Journal of Modern Applied Statistical Methods

Volume 11 | Issue 1 Article 3

5-1-2012

The Impact of Violating Factor Scaling Method Assumptions On Latent Mean Difference Testing in Structured Means Models

Dandan Wang
The University of Texas at Austin

Tiffany A. Whittaker
The University of Texas at Austin

S. Natasha Beretvas
The University of Texas at Austin

Recommended Citation
Wang, Dandan; Whittaker, Tiffany A.; and Beretvas, S. Natasha (2012) "The Impact of Violating Factor Scaling Method Assumptions On Latent Mean Difference Testing in Structured Means Models," *Journal of Modern Applied Statistical Methods*: Vol. 11 : Iss. 1 , Article 3.
DOI: 10.22237/jmasm/1335844920

Journal of Modern Applied Statistical Methods
May 2012, Vol. 11, No. 1, 24-41

1538 – 9472/12/$95.00

The Impact of Violating Factor Scaling Method Assumptions On Latent Mean Difference Testing in Structured Means Models

Dandan Wang Tiffany A. Whittaker S. Natasha Beretvas
The University of Texas at Austin,
Austin, TX

Type I error rates and power of the likelihood ratio test and bias of the standardized effect size measure associated with the latent mean difference in structured means modeling are examined when violating the assumptions underlying the two available factor scaling methods under various conditions. Implications and recommendations are discussed.

Key words: Structured means modeling, latent mean difference testing, non-invariant factor loadings, factor scaling methods, reference indicator.

Introduction

Many social science studies focus on comparing outcomes for groups categorized by observed variables such as gender, race or treatment group membership. Structural equation modeling (SEM) and, more specifically, structured means modeling (SMM; Sörbom, 1974) may be used to compare, for example, male and female high school students' latent variable means on math anxiety. The SMM approach is a multiple-group confirmatory factor analysis (MG-CFA) model in which the mean structure is incorporated into the model for testing the difference in latent variable means across groups.

As with traditional CFA techniques, each latent variable must be assigned a scale of measurement in SMM. This may be accomplished by either constraining one loading per factor to a value of one across groups or constraining each factor's variance to a value of one across groups (Kline, 2011). Both factor scaling methods require meeting certain assumptions. For example, the reference indicator (RI) strategy involves an assumption that the RI has invariant factor loadings across groups. The factor-variance scaling method, by contrast, is based on an assumption that the factor variances are equal across groups. To date, no published study has examined the effect on latent mean comparisons of constraining unequal factor loadings or unequal factor variances to a value of one across groups. The focus of this study is to investigate the impact of violating the assumptions underlying two factor scaling methods on the latent mean difference test and the standardized effect size measure associated with the latent mean difference in SMM.

Structured Means Model and Testing Latent Mean Differences

A single-factor, p-indicator structured means model can be expressed in matrix notation using the following measurement equation:

$$\boldsymbol{x}_g = \boldsymbol{v}_g + \Lambda_g \boldsymbol{\xi}_g + \boldsymbol{\delta}_g, \qquad (1)$$

Dandan Wang received her Ph.D. in Quantitative Methods in the Department of Educational Psychology at the University of Texas at Austin. She currently works at International Baccalaureate. Email her at: dwvv4@utexas.edu. Tiffany Whittaker is an Assistant Professor of Quantitative Methods in the Department of Educational Psychology. Email her at: t.whittaker@austin.utexas.edu. S. Natasha Beretvas is a Professor of Quantitative Methods in the Department of Educational Psychology. Her interests are in multilevel and meta-analytic modeling techniques. Email her at: tasha.beretvas@mail.utexas.edu.

where g (g = 1, 2, …, G) represents group membership, $\boldsymbol{x}$ is a vector containing p × 1 observed variable scores, $\boldsymbol{\nu}$ is a p × 1 vector containing indicator variable intercepts, $\boldsymbol{\Lambda}$ is a p × 1 vector of factor loadings that relates the observed indicator variables to the latent variable, $\boldsymbol{\xi}$ is a latent variable score and $\boldsymbol{\delta}$ is a p × 1 vector of normally distributed random measurement errors associated with the observed indicator variables.

In a structured means model (SMM), certain constraints are imposed in order to validly compare latent means across groups as well as to ensure model identification. The factor loadings and observed indicator variables' intercepts are constrained to be equal across groups in SMM. This allows latent mean differences detected between groups to be attributed to actual differences in the constructs as opposed to differences in the measurement of the constructs across the groups (Rock, Werts & Flaugher, 1978; Sörbom, 1974). In addition, it is not possible to estimate the intercepts of all observed indicator variables separately across groups as this would result in the under-identification of the means portion of the model (Sörbom, 1974).

Under the assumption of factor loading and intercept invariance, and assuming that the mean of the measurement errors within each group is equal to zero, the expected values of observed variables in each group can be expressed in matrix notation as:

$$\boldsymbol{E}\left[\boldsymbol{x}_g\right] = \boldsymbol{\mu}_g = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\kappa}_g, \qquad (2)$$

where $\boldsymbol{\kappa}_g$ is the latent variable mean for group g, $\boldsymbol{\nu}$ is a p × 1 vector containing the observed variables' intercepts which are invariant across groups, and $\boldsymbol{\Lambda}$ represents a p × 1 vector containing invariant factor loadings (Yoon & Millsap, 2007). In addition, assuming that the measurement errors are uncorrelated and that the measurement errors are uncorrelated with the latent variable in each group, g, the covariances among observed variables in each group can be expressed in matrix notation as:

$$\begin{aligned}\boldsymbol{E}\left[\left(\boldsymbol{x}_g - \boldsymbol{\mu}_g\right)\left(\boldsymbol{x}_g - \boldsymbol{\mu}_g\right)'\right] &= \boldsymbol{\Sigma}_g \\ &= \boldsymbol{\Lambda}\boldsymbol{\Phi}_g\boldsymbol{\Lambda}' + \boldsymbol{\Theta}_g\end{aligned} \qquad (3)$$

where $\boldsymbol{\Phi}$ is the latent variable variance, $\boldsymbol{\Theta}$ is a $p \times p$ diagonal matrix containing p measurement error variances associated with the observed indicator variables and $\boldsymbol{\Lambda}$ represents a p × 1 vector containing invariant factor loadings (Sörbom, 1974; Yoon & Millsap, 2007).

If factor loading and/or intercept invariance is not supported by means of model fit assessment and/or model comparisons, Byrne, Shavelson and Muthén (1989) suggested that some of the constraints may be relaxed and that partial invariance suffices when using the SMM approach to compare latent means across groups. It is important to note that there exists a debate concerning the issue of partial measurement invariance and the meaningful interpretation of latent mean differences in SMM. Some contend that strict invariance is required for valid interpretations of latent mean differences (Meredith, 1993), whereas others maintain that strong invariance is sufficient (Hancock, 1997; Thompson & Green, 2006). (In addition, some researchers have argued – and demonstrated – that meaningful interpretations of latent mean differences may be rendered under partial factor loading and partial intercept invariance; this debate is beyond the scope of this work but the interested reader is referred to: Byrne, Shavelson & Muthén, 1989; Muthén & Christoffersson, 1981; Steenkamp & Baumgartner, 1998.)

