage rates were near the nominal 0.95 level across sample sizes.

With respect to the difference between group factor means, power rates exceeded 0.7 even for the smallest difference of 0.2. There was an increase in the amount of negative bias as the group mean difference increased, indicating that, for the group whose mean was larger, the unconstrained model provided a slight underestimate. It should be noted, however, that the largest bias value was -0.038, indicating that the factor mean estimate was approximately 0.76 when in the population the value was 0.80. This increase in bias with a greater group mean difference was also associated with an increase of MSE and a decrease in the coverage rate to 0.90, below the nominal 0.95 rate. The power for testing group factor mean differences was not significantly influenced by loading PI or whether it was modeled correctly.

Factor Variances: Type I Error and Power

ANOVA results showed that the interaction between loading PI and its modeling was the only term significantly related to the Type I error rate for testing factor variance invariance ($F_{2,36} = 22.7$, $p < 0.001$, $\eta^2 = 0.59$). Table 6 includes the Type

I error rates, bias, SD, MSE, and coverage rates for the unconstrained and constrained models when testing factor variance invariance. The Type I error rate for testing group variance invariance was at the nominal rate when full factor loading invariance held, or when there was loading PI and it was modeled correctly. However, when the loadings were PI but modeled as fully invariant, the Type I error rate was inflated to 0.30. In addition, bias in the factor variances for both the unconstrained and constrained models was much lower in the fully invariant or partial correct conditions, than in the partial incorrect. This increase in bias was associated with inflated MSE and lower coverage rates for both the constrained and unconstrained models, though SD was not impacted by the loading invariance condition.

The ANOVA results showed that the 3-way interaction of N by loading PI by modeling of PI ($F_{2,86} = 53.4$, $p < 0.001$, $\eta^2 = 0.55$), and the main effect of factor variance difference ($F_{4,86} = 4.7$, $p = 0.012$, $\eta^2 = 0.10$) were the only statistically significantly related terms to the power for detecting noninvariant group variances. Power, bias, SD, MSE, and coverage rates by N, loading invariance, and modeling conditions appear in Table 7. Power increased with sample size and was slightly higher (approximately 0.04) in the full versus partial invariance conditions for the two smaller sample sizes. Power for the PI condition modeled incorrectly should not be interpreted given the observed inflated Type I error rate. With regard to the estimated factor variances, larger positive bias was observed in the unconstrained group in the incorrect PI condition than either when full invariance held, or in the correctly modeled PI condition. This positive bias indicates that the variance for the group simulated to have the larger value was overestimated when the factor loadings were noninvariant but constrained to be equal across groups.

On the other hand, for the constrained model (where variances for the two groups were held equal), the variance estimate for the group with the larger population value was negatively biased. The MSE for the constrained group was

Table 6. Factor variance invariance testing Type I error rate, parameter bias, parameter estimate standard deviation, MSE, and coverage rates for testing covariance invariance by partial loading invariance and modeling conditions: Unconstrained parameters/Constrained parameters

Loading invariance	Type I error	Bias	SD	MSE	Coverage
Full	0.05	-0.002 / 0.004	0.14 / 0.14	0.15 / 0.16	0.94 / 0.92
Partial correct	0.05	-0.001 / 0.002	0.15 / 0.14	0.15 / 0.16	0.94 / 0.93
Partial incorrect	0.30	0.080 / -0.069	0.14 / 0.13	0.21 / 0.20	0.77 / 0.72

Table 7. Factor variance invariance testing power, parameter bias, parameter estimate standard deviation, MSE, and coverage rates for testing covariance invariance by sample size (*N*), partial loading invariance and modeling conditions: Unconstrained parameters/Constrained parameters

N	Loading Invariance	Power	Bias	SD	MSE	Coverage
150 / 150	Full	0.63	0.001 / -0.308	0.21 / 0.21	0.25 / 0.46	0.95 / 0.53
	Partial correct	0.59	-0.033 / -0.337	0.22 / 0.21	0.26 / 0.46	0.93 / 0.50
	Partial incorrect	**0.83***	0.103 / -0.359	0.20 / 0.21	0.36 / 0.47	0.86 / 0.55
500 / 500	Full	0.91	0.003 / -0.328	0.11 / 0.12	0.25 / 0.47	0.94 / 0.51
	Partial correct	0.87	-0.002 / -0.340	0.11 / 0.12	0.24 / 0.46	0.94 / 0.52
	Partial incorrect	**0.98***	0.097 / -0.361	0.12 / 0.11	0.36 / 0.47	0.87 / 0.54
1000 / 1000	Full	0.97	0.003 / -0.325	0.08 / 0.08	0.24 / 0.46	0.93 / 0.55
	Partial correct	0.96	0.003 / -0.338	0.08 / 0.08	0.24 / 0.47	0.95 / 0.54
	Partial incorrect	**0.99***	0.105 / -0.362	0.07 / 0.08	0.35 / 0.48	0.86 / 0.54

