Educational and Psychological Measurement
http://epm.sagepub.com/

Testing Measurement Invariance Using MIMIC: Likelihood Ratio Test With a Critical Value Adjustment

Eun Sook Kim, Myeongsun Yoon and Taehun Lee

Educational and Psychological Measurement published online 6 December 2011
DOI: 10.1177/0013164411427395

Educational and Psychological
Measurement
XX(X) 1–24
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164411427395
http://epm.sagepub.com

Testing Measurement Invariance Using MIMIC: Likelihood Ratio Test With a Critical Value Adjustment

Eun Sook Kim[1], Myeongsun Yoon[2], and Taehun Lee[3]

Abstract

Multiple-indicators multiple-causes (MIMIC) modeling is often used to test a latent group mean difference while assuming the equivalence of factor loadings and intercepts over groups. However, this study demonstrated that MIMIC was insensitive to the presence of factor loading noninvariance, which implies that factor loading invariance should be tested through other measurement invariance testing techniques. MIMIC modeling is also used for measurement invariance testing by allowing a direct path from a grouping covariate to each observed variable. This simulation study with both continuous and categorical variables investigated the performance of MIMIC in detecting noninvariant variables under various study conditions and showed that the likelihood ratio test of MIMIC with Oort adjustment not only controlled Type I error rates below the nominal level but also maintained high power across study conditions.

Keywords

MIMIC, measurement invariance, differential item functioning, measurement bias

With increasing attention to measurement bias across groups, testing measurement invariance has become a common practice before using a measure in social science. Measurement invariance holds when individuals have identical probabilities to exhibit

[1]University of South Florida, Tampa, USA
[2]Texas A&M University, College Station, USA
[3]University of California, Los Angeles, USA

Corresponding Author:
Eun Sook Kim, Department of Educational Measurement and Research, University of South Florida, 4202 E. Fowler Avenue, EDU 105, Tampa, FL 33620-7750, USA
Email: ekim3@usf.edu

the observed outcomes given the common factor irrespective of the levels of a variable other than the common factor (Barendse, Oort, & Garst, 2010; Mellenbergh, 1989; Meredith, 1993; Millsap & Yun-Tein, 2004):

$$P(Y_{ij} = y \mid \eta_i, G) = P(Y_{ij} = y \mid \eta_i). \tag{1}$$

Given the factor score (η_i), the conditional probability of the response of the *i*th person on the *j*th variable (Y_{ij}) is independent of a variable *G*. With the interest in a group difference, *G* often indicates group membership, although *G* could be any variable of interest.

Among numerous methods of measurement invariance testing (Millsap & Everson, 1993), item response theory (IRT) has long been used to identify the items of measurement noninvariance, which is specifically called differential item functioning (DIF). Measurement invariance has also been assessed with the techniques under structural equation modeling (SEM). Although IRT and SEM approaches to measurement invariance testing were developed separately, both methods are closely related to each other and, in fact, the parameters of IRT (specifically, the two-parameter logistic model) can be easily converted to SEM parameters (Lord & Novick, 1968; Takane & de Leeuw, 1987; Wirth & Edwards, 2007). A large volume of literature is devoted to the comparison of IRT and SEM with regard to measurement invariance testing (e.g., Meade & Lautenschlager, 2004; Raju, Laffitte, & Byrne, 2002; Stark, Chernyshenko, & Drasgow, 2006; Willse & Goodman, 2008). Under SEM, along with multiple group confirmatory factor analysis (CFA), multiple-indicators multiple-causes (MIMIC; Joreskog & Goldberger, 1975) modeling has often been employed for measurement invariance testing (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). With the increasing interest in measurement invariance and the use of MIMIC modeling to test measurement invariance, the purpose of this study is to investigate the behaviors of MIMIC modeling with both continuous and categorical data under various simulation conditions.

MIMIC Modeling for Continuous Variables

The MIMIC model, in general, allows causal indicators of factors as well as effect indicators. For measurement invariance testing across groups, the MIMIC model includes dummy-coded grouping variables (X_i) as causal indicators (Kaplan, 2009; Thompson & Green, 2006). For the simplicity of discussion, we employed a single causal indicator for two groups (focal and reference groups).

$$Y_{ij} = \lambda_j \eta_i + \varepsilon_{ij}, \tag{2.1}$$

$$\eta_i = \gamma X_i + \zeta_i, \tag{2.2}$$

where the observed score of an individual i for a variable j, Y_{ij} is related to the common factor, η_i with the factor loading of the variable, λ_j. The unique factor scores or residuals are denoted by ε_{ij}. Equation (2.1) represents the relationship between an observed variable and the latent factors. In Equation (2.2), X_i denotes a dummy variable indicating group membership, γ is the path coefficient of the grouping variable on the latent factor, and ζ_i is the disturbance of the latent factor (see Figure 1 without a dotted line). Because the expected value of the disturbance of the latent factor equals zero, the expected value of the latent factor is expressed as

$$E(\eta_i) = \gamma E(X_i). \tag{3}$$

Therefore, with a dummy-coded grouping variable (X_i) γ represents the group difference in latent factor means (Thompson & Green, 2006). That is, the latent factor mean of the focal group ($X_i = 1$) is γ units higher (or lower) than that of the reference group ($X_i = 0$).

MIMIC Modeling for Ordered-Categorical Variables

When the variable of concern (Y_{ij}) is ordered-categorical (e.g., dichotomous or polytomous), Y_{ij} is construed as the manifestation of the underlying latent variable (Y_{ij}^*) that is inherently continuous and multivariate-normally distributed. The latent response variate Y_{ij}^* is related to the latent factor (η) in the same way as continuous variables are

$$Y_{ij}^* = \lambda_j \eta_i + \varepsilon_{ij}, \tag{4.1}$$

$$\eta_i = \gamma X_i + \zeta_i. \tag{4.2}$$

The relationship between the observed categorical responses and the latent response variates is expressed as follows with the threshold structure:

$$Y_{ij} = c, \quad \text{IF } \nu_{jc} < Y_{ij}^* \leq \nu_{j(c+1)}, \tag{5}$$

where ν_{jc} indicates the threshold of the jth item with C ordered-categorical responses ($\nu_{j0} = -\infty$; $\nu_{j(c+1)} = \infty$; $c = 0, 1, \ldots, C - 1$). When the number of response categories is C, $C - 1$ thresholds are determined. For example, with four possible response categories (0, 1, 2, 3), the number of thresholds is three. Any latent response variate (Y_{ij}^*) that falls between the threshold for a response category (c) and the threshold for the next higher response category ($c + 1$) is manifested as a response category, c. In other words, when a response variate meets the threshold for a response category 2 but does not exceed the threshold for a response category 3, the observed score will be 2. To test measurement invariance and the latent group mean difference, a dummy-coded grouping variable X_i is introduced as a causal indicator of the latent factor η as in the continuous model (see Equation 4.2). The MIMIC model that incorporates the threshold structure with latent response variates is illustrated in Figure 2.

