2015

Within-level Group Factorial Invariance with Multilevel Data: Multilevel Factor Mixture and Multilevel MIMIC Models

Eun Sook Kim
University of South Florida, ekim3@usf.edu

Myeongsun Yoon
Texas A & M University - College Station

Yao Wen
University of Wisconsin - Milwaukee

Wen Luo
Texas A & M University - College Station

Oi-man Kwok
Texas A & M University - College Station

Running head: WITHIN-LEVEL GROUP FACTORIAL INVARIANCE

Within-level Group Factorial Invariance with Multilevel Data: Multilevel Factor Mixture and Multilevel MIMIC Models

Eun Sook Kim[1], Myeongsun Yoon[2], Yao Wen[3], Wen Luo[2], & Oi-man Kwok[2]

[1]University of South Florida [2]Texas A&M University [3]University of Wisconsin, Milwaukee

Corresponding author:

Eun Sook Kim, Ph.D.

Department of Educational Measurement and Research

University of South Florida

4202 E. Fowler Ave. EDU105

Tampa, FL 33620-7750

Phone: (813) 974-7692

Email: ekim3@usf.edu

Fax: (813) 974-4495

Abstract

This study suggests two approaches to factorial invariance testing with multilevel data when the groups are at the within level: multilevel factor mixture model for known classes (ML FMM) and multilevel multiple indicators multiple causes model (ML MIMIC). The adequacy of the proposed approaches was investigated using Monte Carlo simulations. Additionally, the performance of different types of model selection criteria for determining factorial invariance or in detecting item noninvariance was examined. Generally, both ML FMM and ML MIMIC demonstrated acceptable performance with high true positive and low false positive rates, but the performance depended on the fit statistics used for model selection under different simulation conditions.

Within-level Group Factorial Invariance in Multilevel Data: Multilevel Factor Mixture Model and Multilevel MIMIC Model

For the last several decades, the area of factorial invariance (or measurement invariance in broader contexts) has received great attention. Not only has extensive methodological work been conducted on this topic but testing factorial invariance has become common practice before comparing latent means in applied research (Raykov, Marcoulides, & Li, 2012). However, there are still unresolved issues in this area: for example, locating a truly invariant variable for a reference variable (French & Finch, 2008), establishing partial invariance (e.g., Millsap & Kwok, 2004), and developing practical criteria in determining a lack of factorial invariance, especially with a mean structure (e.g., Fan & Sivo, 2009). With advances in methodology, issues involved in certain analytic methods arise in addition to the previously mentioned general concerns related to factorial invariance testing. For example, special model specification issues occur in factorial invariance testing with multilevel data. This study is particularly interested in model specification issues related to testing factorial invariance for within-level groups in multilevel modeling.

It is well known among social scientists that a single level statistical approach to multilevel data underestimates standard errors in statistical significance testing, which may lead to incorrect statistical inference, that is, Type I error. Recently, Kim, Kwok, and Yoon (2012) studied both within-level and between-level factorial invariance testing. In evaluating weak factorial invariance across between-level groups, they conducted multilevel confirmatory factor analysis for multiple groups (i.e., multilevel multigroup CFA; Muthén, 1989). Their simulation results supported the suitability of multilevel multigroup CFA with acceptable power and adequate Type I error control whereas the single-level multigroup CFA yielded inflated Type I

error rates as a function of the intraclass correlation (ICC) and cluster size. Thus, they recommended multilevel multigroup CFA for factorial invariance testing with multilevel data. In conducting within-level factorial invariance testing, however, Kim et al. suggested a design-based approach (Muthén & Satorra, 1995; Wu & Kwok, 2012) using the TYPE = COMPLEX option in Mplus (Muthén & Muthén, 2012) because constructing multigroup multilevel models for within-level groups is not feasible when the group indicator (e.g., females and males within schools) is crossed across higher-level clusters. For each within-level group they generated a single factor model at both the within- and between-levels with an identical set of factor loadings for both levels. When noninvariance was simulated in one of the factor loadings in the within-level groups, the design-based approach perfectly detected the violation of weak factorial invariance (power = 1.0). Under complete invariance, Type I error rates were around the nominal level (.04 ~ .07).

The design-based approach to multilevel data is in fact a single level CFA, but corrects the underestimated standard errors of parameter estimates due to data dependency. Of note is that the design-based approach implicitly assumes factorial invariance across levels (i.e., cross-level factorial invariance) by constructing a single-level CFA model. Cross-level factorial invariance[1] is referred to as the equivalence of factor loadings with the same number of factors across the between and within levels (Dedrick & Greenbaum, 2011). However, cross-level factorial invariance is not always warranted for the application of the design-based approach. Wu and Kwok (2012) studied the performance of the design-based approach in a single group context when the between-level factor structure was different from the within-level factor structure. When cross-level factorial invariance was violated especially with a simple within factor structure (e.g., a one-factor model) and a complex between factor structure (e.g., a two-factor

model), the design-based approach showed poor model fit to the data and yielded biased estimates of both fixed and random effects.

Given the stringent assumptions of the design-based approach, the purpose of this study is to propose two potential approaches to factorial invariance testing with multilevel data when the groups are at the within level: multilevel factor mixture model for known classes (ML FMM) and multilevel multiple indicators multiple causes model (ML MIMIC). To this end, we conducted three Monte Carlo studies. First, we investigated the adequacy of ML FMM in testing weak and strong factorial invariance across within-level groups at the scale level. Next, the performance of ML MIMIC in detecting a noninvariant item at the item level was examined. Note that ML FMM and ML MIMIC were used for slightly different purposes: the first to establish weak or strong factorial invariance, the latter to detect a particular noninvariant variable. We do not purport to compare the two methods but rather we explore their performance under the circumstances in which they are typically employed for factorial invariance testing. In the third study, the proposed multilevel approach to factorial invariance testing, specifically, ML FMM was compared to the single level design-based approach when between and within factor structures were not identical. Throughout these Monte Carlo studies, different types of model selection criteria were examined with respect to their performance in determining a level of factorial invariance (e.g., weak invariance) or in detecting a noninvariant item. Finally, the proposed methods of factorial invariance testing across within-level groups were illustrated with the mathematics self-efficacy measure from the Programme for International Student Assessment (PISA) 2003 data (OECD, 2005).

Multilevel factor mixture model for known classes

Factor mixture modeling is used to analyze unobserved population heterogeneity and identify latent classes underlying the observed data. In addition, factor mixture models can be used for observed classes, which are, in effect, analogous to multiple group analysis because all latent classes identified in the model correspond to the observed groups. However, multiple group analysis under the factor mixture framework expands modeling capabilities beyond the conventional SEM. Some of the modeling issues as discussed in the previous section (modeling factorial invariance across within-level groups in multilevel data) can be solved by incorporating categorical latent variables into the analysis.

