Behavior Research Methods (2019) 51:2688–2699
https://doi.org/10.3758/s13428-018-1124-6

Impact of error structure misspecification when testing measurement invariance and latent-factor mean difference using MIMIC and multiple-group confirmatory factor analysis

Seang-Hwane Joo[1] · Eun Sook Kim[2]

Published online: 21 September 2018

Abstract
When multiple groups are compared, the error variance–covariance structure is not always invariant between groups. In this study we investigated the impacts of misspecified error structures on testing measurement invariance and the latent-factor mean difference between groups. A Monte Carlo study was conducted to examine how measurement invariance and latent mean difference tests were affected when heterogeneous error structures were misspecified as being invariant across groups. Multiple-group confirmatory factor analysis (MGCFA) and the multiple-indicator multiple-causes model (MIMIC) were employed in the present study. The rejection rates of both metric and strict invariance in measurement invariance testing, as well as the estimation accuracy and statistical inference of the factor mean difference, were investigated under error structure misspecification. In addition, sensitivity of the model fit indices to error structure misspecification was examined. Overall, misspecification of the error structure affected testing for metric but not scalar invariance. Metric invariance was often rejected, especially when error covariance in one group was ignored. In contrast, MGCFA and MIMIC performed comparatively well at detecting latent-factor mean differences between groups, with acceptable power and well-controlled Type I errors. The practical implications of these findings are discussed, as well as recommendations.

Keywords Measurement invariance · Error structure misspecification · Model fit sensitivity · MIMIC · MGCFA

Introduction

When multiple groups are compared using factor models, researchers are often interested in the group mean difference. However, prior to any multiple-group analysis, measurement invariance should hold across the groups (DeShon, 2004). Mellenbergh (1989) developed a mathematical expression of measurement invariance concepts using conditional probability:

$$P\left(Y_{ij} = y|\eta_i, G\right) = P\left(Y_{ij} = y|\eta_i\right), \quad (1)$$

where η_i is the factor score for ith examinee, Y_{ij} is the ith examinee's response score for jth item, and G is a group membership. Equation 1 states that the probability of a response of ith examinee to the jth item, conditioned on the latent-factor score, is independent of group membership G. In other words, if measurement invariance holds, examinees with the same latent-factor scores are expected to have the same probability of endorsing a response on the measure, regardless of their group membership. Multiple-group confirmatory factor analysis (MGCFA) is perhaps the most widely used method for testing measurement invariance among applied researchers, due to its flexibility and convenience.

Alternatively, multiple-indicator multiple-causes modeling (MIMIC; Jöreskog & Goldberger, 1975) has been employed for detecting measurement invariance and testing for latent mean differences (e.g., Fleishman, Spector, & Altman, 2002; McCarthy, Pedersen, & D'Amico, 2009; Muthén, Kao, & Burstein, 1991; Rubio, Berg-Weger, Tebb, & Rauch, 2003; Woods, Oltmanns, & Turkheimer, 2009). MIMIC for measurement invariance testing has several advantages in model specification. For example, multiple categorical variables (e.g., ethnicity or socioeconomic status) and their interaction terms can be tested simultaneously (e.g., Ainsworth, 2008; Fleishman et al., 2002), and the measurement invariance of a continuous covariate can be investigated (Barendse, Oort, & Garst, 2010).

✉ Seang-Hwane Joo
seanghwane.joo@kuleuven.be

[1] KU Leuven, Kortrijk, Belgium

[2] University of South Florida, Tampa, FL, USA

Measurement invariance, in general, is tested with a sequence of increasingly restrictive models. The sequence begins with the equality of confirmatory factor model configurations across groups (configural invariance), then moves across equality of the factor loadings (metric or weak invariance), intercepts (scalar or strong invariance), and error variances of the observed variables (strict invariance). Homogeneity of factor loadings, intercepts, and error variances across groups (strict invariance) is a necessary condition to enable common factor models to obtain some certainty that measurement invariance holds for multiple groups (e.g., DeShon, 2004; Meredith, 1993).

However, measurement invariance studies have suggested that the equality constraints for factor loadings and intercepts across groups are sufficient for multiple-group analysis (e.g., Jöreskog & Sörbom, 1993; Marsh, 1994; McArdle, 1998; Sörbom, 1974), because the difference in error variances affects only the reliability of observed variables (Little, 1997). In addition, when latent variables are compared across groups, measurement errors are taken into account for the latent variables (Marsh, Nagengast, & Morin, 2013). Thus, the invariance of error variances is often not considered if group mean differences in the latent factors or observed scores are of concern, as long as strong invariance holds for the data.

On the other hand, some scholars have raised concerns about possible impacts of heterogeneous error variances in multiple-group analysis. Lubke and colleagues took admission decisions based on observed scores as an example (Lubke, Dolan, Kelderman, & Mellenbergh, 2003). If heterogeneous error variances truly exist between groups, incorrect admission decisions could be made more frequently for the group with the larger error variance. Heterogeneity of error variances could also mislead interpretation of the results of measurement invariance testing using likelihood ratio (LR) test or model fit indices, because inflated chi-squares or poor model fit could occur when the model is misspecified (i.e., if homogeneous error variances are assumed when there is heterogeneity). Consequently, the measurement invariance test could mislead toward noninvariance, although measurement invariance does hold.

In addition, it is not uncommon to observe correlated error structures in practical situations (e.g., Heene, Hilbert, Freudenthaler, & Bühner, 2012; Lubke et al., 2003). Correlated errors could occur if item contents were overlapped or items were logically dependent upon one another. If the item contents were multidimensional but a unidimensional model was chosen, correlated error structure could also occur because unexplained residuals would be correlated with the unspecified factor. In the context of multiple-group analysis, error covariances could be present in one of the groups compared (Lubke & Dolan, 2003). For example, in cross-cultural studies, respondents in one culture might interpret negatively worded items differently, and the errors of the negatively worded items would possibly be correlated in this cultural group. Previously, researchers have empirically examined the impact of correlated error structures for confirmatory factor analysis (CFA) and have reported that correlated errors could lead to bias in the factor loadings and reliability estimates (e.g., Heene et al., 2012; Raykov, 2001; Shevlin, Miles, Davies, & Walker, 2000). Because an error covariance can indicate the presence of an additional factor, either substantive or nuisance, not specified in the model, ignoring error covariance in one group could be consequential in multiple-group analysis.

In this study, heterogeneity in either error variances or error covariances is considered as a heterogeneous error structure across groups. Of note is that an error covariance present in one group but not in the other group would be considered a violation of configural invariance, because the configuration of the CFA model would not be homogeneous across groups. On the other hand, heterogeneity in the error variances would be considered a violation of strict invariance, given equality of the configurations, factor loadings, and intercepts. Strict invariance indicates homogeneous error variances and covariances in addition to equal factor loadings and intercepts.