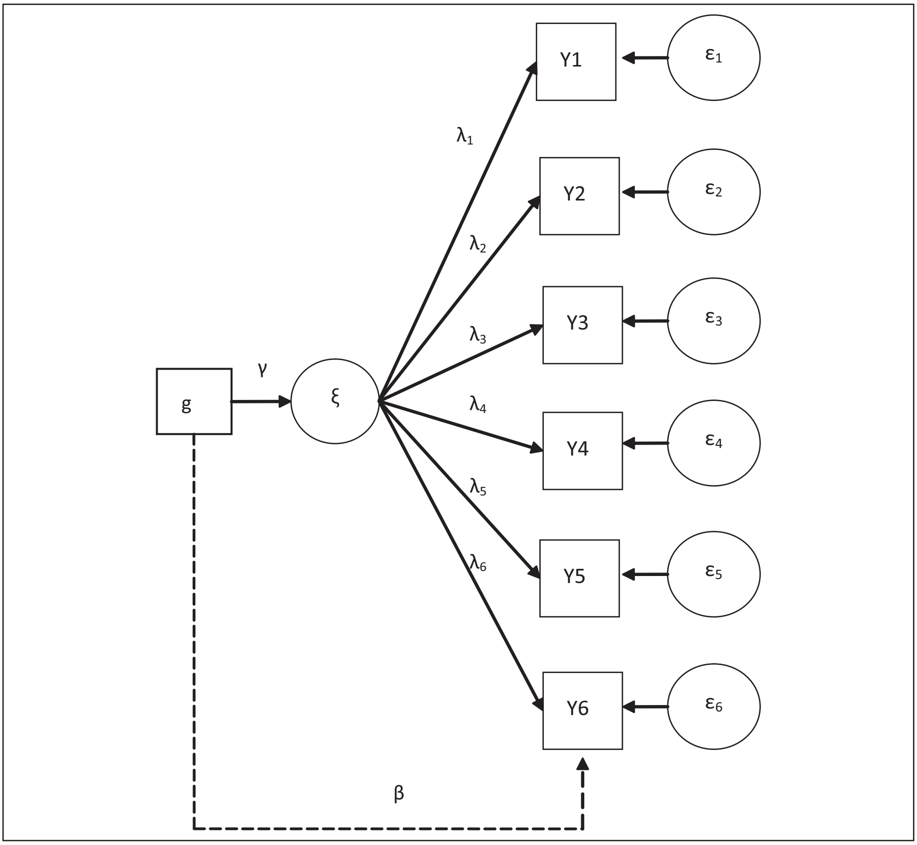

Figure 1. The MIMIC model with a grouping variable as a covariate for continuous data. *g* denotes a grouping variable

Measurement Invariance Testing in MIMIC Modeling

To identify a noninvariant variable, a direct path from a grouping variable to an observed variable is tested in the model (in Figures 1 and 2 with a dotted line for continuous and categorical variables, respectively). The model with the direct path from the grouping variable to the measured variable can be rewritten as

$$Y_{ij} = \lambda_j \eta_i + \beta_j X_i + \varepsilon_{ij},$$
$$\eta_i = \gamma X_i + \zeta_i, \tag{6}$$

where β_j is a path coefficient of the grouping variable in relation to the *j*th observed variable (Finch, 2005; Kaplan, 2009). The β_j coefficient represents the group effect on an observed variable controlling for the effect of the latent factor. Thus, this model allows the statistical significance test on the group difference of an observed variable

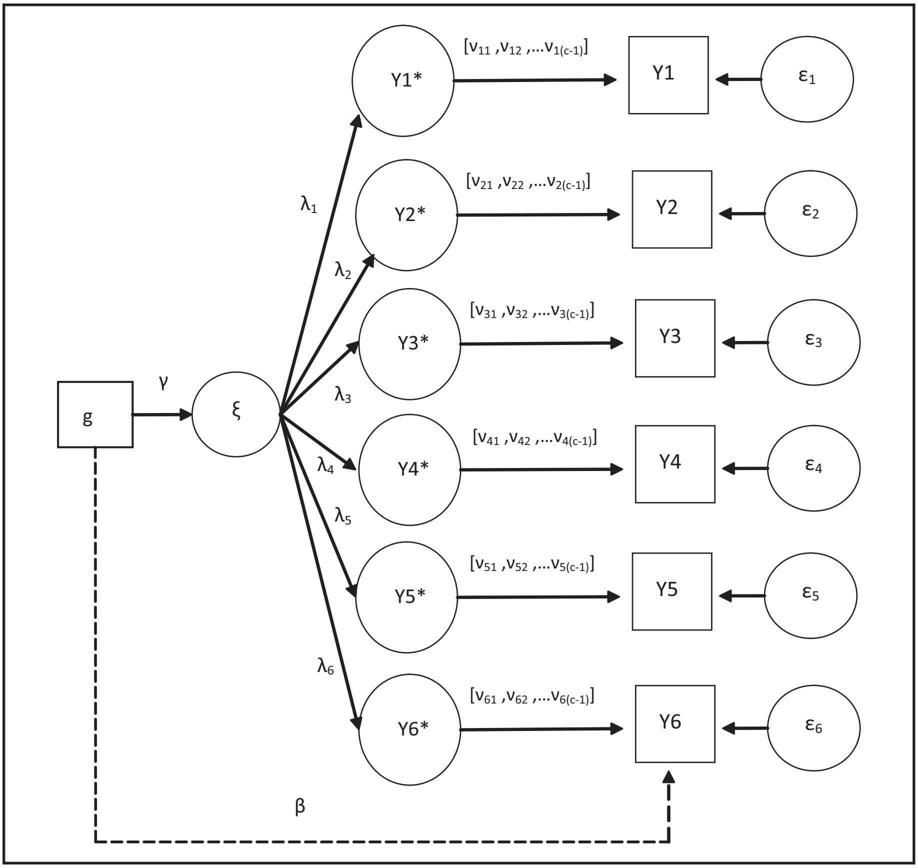

Figure 2. The MIMIC model with a grouping variable as a covariate for categorical data
Note. g denotes a grouping variable. *c* is the number of response categories of an item.

(β_j) as well as on the group difference of the latent means (γ). Otherwise stated, the statistically significant β_j coefficient indicates the violation of measurement invariance across groups (*noninvariance* of the intercept) of the *j*th variable.

Strengths of MIMIC Modeling

MIMIC modeling allows the assessment of measurement invariance and latent mean difference across groups by incorporating grouping variables as covariates instead of testing separate models for each group as in multiple-group CFA. Thus, MIMIC modeling can easily facilitate measurement invariance tests on multiple background variables and their interactions (e.g., gender, race, and gender by race) as well as on more than two groups per grouping variable of interest (e.g., four different race groups;

Ainsworth, 2008; Fleishman et al., 2002). Of note is that the covariates (X_i in Equation 6) of which researchers purport to test invariance across levels can be any types of variables (e.g., continuous or categorical), which is another advantage of MIMIC modeling over multiple group CFA (Barendse et al., 2010).

MIMIC Testing Uniform Measurement Noninvariance

MIMIC modeling has a couple of downsides that MIMIC users for measurement invariance testing should be aware of. First, MIMIC modeling, developed by Joreskog and Goldberger (1975) and disseminated for measurement invariance testing by Muthén (e.g., 1989), tests uniform measurement noninvariance (Woods & Grimm, 2011). That is, factor loadings are assumed to be invariant across groups and are not explicitly tested for invariance. In Equation (6), the path coefficient, β_j, tests group invariance controlling for the latent factor effects ($\lambda_j \eta_i$) assuming the factor loading is invariant across groups (i.e., estimating only one set of factor loadings across groups, not for each group). Considering that this assumption of factor loading invariance is often violated in practice, the equivalence of factor loadings should be tested rather than simply assumed as in the current MIMIC modeling. Hence, when there is lack of invariance in factor loadings, the performance of MIMIC modeling is of question. This study inspected the behaviors of MIMIC modeling under the noninvariance of factor loadings in addition to other sources of noninvariance.

It should be noted that MIMIC model with the interaction between the latent factor (η_i) and the group membership indicator (X_i) allows researchers to test nonuniform measurement bias (Barendse et al., 2010; Barendse, Oort, Werner, Ligtvoet, & Schermelleh-Engel, 2011; Woods & Grimm, 2011). However, this study limits the scope to uniform MIMIC modeling because MIMIC as a test for uniform noninvariance is currently used among applied researchers.

Type I Error Inflation in Measurement Invariance Testing With MIMIC

Type I error inflation in measurement invariance testing. Previous simulation studies on MIMIC as a measurement invariance test consistently reported high Type I error rates over the nominal level (e.g., Finch, 2005). In measurement invariance literature, Type I error and false positive are interchangeably used generally indicating the false detection of invariant variables as DIF. The proportion of false-positive cases across simulation replications is often defined as a Type I error rate or false-positive rate. With the Type I error inflation, invariant variables could be overly detected as DIF when MIMIC is used for measurement invariance testing, which likely blemishes the adequacy of MIMIC as a measurement invariance testing technique.