In factor mixture models, the observed random variable *y* is modeled conditional on the latent class variable, *C* (*C* = 1, 2, …, *c*) (Asparouhov & Muthén, 2008; Lubke & Muthén, 2005):

$$[\mathrm{y}_i|C_i = c] = \nu_c + \Lambda_c\eta_i + \Gamma_{yc}\mathrm{x}_i + \varepsilon_{ci}, \quad (1)$$

$$[\eta_i|C_i = c] = \mu_c + \mathrm{B}_c\eta_i + \Gamma_{\eta c}\mathrm{x}_i + \zeta_{ci}. \quad (2)$$

In Equation 1, Λ is the factor loading matrix expressing the relations of the observed outcome variables *y* with the latent variables η, Γ_y is the pattern coefficients of *y* regressed on observed covariates *x*, and ν and ε are intercepts and residuals, respectively. Equation 2 shows the relations of the endogenous latent variables η to exogenous latent variables η (B) and to observed covariates *x* (Γ_η) with μ and ζ as intercepts and residuals, correspondingly. Conditional on the latent classes, *y* is assumed to be multivariate normally distributed.

Because the latent classes are unordered categories, the latent class membership can be defined as a multinomial variable (or a binary variable for two classes) by

$$\ln\left[\frac{P(C_{ij}=c|\mathrm{x}_{ij})}{P(C_{ij}=k|\mathrm{x}_{ij})}\right] = \lambda_{cj} + \Gamma_{cj}\mathrm{x}_{ij} \quad (3)$$

where the log odds of the probability of being in a specific latent class *c* over a reference class *k* is a function of the regression coefficients of the covariates (Γ_{cj}) and the intercepts (λ_{cj}). With the

known class option, instead of modeling the latent class membership using Equation 3, the unknown classes are replaced by observed groups. The sample units with the known class membership are called training data (Muthén, 2002). When all sample units fall into training data, the latent class membership is simply the observed group membership.

General discussions and information related to multilevel common factor models are applicable to multilevel factor mixture models for known classes. By allowing random effects across *j* clusters, the total covariance matrix (Σ_T) of the observed outcome vector (y_{ij}) is decomposed into within- and between-cluster components:

$$\Sigma_T = \Sigma_w + \Sigma_B \quad (4)$$

where subscripts *W* and *B* denote within and between, respectively. Σ_w represents the variability of individuals within cluster; Σ_B corresponds to the variability across clusters. Correspondingly, the linear relations between the latent factors and the observed outcomes are expressed by within and between components:

$$y_{ij} = \nu + \Lambda_W \eta_{W_{ij}} + \varepsilon_{W_{ij}} + \Lambda_B \eta_{B_j} + \varepsilon_{B_j} \quad (5)$$

where residuals at each level are assumed to be normally distributed with a mean of zero and independent of the latent variables and the residuals at the other level:

$$\varepsilon_{W_{ij}} \sim N\left(0, \theta_{\varepsilon_W}\right), \varepsilon_{B_j} \sim N\left(0, \theta_{\varepsilon_B}\right), cov(\eta, \varepsilon) = 0, cov(\varepsilon_W, \varepsilon_B) = 0.$$

Under these assumptions, the variance covariance matrix at each level is derived as follows:

$$\Sigma_w = \Lambda_W \Phi_W \Lambda'_W + \Theta_W, \ \Sigma_B = \Lambda_B \Phi_B \Lambda'_B + \Theta_B \quad (6)$$

where Φ_W and Φ_B denote the variance covariance matrix of latent variables at the within and between levels, respectively.

Factorial invariance testing across within-level groups requires a separate common factor model for each group at the within level. For each latent class representing a respective observed group, a factor mixture model is constructed as:

$$\left[\mathrm{y}_{ij}\middle|C_{ij}=c\right]=\nu_c+\Lambda_{Wc}\eta_{W_{ij}}+\varepsilon_{W_{ij}}+\Lambda_B\eta_{B_j}+\varepsilon_{B_j}. \quad (7)$$

Thus, factorial invariance refers to the equivalence of a set of parameters across groups including $\nu_1 = \nu_2 = \cdots = \nu_c$ and $\Lambda_{W1} = \Lambda_{W2} = \cdots = \Lambda_{Wc}$. Note that in the current multilevel SEM framework, intercepts are estimated at the between level only because the within-level model is analyzed with deviation scores from cluster means. Although the mean structure is incorporated for group comparison at the within level, the intercepts of observed variables (y) are estimated at the between level only and so is the intercept invariance.

Considering that there is no specific term for intercepts at the within level, Ryu (2014) suggested a method for within-level factorial invariance testing using multigroup single-level CFA. To estimate the intercepts at the within level, she specified within and between models separately for each group using Muthén's maximum likelihood estimation (MUML, Muthén, 1994) and conducted a 2**k* group single-level CFA where *k* equals the number of comparison groups. On the other hand, ML FMM for known classes allows for the free estimation of the intercepts across within-level groups at the between level. Thus, the mean or intercept difference across within-level groups can be tested at the between level under ML FMM.

Multilevel MIMIC Model

MIMIC modeling (Jöreskog & Goldberger, 1975; Muthén, 1989) is one of several methods used to test factorial invariance and population heterogeneity (e.g., Kim, Yoon, & Lee, 2012). MIMIC models employ an observed variable as a covariate of latent factors. The presence of the vector of observed covariates (*x*) in common factor models indicates MIMIC modeling.

$$y_i = \Lambda\eta_i + \Gamma_y x_i + \varepsilon_i, \quad (8)$$

$$\eta_i = \Gamma_\eta x_i + \zeta_i. \quad (9)$$

The inclusion of a grouping covariate allows testing group differences via a regression-type analysis. Thus, a regression coefficient (Γ_η) associated with the grouping covariate x in Equation 9 represents the effect of group membership on the corresponding latent factor (η), specifically, factor mean difference between groups when the grouping covariate is dummy coded. The inclusion of the effect of a grouping covariate on the observed outcome variable ($\Gamma_y x_i$) over and above the effect of the grouping covariate on the latent factor (Γ_η) in Equation 8 allows factorial invariance testing of the intercept (also called uniform invariance). In other words, the regression coefficient of x on y (Γ_y) indicates the status of intercept invariance of the corresponding variable y between groups.

Nonuniform invariance can be tested using MIMIC modeling by including an interaction between the observed covariate and the latent factor (Barendse, Oort, Werner, Ligtvoet, & Schermelleh-Engel, 2012; Woods & Grimm, 2011). Then, Equation 8 expands as

$$y_i = \Lambda\eta_i + \Gamma_y x_i + \Gamma_{\eta y}\eta_i x_i + \varepsilon_i. \quad (10)$$

The regression coefficient of the observed covariate (Γ_y) represents intercept invariance or uniform invariance; the regression coefficient of the interaction ($\Gamma_{\eta y}$) represents factor loading invariance or nonuniform invariance. There are different ways to create an interaction term between a latent factor and an observed covariate[2]. In this study, we adopted the XWITH statement in Mplus to model the interaction of a latent variable with an observed covariate.