Although the issue of noninvariant error variance–covariance structure (or, interchangeably, simply *error structure*) in multiple-group analysis has been raised, the impact of such heterogeneity on measurement invariance testing and latent-factor mean difference testing has not been systematically investigated to date. Hence, a Monte Carlo study was needed so as to empirically examine the extent to which misspecifying the error structure affects testing measurement invariance and latent-factor mean differences with commonly used multiple-group analysis models—namely, MGCFA and MIMIC. Given that heterogeneity in the error structure is potentially more problematic with MIMIC, because MIMIC does not have the flexibility to specify different error variances or covariances across groups, comparing MGCFA and MIMIC directly would be worthwhile.

MGCFA and measurement invariance

In a single-group confirmatory factor model, continuous random variables Y are regressed on continuous latent variables η. Given that $i = 1, \ldots, I$ for examinees and $j = 1, \ldots, J$ for items, the single-group confirmatory factor model can be represented as follows:

$$Y_{ij} = \nu_j + \sum_{k=1}^{K} \lambda_{jk}\eta_{ik} + \varepsilon_{ij}, \tag{2}$$

where ν_j, λ_{jk}, η_{ik}, and ε_{ij} denote the intercepts, factor loadings, latent factors, and residuals, respectively. In Eq. 2, there are a total of K latent factors, $k = 1, \ldots, K$. Additionally, ε_{ij} is assumed to follow a multivariate normal distribution with mean vector $\mathbf{0}$ and diagonal matrix $\boldsymbol{\Theta}$. The diagonal matrix, $\boldsymbol{\Theta}$, implies an uncorrelated error structure of the confirmatory factor model. To incorporate the measurement invariance

concept with MGCFA, suppose there are a total of G groups, denoted as $g = 1, \ldots, G$. Also, let the expected values of the random variables Y_{ij} and η_{ik} in vector form be denoted as $\boldsymbol{\mu_g}$ and $\boldsymbol{\alpha_g}$, respectively, for group g. The covariance matrices of the random variables Y_{ij} and η_{ik} are denoted $\boldsymbol{\Sigma_g}$ and $\boldsymbol{\Psi_g}$, respectively, for group g. Then, the mean and covariance of Y for group g can be represented in matrix forms as follows, with the consideration of measurement invariance satisfied:

$$\boldsymbol{\mu_g} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\alpha_g}, \tag{3}$$

$$\boldsymbol{\Sigma_g} = \boldsymbol{\Lambda}\boldsymbol{\Psi_g}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \tag{4}$$

where $\boldsymbol{\Theta}$ is a diagonal matrix of the variance components of errors, and $\boldsymbol{\Lambda}$ represents the factor loading matrix with respect to latent factors. Equations 3 and 4 imply that (a) the factor loadings are equal across groups ($\boldsymbol{\Lambda_g} = \boldsymbol{\Lambda}$), (b)the intercepts are equal across groups ($\boldsymbol{\nu_g} = \boldsymbol{\nu}$), and (c) the residual covariance matrices are equal across groups ($\boldsymbol{\Theta_g} = \boldsymbol{\Theta}$). When those conditions are satisfied in MGCFA, it is considered that measurement invariance or factorial invariance holds across groups $g = 1, \ldots, G$ (Meredith, 1993). Then, differences in the observed means across groups ($\boldsymbol{\mu_g}$) are due solely to differences in the factor means across groups ($\boldsymbol{\alpha_g}$); differences in observed variance–covariance ($\boldsymbol{\Sigma_g}$) are due solely to differences in the factor variance–covariances across groups ($\boldsymbol{\Psi_g}$).

MIMIC and measurement invariance

Alternatively, in MIMIC, group variables are considered causal indicators of factors. These causal indicators are coded as dummy variables (X_i), and the effects of the variables can be detected according to this model. For simplicity of the discussion, a single causal indicator for two groups (i.e., reference and focal groups) is included.

$$Y_{ij} = \nu_j + \lambda_j\eta_i + \varepsilon_{ij}, \tag{5}$$

$$\eta_i = \gamma X_i + \zeta_i. \tag{6}$$

Equation 5 represents the measurement relationships between an observed variable and a latent factor in MIMIC. In Eq. 6, X_i denotes a dummy variable that indicates group membership, γ denotes an effect or path coefficient of the group variable on the latent factor, and ζ_i represents the disturbance of the latent factor. Given that the mean of the disturbance term for the latent factor is 0, the expected value of the latent factor can be represented as follows:

$$E(\eta_i) = \gamma E(X_i), \tag{7}$$

where γ indicates the group difference in the latent-factor means with a dummy-coded grouping variable, X_i (Thompson & Green, 2006). In other words, the latent-factor mean for the focal group ($X_i = 1$) is γ units higher (or lower) than that of the reference group ($X_i = 0$). Also note that the residual variance (Var[ε_{ij}]) is assumed to be independent (i.e., Cov[ε_{ij}] = 0) and homogeneous, regardless of the multiple groups.

Measurement invariance testing using MIMIC can be performed by adding a direct path from the grouping variable to the observed variables (Kim, Yoon, & Lee, 2012):

$$Y_{ij} = \nu_j + \lambda_j\eta_i + \beta_j X_i + \omega_j\eta_i X_i + \varepsilon_{ij}, \tag{8}$$

$$\eta_i = \gamma X_i + \zeta_i. \tag{9}$$

Statistical significance of the direct path from the grouping variable to the observed variable (β_j) indicates that the intercept of item j is not invariant across groups; this is referred to as *uniform noninvariance* (Woods & Grimm, 2011).

Similarly, nonuniform measurement noninvariance can be tested using MIMIC, by adding a path from an interaction term between the latent factor and the grouping variable to the observed variables ($\eta_i X_i$, in Eq. 8). Statistical significance of the path from the interaction term to the observed variable (ω_j) implies factor loading noninvariance or nonuniform noninvariance of the associated item (e.g., Barendse et al., 2010; Barendse, Oort, Werner, Ligtvoet, & Schermelleh-Engel, 2012; Woods & Grimm, 2011).

Noninvariance of the error structure in MIMIC and MGCFA

The issue of noninvariance in error variance–covariance can be capitalized on with MIMIC, because MIMIC inherently assumes strict invariance. Relaxing the equality of factor loadings and intercepts between groups is possible, as we explained earlier. However, relaxing the equality of residual variances–covariances between groups is challenging, which is one of the major limitations of MIMIC. Previously, Kim, Yoon, and Lee (2012) investigated the performance of MIMIC when strict invariance was incorrectly specified in the presence of factor loading noninvariance. Their study concluded that MIMIC could not detect noninvariance of the factor loadings properly, and they recommended using MIMIC only when factor loading invariance is established, unless the factor loading equality is relaxed by including the Factor × Grouping variable interaction. A question remained unsolved, however: How does MIMIC behave in the presence of the residual variance–covariance noninvariance when MIMIC assumes strict invariance? Moreover, sensitivity of the model fit indices to violation of the strict invariance assumption of MIMIC is worth investigating when MIMIC is misspecified for error structures between groups. MGCFA, on the other hand, has greater flexibility in modeling the error structures between groups. That is,

in MGCFA the error variance–covariance matrix can be freely estimated between groups ($\Theta_1 \neq \Theta_2$).