A number of simulation studies reported high Type I error rates when MIMIC modeling was used to detect noninvariant variables. Oort (1998) studied restricted factor analysis (RFA), and reported Type I error rates between 0.15 and 0.20. RFA is equivalent to MIMIC except that correlation between a factor and a covariate (η_i and

X_i, respectively, in Equation 6) is specified in RFA instead of a causal effect (Barendse et al., 2010). In Finch's (2005) study using MIMIC modeling, Type I error rates ranged from .08 to .22 (mean of .12) depending on the simulation conditions. Navas-Ara and Gomez-Benito (2002) reported a Type I error rate of .36. In the study of Wang, Shih, and Yang (2009), MIMIC modeling showed the false-positive rates as high as .48.

Approaches to Type I error inflation of MIMIC. With the report of Type I error inflation, a body of literature contributed to explain and control for the Type I error inflation. Navas-Ara and Gomez-Benito (2002) used scale purification with categorical items comparing six different DIF detection techniques (e.g., RFA, IRT-based indices) and reported the improvement of Type I error rates (.07). The scale purification is an iterative process in which biased items detected in the initial analysis are eliminated, and the bias detection procedure is repeated with unbiased items to identify remaining bias until no item is detected as noninvariance. When Wang et al. (2009) applied scale purification procedures to MIMIC modeling, MIMIC with scale purification yielded lower Type I error rates (less than .10 for most study conditions) compared with standard MIMIC although the Type I error rate of MIMIC with scale purification was still high (e.g., 0.24) in the conditions of the 40% DIF contamination. In case of the likelihood ratio (LR) test, Stark et al. (2006) suggested the Bonferroni correction of critical values, pointing out the chi-square statistic inflation in a misspecified baseline model and subsequently Type I error rate elevation. On the other hand, Oort (1992, 1998) developed a formula to adjust the critical value to control the chi-square inflation in the use of modification indices. When the adjustment was applied to the iterative procedures using modification indices, the Type I error rates were reported under the nominal level.

Oort adjustment to control the Type I error inflation. For a statistical strategy to control the Type I error inflation, we adopted Oort adjustment in the LR test. Oort (1992, 1998) developed a formula originally to adjust a modification index (MI) taking into account practically inevitable model misspecification errors that likely render unreliable chi-square distribution. We applied the Oort adjustment formula in the LR test considering that the MI is comparable with the chi-square difference with one degree of freedom in the LR test because the MI is an approximation of the change in chi-square when a constrained parameter in the original model is freely estimated (Brown, 2006). In the LR test with two nested models, a model in which a variable or a set of parameters (e.g., factor loadings of all variables) are freely estimated is compared with a baseline model with invariance constraints. The statistical significance of the chi-square fit difference between two rival models indicates the lack of invariance on the tested variable or the tested set of parameters (e.g., the violation of factor loading invariance or weak invariance).

However, as Oort (1992, 1998) stated, chi-square difference is possibly inflated with unavoidable model misspecification errors, which in turn leads to Type I error inflation in the LR tests. Accordingly, Oort adjusted a chi-square critical value by incorporating the chi-square and degrees of freedom of a baseline model:

$$K' = (\frac{\chi^2{}_0}{K + df_0 - 1}) * K, \tag{7}$$

where K' is the adjusted critical value, K is the original critical value chosen from the chi-square distribution given the degree-of-freedom difference between models, χ_0^2 is the chi-square value of a baseline model, and df_0 is the corresponding degrees of freedom. From the speculations that the inflation of chi-square difference between the full-invariance MIMIC model (i.e., baseline model) and the relaxed model testing DIF (i.e., augmented model) plausibly results in Type I error inflation, this simulation study reevaluated the Oort adjustment in the LR test using MIMIC modeling. The Oort adjustment was compared with the Bonferroni correction. Stark et al. (2006) suggested Bonferroni correction to lower the inflated Type I error rates in the LR tests. French and Finch (2008) applied the Bonferroni correction in their simulation study locating invariant reference variables in multiple-group CFA.

To sum up, cognizant of the limitations of MIMIC as a uniform measurement invariance test, we explored the overall performance of MIMIC in detecting DIF with a focus on statistical strategies to control the Type I error inflation under a variety of research situations, including different data types and different locations of noninvariance through Monte Carlo simulations.

Method

Simulation Conditions

Simulation conditions included data type (continuous, dichotomous, or polytomous), location of noninvariance (factor loading, intercept, or both for continuous data and factor loading, threshold, or both for categorical data), magnitude of noninvariance (small or large), number of noninvariant items (one or two out of six), and sample size (200, 400, 1,000, and 2,000). The total 144 ($3 \times 3 \times 2 \times 2 \times 4$) conditions were included in the simulation. In addition, full invariance conditions (data type by sample size) were simulated to establish a baseline of the study. For each condition, 500 replications were generated.

Data type. The performance of MIMIC modeling for continuous variables was compared with that for categorical variables. Dichotomous variables have a single threshold with two response categories, whereas polytomous variables in this study take five ordered response categories that yield four thresholds.

Location of noninvariance. The location of noninvariance varies to factor loadings only, intercepts/thresholds only, or both. In the previous studies on MIMIC modeling in testing uniform measurement bias, the source of noninvariance was not considered as a simulation condition. However, this study purported to examine the behaviors of MIMIC with different sources of noninvariance in the model, including factor loading noninvariance.

Magnitude of noninvariance. The magnitude of noninvariance was manipulated with a small or large difference. For the factor loading noninvariance, 0.2 and 0.4 were subtracted from the factor loadings of the reference group for small and large effect sizes, respectively. In terms of intercept or threshold noninvariance, approximately 0.3 for small difference and 0.6 for large difference were added to the intercepts or thresholds of the reference group.

Number of noninvariant items. Concerning the number of noninvariant items, two conditions of noninvariance contamination were simulated: only one noninvariant variable (about 17% contamination) and two noninvariant variables (about 33% contamination) out of six variables. The noninvariance contamination was less than 50% because it is more likely that the majority of variables are invariant across groups. *Y*5 was simulated as a noninvariant variable, and *Y*2 was added as a noninvariant item for the two-DIF conditions. This study also included the condition in which all six variables were invariant across groups to establish basal Type I error rates.

Sample size and group size. Two balanced groups with size 100, 200, 500, and 1,000 each were examined in this study. Although in many research settings two groups may be disproportionate (e.g., 90% Caucasian and 10% African American), studies in which two group sizes are roughly equal are not uncommon (e.g., boys and girls, primary school students and secondary school students, etc.). Woods (2009) studied an optimal sample size for MIMIC modeling in the detection of DIF and found that the focal group sample size smaller than 100 (e.g., 25, 50, or 100) yielded very low power in detecting the DIF items. Thus, this study included the minimum sample size as low as 100 per group.

Data Generation

In generating three types of data (continuous, dichotomous, and polytomous variables) for two groups (reference and focal groups), we used Mplus 5.2 (Muthén & Muthén, 2008). Six variables (*Y*1-*Y*6) loaded on a single factor under the unidimensionality assumption.

The parameter values used for the reference group data generation are presented below. The parameters of intercepts (τ) were specified for continuous variables, whereas a set of thresholds (ν) were specified for dichotomous and polytomous variables. The same values of intercept (τ) were used for the thresholds (τ) of dichotomous variables.

Y	λ	τ or ν
*Y*1	.9	–0.15
*Y*2	.7	0.25
*Y*3	.6	0.15
*Y*4	.8	–0.25
*Y*5	.7	–0.10
*Y*6	.6	0.10

For the polytomous data with five response categories, a set of thresholds were added as follows:

Y	ν_1	ν_2	ν_3	ν_4
$Y1$	–0.05	0.35	0.75	1.05
$Y2$	–0.80	–0.40	0.00	0.40
$Y3$	–0.55	–0.05	0.45	0.85
$Y4$	0.05	0.50	0.85	1.15
$Y5$	–0.50	–0.10	0.25	0.65
$Y6$	0.15	0.40	0.70	1.25

The factor mean and variance of the reference group were 0.0 and 1.0, respectively. The corresponding parameters of the focal group were assigned as 0.5 and 1.0, respectively. The residual variances were homogeneous across groups as 0.3. The parameter values, the magnitude of DIF, and sample size were selected with the reference to the previous simulation studies on the similar research conditions (Kim & Yoon, 2011; Meade & Lautenschlager, 2004; Muthén & Asparouhov, 2002; Stark et al., 2006; Yoon, 2008; Yoon & Millsap, 2007).