The means portion of the model is estimated through use of a constant which is modeled to have direct effects on the latent and observed indicator variables. The constant's direct effect on a latent variable represents the latent variable mean and its effects on observed indicator variables represent observed variables' intercepts (Kline, 2011). An added constraint must also be imposed for identification of the means portion of the model. Namely, the latent mean in one group, treated as the reference group, must be constrained to zero whereas the latent means in the remaining G – 1 comparison groups are estimated. Therefore, the test of the latent mean of the G – 1 comparison groups corresponds to a test of the latent mean difference between each of the *G* – 1 comparison group's and the reference group's latent mean (Hancock, 1997). For simplicity, a two-group comparison of the latent means in a single-factor model will be assumed for the ensuing discussion of latent mean comparisons in SMM.

Because the expected value of the latent variable is given by:

$$\boldsymbol{E}(\boldsymbol{\xi_g}) = \boldsymbol{\kappa_g}, \tag{4}$$

the latent variable mean in the reference group for a two-group comparison of a single-factor model is:

$$E\left(\xi\right) = \kappa_1 = 0, \tag{5}$$

and the latent variable mean in the comparison group (group two) is:

$$E\left(\xi\right) = \kappa_2 - \kappa_1 = \kappa_2. \tag{6}$$

Accordingly, the null hypothesis tested in SMM is that the two groups' means are equal:

$$H_0 : \kappa_1 = \kappa_2. \tag{7}$$

The *z* test statistic is commonly used to evaluate the statistical significance of the latent mean difference estimate in SMM. If the *z* test statistic associated with the estimated latent mean difference is statistically significant, then it is inferred that there is a significant difference between the two groups' latent variable means. Unfortunately, the *z* test statistic is not invariant to the choice of factor scaling method. Consequently, the likelihood ratio test, LRT_κ, has been suggested to evaluate the statistical significance of the latent mean difference estimate in SMM because it is invariant to factor scaling procedures (Gonzalez & Griffin, 2001; Hancock, Lawrence & Nevitt, 2000).

When calculating the LRT_κ, two models must be estimated: The parameter of interest (the latent variable mean difference) is freely estimated in group two in one model but is constrained to be equal to the reference group's latent variable mean value of zero in the second model. The LRT_κ is calculated as the difference between the two nested models' Chi-square $\left(\chi^2\right)$ statistics, and the LRT_κ is asymptotically distributed as a non-central χ^2 statistic:

$$\mathrm{LRT}_\kappa = \Delta\chi^2 = \chi^2{}_{restricted} - \chi^2{}_{baseline\ model}, \tag{8}$$

where $\chi^2{}_{restricted}$ is the χ^2 statistic associated with the model in which the latent variable mean difference is constrained to a value of zero and $\chi^2{}_{baseline\ model}$ is the χ^2 statistic associated with the model in which the latent mean is freely estimated in group two. The LRT_κ has corresponding degrees of freedom equal to the difference in the degrees of freedom associated with each model and is calculated to evaluate whether there is a statistically significant drop in model fit when constraining a particular parameter (the latent variable mean difference) to zero. A significant LRT_κ indicates that the parameter of interest differs significantly from zero.

Although the LRT_κ may be used to evaluate whether there is a statistically significant difference between two groups' latent means, it does not provide any information about the practical significance of the latent mean difference. Hancock (2001) suggested using a standardized effect size measure, $\hat{\delta}_\kappa$, to describe the practical difference between two groups' latent means. When using the SMM approach, the standardized latent mean difference effect size, $\hat{\delta}_\kappa$, is estimated as follows:

$$\begin{aligned} \hat{\delta}_\kappa &= \left(\hat{k}_1 - \hat{k}_2\right) / \hat{\phi}^{1/2} = \hat{k}_2 / \hat{\phi}^{1/2}, \\ \hat{\delta}_k &= \left|\hat{k}_1 - \hat{k}_2\right| / \hat{\phi}^{1/2} = \hat{k}_2 / \hat{\phi}^{1/2} \end{aligned} \tag{9}$$

where $\hat{k}_1$ and $\hat{k}_2$ represent groups one and two latent mean estimates, respectively. For model identification purposes, the latent mean of the reference group (group one) is typically constrained to a value of zero, but the latent mean of the comparison group (group two) is estimated (Hancock, 1997), resulting in the rightmost expression of Equation 9. The pooled factor variance estimate, $\hat{\phi}$, is determined as follows:

$$\hat{\phi} = (n_1\hat{\phi}_1 + n_2\hat{\phi}_2) / (n_1 + n_2) \qquad (10)$$

where $\hat{\phi}_1$ and $\hat{\phi}_2$ are the estimated factor variances for groups one and two, respectively, and n_1 and n_2 represent the sample sizes for groups one and two, respectively. It is important to note that the calculation and use of the pooled factor variance involves an assumption of homogeneity of the two groups' factor variances. The interpretation of $\hat{\delta}_\kappa$ is similar to that in conventional univariate analyses. For example, $\hat{\delta}_\kappa = 0.5$ can be interpreted as indicating that the two groups' latent mean estimates differ by half a standard deviation (Hancock, 2001).

Factor Scaling Method Implications and Related Research

Both factor scaling methods may be used to scale the latent variable in a structured means model and both involve strict assumptions. For example, the factor-variance-based scaling method is grounded on the assumption that the factor variance is invariant across groups. If the factor variances are not equal across groups, the scale of the factor loadings will be changed, possibly making truly invariant factor loadings falsely appear non-invariant across groups. This could also make the metric invariance test less accurate (Cheung & Rensvold, 1999; Kline, 2011; Yoon & Millsap, 2007).

The reference indicator (RI) strategy is based on the assumption that the RI's loading is invariant across groups. If this assumption is violated, all other factor loadings in a structured means model will be rescaled. Constraining the RI's non-invariant factor loadings to a value of one across groups will result in different metrics for the two groups' factor loadings and can lead to incorrect inferences about the changed loadings' invariance; therefore, it is important to select an item that has invariant factor loadings across groups to serve as the RI in a structured means model (Johnson, Meade & DuVernet, 2009).

Although assumptions associated with the two factor scaling methods are important, researchers have not examined the issue to a great extent. For example, Johnson, Meade and DuVernet (2009) conducted a literature review of studies published between 2005 and 2007 that involved measurement invariance (MI) tests; only 17 out of 153 studies referenced Cheung and Rensvold's (1999) study in which a new technique to select invariant item sets to serve as the RI was recommended. Most research simply assumed that the selected RI variable had invariant factor loadings across groups; consequently, it is essential that the impact of violating the assumptions associated with the two factor scaling methods be inspected in order to better inform applied users of SMM.