The MIMIC model can be extended to a multilevel framework by including subscript *j* for clusters, which indicates that a certain effect of interest can vary across clusters. When factorial invariance is tested across within-level groups, a grouping covariate and its interaction

with the within-level factor are entered at the within level. Thus, a multilevel MIMIC model is expressed as

$$y_{ij} = \nu + \Lambda_W \eta_{W_{ij}} + \Gamma_y x_{ij} + \Gamma_{\eta y} \eta_{W_{ij}} x_{ij} + \Lambda_B \eta_{B_j} + \varepsilon_{W_{ij}} + \varepsilon_{B_j} \quad (11)$$

where x_{ij} is an observed covariate indicating group membership of individual *i* in cluster *j*. The within-level regression coefficient Γ_y represents the difference or noninvariance in intercepts between groups. Again factor loading invariance can be examined by testing the regression coefficient of the interaction between the observed covariate and the latent factor ($\Gamma_{\eta y}$) at the within level. In this exposition of multilevel common factor models, of note is that the two regression coefficients representing factorial invariance (Γ_y and $\Gamma_{\eta y}$) are estimated as fixed effects under the assumption that the status of factorial invariance between groups does not vary across clusters, which is realistic with a reasonably developed measure.

Study 1: Testing Scale-Level Factorial Invariance Using Multilevel Factor Mixture Model

Method

Data Generation

The adequacy of ML FMM for scale-level factorial invariance testing was examined through a Monte Carlo simulation. To create population heterogeneity, two sets of data were generated separately and then combined for data analyses. Each set of data corresponded to each group of the study, and the group size was balanced across simulation conditions. The same sets of clusters were generated in both groups so that two groups existed within cluster. Within- and between-level factor structures were identical. That is, each group had eight variables that loaded on a single factor with factor loadings .3 to .9 and residual variances .25 for both within- and between-levels. Intercepts and factor means were simulated at zero at both levels for both groups.

In data generation the following design factors were considered in investigating the performance of ML FMM under various circumstances researchers possibly encounter in practice.

Simulation Design Factors

The design factors of this simulation study include: (a) intraclass correlation, (b) number of clusters, (c) cluster size, (d) location of noninvariance, and (e) size of noninvariance (Hox & Maas, 2001; Kim, Kwok, et al., 2012; Yoon & Millsap, 2007). Three levels of ICC were simulated by varying the between-level factor variance while fixing the within-level factor variance at 1.00. The between-level factor variances of 0.10, 0.25, and 0.50, yield three ICC levels: .09, .20, and .33, respectively. Such ICCs are commonly observed in educational research. It is of note that the ICCs are computed on the basis of latent factor variance as follows:

$$ICC_\eta = \Phi_B/(\Phi_B + \Phi_W) \quad (12)$$

where Φ_B and Φ_W are two components of the total latent factor variance (Φ_T).

Two cluster size (CS) levels were examined: 10 and 20. Accordingly, the balanced group size within cluster was 5 and 10. Cluster sizes between 5 and 50 are commonly used in multilevel simulation studies (Finch & French, 2011; Hox & Maas, 2001; Jak, Oort, & Dolan, 2013). The number of clusters (CN) had three levels: 60, 100, and 160. The combination of the number of clusters and cluster size produced total sample sizes that varied between 600 and 3200. The noninvariance across within-level groups was simulated for either the factor loading or intercept. Only a single item (y7) out of eight was noninvariant. The size of noninvariance associated with the factor loading consisted of two levels: 0.25 and 0.50; the size of noninvariance associated with intercept was also 0.25 and 0.50, representing small and large noninvariance (i.e., differential item functioning or DIF), respectively (French & Finch, 2008; Kim, Kwok, et al., 2012). Other than the parameter of noninvariance, the two groups have

identical population parameters. When noninvariance is simulated to evaluate power to detect noninvariance, 3 ICC x 3 CN x 2 CS x 2 DIF location x 2 DIF size, 72 conditions are created. In addition, data under complete invariance conditions (3 ICC x 3 CN x 2 CS = 18) are generated to examine Type I error when there is no DIF variable. For each condition, 1000 replications are generated. Each replication is analyzed for factorial invariance using ML FMM following the factorial invariance testing procedures explicated in the next section. Data generation and all subsequent analyses are conducted with Mplus version 6.11.

Scale-Level Factorial Invariance Testing Procedures

The simulated data were fitted to the proposed model. In ML FMM, a single-factor CFA model with eight indicators was constructed at both within and between levels and for both groups as known classes. For identification purposes, the factor variance at each level was fixed at the corresponding population parameter. Although this identification strategy is not realistic in practice without knowing population variances, we adopt this strategy with known variances to recover the parameter estimates in the scale of population parameters. The factor mean of one group at the between level was also fixed at zero for identification.

Factorial invariance is often established hierarchically: configural invariance (i.e., the equivalence between groups in the number of factors and the pattern of indicator loadings on each factor), weak or metric invariance (the equality of factor loadings, in addition), strong or scalar invariance (the equality of intercepts added), and strict invariance (additional residual variance equality). For factor loading noninvariance conditions in this study, weak invariance was examined because configural invariance was already established in the simulation. To test weak invariance a model with all factor loadings constrained equal between classes was compared to a model with such a constraint removed (i.e., factor loadings were freely estimated

except the first variable for identification). Two competing models were evaluated based on model selection criteria (see the following section for details). Strong invariance was tested similarly. For model estimation, robust maximum likelihood (MLR), the Mplus default, was used.

Given the identification strategy adopted in this study with knowledge of population variances and the invariance of the first observed variable, the aforementioned hierarchical approach to factorial invariance is appropriate. The limitations of this approach in real research settings were discussed in Raykov, Marcoulides, and Li (2012). Because one of the variables, called a reference variable, is conventionally constrained equal between groups for identification purposes, the invariance of the full set of variables cannot be tested in the widely used factorial invariance testing procedures (i.e., from configural to strict invariance) unless the reference variable is known to be invariant without testing. Interested readers are encouraged to read Raykov et al.'s full discussions on this matter.

Model Selection Criteria

For factorial invariance testing, we conducted likelihood ratio tests comparing two nested models (constrained and relaxed ones) under the null hypothesis of no difference between the two models. When MLR is utilized for model estimation, Satorra-Bentler scaled likelihood ratio (SB LR) tests are recommended for model comparison (for details see Satorra & Bentler, 1994). When the SB LR yields a negative chi-square difference or negative likelihood ratio, additional adjustment is required to ensure the positive chi-square statistic (Asparouhov & Muthén, 2012a; Satorra & Bentler, 2010). The SB LR is asymptotically chi-square distributed with the difference in degrees of freedom between the two models. When the null hypothesis is rejected, the relaxed model is selected whereas the constrained model is favored otherwise.