* Power rates in bold are associated with conditions in which the Type I error rate was inflated

Table 8. Factor variance invariance testing power, parameter bias, parameter estimate standard deviation, MSE, and coverage rates for intercept by intercept difference: Unconstrained parameters/Constrained parameters

Difference	Power	Bias	SD	MSE	Coverage
0.33	0.70	0.03 / -0.17	0.14 / 0.13	0.28 / 0.39	0.92 / 0.49
0.66	0.91	0.03 / -0.34	0.13 / 0.14	0.28 / 0.45	0.92 / 0.53
1.00	0.98	0.04 / -0.51	0.13 / 0.14	0.29 / 0.53	0.91 / 0.59

greater in the incorrect PI condition than for either full loading invariance or correct modeling of loading PI. However, for the constrained model, there was no notable difference in MSE across these conditions. Parameter coverage was also well below the nominal 0.95 rate for the constrained model, while in the unconstrained case coverage was at or near the nominal rate except when factor loadings were PI but not modeled as such.

Table 8 includes the power, bias, SD, MSE, and coverage rates by level of variance difference. Power for detecting unequal group variances increased as the degree of that difference increased, which would be expected. For the unconstrained model, the amount of bias, the SD, MSE, and coverage rates were essentially the same across differences in group variances. In contrast, for the constrained model, the amount of bias increased concomitantly with differences in the magnitude of the group factor variances. Again, this result is expected when one considers that for the constrained model, one variance is simulated to be 1.00,

while the other is simulated to be 1.33, 1.66, or 2.00 in the simplest two factor case. Thus, when the variances are constrained to be equal, the magnitude of bias should increase along with increases in the difference between group variances.

Discussion

The use of MGCFA for testing FI and latent mean differences will continue to grow as modeling of outcomes over time and across diverse groups takes greater advantage of the rapid methodological changes in latent variable modeling. This increased use leads to a need for focusing on appropriate practices when assessing invariance, especially in the presence of PI, be it for the traditional MGCFA or other models with grouping variables and measurement models (e.g., latent profile analysis). Accurate invariance testing rests upon the assumption that MGCFA works well across many data conditions. Yet, heretofore, empirical research in this area has focused primarily on the performance of tests for factor loading invariance. In particular, there is a pressing need to provide researchers with useful information on how PI affects observed composite scores on assessments. Perhaps more importantly, researchers and practitioners need to be provided assistance in understanding how decisions made about groups and individuals using such scales are impacted by PI. Our current work attempts to address these issues in the FI and MGCFA research domains by examining the accuracy for assessing invariance at all levels of the CFA model across groups under various levels of factor loading invariance. The results allowed us to draw a few main conclusions.

First, when factor loading PI is correctly modeled, invariance testing on the other model parameters was not adversely influenced regardless of how large the group differences in factor loadings were. However, and second, if modeling of such differences is done incorrectly and the degree of group loading difference is 0.50 or 1.00, then invariance testing for other model parameters will suffer from Type I error inflation of 0.8 or higher. When the degree of loading difference is 0.25, and this lack of invariance is not modeled correctly, there will also likely be Type I error inflation for testing the invariance of error variances and factor variances and, to a much smaller extent, factor covariances. Thus, careful attention must be paid to the correct modeling of partial factor loading invariance when researchers are interested in assessing invariance of the latent model variance and covariance structures. Third, PI of the factor loadings with group differences of 0.25 had no impact on testing the invariance of intercepts or factor means, whether the factor loading differences were correctly modeled or not. This result was surprising given the distortion of Type I error in other measurement invariance

studies where one level of non-invariance influenced another (French & Finch, 2008b) and the distortion of mean differences under a lack of factorial invariance(Millsap, 2011; Steinmetz, 2013). However, considered in light of the Type I error inflation for mean and intercept differences associated with factor loading differences of 0.50 and 1.00, it would appear that the impact of PI is simply not felt until group differences on the loadings reaches a critical juncture. Fourth, model complexity, defined as the number of factors and indicators, only influenced invariance testing for error variances. Additionally, this was only a concern when the PI of factor loadings was model incorrectly. Under such conditions, more factors were associated with greater Type I error rates and bias, and lower coverage.