Purpose of the study

The purpose of the present study was to investigate the impact of misspecified error structure on measurement invariance testing and latent-factor mean estimation when MGCFA and MIMIC are used for multiple-group analysis. More specifically, we examined the performance of both metric and scalar invariance tests following typical measurement invariance procedures under conditions in which either error variance or error covariance was heterogeneous across groups. For each type of misspecification, we further examined the accuracy of the latent-factor mean estimations using MGCFA and MIMIC under the assumption of strict invariance. Finally, we investigated the sensitivity of the model fit indices to the misspecification of error structure.

Method

Simulation design

A simulation study was conducted to investigate how the heterogeneity in residual variance–covariance made an impact on testing configural, metric, and scalar invariance with MGCFA and MIMIC. The simulation conditions included manipulations of (1) the type of error structure misspecification (heterogeneous error variances vs. heterogeneous error covariances), (2) the size of the heterogeneity (small vs. large), (3) the number of heterogeneous items (one vs. two), (4) the sample size for each group (100, 200, 500, 1,000), and (5) the size of the population latent-factor mean differences (0, .1, .5). A total of 96 (2 × 2 × 2 × 4 × 3) conditions were included, and 1,000 replicated data were generated for each condition.

Data generation

Data were generated on the basis of a unidimensional single-latent-factor model with six observed variables (Y1–Y6) for two groups. The generating parameters for the reference group are presented in Table 1. The factor loadings and intercepts were simulated as being homogeneous between groups. The latent-factor variance was fixed at 1 for both the reference and focal groups, whereas the latent-factor mean was 0 for the reference group and 0, .1, or .5 for the focal group, depending on the simulation condition. For the reference group, the generating parameter values for residual variances were computed with the parameterization such that $\lambda^2 + \theta^2 = 1$. As a result, the reliability coefficient (ω) for the reference group was .87. For the focal group, because the residual variances were manipulated across conditions, the $\lambda^2 + \theta^2 = 1$ parameterization was not applied. Also, the reliability coefficient was varied ranging from .84 to .86, depending on the size and number of misspecifications (more severe misspecification resulted in a smaller reliability). All items were generated as continuous variables and were assumed to follow a multivariate normal distribution. The generating parameters were based on previous MGCFA and MIMIC studies (e.g., Kim et al., 2012).

Table 1 Generating parameter values for the reference group

Items	Loadings	Intercepts	Error Variances
Y1	.90	– .15	.19
Y2	.70	.25	.51
Y3	.60	.15	.64
Y4	.80	– .25	.36
Y5	.70	– .10	.51
Y6	.60	.10	.64

Simulation conditions

Type of error structure misspecification We manipulated error structure misspecifications in two ways: (a) heterogeneity of *error variances* and (b) heterogeneity of *error covariances* between the reference and focal groups. For the *heterogeneous error variance* condition, the focal group had higher error variances than did the reference group. For the *heterogeneous error covariance* conditions, the focal group's errors were correlated. We included this type of heterogeneity because, in practice, errors are not always independent (e.g., Heene et al., 2012; Lubke et al., 2003), and error covariances could occur only in one of the groups. To the authors' best knowledge, few studies have investigated the impact of heterogeneous variance–covariance in MGCFA (e.g., Green & Hershberger, 2000; Lubke & Dolan, 2003), and no studies have examined the performance of MIMIC under such conditions.

Size of heterogeneity We considered two levels of the size of heterogeneity: small and large. For the heterogeneous error variance condition, the focal group had error variances .2 (small) or .4 (large) higher than those in the reference group. Similarly, for the heterogeneous error covariance condition, the error covariance in the focal group was included at two levels, .2 (small) or .4 (large), which correspond to correlations of about .4 and .8, respectively. It should be noted that small covariance (.2) is more commonly observed in applied research, and the generated large covariance (.4) was considered to be more extreme conditions than usual.[1]

[1] In the multilevel CFA studies Kim, Dedrick, Cao, and Ferron (2016) reviewed, the error correlations ranged from .20 to .65 (n = 22), with one outlier (.08).

Number of heterogeneous items We included a scenario in which the number of heterogeneity was one or two. Note that the total number of items considered in this study across conditions was six (Y1–Y6). Thus, for the heterogeneous error variance conditions, 17% or 33% of the variables were not invariant in conditions of one or two heterogeneity, respectively; for the heterogeneous error covariance conditions, 33% (one pair) or 67% (two pairs) of the items were involved in error covariances. For the condition in which the heterogeneity number was one, Y2 was selected as the heterogeneous item when error variances were heterogeneous, and Y2 and Y3 were selected as the correlated error items when error covariances were heterogeneous. When the heterogeneity number was two, Y2 and Y5 were selected as the heterogeneous error variance items, and two pairs (Y2 and Y3, Y4 and Y5) were selected as the heterogeneous covariance items.

Group size Four levels of group size were considered: 100, 200, 500, and 1,000 in each group. A balanced group design (i.e., equal sample sizes for the reference and focal groups) was considered across simulation conditions.

Size of the population factor mean difference The effect size of the population factor mean difference was manipulated to have three levels in this study—0, .1, and .5—which represent no, small, and large factor mean differences, respectively. These effect sizes had commonly been used in previous simulation studies (e.g., Barendse et al., 2010; Kim et al., 2012). For factor mean difference testing, Type I errors were estimated when no mean difference was generated in the population; power was estimated with small and large mean differences.