We also examined alternative model fit statistics: Akaike Information Criterion (AIC; Akaike, 1987), Bayesian Information Criterion (BIC; Schwarz, 1978), and sample-size adjusted BIC (ssBIC, Sclove, 1987). These model fit statistics can be used to compare the nonnested models. When two models are compared, the model associated with the smaller alternative model fit statistic is considered a better model. Recent studies have examined the performance of such model fit statistics in mixture modeling contexts (e.g., Chen, Kwok, Luo, & Willson, 2010; Henson, Reise, & Kim, 2007). However, the performance in the context of multilevel factorial invariance testing is unknown. In summary, we considered the SB LR tests, AIC, BIC, and ssBIC for model selection in determining factorial invariance for the currently conducted studies.

Analysis of Simulation Results

To evaluate the adequacy of the proposed models as factorial invariance tests, we investigated the proportions of inadmissible solutions, true positive (TP) rates, and false positive (FP) rates throughout the studies. In addition, bias for the parameter estimate of interest (i.e., magnitude of noninvariance) was evaluated when the model was correctly specified. Due to space limits inadmissible solutions are reported in Study 1 only.

TP refers to the detection of noninvariance in testing weak factorial invariance when the factor loading is not invariant and the detection of the violation of strong invariance under intercept noninvariance conditions at $\alpha = .05$. Accordingly, the TP rate is defined as the proportion of replications in which the level of noninvariance is correctly determined. On the other hand, the FP rate is computed as the proportion of replications in which weak invariance is falsely rejected when factorial invariance holds in the true population.

Raw bias $B(\theta)$ and relative bias $RB(\theta)$ of parameter estimates are estimated as follows:

$$B(\theta) = R^{-1}\sum_{r=1}^{R}(\hat{\theta}_r - \theta) \text{ and } RB(\theta) = R^{-1}\sum_{r=1}^{R}\frac{(\hat{\theta}_r - \theta)}{\theta}$$

where $\hat{\theta}_r$ is the parameter estimate for replication *r*, θ denotes the population parameter, and *R* is the total number of replications. When raw bias is close to zero and relative bias is below .05 (Hoogland & Boomsma, 1998), the parameter estimate of interest is considered unbiased. The parameter estimate evaluated for bias is the size of noninvariance in the factor loading or the intercept of one DIF variable (y7). In ML FMM, the group difference in factor loadings (or intercept) of y7 corresponds to factor loading (or intercept) noninvariance.

Results

Inadmissible Solutions

Inadmissible solutions include cases in which there are error messages or the statistics of interest such as the AIC, BIC, or ssBIC are not produced in the output. The inadmissible solution rates of ML FMM were zero or near zero in many cases and the highest rate was 1.6% (16 out of 1000 replications) indicating that ML FMM generally found mathematically proper solutions.

The SB LR tests occasionally yielded negative likelihood ratios that required an adjustment proposed by Satorra and Bentler (2010) to ensure positiveness. The negative SB LR rates were quite small for intercept noninvariance (0% ~ 9%) or in complete invariance conditions (4% ~ 7%) of ML FMM. However, factor loading noninvariance conditions yielded sizable negative SB LR rates for both small DIF (8% ~ 26%) and large DIF (11% ~ 56%). In general, the rates were positively associated with the ICC. In this study the adjustment for positive SB LR (Asparouhov & Muthén, 2012a, Mplus webnote 12) was not applied[3]. Accordingly, it should be noted that the results of SB LR tests reported in the subsequent sections are based only on the cases with the positive SB LR.

True Positive and False Positive Rates

The FP rates varied considerably depending on fit statistics. The FP rates for the SB LR tests ranged between .08 and .17 above the nominal level across simulation conditions. For the AIC, the FP rates were unacceptably high with the minimum of .12 and maximum of .59. The FP rates for the AIC also depended on the ICC and cluster size in a positive direction. The BIC and ssBIC reasonably controlled the FP rates around or below .05. For the BIC, the FP rates were almost zero regardless of the ICC and sample size. The FP rates of the ssBIC were in general about .05, but showed slight inflation with large ICCs and large cluster size conditions (e.g., .16 when CN = 60, CS = 20 under large ICC).

The TP rates markedly depended on the size of noninvariance in relation to the location of noninvariance. When simulated noninvariance was large, the TP rates were moderate to high irrespective of fit indices whereas the power to detect the small size noninvariance dropped considerably across fit indices, particularly for factor loading noninvariance. When the noninvariance of factor loading was small, unacceptably low TP rates were observed for the BIC even with very decent sample sizes (e.g., TP = .13 when CN = 100, CS = 20 with medium ICC). In many small sample size conditions, the BIC showed near zero TP rates. The TP rates of the ssBIC also deteriorated notably when DIF was small for the factor loading in combination with small sample size. On the other hand, the TP rates of the AIC and SB LR were relatively high with small DIF for the factor loading. When the factor loading was noninvariant, the TP rates were slightly negatively related to the ICC for all fit statistics examined (for the SB LR, when CN = 60, CS = 10, and DIF = large, TP = .95 under small ICC; TP = .89 under large ICC). This slightly negative relation of the TP rates with the ICC was not seen in the intercept noninvariance conditions. Overall, the TP rates of intercept noninvariance were higher than those of factor loading noninvariance with less impact of DIF size. In summary, when noninvariance was large,

the BIC and ssBIC showed good performance with acceptable Type I error control and sufficient power. In testing strong invariance, the ssBIC is particularly recommended. However, to detect small size noninvariance of factor loadings, the SB LR and AIC are potentially better options. The TP and FP rates of ML FMM for the small ICC conditions are presented in Table 1[4].

Bias of Parameter Estimates

The estimated size of factor loading noninvariance appears biased in ML FMM (see online supplement Table 3s). The bias ranged from -.093 to -.033 for large DIF and from -.049 to -.015 for small DIF, which led to relative bias between -.195 and -.064 irrespective of DIF size. This magnitude of relative bias was above the cutoff in absolute value (i.e., .05). Two salient points emerged in factor loading noninvariance conditions. First, the size of factor loading noninvariance was consistently underestimated across simulation conditions. Second, underestimation became more serious as the ICC increased. For example, when noninvariance was simulated large in factor loading, the relative bias associated with a large ICC was between -.186 and -.154 as opposed to -.072 and -.065 with a small ICC. In contrast, the size of intercept noninvariance was unbiased regardless of simulation conditions.

Study 2: Testing Item-Level Factorial Invariance Using Multilevel MIMIC Model

Method

Item-Level Factorial Invariance Testing Procedure

ML MIMIC was applied to the data generated in Study 1 (except the complete invariance conditions) to identify any noninvariant variable at the item level. Instead of modeling a separate CFA for each group, ML MIMIC constructs a single model with the grouping variable as a covariate, assuming strict invariance between groups. This constrained model was compared to a relaxed model with two directional effects of a selected variable that tested the invariance of the

intercept and factor loading of the variable (Γ_y and $\Gamma_{\eta y}$ in Equation 11). These two competing models were evaluated with respect to the same model selection criteria employed in Study 1. The selection of the relaxed model with the covariate and interaction terms indicates the noninvariance of the variable. This process is repeated for all variables in the model. ML MIMIC with the factor and observed covariate interaction requires the use of the integration algorithm with robust maximum likelihood estimation.