Previous studies germane to this study include those in which the effect of partial metric invariance on latent mean difference testing was assessed. For example, Kaplan and George (1995) conducted a population study to assess the power to detect latent mean differences between two groups in the SMM approach while manipulating the magnitude of the latent mean difference, group sample size ratio, frequency of non-invariant factor loadings, factor loading size, factor loading pattern and the number of observed indicators per factor. Because factor loadings were varied in the study, the determinant of the covariance matrix [det(**Σ**)] was also varied. The determinant corresponds to the generalized variance, which indicates the amount of variance shared among a set of variables. Two combinations of ratios of larger sample size (n_{Larger}) to generalized variance [det(**Σ**)] conditions were examined, including a positive, n_{Larger}:det(**Σ**), condition in which the group with the larger sample size was paired with the larger generalized variance, and a negative, n_{Larger}:det($\mathbf{\Sigma}$), condition in which the group with the larger sample size was paired with the smaller generalized variance.

The findings demonstrated that when the magnitude of the latent mean difference increased, the power of the latent mean difference test increased and the sample size ratio between the two groups tended to influence the power of the latent mean difference test. When the group sample sizes were equal, the power of the latent mean difference test was less affected by non-invariant factor loadings. However, when unequal sample sizes were present, the power associated with the latent mean difference test was low even though factor loading invariance held. A large drop in power was observed as the group sample size ratio increased which was observed in both positive and negative conditions. Nonetheless, higher power always occurred in the positive n_{Larger}:det($\mathbf{\Sigma}$) condition as compared to the negative n_{Larger}:det($\mathbf{\Sigma}$) condition. Finally, the power of the latent mean difference test increased when the model consisted of more indicator variables per factor.

Hancock, Lawrence and Nevitt (2000) conducted both a Monte Carlo simulation study and a population study investigating how partial metric invariance affected the Type I error rates and the power, respectively, of the latent mean difference test between two groups using SMM, multiple-indicator multiple-cause (MIMIC) modeling, and MANOVA approaches. Manipulated conditions included latent mean difference magnitude, total sample size, group sample size ratio, frequency of non-invariant factor loadings, factor loading size and factor loading pattern. Factor loading pattern manipulations resulted in four scenarios: (1) metric invariance with equal factor loadings across and within two groups; (2) metric invariance with equal factor loadings across two groups but unequal within groups; (3) metric non-invariance with approximately equivalent generalized variances for the two groups; and (4) metric non-invariance with different generalized variances for the two groups.

Type I error rates of the latent mean difference tests in all three approaches were well controlled under metric invariance, approximately equivalent group generalized variance, and equal group sample size conditions. When both the sample size and the generalized variance were unequal between the two groups, however, Type I error rates of the latent mean difference test in the three approaches varied. The SMM approach was the only one in which Type I error rates were well controlled under all manipulated conditions. The Type I error rates when using the MIMIC approach were too low under the negative n_{Larger}:det($\mathbf{\Sigma}$) condition (larger sample size paired with smaller generalized variance) and were too high under the positive n_{Larger}:det($\mathbf{\Sigma}$) condition (larger sample size paired with larger generalized variance). The opposite pattern of Type I error rates were observed when using the MANOVA approach.

The power of the latent mean difference test in the three approaches increased when the sample size, magnitude of the factor loadings, and magnitude of the latent mean difference increased. When the sample size ratio between the two groups became larger, the power of the latent mean difference test in the three approaches decreased. Overall, the power of the latent mean difference test when using the MIMIC technique tended to be approximately equal to, or marginally higher than, the power when using the SMM technique, but the power associated with the MANOVA approach was the lowest. When different generalized variances were paired with unequal sample sizes, results indicated that the SMM approach had greater power in the negative n_{Larger}:det($\mathbf{\Sigma}$) condition whereas the MIMIC approach had greater power in the positive n_{Larger}:det($\mathbf{\Sigma}$) condition.

Hancock et al. (2000) reported that both SMM and MIMIC approaches were acceptable under equal group sample sizes. The SMM approach, however, was recommended under unequal group sample sizes. The choice of the SMM approach was based on its flexibility in accommodating non-invariant factor loadings. Additionally, the SMM approach had satisfactory power without sacrificing the Type I error rate. In contrast, the MIMIC approach's slightly higher power was marred by the potential cost of Type I error inflation (Hancock, et al., 2000).

Previous studies investigating the effects of partial measurement invariance on latent mean difference detection under various conditions have found that group sample size ratio, factor loading pattern, loading difference magnitude, and latent mean difference magnitude can affect both or either the Type I error rate and power of latent mean difference tests. However, previous simulation studies have not devoted much attention to the assumption underlying the RI strategy and – to the authors' knowledge – no published study has investigated the effect of violating the assumption underlying the factor-variance scaling method. The purpose of this Monte Carlo simulation study is to investigate the performance of the likelihood ratio test (LRT_{κ}) and the standardized latent mean difference effect size measure $\left(\hat{\delta}_{\kappa}\right)$ when violating the assumptions fundamental to the two factor scaling methods and using the SMM approach to test latent mean differences.

Methodology

The impact of violating the assumptions associated with the two factor scaling methods on the performance of the LRT_{κ} and $\hat{\delta}_{\kappa}$ were examined under varied conditions, including group sample size ratio, factor loading pattern, loading difference magnitude, latent mean difference magnitude and group factor variance ratio. For each generated sample of data, two factor scaling methods (constraining one loading per factor to a value of one for both groups and assigning a value of one to each factor's variance for both groups) were implemented. The performance of the LRT_{κ} was evaluated via an assessment of its Type I error rates and power under specified conditions. The performance of the $\hat{\delta}_{\kappa}$ in terms of the parameter bias and relative parameter bias under certain conditions was also evaluated.

For simplicity, two groups' latent variable means were compared using the SMM approach. The model used for data generation and estimation was a simple, single-factor model with six observed indicator variables. The choice of the six observed indicator variables was based on designs of previous simulation studies (e.g., Kaplan & George, 1995) and reflects what has been found in applied research (Hinkin, 1995). The values of all invariant factor loadings were set to 0.4 to represent factor loadings commonly observed in applied studies (Enders & Finney, 2003) and because a relatively large loading difference value across groups was included and resulted in markedly large non-invariant factor loadings.

All observed variable intercepts were set to zero across groups in the generating models. Residual variances associated with the observed variables were calculated as one minus the squared condition-specific standardized factor loading. Error covariances were not modeled in the generating or estimating models. Total sample size was 500 and was not varied. This sample size was used because it is in the range of sample sizes utilized in previous simulation research in which adequate power was obtained (e.g., Hancock, et al., 2000) and permits the examination of reasonably disparate group sample sizes.