Measurement invariance tests

A series of measurement invariance tests (configural, metric, and scalar) were conducted with MGCFA and MIMIC under the simulated conditions. Of note is that the measurement invariance tests were conducted only for the conditions in which the latent-factor mean difference was .1. We chose these conditions because the size of the latent-factor mean difference (0, .1, or .5) does not have an impact on measurement invariance testing as long as the factor means are correctly specified (i.e., allowed to be different across groups). For the measurement invariance test using MGCFA, we used likelihood ratio (LR) tests for nested models. That is, a configural-invariance model in which the factor loadings, intercepts, and residual variances were relaxed between groups was compared to a metric-invariance model in which the factor loadings were constrained to be equal between groups. Similarly, scalar invariance was tested by comparing the metric-invariance model and the scalar-invariance model, in which the intercepts were additionally constrained to be equal. It should be kept in mind that MGCFA was a correctly specified model for the heterogeneous error variance condition, because error variances were allowed to be different between groups. However, for the heterogeneous error covariance condition, it is a misspecified model, because MGCFA assumes independent error structures for both groups. For the measurement invariance test using MIMIC, a configural-invariance model was constructed by including two paths (β_j and ω_j in Eq. 8) for all items except the first one for identification: a path from the grouping covariate to each item, and a path from the interaction between the grouping covariate and the latent factor to each item. Then, the metric-invariance model was constructed by constraining all path coefficients from the interaction to the items (ω_j) at zero. The scalar-invariance model was constructed with additional zero constraints on the paths from the grouping covariate to all items (β_j). Note that when configural-, metric-, and scalar-invariance MIMIC models were fitted, the latent-factor mean difference between two groups (γ) was also simultaneously estimated. Similar to the LR tests in MGCFA, these nested models were compared sequentially (i.e., configural vs. metric, metric vs. scalar) to determine the measurement invariance. We used the Satorra–Bentler correction (Satorra & Bentler, 2001) for the LR tests in MIMIC, because robust maximum likelihood (MLR) was used for the model estimation.

The MIMIC model with Factor * Covariate interaction is often estimated with MLR rather than maximum likelihood (ML) estimation because a numerical integration algorithm is required (i.e., TYPE = RANDOM and ALGORITHM = INTEGRATION in Mplus). Because we generated the response data from a multivariate normal distribution, we did not expect any substantial difference between MLR and ML. In the preliminary study, we compared the performance of ML and MLR and did not find any notable differences for MIMIC (e.g., the ML and MLR outputs with several replications were identical). Thus, the choice between ML and MLR would not impact the results of this study.

In addition to the LR tests, measurement invariance was evaluated with Wald tests in MIMIC. In the configural-invariance model, the statistical significance of a ω_j path coefficient indicates the lack of nonuniform or metric invariance of the tested item; the statistical significance of a β_j path coefficient indicates a violation of uniform or scalar invariance of the tested item. For parameterization of the models, we fixed the first observed variable to be equal across groups, and parameters of the other items were allowed to be estimated. Of note is that MIMIC was a misspecified model for heterogeneity of both error variances and covariances, because MIMIC does not allow for modeling heterogeneity in errors.

Latent-factor mean tests

In addition to measurement invariance tests, in the present study we further explored the accuracy of the latent-factor mean differences across groups when the error structure was

misspecified. To create misspecification in the error structure, we constrained the error structures to be equal between groups when they were heterogeneous in the population. In other words, the latent-factor mean difference was tested assuming that strict invariance was satisfied; that is, the error structures were constrained to be equal between groups (i.e., with equal error variances and no error covariances), in addition to the equality of factor loadings and intercepts. For MIMIC, a grouping covariate was included in the model to test a latent-factor mean difference (γ in Eq. 9). Because strict invariance was assumed, no other paths from the grouping variable (i.e., $\beta_j X_i$ and $\omega_j \eta_i X_i$ in Eq. 8) were included in the model. In this model the heterogeneous error structure was the only source of model misspecification, because the factor loadings and intercepts were generated to be equal in the population. The statistical significance of the γ coefficient indicated a statistically significant latent group mean difference. To investigate the behaviors of MGCFA in factor mean difference testing under the same error structure misspecification, a strict-invariance model was constructed (i.e., strict-invariance MGCFA), as in MIMIC. Then the latent-factor mean difference was evaluated by testing the statistical significance of the second group's latent-factor mean, because the first group's mean was constrained to be 0 and the second group's mean represented the mean difference between the groups. Thus, both strict MGCFA and MIMIC were considered as incorrectly specified models. Strict-invariance MGCFA was examined because it has theoretical similarities to MIMIC (Woods, 2009), and it was worthwhile to compare their performance in terms of parameter estimation and model fit indices. For both MGCFA and MIMIC, maximum likelihood estimation was used for the model estimation.

In addition to the strict-invariance models, we fitted the correctly specified MGCFA in order to establish the baseline results for latent-factor mean estimation and model fit index sensitivity. As this was a correctly specified model, the factor loadings and intercepts were constrained to be equal between groups. However, the error variances were freely estimated. For the heterogeneous error covariances conditions, errors were additionally allowed to be correlated for the designated items for one group, as had been generated in the population. Mplus 7 (Muthén & Muthén, 2012) was used for both generating the data and fitting the models. The Mplus program code for the study can be obtained from the authors upon request.

Simulation analysis

Rejection rates We investigated how error structure misspecification affected tests for metric and scalar invariance when the factor loadings and intercepts were invariant in the population. For simulation outcomes, we examined the rejection rates of metric and scalar invariance. The rejection rate of metric invariance was computed as the proportion of replications in which metric invariance was rejected at alpha .05; the rejection rate of scalar invariance, as the proportion of replications in which scalar invariance was rejected. For the Wald test of MIMIC, the rejection rates were computed as the proportions of replications in which any of the tested path coefficients across five items (Y2–Y5) was flagged with statistical significance at alpha .01. The significance level was adjusted to .01 (.05/5 items) in order to control for experiment-wise Type I errors.

Relative bias, RMSE, and SE The relative biases of the latent-factor mean difference were computed for MGCFA and MIMIC, to investigate the accuracy of latent-factor mean difference estimation. In addition, the root mean squared error (RMSE) and standard error (*SE*) of the parameter estimates were examined. The relative bias and RMSE were computed as

$$\text{Relative Bias} = R^{-1}\sum_{i=1}^{R}\frac{\hat{\theta}_i - \theta}{\theta}, \tag{10}$$

$$\text{RMSE} = \sqrt{R^{-1}\sum_{i=1}^{R}\left(\hat{\theta}_i - \theta\right)^2}, \tag{11}$$

where θ and $\hat{\theta}_i$ represent the generated and estimated parameters for *i*th replication, respectively, and R represents the total number of replications (i.e., R = 1,000). For the conditions in which the generating parameter was 0, we simply computed bias in the traditional fashion (i.e., $R^{-1}\sum_{i=1}^{R}\hat{\theta}_i - \theta$). A relative bias above .05 is considered as representing biased estimates of the parameters (Hoogland & Boomsma, 1998). The *SE* was obtained by averaging the standard errors across 1,000 replications.

Power and Type I error We also evaluated the statistical inference of the latent-factor mean estimates. When the true population effect size was 0, the Type I error rates of the latent-factor mean difference tests were computed for MGCFA and MIMIC. When the true population effect size was .1 or .5, statistical power was computed. The level of significance of the test was set at .05 across conditions. Power and Type I error were computed by taking the proportions of a statistically significant group mean difference over replications.