Analysis of Simulation Results

The same set of simulation outcomes used for ML FMM was evaluated for ML MIMIC. However, the definitions of TP and FP rates in Study 2 were slightly different from those in Study 1 because factorial invariance was tested at the item level in Study 2. The TP rate is computed as the proportion of cases in which the simulated noninvariant variable is correctly identified as functioning differentially (i.e., DIF); the FP rate is defined as the proportion of cases in which any of invariant variables is falsely detected as DIF. It should be noted that the significance level was adjusted with a Bonferroni correction ($\alpha = .05/8 = .00625$) for the ML MIMIC analysis because the LR test was conducted for each variable (i.e., eight times per replication). Further, when FP rates were inflated due to baseline model misspecification in the LR tests, we considered Oort adjustment (1998). For bias, the estimates of two directional effects on y7 (Γ_y and $\Gamma_{\eta y}$ in Equation 11) were compared to the corresponding population parameters. The results of bias are similar to those of ML FMM and not reported in the paper (see Table 3s).

Results

True Positive and False Positive Rates

When there was noninvariance for the factor loading only, the FP rates were different across the fit statistics as presented in Table 2. For the BIC, the FP rates were all below .02. The ssBIC also controlled the FP rates close to zero except in a couple of conditions of large DIF (.09 for CN = 160, CS = 20, ICC = medium; .12 for CN = 160, CS = 20, ICC = large). In contrast, the SB LR and AIC frequently misidentified the invariant variables as noninvariant (FP = .12 ~ .30 for the SB LR; FP = .02 ~ .45 for the AIC). In addition, the behaviors of these two fit statistics were positively associated with the ICC, number of clusters, and cluster size. In summary, the FP rates of the BIC and ssBIC in detecting factor loading noninvariance were generally close to zero; those of the SB LR and AIC were overall inflated, especially in the large sample size and large ICC conditions. In regard to the TP rates, all fit statistics generally showed decent performance in correctly detecting a noninvariant variable (e.g., TP rates of the ssBIC were above .96 across all simulation conditions). Interestingly, the serious power reduction in the BIC and ssBIC with small size DIF was not obvious in item-level factorial invariance tests using ML MIMIC although the TP rates slightly decreased as the DIF size became smaller.

When there was noninvariance in the intercept, factorial invariance testing with ML MIMIC showed appreciable inflation in the FP rates (see Table 2 for the small ICC conditions and Table 2s for the complete conditions). All four fit statistics investigated in this study yielded inadequately high FP rates. For example, the FP rates of the AIC ranged from .40 to .88 for large intercept noninvariance; those of the BIC were between .06 and .56. Simulation factors – DIF size, ICC, number of clusters, and cluster size were positively related to the FP rates. Although the TP rates of all four fit statistics were fairly high in most conditions, ML MIMIC appeared inadequate for detecting noninvariance in the intercept with highly inflated FP rates. That is, an invariant variable is likely to be misidentified as DIF. However, high FP rates of the SB LR tests

in ML MIMIC are not surprising because it is reported that the equality-imposed baseline MIMIC model with DIF (namely, a misspecified baseline model) possibly leads to Type I error inflation in factorial invariance testing (Kim, Yoon, et al., 2012).

To control the Type I error inflation due to an inflated baseline model chi-square and inflated chi-square difference, we applied Oort-adjusted critical chi-square to the likelihood ratio tests when testing factorial invariance of the intercept using uniform ML MIMIC that does not include the interaction term[5]. The results of uniform ML MIMIC with Oort adjustment are presented in Table 2 for the small ICC conditions (see Table 2s for the complete conditions). As expected, the FP rates were controlled below or around .05 while the TP rates were sufficiently high across the ICC and sample size levels. To sum up, when nonuniform invariance is tested with a factor by observed variable interaction, the BIC and ssBIC could be reasonable indices in identifying a variable with factor loading noninvariance. Of the two measures, the ssBIC is more susceptible to Type I error when sample size is large (e.g., 3200) whereas power loss is substantial with the BIC when sample size is small (e.g., 600). In detecting the intercept noninvariance of a variable, the SB LR with Oort adjustment is strongly recommended given the Type I error inflation in ML MIMIC with the constrained baseline model.

Study 3: Multilevel Factor Mixture Model vs. Design-Based Multigroup CFA

Method

In Study 3, ML FMM was compared to design-based multigroup CFA (MG CFA) in terms of the detection of factorial noninvariance when between and within factor structures were different. Wu and Kwok (2012) observed bias and poor model fit of the design-based approach to multilevel data under complex (i.e., two factors) between and simple (i.e., one factor) within factor structures. Based on this finding, a one-factor model with eight indicators at the within

level and a two-factor model with four indicators each at the between level were simulated for two within-level groups. The correlation between the two factors at the between level was set at .30 (Wu & Kwok, 2012). All eight observed variables had .80 factor loadings with residual variances of .36 at the within level. A factor loading of .80 is considered moderately high with a factor variance of 1.00. The between-level factor loadings were 1.20 with residual variances 0 assuming no measurement error at the between level (e.g., Heck & Thomas, 2009; Jak et al., 2013; Ryu, 2014). Zero residual variance at the between level is also the default of ML FMM specification in Mplus 7 (Muthén & Muthén, 2012).

The simulation design factors of Study 1, with the exception of noninvariance size, were considered in Study 3 (3 ICC x 3 CN x 3 CS x 2 DIF location). Only small size noninvariance (i.e., .25) of the factor loading or intercept was simulated. A bigger cluster size of 50 was included in addition to 10 and 20. Complete invariance conditions (3 ICC x 3 CN x 3 CS) were also included to evaluate the FP rates. Per condition 500 replications were generated with Mplus 7 (Muthén & Muthén, 2012). For factorial invariance testing two data analytic models were employed: ML FMM and design-based MG CFA. Note that in design-based MG CFA, an overall one-factor model with eight indicators was specified ignoring the complex factor structure at the between level (a two-factor model with four indicators per factor). Weak (or strong) factorial invariance between two groups at the within level was evaluated by comparing two competing models using the SB LR tests, AIC, BIC, and ssBIC as explained in Study 1.