Fitted Models

To test measurement invariance, two models under the nested condition were constructed according to Equation (2) (or Equation 4) and Equation (6) for the LR test. The MIMIC model with a direct path from the grouping indicator to each variable (for continuous data, an augmented model with β_j as shown in Equation 6 and Figure 1 with the dotted path) is compared with the model in which the corresponding path parameter (β_j) was constrained at zero assuming invariance (a baseline model as presented in Equation 2, i.e., a model without the dotted path in Figure 1). The statistical significance of the chi-square difference given degrees of freedom between two models (in this study, $df = 1$) indicates the direct effect of group membership on the tested variable in favor of the augmented model. In other words, the tested variable is considered *non*invariant over groups. In the LR test, two critical value adjustment strategies were employed and compared with no adjustment conditions. For the Bonferroni correction, critical p value .008 (= .05/6) was adopted because six LR tests were performed for each replication. For Oort adjustment, an adjusted chi-square critical value was computed for each simulation replication with Equation (7).

For model identification, the factor variance was fixed at 1. This identification strategy allows freely estimating the factor loadings of all observed variables instead of constraining one of the factor loadings at 1. For model estimation, we used maximum likelihood for continuous data and weighted least squares with robust mean and variance with theta parameterization for categorical data that are the defaults of the Mplus program, respectively.

Data Analytic Procedures

Sensitivity of model fit indices under measurement noninvariance. To investigate the behaviors of MIMIC modeling under measurement noninvariance, we evaluated a model fit of the baseline model, which is a misspecified MIMC model with one or two items of noninvariance. The purpose of this investigation is to examine the sensitivity of model fit indices to the violation of the strict invariance assumption of MIMIC modeling expressed in Equation (2) (or Equation 4). The presence of noninvariance across groups should lead to the lack of fit of the MIMIC model. We examined a chi-square fit statistic and the following goodness-of-fit indices: (a) the weighted root mean square residual (WRMR) for categorical items and the standardized root mean square residual (SRMR) for continuous items; (b) comparative fit index (CFI); and (c) the root mean square error of approximation (RMSEA). Recommended cutoff values of these goodness-of-fit indices for a good model fit are CFI ≥ .95, RMSEA ≤ .05, SRMR ≤ .05, and WRMR ≤ 1.00 (Browne & Cudeck, 1993; Hu & Bentler, 1999; Yu, 2002) in addition to statistically nonsignificant chi-square ($p \geq .05$). We examined the sensitivity of each fit statistic to the model misspecification due to measurement noninvariance. The sensitivity of a fit index was defined as the proportion of the replications in which the fit index did not meet the cutoff of a good fit. In other words, each fit index off the given range was considered as a correct indication of model misfit under the presence of noninvariant variables.

Power and Type I error rates. We examined power and Type I error rates to explore the performance of MIMIC modeling in detecting noninvariant variables. In this study, the power rate is defined as the proportion of the cases in which the LR test detected the noninvariant item or items correctly over 500 replications. Note that for two-DIF conditions, only when both noninvariant variables were correctly identified as noninvariance, the case was counted for power. For the Type I error rate, the proportion of the cases in which the LR test falsely detected an invariant item as DIF was computed across all invariant items over 500 replications.

Raw bias. In addition to power and Type I error rates, this study examined the raw bias of two-parameter estimates of MIMIC modeling. The parameter estimates of interest are the estimate of latent group mean difference (γ, a direct path from the group indicator to the latent factor) and the DIF estimate (β_{Y5}, a direct path from the group indicator to the DIF item, $Y5$). The bias was investigated only for the correctly specified models (i.e., the augmented model of the LR test in which the effect of group membership X on the noninvariant variable $Y5$ was freely estimated for difference in the one-DIF conditions).

The raw bias of each simulation condition, $B(\theta_c)$ was calculated as

$$B(\theta_c) = R^{-1}\sum\nolimits_{r=1}^{R}(\hat{\theta}_{rc} - \theta_c), \tag{8}$$

where $\hat{\theta}_{rc}$ denotes the parameter estimate for replication r in condition c, θ_c is the population parameter for θ in condition c, and R denotes the total number of replications (i.e., 500 in this study). For the γ estimates, the population parameter (θ_c) was 0.5, which was the difference of factor means between reference and focal groups (reference group factor mean = 0.0; focal group factor mean = 0.5). For the β estimates of *Y*5, the magnitude of DIF of the intercept/threshold was considered as the population parameter because β_j represents the group difference in the intercept or threshold of the *j*th variable (0.3 and 0.6 for small and large DIF, respectively).

Results

Simulation Baseline Check

We established the basal Type I error rate by examining the conditions without noninvariance. When all six variables were invariant across groups, the Type I error rate was evaluated at the critical *p* value of .05. The examined Type I error rates were simply .05, equal to the predetermined significance level across all simulation conditions (i.e., four levels of sample size conditions by three types of data; .04 only for polytomous data with sample size 100 per group). The simulation baseline check supported the adequacy of the simulation.

The Sensitivity of Model Fit Index to Noninvariance

This study examined a set of commonly reported fit indices of MIMIC modeling when noninvariance was present (Table 1). First of all, when factor loadings were noninvariant across groups, none of the studied fit indices could detect the model misspecification. In other words, the fit indices consistently and falsely supported a good fit of the factor loading–noninvariant MIMIC model. The sensitivity was virtually zero for most simulation conditions. Thus, in the report of each fit index below, the conditions of factor loading noninvariance were not included. Because the results of one-DIF conditions were similar to those of two DIFs, we reported the results of two-DIF scenarios only.

Chi-square fit statistic. Chi-square fit statistics showed high sensitivity to the model misspecification due to measurement noninvariance irrespective of simulation conditions. Even for small sample size conditions, chi-square was usually below the critical value rejecting the null hypothesis of a good model fit. The average *p* value of chi-square goodness-of-fit testing was .00 for most conditions.

Comparative fit index. CFI detected the noninvariance only when the magnitude of DIF was large. When DIF was small, CFI was not able to capture the noninvariance between groups with a sensitivity of near zero regardless of sample size. For CFI, the DIF size appeared to be a critical factor in detecting the model misspecification of measurement noninvariance. CFI was more sensitive to the noninvariance in continuous and polytomous data than in dichotomous data. The average CFI was .92 for most large DIF conditions, whereas the average CFI was .97 for most small DIF conditions regardless of sample size. Hence, a more liberal cutoff of a good fit (i.e., CFI >.90) possibly fails to detect a model misfit with noninvariance between groups.