Manipulated Conditions: Group Sample Size Ratio

Three group sample size ratio conditions (n_1: n_2) were used when generating the data. The equal sample size condition (1:1) served as a baseline condition in which the sample size in each group was equal to 250. Two unequal sample size ratio conditions (1:4 and 4:1) were also used to generate the data: data in the 1:4 condition were generated such that the sample size was 100 and 400 in group one and in group two, respectively, and data in the 4:1 condition were generated in which the two groups' sample sizes were reversed.

Manipulated Conditions: Factor Loading Pattern

Five factor loading patterns were manipulated in this study. In the equal factor loading pattern condition, all factor loadings were generated to be invariant across groups to serve as a baseline condition. In the 1st loading unequal pattern condition, the RI's (here, the first observed indicator variable's) factor loading was set to be higher in group two than in group one by the condition-specific factor loading difference. In the 2nd loading unequal pattern condition, the factor loading of a non-RI variable (here, the second observed indicator variable) was set to be higher in group two than in group one by the condition-specific factor loading difference. In the all lower pattern condition and the mixed pattern condition, both the RI and the second observed indicator variable had non-invariant factor loadings across groups in the generating models. In the all lower pattern condition, both of the non-invariant factor loadings had lower true values in group one. In the mixed pattern condition, the true factor loading value for the RI was higher in group one and the true factor loading value for the second observed indicator variable was higher in group two.

Manipulated Conditions: Loading Difference Magnitude

Two factor loading difference values ($|\lambda_1 - \lambda_2| = 0.1$ and $|\lambda_1 - \lambda_2| = 0.4$) were investigated in the current simulation study to represent small and large differences. These two values are in the range of factor loading difference values investigated in previous simulation research (Hancock, et al., 2000; Kaplan & George, 1995). These factor loading differences were added to the invariant factor loading value of 0.4, resulting in factor loading non-invariance across groups (with loading values of 0.5 and 0.8, respectively).

Manipulated Conditions: Latent Mean Difference Magnitude

This study considered two latent mean difference values ($\kappa_2 - \kappa_1$ = 0 and $\kappa_2 - \kappa_1$ = 0.5). The condition of equal latent means ($\kappa_2 - \kappa_1$ = 0) across groups was included because this permits an assessment of the Type I error rates associated with the LRT_κ and the performance of the $\hat{\delta}_\kappa$ in terms of parameter bias. Scenarios with unequal latent means across groups were also investigated in order to assess the power of the LRT_κ and the performance of the $\hat{\delta}_\kappa$ in terms of relative parameter bias. A moderately large latent mean difference value ($\kappa_2 - \kappa_1$ = 0.5) was included because previous simulation studies found sufficient power with this latent mean difference value (Hancock, et al., 2000; Kaplan & George, 1995).

Manipulated Conditions: Group Factor Variance Ratio

In the simulation study, three factor variance ratio conditions $(\Phi_1 : \Phi_2)$ were considered. In the first factor variance ratio condition, the factor variances for the two groups were set to be equal (1:1). In the second and third factor variance ratio conditions, the factor variances for the two groups were set to be unequal with a ratio of 0.8:1.2 and 1.2:0.8. These two unequal factor variance conditions represent a realistic yet moderate difference (Kim, Cramond & Bandalos, 2006) between the two groups' factor variances which provides a starting point for this line of research.

Data Generation

Raw data for the two groups were generated in SAS (Version 9.2; SAS Institute Inc., 2008) according to the specified population parameters for a single-factor, six-indicator CFA model using the Kaiser and Dickman (1962) matrix decomposition procedure (Fan & Fan, 2005). Thus, each generated sample of data consisted of $n_1 \times 6$ and $n_2 \times 6$ matrices for group one and group two, respectively, where n_1 and n_2 represent the condition-specific sample size for each of the two groups. One thousand (1,000) raw data sets were generated for each of the 162 combinations of manipulated conditions. After raw data for the two groups were generated, SAS 9.2 was programmed to call DOS to run M*plus* (Version 6.1; Muthén & Muthén, 2010), as described by Gagné and Furlow (2009), to estimate the models. Maximum likelihood (ML) estimation was used to estimate all model parameters.

When estimating the model parameters, cross-group constraints were imposed on all factor loadings and observed variable intercepts whereas error variances were freely estimated in both groups. When using the RI strategy to scale the factor, the RI's loading was constrained to a value of one in both groups. Two different structured means models were estimated. The traditional structured means model was estimated in which the latent mean of group one was constrained to be equal to zero but the latent mean of group two was estimated freely (the $\text{SMM}_{\kappa*}$ model) and another model in which the latent means for both groups were constrained to zero (the SMM$_{\kappa 0}$ model) was estimated. Also, two factor scaling methods were used to set the scale of the latent variable for each generated data set. When using the RI strategy, the first factor loading was constrained to a value of one across groups, all other factor loadings were constrained to be equal across groups, and factor variances for the two groups were freely estimated. When the factor-variance-based scaling method was implemented, the factor variance was instead constrained to a value of one across groups and all factor loadings were estimated yet constrained to be equal across groups. Thus, for each generated data set, four models (two factor scaling methods × two latent mean constraints) were estimated. It is important to note that the models using the factor-variance scaling method had one degree of freedom more than the models using the RI strategy.

Data Analysis

The χ^2 statistic associated with each estimated model from each replication was saved to calculate the LRT_{κ} (see Equation 8) between the two estimated models (SMM$_{\kappa *}$ and SMM$_{\kappa 0}$) when using each of the two factor scaling methods. The performance of the LRT_{κ} was evaluated by summarizing its Type I error rates and power. Type I error rates of the LRT_{κ}, defined as the proportion of incorrect rejections of the null hypothesis $\left(H_0 : \kappa_1 = \kappa_2\right)$ out of the 1,000 replications in equal latent mean conditions $\left(\kappa_2 - \kappa_1 = 0\right)$, were evaluated using Bradley's (1978) liberal criterion of α ± 1/2α (where α = 0.05) such that rates less than 2.5% were considered overly conservative and rates greater than 7.5% were considered overly liberal.

The power of the LRT_{κ} is defined as the proportion of correct rejections of the null hypothesis $\left(H_0 : \kappa_1 = \kappa_2\right)$ out of the 1,000 replications in unequal latent mean conditions $\left(\kappa_2 - \kappa_1 = 0.5\right)$. A minimum power criterion of 0.8 is traditionally recommended as a reasonable level of power (Cohen, 1988), whereas others have recommended a minimum power criterion of 0.95 as a more appropriate level of power (Cashen & Geiger 2004; Rossi, 1990). In this study, a minimum power criterion of 0.9 was selected to gauge the adequacy of the power associated with the LRT_{κ} as a compromise between the traditional and more stringent power recommendations.