Results

True Positive and False Positive Rates

Compared to the design-based MG CFA, ML FMM showed slightly superior performance across simulation outcomes irrespective of design factors (see Table 3 for the CN =

60 conditions; see Tables 4s and 5s for the complete conditions). The FP rates of ML FMM were controlled about or below .05 regardless of simulation conditions for all fit statistics investigated. The FP rates of the BIC and ssBIC were zero for all conditions. Overall, no design factor appeared related to the FP rates for ML FMM. Additionally, the TP rates of the SB LR and AIC were exactly 1.00 in most simulation conditions. When design-based MG CFA was used for factorial invariance testing, the FP rates of the SB LR and AIC were somewhat inflated ranging from .06 to .15 and .05 to .12, respectively. Again, the BIC and ssBIC yielded zero or near zero FP rates across simulation conditions. The TP rates in detecting the violation of weak or strong invariance were fairly decent with the SB LR and AIC, but less optimal than those of ML FMM. Importantly, the negative impact of ICC on the performance of the design-based MG CFA was evident. As an example, for the AIC the FP rates increased from .07 to .12 and the TP rates of factor loading noninvariance decreased from .92 to .01 as the ICC increased from small to large when CN = 60 and CS = 10.

As observed in the small DIF conditions of Study 1, the BIC and ssBIC showed extremely low TP rates when sample size was small irrespective of the method used. The low power was more serious with the BIC than the ssBIC and with the design-based approach than ML FMM. For design-based MG CFA, the BIC showed poor performance even with large sample sizes over 3000, especially under large ICC conditions. In contrast, the TP rates of the SB LR tests were 1.00 in nearly all simulation conditions. The AIC with ML FMM also showed high TP rates of over .95 across simulation conditions. In summary, for Study 3 in which between-level residual variances were simulated at zero, all four fit indices of ML FMM were able to determine the level of factorial invariance well especially when sample size was sufficiently large (i.e., over 3000).

Bias of Parameter Estimates

Similar to the patterns observed in Study 1, the size of noninvariance in the factor loading was underestimated as a function of ICC although the degree of underestimation depended on the analytic method used as shown in Table 4 (see also Table 6s). For ML FMM the underestimation was not very serious and was below the cutoff of relative bias (.05; Hoogland & Boomsma, 1998) when the cluster size equaled 50. Relative bias of design-based MG CFA was generally higher than that of ML FMM. The association of relative bias with ICC was also apparent. That is, relative bias of design-based MG CFA was about -.06, -.13, and -.26 for small, medium, and large ICC, respectively. In contrast, bias in the size of intercept noninvariance was literally zero in both ML FMM and design-based MG CFA.

Demonstration

Factorial invariance testing is demonstrated with the PISA 2003 data (OECD, 2005). Five items measuring mathematics self-efficacy (see online supplement[6] Appendix Bs) were selected to fit ML FMM and ML MIMIC. There were 4270 students nested within 159 schools in Turkey. The number of students per school ranged from 2 to 35 (mean = 27, *SD* = 7.94). Gender was a within-level grouping variable (1855 females, 2415 males). This research setting is similar to the simulation condition of CN = 160 and CS = 20 although the total sample size is larger in the PISA data.

Multilevel Factor Mixture Model for Known Classes

ICCs of the observed variables ranged from .06 to .11. Three nested models were evaluated to determine weak and strong invariance (see Appendix Cs for the parameter estimates and fit indices of the three models). In the configural invariance model, factor loadings and residual variances at the within-level and intercepts at the between-level were freely estimated

between groups. For identification, the factor loading and intercept of the first variable were constrained equal across groups (see Appendix As for Mplus syntax). Next, within-level factor loadings were constrained to be the same for boys and girls to test weak invariance and in addition, between-level intercepts were constrained to be the same across groups to test strong invariance.

When evaluating factorial invariance between boys and girls, we applied the SB LR tests, AIC, BIC, and ssBIC. All four indices supported weak invariance but rejected strong invariance (Appendix Cs). Given large sample sizes over 3000 (i.e., 4270) and near zero residual variances at the between-level (obtained from a ML MIMIC without an interaction term or numerical integration), all four indices were informative of factorial invariance. Thus, we conclude that the PISA 2003 mathematics self-efficacy measure meets the weak invariance assumption but fails to meet strong invariance.

Multilevel MIMIC Model

For each item, a set of nested ML MIMIC models were compared to test factorial invariance. First, two direct effects on each observed variable were constructed to test factorial invariance of the intercept and factor loading (e.g., y1 regressed on gender and interaction). This model with two direct effects on each observed variable was compared to a model in which both effects were constrained to be zero assuming invariance of the intercept and factor loading of the variable (see Mplus syntax in Appendix As). Again the SB LR, AIC, BIC, and ssBIC were evaluated in model selection.

As presented in Appendix Ds, item 4 and item 5 were flagged as DIF by all four fit indices when we tested both the factor loading and intercept. Then, the factor loading and intercept were tested separately to locate the source of noninvariance for item 4 and item 5. For

factor loadings, all four indices were evaluated but the BIC and ssBIC received more weight based on the simulation results (high Type I error with the SB LR and AIC). Because (a) the SB LR, BIC, and ssBIC supported the invariance of item 4; the BIC and ssBIC supported invariance of item 5, and (b) the size of noninvariance of item 4 was 0.056; the size of noninvariance of item 5 was 0.127 (both smaller than what we simulated as small DIF [0.25]), we concluded that all five item factor loadings were invariant between boys and girls. For intercepts, both item 4 and item 5 were detected as noninvariant by all indices. However, we observed substantially high Type I error in detecting intercept noninvariance for all indices. Thus, we conducted regular ML MIMIC without the interaction and applied Oort adjustment to control Type I error inflation when conducting SB LR tests. The results supported the invariance of item 4 but item 5 was detected as DIF in the intercept, $\chi^2(1) = 22.62$ for item 4, $\chi^2(1) = 86.18$ for item 5 with the adjusted critical value of 47.38. In addition, the DIF size in the intercept of item 4 was 0.107 and that for item 5 was 0.202.

In summary, weak factorial invariance (i.e., equality of the factor loadings of the items) was supported. However, strong invariance was rejected. The item-level factorial invariance testing revealed that item 5 exhibited DIF with the noninvariance in the intercept with a magnitude of 0.202.

Discussion

Multilevel Factor Mixture Model

When ML FMM was utilized for within-level factorial invariance testing, the behaviors of the BIC and ssBIC were heavily influenced by the degree of noninvariance. With complete invariance, both BIC and ssBIC generally controlled the FP rates below the nominal level. In particular, the BIC showed an over-control of FP rates (e.g., .00 ~ .01 in ML FMM). When the

size of factor loading noninvariance was large, both fit statistics correctly detected the violation of weak or strong factorial invariance almost all the time. However, neither fit statistic was sensitive enough to detect small DIF involving the factor loading when the total sample size was small. This insensitivity to small DIF was more striking with the BIC where TP rates across simulation conditions were below 20% except for a sample size of 3200 (i.e., the largest sample size in Study 1). The BIC prefers a parsimonious model and penalizes a model with a larger number of free parameters (e.g., a relaxed model allowing noninvariance in factorial invariance testing). When amplified by sample size, the penalty of the BIC on the free estimation of small noninvariance appears excessive in comparison to the improvement in log likelihood due to free estimation. In Studies 1 and 3, when there was only one factor loading with small DIF, relaxing seven factor loadings ($\Delta df = 7$) did not improve the log likelihood appreciably while the penalty for seven degrees of freedom was severe. In ML MIMIC (Study 2) when testing the factor loading and intercept of a single variable ($\Delta df = 2$), the BIC and ssBIC showed decent TP rates for small DIF.