Table 1. The Sensitivity of Fit Indices of MIMIC Models With Two DIF Variables

			Dichotomous				Polytomous				Continuous			
			p of χ^2	CFI	RMSEA	WRMR	p of χ^2	CFI	RMSEA	WRMR	p of χ^2	CFI	RMSEA	SRMR
Both	Small	100	.41	.03	.58	.03	.98	.28	.99	.00	.74	.02	.82	.01
		200	.80	.01	.67	.16	1.00	.37	1.00	.31	.97	.01	.93	.00
		500	1.00	.00	.83	.90	1.00	.37	1.00	1.00	1.00	.00	1.00	.00
		1,000	1.00	.00	.92	1.00	1.00	.45	1.00	1.00	1.00	.00	1.00	.00
	Large	100	.99	.70	.99	.69	1.00	1.00	1.00	.93	1.00	.87	1.00	.83
		200	1.00	.87	1.00	.99	1.00	1.00	1.00	1.00	1.00	.94	1.00	.83
		500	1.00	.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.93
		1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.97
Intercept	Small	100	.31	.01	.47	.01	.64	.00	.77	.00	.83	.04	.91	.02
		200	.67	.00	.52	.08	.95	.00	.87	.00	1.00	.01	.98	.00
		500	1.00	.00	.58	.66	1.00	.00	.99	.20	1.00	.00	1.00	.00
		1,000	1.00	.00	.63	1.00	1.00	.00	1.00	1.00	1.00	.00	1.00	.00
	Large	100	.91	.21	.96	.23	1.00	.63	1.00	.06	1.00	.93	1.00	.73
		200	1.00	.30	.99	.86	1.00	.81	1.00	.87	1.00	.98	1.00	.72
		500	1.00	.28	1.00	1.00	1.00	.98	1.00	1.00	1.00	1.00	1.00	.77
		1,000	1.00	.30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.77
Factor loading	Small	100	.06	.00	.11	.00	.04	.00	.08	.00	.08	.00	.14	.00
		200	.05	.00	.02	.00	.06	.00	.02	.00	.10	.00	.03	.00
		500	.08	.00	.00	.00	.06	.00	.00	.00	.14	.00	.00	.00
		1,000	.12	.00	.00	.00	.12	.00	.00	.00	.29	0.00	.00	.00
	Large	100	.07	.00	.15	.00	.07	.00	.15	.00	.21	.00	.33	.00
		200	.11	.00	.04	.00	.16	.00	.08	.00	.46	.00	.23	.00
		500	.15	.00	.00	.00	.36	.00	.00	.00	.89	.00	.10	.00
		1,000	.42	.00	.00	.02	.72	.00	.00	.00	1.00	.00	.04	.00

Note. MIMIC = multiple-indicators multiple-causes; DIF = differential item functioning; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; WRMR = weighted root mean square residual. "Both," "intercept," and "factor loading" are the location of noninvariance. "Small" and "large" are the magnitude of noninvariance. The sample size per group includes 100, 200, 500, and 1,000. The p of χ^2 is the p value of chi-square fit statistic. The sensitivity is defined as the proportion of cases in which each model fit index did not meet the cutoff of a good fit: p of $\chi^2 \geq .05$. CFI $\geq .95$, RMSEA $\leq .05$, SRMR $\leq .05$, WRMR $\leq .10$.

Root mean square error of approximation. RMSEA on average of 500 replications ranged from .07 to .15, which exceeded the cutoff of a good fit, indicating model misspecification. That is, across all simulation conditions, RMSEA was quite sensitive to the presence of noninvariance between groups.

Standardized root mean square residual. The average SRMR for small noninvariance was .03, indicating a good fit failing to detect the model misspecification due to noninvariance. On the other hand, SRMR was greater than .05 on average when noninvariance was large. Thus, SRMR with the cutoff of .05 could inform researchers of model misspecification due to measurement noninvariance only when the DIF was large. As observed in CFI, SRMR appeared to be an adequate indicator of model misfit only for large DIF.

Weighted root mean square residual. With the cutoff of 1.00 (Yu, 2002) WRMR showed model misfits for large DIF or large sample size conditions. On average, WRMR ranged from 0.53 (small DIF in intercept with sample size 100 per group) to 3.51.

Overall, when the magnitude of DIF was large, all examined model fit indices performed reasonably, indicating model misfits due to the presence of noninvariance in intercepts/thresholds. However, CFI and SRMR were not sensitive to small DIF showing a good fit regardless of sample size and data types.

Power and Type I Error Rates in Measurement Invariance Testing

The power rates. The power and Type I error rates of measurement invariance testing are presented in Tables 2 and 3. As observed in model fit evaluation, factor loading noninvariance was not detected reasonably in most simulation conditions. When a single variable was noninvariant across groups, the power rates improved with large sample and large DIF (Table 2). However, overall power rates were considerably low especially with two DIFs (e.g., near zero in most two-DIF conditions as presented in Table 3). On the contrary, for intercept/threshold noninvariance and noninvariance in both, the power rates were simply 1.00 unless sample size and DIF magnitude were small.

Comparing Bonferroni correction and Oort adjustment with no adjustment, we did not find any prominent difference in terms of power. Bonferroni correction degraded power slightly as expected. The performance of Oort adjustment with respect to power is worthy of further note because Oort adjustment in some cases improved and in other cases lowered the power rates slightly. For example, comparing one-DIF conditions (Table 2) to two-DIF conditions (Table 3), we observed relatively lower power rates for two-DIF conditions when Oort adjustment was applied. The lower power rates of Oort adjustment with multiple DIFs will be discussed later. Comparing three data types, MIMIC detected the noninvariance of continuous variables best followed by polytomous and dichotomous data.

The Type I error rates. As reported in previous studies, the Type I error rates were substantial throughout simulation conditions under no adjustment conditions. The

Table 2. The Power and Type I Error Rates of the MIMIC LR Tests: One DIF Variable

			Dichotomous						Polytomous						Continuous					
			No		Bonferroni		Oort		No		Bonferroni		Oort		No		Bonferroni		Oort	
			Pw	Er	Pw	Er	Pw	Er	Pw	Er	Pw	Er	Pw	Er	Pw	Er	Pw	Er	Pw	Er
Both	Small	100	.55	*.08*	.30	*.02*	.58	*.06*	.96	*.11*	.81	*.02*	.95	*.03*	.84	*.08*	.58	*.02*	.81	*.04*
		200	.84	*.10*	.66	*.03*	.85	*.05*	1.00	*.18*	.99	*.06*	1.00	*.02*	.99	*.11*	.92	*.02*	.98	*.03*
		500	1.00	*.17*	.97	*.05*	1.00	*.02*	1.00	*.38*	1.00	*.16*	1.00	*.00*	1.00	*.20*	1.00	*.07*	1.00	*.01*
		1,000	1.00	*.27*	1.00	*.10*	1.00	*.00*	1.00	*.64*	1.00	*.38*	1.00	*.00*	1.00	*.36*	1.00	*.15*	1.00	*.00*
	Large	100	1.00	*.14*	.95	*.05*	.99	*.02*	1.00	*.27*	1.00	*.09*	1.00	*.00*	1.00	*.13*	1.00	*.04*	1.00	*.01*
		200	1.00	*.23*	1.00	*.08*	1.00	*.00*	1.00	*.47*	1.00	*.24*	1.00	*.00*	1.00	*.21*	1.00	*.07*	1.00	*.00*
		500	1.00	*.47*	1.00	*.23*	1.00	*.00*	1.00	*.84*	1.00	*.63*	1.00	*.00*	1.00	*.41*	1.00	*.20*	1.00	*.00*
		1,000	1.00	*.74*	1.00	*.50*	1.00	*.00*	1.00	*.98*	1.00	*.93*	1.00	*.00*	1.00	*.66*	1.00	*.44*	1.00	*.00*
Intercept	Small	100	.54	*.08*	.30	*.02*	.57	*.06*	.80	*.08*	.62	*.01*	.82	*.04*	.93	*.11*	.79	*.02*	.91	*.04*
		200	.84	*.10*	.65	*.02*	.85	*.05*	.99	*.13*	.94	*.03*	.97	*.03*	1.00	*.17*	.99	*.05*	1.00	*.02*
		500	1.00	*.16*	.98	*.05*	.99	*.02*	1.00	*.27*	1.00	*.09*	1.00	*.01*	1.00	*.34*	1.00	*.14*	1.00	*.00*
		1,000	1.00	*.28*	1.00	*.10*	1.00	*.01*	1.00	*.47*	1.00	*.22*	1.00	*.00*	1.00	*.56*	1.00	*.33*	1.00	*.00*
	Large	100	.98	*.13*	.92	*.04*	.98	*.04*	1.00	*.21*	1.00	*.06*	1.00	*.01*	1.00	*.25*	1.00	*.09*	1.00	*.01*
		200	1.00	*.21*	1.00	*.07*	1.00	*.02*	1.00	*.38*	1.00	*.16*	1.00	*.00*	1.00	*.44*	1.00	*.22*	1.00	*.00*
		500	1.00	*.42*	1.00	*.20*	1.00	*.00*	1.00	*.72*	1.00	*.48*	1.00	*.00*	1.00	*.78*	1.00	*.57*	1.00	*.00*
		1,000	1.00	*.69*	1.00	*.44*	1.00	*.00*	1.00	*.95*	1.00	*.83*	1.00	*.00*	1.00	*.95*	1.00	*.84*	1.00	*.00*
Factor loading	Small	100	.06	*.05*	.01	*.01*	.09	*.07*	.07	*.04*	.02	*.00*	.09	*.07*	.09	*.05*	.02	*.01*	.10	*.06*
		200	.09	*.05*	.02	*.01*	.12	*.07*	.10	*.05*	.03	*.01*	.12	*.07*	.14	*.05*	.03	*.01*	.16	*.07*
		500	.11	*.05*	.03	*.01*	.16	*.07*	.12	*.05*	.03	*.01*	.16	*.07*	.24	*.05*	.07	*.01*	.27	*.06*
		1,000	.19	*.06*	.06	*.01*	.21	*.07*	.28	*.05*	.10	*.01*	.29	*.05*	.44	*.06*	.22	*.01*	.45	*.06*
	Large	100	.14	*.05*	.04	*.01*	.15	*.07*	.13	*.05*	.04	*.01*	.14	*.07*	.19	*.05*	.07	*.01*	.22	*.06*
		200	.20	*.05*	.08	*.01*	.22	*.07*	.26	*.06*	.09	*.01*	.25	*.06*	.36	*.05*	.16	*.01*	.37	*.05*
		500	.43	*.06*	.17	*.01*	.44	*.05*	.52	*.06*	.26	*.01*	.52	*.05*	.70	*.06*	.45	*.01*	.69	*.04*
		1,000	.75	*.07*	.47	*.02*	.73	*.05*	.83	*.09*	.59	*.02*	.80	*.03*	.93	*.08*	.83	*.02*	.93	*.02*