Model fit indices In addition to measurement invariance and latent-factor mean estimation, we also investigated the sensitivity of the model fit indices of the strict-invariance models. To investigate the sensitivity of the model fit indices of MGCFA and MIMIC to misspecification of the error structure, commonly used model fit indices—namely, the chi-square (χ^2), root mean square error of approximation (RMSEA), comparative fit index (CFI), and standardized root mean residual (SRMR) statistics—were examined. We

applied the Hu and Bentler (1999) criteria for RMSEA, SRMR, and CFI in the present study. That is, RMSEA greater than .06, SRMR greater than .08, and CFI less than .95 were considered as representing poor model fit. Also, a chi-square *p* value less than .05 was considered a poor model fit. We also examined the fit of the configural-invariance models using the same criteria, because configural invariance was violated in the heterogeneous error covariance conditions. Note that evaluation of the configural-invariance model fit was conducted only with MGCFA, because MIMIC does not have the flexibility to relax the factor loadings, intercept, and error variances simultaneously across groups.

Results

Simulation check with the correctly specified MGCFA model

To establish the baseline results for the latent-factor mean difference and model fit indices, we first fitted the correctly specified MGCFA. The results showed that the latent-factor group mean differences were estimated accurately and the corresponding power and Type I error rates were well established. The power rates reached up to 1.00, and Type I error rates were controlled across conditions, ranging from .05 to .07. With regard to model fit indices, the results overall showed good model fit. The average model fit values ranged from .01 to .03 (RMSEA), .01 to.04 (SRMR), .98 to 1 (CFI), and .45 to .49 (chi-square *p* value). The result table for the correctly specified model is not reported, but it is available upon request.

Measurement invariance tests

Table 2 shows the rejection rates of the measurement invariance tests with MGCFA and MIMIC. When error variances were heterogeneous, MGCFA controlled for the rejection rates around .05, as expected, because the error variances were allowed to be different in MGCFA (i.e., correctly specified model). On the other hand, MIMIC was a misspecified model, because it estimated a single set of error variances for both groups. The rejection rates of MIMIC were slightly inflated when metric invariance was tested with a large sample size under large error structure heterogeneity (e.g., .13 for the SB LR tests when the group size was 1,000 and there were two large-size misspecified items). The rejection rates of the scalar-invariance test were .05 or less across conditions.

When error covariances were heterogeneous, both MGCFA and MIMIC were misspecified models. Both MGCFA and MIMIC showed relatively high rejection rates when metric invariance was tested. As the sample size and the number and size of error covariance increased, the rejection rates increased considerably. In contrast, the rejection rates of the scalar invariance tests were generally low, with values less than or around .05, with some exceptions (i.e., when the group size was 1,000 and the number of misspecified items was two of large magnitude).

Latent-factor group mean difference

The results of latent-factor mean difference tests are presented in Table 3. The table includes relative bias, RMSE, and Type I error and power rates for MIMIC. Because relative bias and RMSE were similar across the three factor mean difference conditions, only those of the large effect size conditions are reported. The *SE* results were very similar to the RMSE results and not included in the table. The results for strict-invariance MGCFA are not included because no notable difference between strict-invariance MGCFA and MIMIC was found. Note that both models were misspecified in terms of the error structure.

As is shown in Table 3, the latent-factor mean difference in MIMIC was estimated accurately with minimal bias, regardless of the error structure misspecifications. The maximum relative bias was – .03, which is less than 5% bias of the population value. In addition, the statistical power and Type I error rates were comparable to those of the correctly specified strong-invariance models. Type I errors were well-controlled, ranging from .05 to .07 across conditions, and power reached 1.00 when the effect size was large. As expected, larger effect size and sample size resulted in higher power across conditions.

Sensitivity of model fit indices

We computed sensitivity by taking the proportions of replications in which the fitted model was flagged as having poor model fit. Thus, a sensitivity rate close to 1.00 indicates that the model fit index was sensitive to the misspecification in the error structure. Table 4 shows sensitivity of the model fit indices for strict-invariance MGCFA and MIMIC across conditions.

Three patterns emerged from the sensitivity rates (see Table 4). In general, the model fit indices for MIMIC showed less sensitivity than did those for MGCFA. In addition, for both strict-invariance MGCFA and MIMIC, the model fit indices were more sensitive to misspecification due to heterogeneous error covariances than to heterogeneous error variances. Among the model fit indices, CFI and SRMR were less sensitive than the RMSEA and chi-square tests. When heterogeneous error variances were present, the RMSEA and chi-square of strict-invariance MGCFA were more sensitive to the misspecification as the size and number of error misspecifications increased, whereas MIMIC consistently showed a good model fit across conditions, and the sensitivity

Table 2 Rejection rates of metric and scalar invariance under error structure heterogeneity

			MGCFA (LR test)				MIMIC (SB Test)				MIMIC (Wald Test)			
			Hetero Var.		Hetero Cov.		Hetero Var.		Hetero Cov.		Hetero Var.		Hetero Cov.	
Size	Num	*N*	Metric	Scalar	Metric	Scalar	Metric	Scalar	Metric	Scalar	Metric	Scalar	Metric	Scalar
Small	One	100	.06	.05	.06	.05	.06	.00	.07	.00	.06	.03	.06	.04
		200	.05	.05	.08	.05	.06	.00	.07	.00	.06	.05	.06	.05
		500	.05	.04	.09	.04	.06	.00	.06	.00	.05	.04	.05	.04
		1,000	.05	.05	.17	.05	.07	.00	.09	.00	.07	.05	.09	.05
	Two	100	.05	.05	.15	.05	.06	.00	.07	.01	.06	.03	.08	.03
		200	.05	.05	.22	.06	.07	.00	.07	.00	.06	.05	.08	.05
		500	.05	.05	.52	.05	.07	.00	.13	.00	.05	.04	.11	.04
		1,000	.05	.04	.85	.05	.08	.00	.22	.02	.07	.05	.20	.05
Large	One	100	.05	.05	.19	.05	.07	.01	.09	.01	.06	.03	.07	.04
		200	.05	.05	.24	.06	.07	.01	.10	.00	.06	.05	.09	.05
		500	.05	.04	.43	.04	.07	.00	.13	.01	.06	.04	.10	.04
		1,000	.05	.05	.72	.06	.09	.01	.30	.03	.09	.05	.22	.06
	Two	100	.05	.05	.98	.07	.07	.00	.15	.01	.06	.03	.12	.03
		200	.05	.05	1.00	.08	.08	.01	.22	.02	.07	.05	.19	.05
		500	.04	.04	1.00	.07	.09	.00	.52	.05	.07	.04	.47	.05
		1,000	.05	.04	1.00	.12	.13	.01	.85	.26	.11	.05	.82	.08