Taken together, these results demonstrate the importance of assessing and correctly modeling the invariance of factor loadings prior to testing for invariance in other model parameters. However, if this is done and if PI in loadings is modeled correctly, the researcher can be confident that it will not impact assessment of other model parameters, even when one group has a majority of loadings that are twice the size of the other groups' loadings. This is likely the best outcome that can be achieved, and yet one that is the most challenging to achieve as it is difficult to be certain that all such group differences in loadings have been correctly modeled. Thus, the applied researcher must be keenly aware of both the steps that they take in testing parameter invariance and how they account for loading PI. This requires an awareness of the correct sequence of steps used in invariance testing, and a knowledge of what software programs with automated functions are doing to account for different levels of invariance as the program systematically tests for full invariance. This latter issue becomes ever more challenging as new versions of software are released on a yearly basis with increasingly automated functions for conducting invariance testing (e.g., Mplus, IRTPRO). This is not to say such automation is entirely negative. Rather, it is a call for clearer documentation of the steps that are being taken and the underlying assumptions and model constraints that are imposed and, perhaps most importantly, users taking the responsibility to understand what they are modeling.

The purpose of this investigation was to provide evidence of MGCFA performance for FI testing under a variety of practical and applied conditions, specifically focused on models where loading PI is present. However, although we estimated many models under various conditions, simulation of exhaustive conditions is not practically possible. Therefore, additional simulation work is encouraged to continue examining MGCFA analyses under an even greater array of conditions (e.g., percent of misspecification, mixed invariant conditions) as there are several problems which remain to be solved in invariance testing (Millsap,

2011). We note that analytic solutions should be sought first or in conjunction with simulation work to aid the understanding of the underlying models and reasons for differences in outcomes (Boomsma, 2013). That said, hopefully this research will inform practice for those engaged in FI analyses to better understand phenomena in their disciplines. The results described here should allow practitioners to make informed decisions in the presence of loading PI. Furthermore, they should inform practice by highlighting the strengths and limitations of MGCFA given certain conditions. We also think these results can stimulate new research surrounding the implementation of the MGCFA and other latent group models. For example, the development of an effect size to capture the magnitude of the difference in parameters and the influence it has on latent mean difference testing would be helpful in allowing power results to be more meaningfully examined (Millsap, 2011). This line of work should lead to accurate decisions about individuals and groups in the presence of partial invariance through MGCFA analysis.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baumgartner, H., & Steenkamp, J. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143-156. doi: 10.1509/jmkr.38.2.143.18840

Bavarian, N., Flay, B.R., Ketcham, P.L., Smit, E., Kodama, C., Martin, M., & Saltz, R.F. (2014). Using structural equation modeling to understand prescription stimulant misuse: A test of the theory of triadic influence. *Drug and Alcohol Dependence, 138*(1), 193-201. doi: 10.1016/j.drugalcdep.2014.02.700

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

Boomsma, A. (2013). Reporting Mote Carlo studies in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(3), 518-540. doi: 10.1080/10705511.2013.797839

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures. *Psychological Bulletin, 105*(3), 456-466. doi: 10.1037/0033-2909.105.3.456

Church, A. T., Alvarez, J. M., Mai, N. T. Q., French, B. F., Katigbak, M. S., & Ortiz, F. A. (2011). Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the Revised NEO Personality Inventory. *Journal of Personality and Social Psychology, 101*(5), 1068-1089. doi: 10.1037/a0025290

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hoboken, NJ: Lawrence Erlbaum Associates.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70*(4), 662-680. doi: 10.1037/0021-9010.70.4.662

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(3), 378-402. doi: 10.1207/s15328007sem1303_3

Finch, W. H., & French, B. F. (2008a). Anomalous Type I error rates for identifying one type of differential item functioning in the presence of another. *Educational and Psychological Measurement, 68*(5), 742-759. doi: 10.1177/0013164407313370

French, B. F., & Finch, W. H. (2008b, March). *When under the influence of non-invariant factor loadings, does computation method of the factor score matter?* Paper presented at the annual meeting of the American Educational Research Association. New York, NY.

French, B. F., & Gotch, C. M. (2013). Sex differences in item functioning in the comprehensive inventory of basic skills-II vocabulary assessments. *Journal of Psychoeducational Assessment, 31*(4), 410-417. doi: 10.1177/0734282912460857

French, B. F., & Mantzicopoulos, P. (2007) An examination of the first/second-grade form of the pictorial scale of perceived competence and social acceptance: Factor structure and stability by grade and gender across groups of econimically disadvantaged children. *Journal of School Psychology, 45*(3), 311-331. doi: 10.1016/j.jsp.2006.07.004

Hancock, G. R. & French, B. F. (2013). Power analysis in structural equation modeling. In G. R. Hancock & R. O. Muller (Eds.), *Structural equation*

modeling: A second course (2nd ed.) (pp. 117-159). Charlotte, NC: Information Age Publishing.