Note. MIMIC = multiple indicators multiple causes; LR = likelihood ratio; DIF = differential item functioning; Pw = power; Er = Type I error. "No," "Bonferroni," and "Oort" mean no correction, Bonferroni correction, and Oort adjustment on the critical values, respectively. "Both," "intercept," and "factor loading" are the location of noninvariance. "Small" and "large" are the magnitude of noninvariance. The sample size per group includes 100, 200, 500, and 1,000. Type I error rates are in italics.

Table 3. The Power and Type I Error Rates of the MIMIC LR Tests: Two DIF Variables

			Dichotomous						Polytomous						Continuous					
			No		Bonferroni		Oort		No		Bonferroni		Oort		No		Bonferroni		Oort	
			Pw	*Er*	Pw	*Er*	Pw	*Er*	Pw	*Er*	Pw	*Er*	Pw	*Er*	Pw	*Er*	Pw	*Er*	Pw	*Er*
Both	Small	100	.18	*.16*	.02	*.05*	.01	*.09*	.64	*.34*	.38	*.13*	.19	*.09*	.48	*.22*	.19	*.07*	.17	*.07*
		200	.57	*.27*	.26	*.11*	.10	*.07*	.97	*.57*	.85	*.30*	.45	*.06*	.89	*.38*	.70	*.18*	.45	*.06*
		500	.96	*.55*	.86	*.30*	.43	*.04*	1.00	*.90*	1.00	*.73*	.76	*.02*	1.00	*.74*	.99	*.50*	.77	*.03*
		1,000	1.00	*.81*	1.00	*.61*	.57	*.02*	1.00	*1.00*	1.00	*1.00*	.89	*.01*	1.00	*.91*	1.00	*.78*	.91	*.01*
	Large	100	.90	*.39*	.70	*.18*	.22	*.06*	1.00	*.84*	1.00	*.62*	.69	*.01*	.99	*.47*	.95	*.25*	.67	*.03*
		200	1.00	*.65*	.99	*.41*	.54	*.03*	1.00	*.98*	1.00	*.93*	.90	*.00*	1.00	*.72*	1.00	*.51*	.93	*.01*
		500	1.00	*.95*	1.00	*.85*	.83	*.00*	1.00	*1.00*	1.00	*1.00*	1.00	*.00*	1.00	*.96*	1.00	*.89*	1.00	*.00*
		1,000	1.00	*1.00*	1.00	*.99*	.94	*.00*	1.00	*1.00*	1.00	*1.00*	1.00	*.00*	1.00	*1.00*	1.00	*.99*	1.00	*.00*
Intercept	Small	100	.13	*.14*	.01	*.04*	.02	*.09*	.41	*.22*	.14	*.07*	.11	*.09*	.64	*.30*	.29	*.13*	.21	*.09*
		200	.46	*.24*	.18	*.09*	.09	*.09*	.85	*.39*	.50	*.18*	.32	*.07*	.95	*.54*	.83	*.31*	.51	*.08*
		500	.96	*.48*	.79	*.23*	.42	*.06*	1.00	*.74*	.99	*.51*	.71	*.05*	1.00	*.87*	1.00	*.71*	.81	*.05*
		1,000	1.00	*.75*	1.00	*.51*	.68	*.03*	1.00	*.96*	1.00	*.85*	.93	*.02*	1.00	*.98*	1.00	*.92*	.95	*.03*
	Large	100	.70	*.31*	.35	*.13*	.19	*.08*	.99	*.65*	.96	*.39*	.66	*.05*	1.00	*.75*	.98	*.53*	.69	*.09*
		200	.98	*.54*	.87	*.29*	.46	*.06*	1.00	*.91*	1.00	*.75*	.89	*.03*	1.00	*.93*	1.00	*.82*	.94	*.06*
		500	1.00	*.89*	1.00	*.72*	.66	*.01*	1.00	*1.00*	1.00	*.99*	1.00	*.01*	1.00	*1.00*	1.00	*1.00*	1.00	*.03*
		1,000	1.00	*.99*	1.00	*.95*	.82	*.00*	1.00	*1.00*	1.00	*1.00*	1.00	*.00*	1.00	*1.00*	1.00	*1.00*	1.00	*.01*
Factor loading	Small	100	.01	*.05*	.00	*.01*	.01	*.08*	.00	*.05*	.00	*.01*	.00	*.08*	.01	*.05*	.00	*.01*	.00	*.06*
		200	.00	*.06*	.00	*.01*	.00	*.07*	.00	*.06*	.00	*.01*	.00	*.07*	.01	*.06*	.00	*.01*	.00	*.07*
		500	.01	*.05*	.00	*.00*	.01	*.06*	.01	*.05*	.00	*.01*	.01	*.07*	.03	*.07*	.00	*.02*	.00	*.06*
		1,000	.03	*.05*	.01	*.01*	.02	*.06*	.01	*.07*	.00	*.01*	.00	*.07*	.09	*.09*	.01	*.02*	.02	*.06*
	Large	100	.01	*.05*	.00	*.01*	.00	*.07*	.00	*.06*	.00	*.01*	.00	*.07*	.04	*.06*	.00	*.01*	.02	*.05*
		200	.01	*.06*	.00	*.01*	.01	*.06*	.01	*.07*	.00	*.01*	.00	*.06*	.11	*.08*	.03	*.02*	.02	*.04*
		500	.05	*.05*	.00	*.01*	.02	*.04*	.02	*.09*	.00	*.02*	.00	*.05*	.37	*.12*	.12	*.04*	.03	*.02*
		1,000	.11	*.07*	.02	*.01*	.03	*.03*	.09	*.15*	.02	*.04*	.00	*.03*	.80	*.21*	.52	*.07*	.07	*.01*

Note. MIMIC = multiple indicators multiple causes; LR = likelihood ratio; DIF = differential item functioning; Pw = power; Er = Type I error. "No," "Bonferroni," and "Oort" mean no correction, Bonferroni correction, and Oort adjustment on the critical values, respectively. "Both," "intercept," and "factor loading" are the location of noninvariance. "Small" and "large" are the magnitude of noninvariance. The sample size per group includes 100, 200, 500, and 1,000. Type I error rates are in italics.