MGCFA = multiplegroup confirmatory factor analysis, MIMIC = multiple-indicator multiple-causes model, LR test = likelihood ratio test, SB test = Satorra–Bentler corrected LR test, Hetero Var = heterogeneous error variance between groups, Hetero Cov = heterogeneous error covariance between groups, Size = size of error structure heterogeneity, Num = number of heterogeneous error variance–covariance, *N* = sample size per group

Table 3 Accuracy of the estimated latent-factor difference and the corresponding power and Type I error for MIMIC

			Heterogeneous error variances					Heterogeneous error covariances				
Size	Num	*N*	Bias	RMSE	Power .5	Power .1	TI	Bias	RMSE	Power .5	Power .1	TI
Small	One	100	.00	.15	.92	.12	.07	.00	.15	.92	.12	.07
		200	.00	.11	1.00	.17	.05	.00	.11	1.00	.17	.05
		500	.00	.07	1.00	.35	.06	.00	.07	1.00	.34	.06
		1,000	.00	.05	1.00	.59	.06	.00	.05	1.00	.59	.06
	Two	100	.00	.15	.92	.12	.07	– .01	.15	.92	.12	.06
		200	.00	.11	1.00	.17	.05	– .01	.10	1.00	.16	.05
		500	.00	.07	1.00	.35	.06	– .01	.06	1.00	.34	.06
		1,000	.00	.05	1.00	.59	.06	– .01	.05	1.00	.58	.07
Large	One	100	.00	.15	.92	.12	.06	.00	.15	.92	.12	.07
		200	.00	.11	1.00	.17	.05	.00	.11	1.00	.17	.05
		500	.00	.07	1.00	.34	.07	.00	.07	1.00	.34	.07
		1,000	.00	.05	1.00	.59	.07	.00	.05	1.00	.59	.07
	Two	100	.00	.15	.92	.12	.07	– .03	.15	.92	.12	.07
		200	.00	.11	1.00	.17	.05	– .03	.11	1.00	.16	.05
		500	.00	.07	1.00	.34	.07	– .03	.07	1.00	.33	.07
		1,000	.00	.05	1.00	.59	.07	– .03	.05	1.00	.57	.06

MIMIC = multiple-indicators multiple-causes model, Size = size of error structure heterogeneity, Num = number of heterogeneous error variance–covariance, *N* = sample size per group, Bias = relative bias, RMSE = root mean squared error, TI = Type I error, Power .1 = power under the effect size .1 conditions, Power .5 = power under the effect size .5 conditions. Bias and RMSE under the effect size .5 conditions are presented in this table

Table 4 Sensitivity of the model fit indices of MGCFA and MIMIC

			Heterogeneous error variances								Heterogeneous error covariances							
			MGCFA (*df* = 34)				MIMIC (*df* = 14)				MGCFA (*df* = 34)				MIMIC (*df* = 14)			
		N	RMSEA	CFI	SRMR	Chisq	RMSEA	CFI	SRMR	Chisq	RMSEA	CFI	SRMR	Chisq	RMSEA	CFI	SRMR	Chisq
Small	One	100	.15	.01	.07	.11	.05	.00	.00	.06	.47	.08	.06	.38	.22	.00	.00	.24
		200	.03	.00	.00	.14	.00	.00	.00	.05	.42	.01	.00	.73	.12	.00	.00	.50
		500	.00	.00	.00	.39	.00	.00	.00	.05	.43	.00	.00	1.00	.03	.00	.00	.93
		1,000	.00	.00	.00	.79	.00	.00	.00	.07	.38	.00	.00	1.00	.00	.00	.00	1.00
	Two	100	.22	.02	.13	.17	.05	.00	.00	.06	.94	.54	.12	.92	.56	.04	.00	.58
		200	.08	.00	.00	.30	.00	.00	.00	.05	.99	.52	.00	1.00	.62	.00	.00	.91
		500	.01	.00	.00	.75	.00	.00	.00	.05	1.00	.59	.00	1.00	.73	.00	.00	1.00
		1,000	.00	.00	.00	.99	.00	.00	.00	.07	1.00	.62	.00	1.00	.80	.00	.00	1.00
Large	One	100	.31	.04	.23	.24	.05	.00	.00	.06	1.00	.99	.34	1.00	.85	.21	.00	.87
		200	.17	.00	.01	.46	.00	.00	.00	.05	1.00	1.00	.01	1.00	.96	.12	.00	1.00
		500	.05	.00	.00	.95	.00	.00	.00	.05	1.00	1.00	.00	1.00	1.00	.04	.00	1.00
		1,000	.01	.00	.00	1.00	.00	.00	.00	.07	1.00	1.00	.00	1.00	1.00	.01	.00	1.00
	Two	100	.60	.18	.54	.51	.05	.00	.00	.06	1.00	1.00	.88	1.00	1.00	.88	.00	1.00
		200	.59	.06	.15	.86	.00	.00	.00	.05	1.00	1.00	.54	1.00	1.00	.97	.00	1.00
		500	.71	.01	.01	1.00	.00	.00	.00	.05	1.00	1.00	.08	1.00	1.00	1.00	.00	1.00
		1,000	.83	.00	.00	1.00	.00	.00	.00	.07	1.00	1.00	.01	1.00	1.00	1.00	.00	1.00

MGCFA = multigroup confirmatory factor analysis; MIMIC = multiple-indicators multiple-causes; *df* = degrees of freedom; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; Chisq = chi-square fit statistic. The sensitivity is defined as the proportion of replications in which model fit indices indicated poor fit. Poor fit criteria for model fit indices: RMSEA ≥ .06, CFI ≤ .95, SRMR ≥ .08, and *p* value of χ^2 ≤ .05

rates were very similar to those of the correctly specified model. When error covariance in one group was ignored, the RMSEA and chi-square tests generally detected the model misspecification, showing higher sensitivity to larger misspecification. On the other hand, CFI only showed high sensitivity rates for the conditions in which the size and number of error misspecifications were large and two, respectively, and SRMR often failed to detect the misspecification, showing good fit across conditions. Especially, both CFI and SRMR in MIMIC were insensitive to the heterogeneity in error variance–covariance when the size and number of error misspecifications were small and one, respectively.

In addition, we also investigated the model fit results for the configural-invariance MGCFA model in this study. Because a heterogeneous covariance structure violates configural invariance due to the unmeasured factor structure in one group, it is worthwhile to examine the model fit of the configural-invariance model when a covariance error structure is present. Similar to in Table 4, RMSEA and chi-square, overall, showed very sensitive results to the heterogeneous covariance structure, whereas CFI and SRMR showed somewhat mixed results. For example, the sensitivity values ranged from .69 to 1.00 for RMSEA, and from .49 to 1.00 for chi-square across simulation conditions. CFI and SRMR, however, only showed sensitive results for conditions in which the size and number of heterogeneous covariances were large and two, respectively (e.g., the sensitivity rates ranged from .88 to 1 for CFI and from .97 to 1 for SRMR). For the other conditions, the sensitivity rates decreased substantially (e.g., the sensitivity rates ranged from .00 to .21 for CFI and from .00 to .11 for SRMR).