The latent mean estimate in group two and the factor variance estimates in both groups were saved from the SMM$_{\kappa *}$ model, which were used to estimate the standardized latent mean difference effect size, $\hat{\delta}_{\kappa}$ (see Equations 9 and 10). The performance of the $\hat{\delta}_{\kappa}$ was examined through an assessment of its parameter bias and relative parameter bias under specific latent mean difference magnitude conditions. In conditions in which the latent mean difference was equal to zero $\left(\kappa_2 - \kappa_1 = 0\right)$, the parameter bias of $\hat{\delta}_{\kappa}$ was calculated as follows:

$$B\left(\hat{\delta}_{\kappa}\right) = \overline{\hat{\delta}}_{\kappa} - 0, \qquad (11)$$

where $\overline{\hat{\delta}}_{\kappa}$ is the mean of the δ_{κ} estimates across the 1,000 replications in each relevant condition (Hoogland & Boomsma, 1998). The relative parameter bias of the $\hat{\delta}_{\kappa}$ was calculated with conditions in which the latent mean difference was equal to 0.5 $\left(\kappa_2 - \kappa_1 = 0.5\right)$ as:

$$RPB\left(\hat{\delta}_{\kappa}\right) = \frac{\overline{\hat{\delta}}_{\kappa} - 0.5}{0.5} \qquad (12)$$

(Hoogland & Boomsma, 1998). According to Hoogland and Boomsma (1998), conditions in which the $\left|B\left(\hat{\delta}_{\kappa}\right)\right|$ or the $\left|RPB\left(\hat{\delta}_{\kappa}\right)\right|$ is less than 0.05 indicates acceptable levels of bias in the $\hat{\delta}_{\kappa}$.

Results

The results describing the performance of the LRT_{κ} are presented first, including Type I error

Table 1: Explanations of Abbreviations Used in the Tables of Results

Abbreviation	Explanation
RI	Reference indicator strategy implemented
FV	Factor-variance-based scaling method implemented
Equal Loading	All factor loadings were equal/invariant across groups
1st Loading Unequal	The first factor loading (RI) was higher in group two than in group one with the condition-specific loading difference
2nd Loading Unequal	The second (non-RI) factor loading was higher in group two than in group one with the condition-specific loading difference
All Lower	Both the first (RI) and second (non-RI) factor loading were higher in group two than in group one with the condition-specific loading difference
Mixed	The first factor loading (RI) was higher in group one than in group two and the second (non-RI) factor loading was higher in group two than in group one with the condition-specific loading difference

rates and power. The results describing the performance of the $\hat{\delta}_{\kappa}$, including parameter and relative parameter bias, are subsequently presented. Table 1 provides the explanations of abbreviations used in all the Tables illustrating the performance of the LRT_{κ} and the $\hat{\delta}_{\kappa}$.

Performance of the LRT_{κ}: Type I Error Rates

Table 2 presents the observed Type I error rates associated with the LRT_{κ} under equal latent mean conditions $(\kappa_2 - \kappa_1 = 0)$. Values above the dashed line in Table 2 are the Type I error rates in the equal/invariant factor loading conditions and, thus, for scenarios in which the covariance structures are appropriately modeled. Values below the dashed line in Table 2 are the Type I error rates in the unequal/non-invariant factor loading conditions and, thus, for scenarios in which the covariance structures are not modeled appropriately. In each design cell, Type I error rates when implementing the RI strategy and when implementing the factor-variance (FV) scaling method are both reported. Employing Bradley's (1978) criterion, Table 2 shows overly conservative Type I error rates (i.e., less than or equal to 2.5%) denoted with boldface and italics; overly liberal rates (i.e., greater than or equal to 7.5%) are underlined.

In the equal/invariant factor loading conditions, all observed Type I error rates when using the RI strategy were within the criterion of 0.05 ± 0.025. Type I error rates did not appear to vary substantially or systematically as a function of group sample size ratio or group factor variance ratio. Upon implementing the factor-variance scaling method, one Type I error rate was found to be overly liberal (0.079) in the condition with the group sample size ratio of 1:4 and the group factor variance ratio of 1.2:0.8.

In the unequal/non-invariant factor loading conditions, the Type I error rates when the RI strategy was implemented were within the acceptable range of 0.05 ± 0.025. When the factor-variance scaling method was used, however, twelve observed Type I error rates were beyond the criterion of 0.05 ± 0.025. These unacceptable Type I error rates all occurred in the unequal sample size conditions such that liberal rates tended to occur in the 4:1 group sample size ratio scenarios and conservative rates tended to occur in the 1:4 group sample size ratio scenarios. Further, the majority (83%) of these unacceptable Type I error rates were found in the large loading difference ($|\lambda_1 - \lambda_2| = 0.4$) magnitude conditions (see Table 2).

Power of the LRT_κ

Table 3 presents the observed power rates associated with the LRT_κ under conditions in which the latent mean difference was unequal across groups $(\kappa_2 - \kappa_1 = 0.5)$. A criterion of 0.90 was used to evaluate the power of the LRT_κ; hence, power rates below 0.90 were deemed too low (see Table 3). In the equal factor loading conditions, three power rates fell below the 0.90 cutoff when the RI strategy was implemented. These occurred in the 1:1 group factor variance ratio with the 4:1 group sample size ratio condition, in the 1.2:0.8 group factor variance ratio with the 1:4 group sample size ratio condition, and in the 0.8:1.2 group factor variance ratio with the 4:1 group sample size ratio condition. Although these values were lower than the cutoff criterion, they were not substantially lower than a value of 0.90, ranging from 0.866 to 0.888. Power tended to be higher in the equal group sample size conditions, but it did not vary substantially or systematically as a function of the group factor variance ratio under the RI strategy.

When the factor-variance scaling method was implemented under equal factor loadings, five out of nine power rates were lower than 0.90. Nonetheless, these values did not deviate substantially from 0.90 (range was from 0.891 to 0.894) and all were found in the unequal sample size conditions. Power rates were higher in the equal sample size conditions than in the unequal sample size conditions. Additionally, power rates when using the factor-variance scaling method did not differ markedly as a function of the group factor variance ratio (see Table 3).

In the unequal factor loading conditions, five power rates were lower than 0.90 when using the RI strategy which all occurred in conditions in which the loading difference was small ($|\lambda_1 - \lambda_2| = 0.1$) and the group sample sizes were unequal (1:4 or 4:1). Again, these power rates were not substantially lower than the cutoff criterion, ranging from 0.890 to 0.898 (see Table 3). Power tended to be marginally higher when the loading difference was large ($|\lambda_1 - \lambda_2| = 0.4$) than when small ($|\lambda_1 - \lambda_2| = 0.1$). Across the three group sample size ratios, power rates were slightly higher when sample sizes were equal across groups than when they were unequal. Further, power rates under the RI strategy did not vary substantially as a function of the group factor variance ratios or the factor loading patterns.