The SB LR tests in ML FMM showed reasonably high TP rates throughout the simulation conditions. However, two major issues should be addressed in the use of the SB LR for within-level factorial invariance testing. First, the FP rates were slightly inflated across the conditions of Study 1. The frequent occurrence of negative SB LR is also an issue. Importantly, the negative SB LR appears associated with the ICC. Of note is that the reported TP rates counted only positive SB LR cases in this study. Although a correction method to warrant the positiveness of the SB LR (Asparouhov & Muthén, 2012a; Satorra & Bentler, 2010) is available, this additional step for positive SB LR could be a demanding and challenging task for applied researchers in conducting factorial invariance testing. In terms of TP rates, the AIC was superior to other fit

statistics with high consistency across simulation conditions though unacceptably high FP rates should be taken into consideration in the use of the AIC.

The excellent performance of ML FMM in Study 3, especially for the SB LR and AIC, is worthy of note in comparison to the results of Study 1. In Study 3, the SB LR and AIC of ML FMM showed proper Type I error control and almost perfect power in detecting the violation of weak or strong factorial invariance throughout the simulation conditions. When simulation settings were compared between Studies 1 and 3, the better performance of ML FMM could be attributed to two prominent features: (a) equally high factor loadings for items especially at the between level (1.2 for all items in Study 3; from .3 to .9 in Study 1), and (b) relatedly, zero residual variance at the between level (.25 in Study 1). It is noted that ML FMM in the current version of Mplus does not allow residual variance of the indicators at the between level assuming perfect relations between the indicators and the latent factors. In Study 1, a considerable amount of residual variance was simulated at the between level with occasionally less ideal factor loadings. When these residual variances were fixed at zero, the performance of ML FMM in testing factorial invariance deteriorated to the extent of model misspecification. In summary, when the between-level factors and indicators are highly related with near zero residuals, ML FMM is expected to show its optimal behavior in detecting factorial noninvariance at the within level. However, substantial unexplained variance present at the between level could lead to a less favorable performance of ML FMM as a method of within-level factorial invariance testing.

Multilevel MIMIC Model

In detecting a noninvariant variable at the item level, ML MIMIC by and large exhibited moderate to high TP rates. However, the FP rates were substantially higher in identifying noninvariance of the intercept. All fit statistics observed in Study 2 including the BIC and ssBIC

failed to control the FP rates properly near the nominal level. This is consistent with the findings of previous studies on single-level MIMIC (e.g., Kim, Yoon, et al., 2012; Oort, 1998). Kim and colleagues explained the high FP rates in detecting intercept noninvariance using MIMIC as due in part to the likelihood ratio tests. MIMIC modeling inherently assumes complete invariance between groups by constructing a single model for both groups. When a MIMIC model has a noninvariant variable (i.e., misspecified), the model chi-square statistic is likely inflated due to model misspecification. In the subsequent likelihood ratio test with a misspecified baseline model, the chi-square difference is also inflated resulting in larger Type I error rates (Yuan & Bentler, 2004).

In such cases, statistical adjustment taking into account model misspecification at the baseline model, for example, the Oort adjustment is expected to reduce the Type I error inflation. However, this supposition was not directly tested for nonuniform ML MIMIC in this study. Alternatively, uniform ML MIMIC without the interaction between a factor and a grouping covariate was run with the same data, and the likelihood ratio was tested with the Oort adjusted chi-square critical value. The results were consistent with the literature: FP rates were controlled near zero irrespective of simulation conditions. Therefore, ML MIMIC can be used for item DIF detection across within-level groups when the chi-square inflation due to the misspecification of a constrained baseline model is properly adjusted. It should be noted that the free baseline approach in the LR test[7] is possible with ML MIMIC and is expected to control the FP rates reasonably well though an invariant reference item between groups should be selected first for identification. An iterative specification search (e.g., Yoon & Millsap, 2007) is also a viable option. These alternative approaches to Type I error control in MIMIC models need further investigation under the framework of multilevel modeling.

Of note is that ML MIMIC has great potential in factorial invariance testing with multilevel data due to its flexibility in modeling within- and between-level covariates. As demonstrated in this study, intercept invariance testing can be easily implemented at the within level by including a within-level grouping covariate. Furthermore, testing factorial invariance for both within- and between-level groups simultaneously will be possible with ML MIMIC whereas such model specification could be challenging in multiple group analysis. The advantages of MIMIC in multilevel modeling has been discussed and illustrated in general multilevel CFA contexts (e.g., Finch & French, 2011). Future research on ML MIMIC for factorial invariance testing under various research circumstances is called for.

In this study, ML FMM was used for the scale-level factorial invariance testing to establish either weak or strong factorial invariance. On the other hand, the performance of ML MIMIC was demonstrated for the item-level factorial invariance testing. However, both ML FMM and ML MIMIC can be used either to establish the level of factorial invariance (e.g., strong invariance) or to identify one or more noninvariant variables.

Multilevel Factor Mixture Model and Design-Based Multigroup CFA

In Study 3, the performance of ML FMM was compared to that of design-based MG CFA when the within model was simple with a single factor but the between model was complex with two factors. Although the behaviors of design-based MG CFA ignoring the complex factor structure at the between level were generally decent and not strikingly different from those of ML FMM, the MG CFA was obviously less optimal with higher FP rates across conditions and lower TP rates when the sample size was small. Moreover, the ICC emerged as a major factor associated with the lower performance of design-based MG CFA. For example, for bias estimates the size of factor loading noninvariance was substantially underestimated as a function

of ICC. When the ICC was large, the relative bias of design-based MG CFA was much higher (about -.26) than the cutoff of .05. Although not reported in the Results, the model fit of design-based MG CFA was usually unacceptable with the CFI below .90, RMSEA above .10, and extremely high chi-square relative to its degrees of freedom. Hence, for within-level factorial invariance testing in multilevel data with different factor structures at between- and within-levels, a multilevel approach that allows level-specific model specifications is recommended over a design-based approach.

Impact of ICC

Across the studies, ICC emerged as a factor associated with the performance of factorial invariance testing with multilevel data. The major simulation outcome related to ICC was bias or relative bias. Interestingly, the relation between ICC and bias depended on the location of noninvariance. When noninvariance was simulated for the factor loading, for both ML FMM and ML MIMIC a positive relation between bias and ICC was observed with the underestimation of factor loading noninvariance becoming more serious as ICC increased. Of particular concern is that even the large size of noninvariance of the factor loading was consistently underestimated. Furthermore, the underestimation of factor loading noninvariance possibly leads to lower TP rates in detecting the factor loading noninvariance as ICC increases. Intercept noninvariance bias was estimated near zero across ICC levels and the negative relation between the TP rates and ICC was not observed.