Discussion

MGCFA and MIMIC are two widely used models for testing measurement invariance and factor mean differences in applied research. Although a number of studies have investigated the efficacy of MGCFA and MIMIC for testing measurement invariance and factor mean differences, little research has been devoted to investigating the impact of misspecification of the error structure. In this study, therefore, we examined the impact of such violations on measurement invariance testing. We also examined the accuracy of latent-factor mean estimation, inference, and sensitivity of the model fit indices for MGCFA and MIMIC when the invariance assumption was violated for error variance–covariance across multiple groups.

Our key findings in this study were as follows. First, misspecification of the error structure—that is, misspecifying heterogeneous error variances and covariances as invariant—

did not make a substantial impact on the estimation and statistical inference of parameter estimates in the mean structure (i.e., factor means and intercepts). Especially, we observed a minimal impact of such misspecification on factor mean difference estimation and testing. Although the fitted models were misspecified for the heterogeneous error variances or covariances, both MGCFA and MIMIC accurately estimated the latent-factor mean difference between the reference and focal groups. Also, statistical inferences such as the power and Type I error for the group mean difference were not affected. Note that we examined conditions in which the population latent-factor mean difference was varied (0, .1, and .5), and no substantial differences were found across conditions. As expected, statistical power increased as the latent-factor mean difference increased. These findings imply that misspecification of the error structures between groups is not of great concern when researchers are estimating and testing latent mean differences using MIMIC or MGCFA.[2]

In addition, we observed the impact of the misspecified error structure on measurement invariance testing. This impact was more evident for metric-invariance testing when error covariance in one group was ignored. The rejection rates reached 100% when the sample size was large and the size of two error covariances was large (.40). The high rejection of metric invariance might have happened because the heterogeneity in error variance and covariance transferred to the factor loading difference when the error structure was constrained to be equal, which appeared as metric noninvariance with little impact on the mean structure (intercepts and means). Although factor loadings were generated to be equal in the population, we observed that the factor loadings were estimated strikingly differently between groups in the configural-invariance model when the error structure was forced to be homogeneous, which could be misinterpreted as an indication of metric noninvariance in applied research settings.

We also examined the sensitivity of model fit indices to misspecification of the error structure. The model fit indices of MIMIC were less sensitive to the error structure misspecification than were those of MGCFA, in general. For both MGCFA and MIMIC, the model fit indices were generally more sensitive to heterogeneous error covariances rather than variances. Among the model fit indices we examined, CFI and SRMR showed relatively less sensitivity to the error structure misspecification than did chi-square and RMSEA. Particularly, SRMR in MIMIC was completely insensitive and always showed good fit across conditions when the error structure was misspecified. The insensitivity of SRMR in MIMIC was also reported in a previous model-fit statistics study (Kim et al., 2012). This finding suggests that if the heterogeneity of error variances–covariances across groups is of concern, it is recommended to use MGCFA to detect a potential misspecification of the error structure. Beyond detecting the error structure misspecification through model fit evaluation, MGCFA is advantageous because it allows researchers to model heterogeneous error structures between groups, whereas modeling different error structures between groups is not feasible in MIMIC.

We additionally conducted a simulation with conditions in which both variance and covariance were heterogeneous across groups. However, the results were almost identical to those in the heterogeneous covariance conditions, which suggests that the misspecification of heterogeneous error covariances is more influential and also is detected better than misspecification of heterogeneous error variances across groups. In terms of measurement invariance, error covariance in one group is considered as a violation of configural invariance, whereas heterogeneous error variance is a violation of strict invariance. The sensitive model-fit results for the configural invariance model supported this finding as well. This finding implies that the lack of configural invariance in the error structure appears more consequential than is the lack of strict invariance, although both turned out not to be related to latent-factor mean estimation and testing.

To increase the generalizability of the study, we conducted an additional simulation. We considered a condition in which reliability was medium ($\omega = .75$). The medium size of reliability was only considered for conditions in which the size of two error variances or covariances was large, because we observed the most evident result in these conditions. We still varied the level of sample size for the additional simulation. From the simulation, we found no substantial pattern between the two levels of reliability. That is, the error structure misspecification significantly affected the metric invariance test but not the scalar invariance test. The rejection rates increased substantially for metric invariance as the sample size increased for both MGCFA and MIMIC, and this pattern was more evident for the heterogeneous covariance condition. Also, the latent-factor mean estimation and inference were accurate and reliable. These results indicate that the findings of the study can be extended to medium-size reliability measurements.

Practically and theoretically, understanding the error structures and specifying correct error structures between groups is essential in multiple-group analysis for two major reasons, even though misspecification does not directly impact the latent and even the observed group mean comparisons. First, error covariance can be interpreted as the presence of an unmeasured factor. In other words, error covariance in one group suggests that this group may have a different factor structure than the other groups. Understanding the unmeasured factor in one group, either substantive or nuisance, is an essential part

[2] We also found that the observed group mean difference was unbiased. The observed group means were computed with the mean composite scores of six observed variables. When we tested the observed group-mean difference using *t* tests, Type I error and power rates were very close to those of MGCFA and MIMIC, showing no impact of heterogeneous error structures on the observed mean comparisons.

of group comparisons. Second, metric invariance is less likely to be supported when the error structure is considerably misspecified between groups. This is problematic because group heterogeneity in error variances–covariances is manifested in factor loadings. For example, in the configural-invariance model, the ignored error covariance in one group was manifested as notable differences in the factor loadings. In practical settings without knowledge of the true population parameters, differences in the factor loadings between groups observed in the configural-invariance model can be misinterpreted as an indication of metric noninvariance rather than of misspecification of the error structure (i.e., violation of configural invariance through error covariance in one group). Problematically, in subsequent metric-invariance testing, metric invariance will possibly be rejected. Moreover, metric invariance is a precondition of scalar invariance; that is, scalar invariance cannot be established without metric invariance. Thus, a group mean comparison may be invalidated because scalar invariance is considered a prerequisite for a valid group mean comparison. That is, even though the misspecification of the error structure does not impact the group mean comparison, applied researchers will be less likely to proceed with the comparison given the violation of metric invariance.