When the factor-variance scaling method was implemented under unequal factor loadings, two observed power rates were lower than the cutoff criterion, although they did not differ substantially from the 0.90 criterion (0.890 and 0.892). These two low power rates were found in conditions in which the loading difference was small ($|\lambda_1 - \lambda_2| = 0.1$) with the 0.8:1.2 group factor variance ratio and 1:4 group sample size ratio (see Table 3). Power when using the factor-variance scaling method was consistent with the power found when using the RI strategy. Specifically, power rates were marginally higher when the loading difference was large than when it was small and when sample sizes were equal across groups than when unequal. In addition, power rates did not vary considerably as a function of group factor variance ratio or the factor loading pattern when implementing the factor-variance scaling method.

Performance of the $\hat{\delta}_\kappa$: Parameter Bias of the $\hat{\delta}_\kappa$

Parameter bias of the standardized latent mean difference effect size measure ($\hat{\delta}_\kappa$) was calculated in conditions in which the true latent mean difference was equal to zero

Table 2: Type I Error Rates Associated with the Likelihood Ratio Test as a Function of Manipulated Conditions $\left(\kappa_2 - \kappa_1 = 0\right)$

Loading Difference	Loading Pattern	Group Sample Size Ratio	Group Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
			RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	0.056	0.057	0.049	0.051	0.058	0.058
		100:400	0.062	0.052	0.060	0.079	0.059	0.044
		400:100	0.068	0.067	0.051	0.037	0.055	0.070
0.1	1st Loading Unequal	250:250	0.047	0.045	0.050	0.049	0.057	0.058
		100:400	0.047	0.043	0.046	0.052	0.068	0.046
		400:100	0.043	0.044	0.062	0.054	0.056	0.068
	2nd Loading Unequal	250:250	0.046	0.046	0.046	0.046	0.050	0.051
		100:400	0.052	0.051	0.047	0.058	0.059	0.047
		400:100	0.038	0.045	0.052	0.040	0.064	0.083
	All Lower	250:250	0.068	0.067	0.057	0.060	0.071	0.071
		100:400	0.056	0.049	0.054	0.064	0.045	0.029
		400:100	0.058	0.070	0.049	0.038	0.051	0.068
	Mixed	250:250	0.048	0.048	0.054	0.055	0.047	0.046
		100:400	0.050	0.049	0.049	0.059	0.052	0.037
		400:100	0.053	0.054	0.056	0.040	0.059	0.081
0.4	1st Loading Unequal	250:250	0.044	0.043	0.058	0.059	0.044	0.042
		100:400	0.048	0.030	0.053	0.044	0.060	***0.019***
		400:100	0.058	0.070	0.054	0.054	0.061	0.096
	2nd Loading Unequal	250:250	0.053	0.050	0.054	0.054	0.055	0.053
		100:400	0.055	0.027	0.053	0.050	0.049	***0.025***
		400:100	0.050	0.064	0.045	0.043	0.051	0.085
	All Lower	250:250	0.055	0.050	0.048	0.044	0.050	0.040
		100:400	0.044	***0.016***	0.066	0.046	0.061	***0.016***
		400:100	0.045	0.082	0.047	0.059	0.051	0.113
	Mixed	250:250	0.055	0.056	0.052	0.052	0.042	0.041
		100:400	0.050	0.041	0.045	0.058	0.040	***0.021***
		400:100	0.038	0.029	0.045	***0.024***	0.046	0.052

Note: Type I error rates less than 0.025 are bold and italicized. Type I error rates greater than 0.075 are underlined. Abbreviations used in this table are described in Table 1.

Table 3: Power Associated with the Likelihood Ratio Test as a Function of Manipulated Conditions $\left(\kappa_2 - \kappa_1 = 0.5\right)$

Loading Difference	Loading Pattern	Group Sample Size Ratio	Group Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
			RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	0.979	0.979	0.983	0.983	0.982	0.982
		100:400	0.906	0.906	0.866	0.892	0.911	0.893
		400:100	0.888	0.891	0.920	0.894	0.873	0.893
0.1	1st Loading Unequal	250:250	0.987	0.988	0.983	0.982	0.980	0.981
		100:400	0.924	0.926	0.927	0.937	0.941	0.921
		400:100	0.898	0.905	0.929	0.916	0.897	0.927
	2nd Loading Unequal	250:250	0.984	0.984	0.988	0.988	0.982	0.982
		100:400	0.925	0.917	0.890	0.911	0.920	0.890
		400:100	0.904	0.912	0.947	0.936	0.912	0.937
	All Lower	250:250	0.991	0.991	0.993	0.993	0.994	0.994
		100:400	0.935	0.928	0.939	0.942	0.953	0.926
		400:100	0.919	0.931	0.938	0.927	0.916	0.929
	Mixed	250:250	0.983	0.984	0.985	0.986	0.986	0.986
		100:400	0.906	0.901	0.890	0.903	0.908	0.892
		400:100	0.911	0.908	0.942	0.924	0.892	0.910
0.4	1st Loading Unequal	250:250	0.998	0.998	1.000	1.000	0.999	0.999
		100:400	0.971	0.967	0.975	0.978	0.989	0.971
		400:100	0.964	0.973	0.973	0.979	0.940	0.962
	2nd Loading Unequal	250:250	0.997	0.997	0.999	0.999	0.998	0.998
		100:400	0.984	0.967	0.976	0.979	0.984	0.966
		400:100	0.965	0.974	0.974	0.974	0.944	0.962
	All Lower	250:250	1.000	1.000	0.999	0.999	1.000	1.000
		100:400	0.993	0.980	0.993	0.989	0.995	0.981
		400:100	0.982	0.995	0.993	0.995	0.971	0.990
	Mixed	250:250	0.995	0.996	0.995	0.997	0.998	0.998
		100:400	0.961	0.957	0.953	0.964	0.972	0.959
		400:100	0.925	0.920	0.931	0.906	0.924	0.932

Note: Power rates below 0.90 are underlined. Abbreviations used in this table are described in Table 1.

$\left(\kappa_2 - \kappa_1 = 0\right)$. A cutoff value of 0.05 was used to evaluate the acceptability of the parameter bias, meaning that absolute parameter bias values less than 0.05 indicated acceptable bias (Hoogland & Boomsma, 1998). No substantial parameter bias was found across conditions examined with bias values ranging from –0.010 to 0.013. Although both negative and positive parameter bias was observed, no clear trend was noticed.

Performance of the $\hat{\delta}_{\kappa}$: Relative Parameter Bias of the $\hat{\delta}_{\kappa}$

In conditions where the true latent mean difference was equal to 0.5 $\left(\kappa_2 - \kappa_1 = 0.5\right)$, relative parameter bias of the $\hat{\delta}_{\kappa}$ was calculated. Table 4 presents the relative parameter bias of the $\hat{\delta}_{\kappa}$ for each combination of manipulated conditions. Following Hoogland & Boomsma (1998), a minimum cutoff of 0.05 was used to represent a substantial degree of bias (see Table 4).