Type I error rates inflated as sample size and the degree of noninvariance increased. In combination of large sample size and large degree of noninvariance, the Type I error rates reached near 100% before critical value adjustment. When Bonferroni correction was applied, the Type I error rates slightly decreased, but they were still unacceptably high. Interestingly, Oort adjustment controlled the Type I error rates about the nominal level (i.e., .05) across conditions. The range of Type I error rates was .00 through .07 for one-DIF conditions, .00 through .09 for two-DIF conditions.

When factor loadings were noninvariant across groups, the Type I error rates were not inflated as much as for noninvariance in intercepts/thresholds. With no critical value adjustment, the Type I error rates were below 10% in most conditions (the highest was 21% with two DIFs). However, these low Type I error rates do not support MIMIC modeling for factor loading noninvariance tests given the low (near zero) power.

Raw Bias

First, the raw bias in the parameter estimates of the latent group mean difference (γ) was examined (see Table 4). Irrespective of simulation conditions, the raw bias of γ was noticeably small ranging from −.001 to .007. Given that the simulated latent group mean difference is .500, the estimates, on average, fall between .499 and .507, which is close to the population parameter. Even in the conditions of factor loading noninvariance, the raw bias of γ was near zero.

When the intercepts were noninvariant across groups, the population parameter β_{Y5} equals the size of simulated intercept or threshold noninvariance (.3 and .6 for small and large DIF, respectively). For continuous data, the estimates of β_{Y5} were very close to the parameter, which yielded negligible raw bias (.000 in most conditions). For categorical data, raw bias was about −.04 for small DIF and −.08 for large DIF regardless of sample size.

For factor loading noninvariance conditions, MIMIC modeling does not estimate factor loading noninvariance explicitly. However, because the parameter β_{Y5} (i.e., intercept noninvariance) equals zero, we reasoned that the raw bias of β_{Y5} reflected the impact of factor loading noninvariance in the model. Entering the parameters in Equation (6), we could derive the effect of factor loading noninvariance mathematically: One group intercept is higher by the magnitude of the difference in $\lambda\eta$ (i.e., $.2\eta$ for small factor loading noninvariance; $.4\eta$ for large factor loading noninvariance). Then, the β_{Y5} estimate will be negatively biased (i.e., $-.2\eta$ and $-.4\eta$) to yield the equal intercepts across groups. In this study, the raw bias was estimated as −.05 and −.10 for small and large DIF conditions, respectively. As presented in Table 4, raw bias of β_{Y5} was close to the calculated effect of factor loading noninvariance.

When the noninvariance was present in both factor loading and intercept, the raw bias was calculated by subtracting the intercept/threshold noninvariance (.3 and .6 for large and small DIF, respectively) from the β_{Y5} estimates. Then, the computed raw bias will be the sum of the raw bias of intercept/threshold noninvariance and the effect of

Table 4. The Raw Bias of Parameter Estimates in Correctly Specified MIMIC Models: Latent Group Mean Difference (γ) and Intercept DIF (β)

			Dichotomous		Polytomous		Continuous	
			γ	β	γ	β	γ	β
Both	Small	100	.000	−.056	.000	−.140	−.001	−.049
		200	.007	−.052	.005	−.140	.003	−.050
		500	.003	−.053	.003	−.138	.004	−.048
		1,000	.005	−.049	.004	−.141	.003	−.050
	Large	100	.001	−.224	.000	−.400	−.001	−.099
		200	.007	−.219	.005	−.395	.003	−.100
		500	.003	−.220	.003	−.394	.004	−.098
		1,000	.005	−.217	.004	−.398	.003	−.100
Intercept	Small	100	.001	−.041	.000	−.039	−.001	.000
		200	.007	−.038	.005	−.036	.003	.000
		500	.003	−.040	.003	−.035	.004	.002
		1,000	.005	−.035	.004	−.038	.003	.000
	Large	100	.001	−.082	.000	−.081	−.001	.000
		200	.007	−.080	.005	−.075	.003	.000
		500	.003	−.083	.003	−.072	.004	.002
		1,000	.005	−.075	.004	−.076	.003	.000
Factor loading	Small	100	.000	−.055	.000	−.051	−.001	−.049
		200	.007	−.047	.005	−.051	.003	−.050
		500	.003	−.052	.003	−.048	.004	−.048
		1,000	.005	−.055	.004	−.053	.003	−.050
	Large	100	.001	−.131	.000	−.119	−.001	−.099
		200	.007	−.122	.005	−.123	.003	−.100
		500	.003	−.129	.003	−.118	.004	−.098
		1,000	.005	−.134	.004	−.122	.003	−.100

Note. MIMIC = multiple indicators multiple causes; LR = likelihood ratio; DIF = differential item functioning. "Both," "intercept," and "factor loading" are the location of noninvariance. "Small" and "large" are the magnitude of noninvariance. The sample size per group includes 100, 200, 500, and 1,000.

factor loading noninvariance. Therefore, we expected that the magnitude of raw bias in this case was close to the raw bias of factor loading noninvariance conditions if the raw bias of intercept/threshold noninvariance was near zero as observed in continuous data (Table 4). Overall, MIMIC modeling with categorical data exhibited larger bias (e.g., near −.40 for large DIF in polytomous data) when both factor loading and intercept were noninvariant.

Discussion

The Sensitivity of Model Fit Index of the Baseline MIMIC Model

When noninvariance are present in a model, researchers expect to observe lack of fit of the model that alerts researchers to check model misspecification, including measurement noninvariance. Although a model misfit does not inform researchers of the source of lack of fit, a misspecified model with measurement noninvariance (i.e., running an invariance-imposed model when a certain type of invariance is violated) should not show a good model fit. It is also important to know which model fit index is sensitive to the presence of measurement noninvariance.

This study observed the insensitivity of model fit indices to the violation of factor loading invariance assumption of the MIMIC model. All examined model fit indices, including chi-square *p*, CFI, RMSEA, and SRMR/WRMR supported a good fit failing to detect the factor loading noninvariance. This finding implies that the good fit of a MIMIC model (in Equations 2 and 4) that inherently assumes full invariance may not guarantee the equivalence of factor loadings over groups and underscores the importance of explicit tests of measurement invariance using either nonuniform MIMIC/RFA or multiple group CFA.

Among the fit indices investigated in this study, chi-square *p* and RMSEA showed the highest sensitivity to the model misspecification with the measurement noninvariance in intercepts/thresholds. CFI and SRMR were not sensitive to the model misspecification because of the presence of *small*-size DIF. In practice, because researchers assessing measurement invariance do not know the magnitude of DIF, it is recommended to refer to RMSEA and WRMR (for categorical data) in addition to chi-square fit statistic. However, when small noninvariance is not of great concern in a study, all fit statistics examined in this study appeared to provide correct information of model lack of fit. The final recommendation about fit indices of the MIMIC model is to use a more conservative cutoff, which means CFI >.95, RMSEA and SRMR <.05 rather than .90 and .08, respectively, if measurement invariance is of interest.

The Power and Type I Error Rates

The previous simulation studies consistently presented high Type I error rates in detecting noninvariant variables with MIMIC modeling (Finch, 2005; Navas-Ara & Gomez-Benito, 2002; Oort, 1998; Wang et al., 2009). So did this simulation study before the Oort adjustment was applied to the chi-square critical values. We observed enormously large chi-square differences in the LR tests not only for the noninvariant variables but also for the invariant variables, especially when sample size and DIF size were large. The inflated chi-square differences presumably lead to the more frequent rejections of the null hypothesis even when the null hypothesis is true, which results in Type I error inflation. This was demonstrated in the no-adjustment conditions of this study.

The chi-square inflation in the LR test using MIMIC modeling may be explained with the misspecification of the baseline model. When the model includes any noninvariant variable and is analyzed with the assumption of invariance across groups, which is inherently done in the baseline MIMIC model, the chi-square fit statistic likely inflates because of the model misspecification error. However, this speculation requires further research for confirmation.