One positive finding is that some model fit indices are sensitive to the error structure misspecification in the configural-invariance model if the size and number of heterogeneity are large, which could lead to applied researchers scrutinizing the source of model misfit. For applied researchers, we recommend thoroughly investigating configural invariance across groups using MGCFA. In this investigation, we recommend using the RMSEA and chi-square model fit indices to uncover possible differences in error structures for multiple groups on the basis of simulation study results. Modification indices could guide researchers to find the source of model misfit when configural invariance is rejected, although future research is called for to investigate the performance of modification indices in detecting error covariance misspecification. Also, if researchers have strong theoretical evidence regarding nonzero error covariance for specific items, we recommend specifying the error covariances using MGCFA and testing for the equality of error structures between the groups. Theoretical consideration will be particularly useful when the size and number of misspecification are small, because the fit indices of the configural-invariance model may not be informative as to the error structure misspecification. Finally, if the research purpose is to test measurement invariance and error covariance is found in only one group, applied researchers can conclude that configural invariance has been violated, and subsequent analyses will not be executed. If testing and estimating the latent group mean difference is of focal interest, we expect a minimal impact of heterogeneous error variance and covariance on the results. However, it should be kept in mind that no statistical impact does not mean that the group mean difference will be theoretically interpretable. Thus, the heterogeneity in the error structure, and particularly error covariance, should be theoretically explained and justified.

There are several limitations to the present study. First, we considered only the one-factor model across simulation conditions. This study was conducted with a relatively simple model, but more complex models including bifactor or two-tier models can be considered in future research. Because more complex models could be more challenging for estimating their parameters, the results might be different in various situations. Second, in this study we did not manipulate different sample sizes for the reference and focal groups (unbalanced design).[3] Given that unbalanced designs are more common in applied educational and psychological settings (e.g., Lubke & Dolan, 2003; Woods, 2009), we recommend investigating the potential impact of unbalanced group sizes being paired with heterogeneous error structures in multiple-group analysis in future studies. Third, although we found that misspecification of the error structure minimally affects the latent-factor mean estimates and inference, the simulation study design was limited, and the results should not be generalized to all situations. More studies should be conducted to explore the impact of the error structure on MGCFA and MIMIC with different simulation conditions. For example, it would be interesting to investigate the impact of error structure misspecification with categorical observed variables. Future research should conduct a study investigating how error structure misspecification can affect measurement invariance tests and the latent mean estimates with the categorical observed variables using weighted least squares means and variance-adjusted estimation, given that categorical variables are prevalent in applied settings.

In conclusion, we investigated the impact of misspecified heterogeneous error variances–covariances in multiple-group analysis, including measurement invariance testing and latent-factor mean difference testing. We found that misspecification in the error variance–covariance structure affects testing and estimating the variance–covariance structure (i.e., factor loadings). In contrast, little impact was observed on the mean structure (i.e., intercepts and latent-factor means). Thus, we found that both the MGCFA and MIMIC approaches robustly estimated the latent-factor mean difference and yielded correct statistical inferences under the misspecification of error variance–covariance between groups. Strict invariance was also mostly supported when it was true in the population. However, metric invariance could be rejected under such

[3] In a preliminary analysis of the simulation, we did not observe any remarkable difference between balanced and unbalanced group size conditions in terms of factor mean difference testing and estimation, and so we did not include unbalanced conditions in the present simulation. However, full investigation of unbalanced designs across different research settings and with various simulation outcomes will be needed.

misspecification. Thus, we suggest that MGCFA be preferred when heterogeneous error structures are of concern, because MGCFA allows for modeling heterogeneous error structures across groups. It should also be kept in mind that the model fit indices of MIMIC are not generally sensitive to misspecified error structures.

References

Ainsworth, A. T. (2008). Dimensionality and invariance: Assessing differential item functioning using bifactor multiple indicator multiple cause models. *Dissertation Abstracts International*, *68*, 6383B.

Barendse, M. T., Oort, F. J., & Garst, G. J. A. (2010). Using restricted factor analysis with latent moderated structures to detect uniform and nonuniform measurement bias: A simulation study. *Advances in Statistical Analysis*, *94*, 117–127.

Barendse, M. T., Oort, F. J., Werner, C. S., Ligtvoet, R., & Schermelleh-Engel, K. (2012). Measurement bias detection through factor analysis. *Structural Equation Modeling*, *19*, 561–579.

DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science*, *46*, 137–149.

Fleishman, J., Spector, W., & Altman, B. (2002). Impact of differential item functioning on age and gender differences in functional disability. *Journals of Gerontology: Series B*, *57*, S275–S284.

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling*, *7*, 251–270.

Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM fit indexes with respect to violations of uncorrelated errors. *Structural Equation Modeling*, *19*, 36–50.

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

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus. *Structural Equation Modeling*, *6*, 1–55. https://doi.org/10.1080/10705519909540118

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*, 631–639.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8 user's guide*. Chicago, IL: Scientific Software International.

Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016). Multilevel factor analysis: Reporting guidelines and a review of reporting practices. *Multivariate Behavioral Research*, *51*, 881–898. https://doi.org/10.1080/00273171.2016.1228042

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, *72*, 469–492.

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

Lubke, G. H., & Dolan, C. V. (2003). Can unequal residual variances across groups mask differences in residual means in the common factor model? *Structural Equation Modeling*, *10*, 175–192.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence*, *31*, 543–566.

Marsh, H. W. (1994). Confirmatory factor models of factorial invariance: A multi-faceted approach. *Structural Equation Modeling*, *1*, 5–34.

Marsh, H. W., Nagengast, B., & Morin, A. J. (2013). Measurement invariance of Big-Five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology*, *49*, 1194.

McArdle, J. J. (1998). Contemporary statistical models of test bias. In J. J. McArdle & R. W. Woodcock (Eds.), *Human abilities in theory and practice* (pp. 157–195). Mahwah, NJ: Erlbaum.

McCarthy, D. M., Pedersen, S. L., & D'Amico, E. J. (2009). Analysis of item response and differential item functioning of alcohol expectancies in middle school youths. *Psychological Assessment*, *21*, 444–449.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143.

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

Muthén, B. O., Kao, C., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, *28*, 1–22.

Muthén, B. O., & Muthén, L. K. (2012). Mplus (Version 7) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Raykov, T. (2001). Bias of coefficient *a* for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, *25*, 69–76.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., & Rauch, S. M. (2003). Validating a measure across groups: The use of MIMIC models in scale development. *Journal of Social Service Research*, *29*, 53–67.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507–514.

Shevlin, M., Miles, J. N. V., Davies, M. N. O., & Walker, S. (2000). Coefficient alpha: A useful indicator of reliability? *Personality and Individual Differences*, *28*, 229–237.

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.

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

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*, 1–27.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, *35*, 339–361.

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2009). Illustration of MIMIC-model DIF testing with the schedule for nonadaptive and adaptive personality. *Journal of Psychopathology and Behavioral Assessment*, *31*, 320–330.