In the equal factor loading conditions, relative parameter bias was acceptable, regardless of the factor scaling method used. Relative parameter bias when using the RI strategy and the factor-variance scaling method showed consistent trends. Negative relative parameter bias values occurred in conditions in which the group factor variance ratio was 1:1 or 0.8:1.2 and positive relative parameter bias values emerged in conditions in which the group factor variance ratio was 1.2:0.8 (see Table 4). Although the relative parameter bias values were in opposite directions, their absolute values did not differ substantially as a function of group factor variance ratio or group sample size ratio.

Unacceptable relative parameter bias was found when implementing the RI strategy in 33 conditions under the unequal factor loading scenarios. Unacceptable and substantial relative parameter bias was found more often in conditions in which the loading difference was large ($|\lambda_1 - \lambda_2| = 0.4$) than when it was small ($|\lambda_1 - \lambda_2| = 0.1$). Relative parameter bias varied as a function of factor loading pattern as well. For example, relative parameter bias was more substantial in the all lower factor loading pattern conditions than in the remaining three factor loading scenarios (i.e., 1st loading unequal, 2nd loading unequal and mixed pattern conditions) whereas relative parameter bias was the least substantial in the mixed pattern scenarios (see Table 4). Relative parameter bias values were the smallest in the 1:4 group sample size ratio scenarios whereas they were more substantial in the 4:1 group sample size ratio scenarios. In addition, no clear trend was exhibited across the three group factor variance ratios when using the RI scaling method. When the factor-variance scaling strategy was implemented, relative parameter bias trends closely resembled those found when using the RI scaling strategy as previously described (see Table 4).

Conclusion

The primary question addressed in this study was whether violating the assumptions underlying the RI strategy and/or the factor-variance scaling method (i.e., using a RI with non-invariant factor loadings or constraining unequal factor variances to a value of one across groups) would affect the testing and description of the latent mean difference across groups. When implementing the RI strategy, the Type I error rates associated with the LRT_{κ} were not adversely affected by factor loading difference magnitude, factor loading pattern, group sample size ratio, or group factor variance ratio. This result indicates that violating the assumption of equivalent reference indicator loadings underlying the RI strategy did not affect Type I error rates associated with the LRT_{κ} for conditions and models examined here. This is consistent with the findings from the study conducted by Hancock, et al. (2000).

Previous research on SMM has not thoroughly investigated the factor-variance scaling method. The study found that when implementing the factor-variance scaling method, group factor variance ratio, group sample size ratio and loading difference magnitude did affect the Type I error rates associated with the LRT_{κ}. More specifically, when factor loadings were non-invariant/unequal, all Type I error rates that

Table 4: Relative Parameter Bias of the Standardized Latent Mean Difference Effect Size Measure as a Function of Manipulated Conditions $(\kappa_2 - \kappa_1 = 0.5)$

Loading Difference	Loading Pattern	Group Sample Size Ratio	Group Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
			RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	-0.008	-0.008	0.012	0.012	-0.004	-0.003
		100:400	-0.011	-0.011	0.004	0.005	-0.013	-0.012
		400:100	-0.006	-0.006	0.004	0.004	-0.004	-0.008
0.1	1st Loading Unequal	250:250	0.017	0.016	0.022	0.024	0.033	0.033
		100:400	0.019	0.019	0.050	0.051	0.013	0.014
		400:100	0.050	0.050	0.049	0.051	0.037	0.037
	2nd Loading Unequal	250:250	0.013	0.013	0.034	0.035	0.008	0.009
		100:400	0.010	0.010	-0.004	-0.003	-0.014	-0.013
		400:100	0.041	0.041	0.050	0.051	0.057	0.057
	All Lower	250:250	0.046	0.045	0.059	0.061	0.037	0.036
		100:400	0.015	0.014	0.042	0.043	0.010	0.010
		400:100	0.066	0.066	0.074	0.076	0.070	0.068
	Mixed	250:250	0.0004	0.0002	-0.013	-0.011	0.015	0.015
		100:400	-0.002	-0.002	-0.017	-0.017	-0.015	-0.014
		400:100	0.009	0.009	0.016	0.017	0.014	0.011
0.4	1st Loading Unequal	250:250	0.111	0.103	0.155	0.154	0.097	0.088
		100:400	0.039	0.034	0.052	0.051	0.043	0.037
		400:100	0.215	0.204	0.229	0.229	0.175	0.155
	2nd Loading Unequal	250:250	0.129	0.121	0.147	0.145	0.092	0.083
		100:400	0.025	0.021	0.077	0.076	0.038	0.032
		400:100	0.198	0.187	0.213	0.214	0.180	0.162
	All Lower	250:250	0.184	0.153	0.230	0.212	0.130	0.091
		100:400	0.060	0.046	0.068	0.059	0.033	0.019
		400:100	0.344	0.313	0.390	0.374	0.299	0.255
	Mixed	250:250	-0.005	-0.006	-0.011	0.008	0.030	0.022
		100:400	0.008	0.006	0.050	0.058	-0.007	-0.012
		400:100	-0.064	-0.058	-0.087	-0.072	-0.015	-0.026

Note: Relative parameter bias values equal to or greater than 0.05 are underlined which represent unacceptable bias. Abbreviations used in this table are described in Table 1.

were deemed unacceptable occurred when group sample sizes were unequal. Additionally, the majority of unacceptable Type I error rates were found in conditions in which the group factor variance ratio was lower in group one than in group two and the loading difference was large. However, when the sample sizes were equal across groups, violating the equal factor-variance assumption did not have any substantial impact on Type I error rates associated with the LRT_{κ}.

Power associated with the LRT_{κ} was affected by group sample size ratio and loading difference magnitude. For example, when factor loadings were either equal or unequal across groups, power rates below the cutoff criterion of 0.90 were found only in the unequal sample size conditions; this finding is consistent with Kaplan and George (1995). Both group sample size ratio and loading difference magnitude influenced the power of the LRT_{κ} when factor loadings were unequal. That is, power was low only in unequal sample size scenarios and low factor loading difference scenarios. These low power rates, nonetheless, were not considerably lower than 0.90 and would in fact be deemed as acceptable if the traditional, less stringent 0.80 power criterion had been used as the benchmark. High power was particularly observed in the large latent mean difference conditions, as would be expected.

Previous studies have not investigated the performance of the standardized effect size measure, $\hat{\delta}_{\kappa}$, under varying conditions, particularly when the assumptions underlying the RI strategy and the factor-variance scaling method are violated. The findings in this study demonstrate that the $\hat{\delta}_{\kappa}$ is not biased in conditions in which there was no latent mean difference between the two groups. Thus, violating the assumptions associated with the RI strategy and the factor-variance scaling method did not have any substantial or systematic impact on the parameter bias of the $\hat{\delta}_{\kappa}$. Further, the parameter bias of the $\hat{\delta}_{\kappa}$ was not affected by loading difference magnitude, group sample size ratio, group factor variance ratio, or factor loading pattern.