Bonferroni correction is one option of critical value adjustment to control Type I error inflation. In this study, each replication went through six LR tests that might inflate Type I error. However, the major cause of Type I error inflation appeared beyond the experimentwise Type I error inflation. As Oort noted (1992, 1998), the inflation more likely originated from the oversensitivity of chi-square fit statistic to model misspecification errors, which becomes more evident with large sample size and large DIF. Therefore, Bonferroni correction will not be an appropriate remedy in this case although it could lower Type I error rates. As the results showed, after the Bonferroni adjustment, the Type I error rates were still considerably high throughout all conditions. Another concern of Bonferroni correction is in the power reduction. Adopting more conservative critical *p* value reduced the power to identify the noninvariance. However, this power shrinkage was less obvious when sample size and the degree of noninvariance were large. To sum up, Bonferroni correction appears not to be an optimal method to suppress the inflated Type I error rates in measurement invariance testing with MIMIC model.

Oort correction does not merely lower the critical value but takes into account the magnitude of baseline chi-square value given degrees of freedom. Therefore, instead of evaluating the model fit with one fixed critical value such as 3.84 (χ^2 with one degree of freedom at $\alpha = .05$), the critical chi-square value was tailored for each model depending on the degree of inflation of chi-square. The results of this simulation study showed that the Oort adjustment worked remarkably well when the Type I error inflation was severe. For example, in the large sample and large intercept DIF condition of continuous variables the Type I error rate dropped from .95 to .00 after the Oort adjustment. Because Oort adjustment tailored the critical chi-square value for each baseline model, it did not lessen the power to detect the noninvariance much when there was only one noninvariant variable.

However, Oort adjustment attenuated power when two variables were noninvariant: the deterioration of power was notable compared with the one-noninvariant-variable cases. Given that only one variable was relaxed at a time in the LR test, even when the less restricted model correctly specified one of the DIF items, another noninvariant variable existed in the model. That is, when the baseline model had two noninvariant variables, the less restricted model in which one of DIF was relaxed for inequality over groups still had one noninvariant variable. That is, both models (baseline model with two DIF variables and augmented model with one DIF variable) in the LR test are incorrectly specified. As explained by Yuan and Bentler (2004), the LR tests between a misspecified baseline model and a misspecified unconstrained model did not yield the same power in detecting DIF items as the LR tests did with only one DIF item. This understanding on the power degradation with more than one noninvariance calls for the iterative procedure of the LR test. Because the Type I error rates were considerably low throughout conditions with Oort adjustment, the detected variable in the first LR test was more likely to be one of the DIF variables. Hence, if this detected item is free to be estimated across groups and if this unconstrained model is used as a baseline for the following LR test, the same high performance of MIMIC is expected as we observed in the one-DIF conditions. We can expect decent results from the iterative LR tests when the detected variables are likely to be noninvariant variables (i.e., when the Type I error rate is low). Therefore, the statistical approach to control for the Type I error rates (e.g., Oort adjustment) will play a critical role in the LR tests using MIMIC modeling.

Raw Bias

Regarding raw bias, overall parameter estimates were fairly accurate if the model was correctly specified (i.e., *Y*5 was specified as DIF). Importantly, the estimates of latent group mean difference (γ) were unbiased even with the factor loading noninvariance. Irrespective of data type, sample size, and DIF size, the parameter estimates were very close to the population parameter, which is very encouraging to the users of MIMIC modeling for latent group mean difference testing. However, special attention is needed for the parameter estimates of β_{Y5}. We observed that the estimates of β_{Y5} included noninvariance of all sorts (i.e., not only intercept noninvariance but also factor loading noninvariance if present). Because the current MIMIC modeling is not capable of separating intercept noninvariance from factor loading noninvariance, the estimate of β_{Y5} should not be interpreted as the estimate of intercept noninvariance only.

If factor loading noninvariance was confounded with intercept noninvariance and manifested as β_{Y5}, why were the power rates so low in the factor-loading-noninvariance-only conditions? When the factor loading noninvariance was converted to β_{Y5}, the magnitude of the converted β_{Y5} in this study was −.05 and −.10 for small (.20) and large (.40) factor loading noninvariance, respectively. Given that the magnitude of small intercept noninvariance was .30 in this study, the converted factor loading noninvariance was very small in size, which is less likely to be detected in the measurement invariance testing.

In summary, whereas the performance of MIMIC modeling with Oort adjustment was decent (i.e., high power and about nominal-level Type I error) in the identification of intercept noninvariance, MIMIC modeling showed poor performance in detecting factor loading noninvariance. Therefore, the researchers interested in measurement invariance testing should be aware of the downsides of the current MIMIC modeling for uniform noninvariance.

Conclusion

Measurement invariance testing is important to establish the validity of a measure across subpopulations of interest. The detection of noninvariant variables is essential to improve test quality and, furthermore, to understand the meaning of noninvariance over groups, and to avoid flawed conclusions based on measurement bias in the use of a test. From the findings of this study, we make two suggestions to the researchers conducting measurement invariance testing. First, the MIMIC model for uniform noninvariance did not detect the factor loading noninvariance properly. Thus, the current MIMIC model should be used only when the factor loading invariance is achieved. In practice, it is recommended to use MIMIC modeling for nonuniform bias or multiple group CFA for measurement invariance testing. Second, for the LR tests, when the chi-square difference between two models is likely to inflate (e.g., the baseline model is possibly contaminated with noninvariant variables in measurement invariance testing), and subsequently Type I error is inclined to inflate, we strongly recommend Oort adjustment to control Type I error inflation in the LR tests.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interests with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Ainsworth, A. T. (2008). Dimensionality and invariance: Assessing differential item functioning using bifactor multiple indicator multiple cause models. *Dissertation Abstracts International, 68*(9), 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*(2), 117-127. doi:10.1007/s10182-010-0126-1612KL

Barendse, M. T., Oort, F. J., Werner, C. S., Ligtvoet, R., & Schermelleh-Engel, K. (2011). *Measurement bias detection through factor analysis*. Manuscript submitted for publication.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternate ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*, 278-295.

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: Psychological Sciences and Social Sciences, 57*, S275-S284.

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling, 15*, 96-113.

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.

Joreskog, 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.

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (2nd ed.). Thousand Oaks, CA: Sage.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, *18*, 212-228.

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

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.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7*, 361-388.

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.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

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

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557-585.

Muthén, B. O., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus Web Notes: No. 4*. Retrieved from http://www.statmodel.com/download/webnotes/CatMGLong.pdf

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. (2008). Mplus (Version 5.2) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Navas-Ara, M. J., & Gomez-Benito, J. (2002). Effects of ability scale purification on the identification of DIF. *European Journal of Psychological Assessment, 18*, 9-15. doi:10.1027//1015-5759.18.1.9

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika, 6*, 150-166.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling, 5*, 107-124.

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

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*(3), 53-67. doi:10.1300/J079v29n03_03

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1292-1306.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

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.

Wang, W.-C., Shih, C.-L., & Yang, C.-C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement, 69*, 713-731.

Willse, J. T., & Goodman, J. T. (2008). Comparison of multiple-indicators, multiple-causes– and item response theory–based analyses of subgroup differences. *Educational and Psychological Measurement, 68*, 587-602. doi:10.1177/0013164407312601

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*, 58-79. doi:10.1037/1082-989X.12.1.58

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*, 1-27. doi:10.1080/00273170802620121

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. 10.1177/0146621611405984

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 & Behavioral Assessment, 31*, 320-330. doi:10.1007/s10862-008-9118-9

Yoon, M. (2008). Statistical power in testing factorial invariance with ordinal measures. *Dissertation Abstracts International, 68*(11), 7705B.

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

Yu, C. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Unpublished manuscript, University of California, Los Angeles.

Yuan, K. H., & Bentler, P. M. (2004). On chi-square difference and *z* tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement, 64*, 737-757. doi:10.1177/0013164404264853