Hesse, K., & Klingberg, S. (2014). Examining the cognitive model of caregiving – A structural equation modeling approach. *Psychiatry Research, 217*(3), 171-176. doi: 10.1016/j.psychres.2014.02.023

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3), 117-144. doi: 10.1080/03610739208253916

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. doi: 10.1080/10705519909540118

Javaras, K. N., & Ripley, B. D. (2007). An "unfolding" latent variable model for Likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association, 102*(478), 454-463. doi: 10.1198/016214506000000960

Maller, S. J., & French, B. F. (2004). Universal Nonverbal Intelligence Test factor invariance across deaf and standardization samples. *Educational and Psychological Measurement, 64*(4), 647-660. doi: 10.1177/0013164404263877

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(1), 60-72. doi: 10.1207/S15328007SEM1101_5

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*(2), 289-311. doi: 10.1007/BF02294510

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Millsap, R. E., & Kwok, O. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Measurement, 9*(1), 93-115. doi: 10.1037/1082-989X.9.1.93

Millsap, R. E., & Meredith, W. (2007). Factorial invariance: historical perspectives and new problems. In R. Cudeck & R. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 131-152). Mahwah, NJ: Erlbaum.

Muthén, L. K., & Muthén, B. O. (2013). *Mplus user's guide, version 7.1*. Los Angeles, CA: Muthén and Muthén.

Poortinga, Y. P. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology, 24*(6), 737-756. doi: 10.1080/00207598908247842

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling: A Multidisciplinary Journal, 8*(2), 287-312. doi: 10.1207/S15328007SEM0802_7

Schriesheim, C. A., & Neider, L. L. (2001). *Equivalence in measurement*. Greenwich, CN: Information Age.

Steenkamp, J. E. M. & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78-90. doi: 10.1086/209528

Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology, 9*, 1-12. doi: 10.1027/1614-2241/a000049

Tam, F. W. (2014). A structural equation model of drug abuse among secondary students in Hong Kong. *International Journal of Mental Health and Addiction, 12*(3), 255-269. doi: 10.1007/s11469-013-9450-5

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*(1), 4-70. doi: 10.1177/109442810031002

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practices, 29*(3), 39-47. doi: 10.1111/j.1745-3992.2010.00182.x

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

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with the TIMSS data. *Practical Assessment, Research,*

& Evaluation, 12(3), 1-26. Retrieved from
http://pareonline.net/getvn.asp?v=12&n=3

      Yoon, M. & Millsap, R. E. (2007). Detecting violations of factorial
invariance using data-based specification searches: A Monte Carlo study.
Structural Equation Modeling, 14(3), 435-463. doi: 10.1080/10705510701301677

Appendix

Sample Mplus Code

Model lack of invariance

```
title:      Model lack of invariance
data:       file is replist.dat;
            type=montecarlo;
variable:   names are y1-y6 group;
            grouping is group (1=g1 2=g2);

model:
            f1 by y1;
            f1 by y2;
            f1 by y3;
            f2 by y4-y6;
            f1@1 f2@1;
            f1 with f2;
            y1-y6;

model g1:   f1 by y2* ;
            [y1*0] (1);
            [y2*0] (2);
            [y3*] ;
            [y4*0] (4);
            [y5*0] (5);
            [y6*0] (6);

model g2:
            f1 by y2* ;
            [y1*0] (1);
            [y2*0] (2);
            [y3*] ;
            [y4*0] (4);
            [y5*0] (5);
            [y6*0] (6);
savedata:   results are diffresults.out;
```

Model total invariance

```
title:      Model total invariance
data:       file is replist.dat;
            type=montecarlo;
variable:   names are y1-y6 group;
            grouping is group (1=g1 2=g2);
model:
```

```
            f1 by y1;
            f1 by y2;
            f1 by y3;
            f2 by y4-y6;
            f1@1 f2@1;
            f1 with f2;
            y1-y6;

model g1:   f1 by y2* ;
            [y1*0] (1);
            [y2*0] (2);
            [y3*0] (3);
            [y4*0] (4);
            [y5*0] (5);
            [y6*0] (6);

model g2:
            f1 by y2* ;
            [y1*0] (1);
            [y2*0] (2);
            [y3*0] (3);
            [y4*0] (4);
            [y5*0] (5);
            [y6*0] (6);

savedata:   results are nodiffresults.out;
```

Test run

```
title:      test run
montecarlo: names are y1-y6;
            nobservations=1000 1000;
            nreps=1000;
            seed=94756;
            ngroups=2;
            repsave=all;
            save=rep*.dat;

model population:
            [y1-y6@0];
            y1-y6@1;
            f1 by y1@1 y2-y3*.6;
            f2 by y4@1 y5-y6*.6;
            f1@1 f2@1;
            f1 with f2@.5;
model population-g2:
f1 by y2*.85;
[y3*0.2];
```

```
model:
          f1 by y1;
          f1 by y2;
          f1 by y3;
          f2 by y4;
          f2 by y5;
          f2 by y6;
          f1@1 f2@1 (1);
          f1 with f2 (2);
          y1-y6 (3);

model g2:    [y3*0.2];
f1 by y2*.85;
output:      tech9;
```