When there was a latent mean difference between the two groups, the $\hat{\delta}_{\kappa}$ was not biased in the baseline conditions in which factor loadings were equal/invariant. However, substantial relative parameter bias of the $\hat{\delta}_{\kappa}$ was found in the partial metric invariance conditions in which factor loadings were unequal. In addition, the relative parameter bias of the $\hat{\delta}_{\kappa}$ in these partial invariance conditions varied as a function of loading difference magnitude, factor loading pattern, and group sample size ratio, regardless of the factor scaling method used. Overall, the relative parameter bias was more unacceptable when the factor loading difference magnitude was large, when the non-invariant factor loadings were higher in group two, and when sample size in group one was larger than sample size in group two.

Study Limitations

The assumptions underlying the RI strategy and the factor-variance scaling method have not been widely investigated in previous simulation studies. Thus, as a starting point for this line of research, this study included a relatively simple model and investigated latent mean difference comparisons under relatively ideal conditions. Due to the preliminary nature of the research, there are several limitations inherent in this study. For example, only a moderately large latent mean difference was included when investigating the power of the LRT_{κ}.

As a result, power associated with the LRT_{κ} was high in these conditions and did not differ systematically as a function of the factor loading pattern or group factor variance ratio. It was found that violating the assumptions underlying the two factor-scaling methods did not influence the power of the LRT_{κ}. However, it is not clear whether the same findings would be obtained with smaller latent mean differences (e.g., 0.1 and 0.3). In future simulation studies, researchers could include smaller latent mean differences and examine how violating the assumptions underlying the two factor-scaling methods may affect the power of the LRT_{κ}.

Neither model size nor model complexity was varied in this study. For simplicity, a two-group, one-factor CFA model with six indicator variables was the true generating model. Future researchers could consider more complex models (for example, more observed indicators and/or additional latent variables) to investigate whether varying the model size and/or model complexity would affect the testing and description of the latent mean difference across groups. Future research that includes models with more observed indicators could likewise investigate more severe loading non-invariance conditions. Further, mean comparisons between more than two groups are not uncommon and, hence, the impact of including more than two groups on latent mean comparisons could be examined in future investigations. In addition, multivariate normal data were generated. Future studies could also explore the implications of violating the assumption of normality when using the LRT_{κ} and the $\hat{\delta}_{\kappa}$ to test and describe groups' latent mean differences.

The results of this study suggest that researchers do not necessarily need to be concerned about violating the assumption underlying the RI strategy given that it does not adversely affect the performance of the LRT_{κ}. The results also suggest that researchers do not necessarily need to be concerned about violating the assumptions underlying either of the two factor scaling methods when using the $\hat{\delta}_{\kappa}$ in order to describe the latent mean difference across groups.

The findings concerning the RI factor scaling method are notable because the assumption underlying the RI strategy may be frequently violated given the difficulty of identifying an item with truly invariant factor loadings (Hancock, Stapleton & Arnold-Berkovits, 2009). Nonetheless, more research is necessary in order to assuredly know that violating the RI assumption does not impact latent mean difference testing and that violating either of the factor scaling method assumptions does not impact latent mean difference descriptions. By contrast, the results clearly indicate that researchers should be aware of the assumption underlying the factor-variance scaling method. In particular, when the sample sizes for the two groups being compared are unequal, constraining unequal factor variances to a value of one across groups is likely to produce overly conservative or liberal Type I error rates associated with the latent mean difference test (LRT_{κ}). Additionally, researchers should cautiously interpret the $\hat{\delta}_{\kappa}$ when factor loadings are non-invariant across groups in combination with unequal group sample sizes, regardless of factor scaling method employed.

References

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

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.

Cashen, L. H., & Geiger, S. W. (2004). Statistical power and the testing of null hypotheses: A review of contemporary management research and recommendations for future studies. *Organizational Research Methods*, *7*, 151-167.

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

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (*2nd Ed.*). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin*, *114*, 174-184.

Enders, C. K., & Finney, S. J. (2003, April). *SEM fit index criteria re-examined: An investigation of ML and robust fit indices in complex models*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Fan, X., & Fan, X. (2005). Using SAS for Monte Carlo simulation research in SEM. *Structural Equation Modeling*, *12*, 299-333.

Gagné, P., & Furlow, C. F. (2009). Automating multiple software packages in simulation research for structural equation modeling and hierarchical linear modeling. *Structural Equation Modeling*, *16*, 179-185.

Gonzalez, R., & Griffin, D. (2001). Testing parameters in structural equation modeling: Every "one" matters. *Psychological Methods*, *6*, 258-269.

Hancock, G. R. (1997). Structural equation modeling methods of hypothesis testing of latent variable means. *Measurement and Evaluation in Counseling and Development*, *30*, 91-105.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, *66*, 373-388.

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2000). Type I error and power of latent mean methods and MANOVA in factorially invariant and noninvariant latent variable systems. *Structural Equation Modeling*, *7*, 534-556.

Hancock, G. R., Stapleton, L. M., & Arnold-Berkovits, I. (2009). The tenuousness of invariance tests within multisample covariance and mean structure models. In *Structural equation modeling: Concepts and applications in educational research*, T. Teo & M. S. Khine (Eds.), 137-174. Rotterdam, Netherlands: Sense Publishers.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, *21*, 967-988.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. *Sociological Methods and Research*, *26*, 329-367.

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling*, *16*, 642-657.

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychomerika*, *27*, 179-182.

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

Kim, K. H., Cramond, B., & Bandalos, D. L. (2006). The latent structure and measurement invariance of scores on the Torrance tests of creative thinking-Figural. *Educational and Psychological Measurement*, *66*, 459-477.

Kline, R. (2011). *Principles and Practice of Structural Equation Modeling (3rd Ed.)*. New York, NY: Guilford Publications.

Lawrence, F. R., & Hancock, G. R. (1998). *Finite sample behavior of the likelihood ratio, Wald, and Lagrange Multiplier tests: Bias and variability in univariate noncentrality parameter estimation*. Paper presented at the annual meeting of the American Educational Research Association, April, San Diego, CA.

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

Muthén, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, *46*, 407-419.

Muthén, L. K., & Muthén, B. O. (2010). M*plus* (Version 6.1) [computer software]. Los Angeles, CA: Muthén & Muthén.

Rensvold, R. B. & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In *Research in management: Vol. 1. Equivalence in measurement*, C. A. Schriesheim & L. L. Neider (Eds.), 21-50. Greenwich, CT: Information Age.

Rock, D. A., Werts, C. E., & Flaugher, R. L. (1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research*, *13*, 403-418.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, *58*, 646-656.

SAS. (2008). SAS (Version 9.2) [computer software]. Cary, NC: SAS Institute Inc.

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

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

Thompson, M. S., & Green, S. B. (2006). Evaluating between-group differences in latent variable means. In *Structural equation modeling: A second course*, G. R. Hancock & R. O. Mueller (Eds.), 119-169. Greenwood, CT: Information Age Publishing, Inc.

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

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.