The estimation of factor loadings is based on the variance-covariance structure while intercepts are estimated on the basis of mean structure (Raykov et al., 2012). The negative relations of ICC with bias and TP rates were apparent only in the factor loading noninvariance conditions because ICC is the ratio of between-level factor variance to total factor variance,

which is derived from the variance-covariance structure. As ICC increased, that is, the proportion of within-level factor variance became smaller relative to total factor variance, it was observed that the factor loading noninvariance estimated at the within level was negatively biased showing lower TP rates. The relation between ICC and factor loading noninvariance needs further investigation especially when the noninvariance is simulated at the between level.

Software Requirement

For the current studies, Mplus was utilized to demonstrate and evaluate the two proposed methods for within-level factorial invariance testing. For alternative SEM software programs, several features are needed to implement the proposed methods. For ML FMM, latent categorical variables should be incorporated in at least two-level models. For ML MIMIC, the interaction between a latent variable and an observed variable is required to test nonuniform invariance. If these features are either readily available or can be implemented in the programs (along with adequate estimation methods), such SEM packages are expected to work for the proposed models to test within-level factorial invariance.

Limitations

Although the simulation conditions of this study were selected to reflect the characteristics of multilevel data common in education and the social sciences, certain conditions were intentionally restricted for the simplicity of discussion (for example, a single noninvariant variable, a single factor model at the within level, etc.). Thus, the findings of this simulation study are applicable to the simulation conditions that were included.

Asparouhov and Muthén (2012b) recently discussed multiple group analysis with multilevel data using Mplus. Specifically, when grouping occurs at the within level, they suggest a two-level factor mixture model with the knownclass option that allows *multiple* random effects

for multiple groups within cluster that can be *correlated*. In Mplus the multiple random effects of within-level groups and the correlation between them are specified at the between level as latent variables. We call for future research expanding the discussions of the use of multigroup multilevel analysis for factorial invariance testing that takes into account the correlation between groups within cluster.

Conclusions

Factorial invariance testing in multilevel data can be challenging due to model complexity. This study addressed modeling issues in factorial invariance testing with multilevel data and proposed two approaches when within-level groups are compared for invariance: multilevel factor mixture model and multilevel MIMIC model. Overall, the Monte Carlo simulation supported the adequacy of the proposed models. In practical evaluations of scale-level factorial invariance ML FMM, the BIC and ssBIC are recommended either (a) when small DIF is not of great concern or (b) when the total sample size is sufficiently large (over 3000 for the BIC; over 2000 for the ssBIC if the research situations are similar to the simulation settings). For strong factorial invariance testing, the ssBIC is particularly recommended. If the relations between indicators and latent factors are substantial with near zero residuals at the between level, the SB LR and AIC are expected to determine the level of factorial invariance with great precision regardless of DIF size. In using ML MIMIC, the invariance of factor loading of each variable (i.e., nonuniform invariance) can be tested by including the factor by group interaction in the model. In this case the BIC and ssBIC are optimal fit statistics in model comparison. However, when the intercept invariance of each variable is tested using ML MIMIC with a constrained baseline model in the LR test, uniform ML MIMIC with Oort adjusted LR tests is recommended.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317-322.

Asparouhov, T., & Muthén, B. (2008). Multilevel mixture models. In G. R. Hancock, & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 27-51). Charlotte, NC: Information Age Publishing.

Asparouhov, T., & Muthén, B. (2012a). *Computing the strictly positive Satorra-Bentler chi-square test in mplus* (Mplus web notes No. 12). Retrieved from http://statmodel.com/examples/webnotes/SB5.pdf

Asparouhov, T., & Muthén, B. (2012b). *Multiple group multilevel analysis* (Mplus web notes No. 16). Retrieved from http://statmodel.com/examples/webnotes/webnote16.pdf

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. doi: 10.1080/10705511.2012.713261

Chen, Q., Kwok, O., Luo, W., & Willson, V. L. (2010). The impact of ignoring a level of nesting structure in multilevel growth mixture models: A Monte Carlo study. *Structural Equation Modeling, 17*(4), 570-589. doi: 10.1080/10705511.2010.510046

Dedrick, R. F., & Greenbaum, P. E. (2011). Multilevel confirmatory factor analysis of a scale measuring interagency collaboration of children's mental health agencies. *Journal of Emotional and Behavioral Disorders, 19*(1), 27-40. doi: 10.1177/1063426610365879

Fan, X., & Sivo, S. A. (2009). Using ΔGoodness-of-fit indexes in assessing mean structure invariance. *Structural Equation Modeling, 16*(1), 54-69. doi: 10.1080/10705510802561311

Finch, W. H., & French, B. F. (2011). Estimation of MIMIC model parameters with multilevel data. *Structural Equation Modeling, 18*(2), 229-252. doi:10.1080/10705511.2011.557338

Heck, R. H., & Thomas, S. L. (2009). *An introduction to multilevel modeling techniques* (2nd ed.). New York, NY: Routledge.

Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling, 14*(2), 202-226. doi: 10.1080/10705510709336744

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. *Sociological Methods & Research, 26*(3), 329-367.

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling, 8*(2), 157-174.

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling, 20*(2), 265-282. doi:10.1080/10705511.2013.769392

Joreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and mulitple causes of a single latent variable. *Journal of the American Statistical Association, 70*(351), 631-639.

Kim, E. S., Kwok, O., & Yoon, M. (2012). Testing factorial invariance in multilevel data: A Monte Carlo study. *Structural Equation Modeling, 19*(2), 250-267. doi: 10.1080/10705511.2012.659623

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC. *Educational and Psychological Measurement, 72*(3), 469-492. doi: 10.1177/0013164411427395

Klein, A. G., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika, 65,* 457–474.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*(1), 21-39. doi: 10.1037/1082-989X.10.1.21

Millsap, R. E., & Kwok, O. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9*(1), 93-115.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*(4), 557-585.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*(3), 376-398.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*(1; ISSU 51), 81-118.

Muthén, B. O., & Asparouhov, T. (2003). Modeling interactions between latent and observed continuous variables using Maximum-Likelihood estimation in M*plus* (M*plus* Web Notes No. 6). Retrieved July 7, 2012, from http://www.statmodel.com/download/webnotes/webnote6.pdf

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

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25,* 267-316.

OECD (2005). *PISA 2003 technical report*. Paris, France: OECD.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling, 5*(2), 107-124.

Raykov, T., Marcoulides, G. A., & Li, C. (2012). Measurement invariance for latent constructs in multiple populations: A critical view and refocus. *Educational and Psychological Measurement, 72*(6), 954-974. doi: 10.1177/0013164412441607

Ryu, E. (2014). Factorial invariance in multilevel confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology, 67*, 172-194. doi:10.1111/bmsp.12014

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75*(2), 243-248.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Ed.), *Latent variables analysis: applications for developmental research* (pp. 399-419). Thousand Oaks, CA, US: Sage Publications, Inc.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), pp. 461-464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*(3), 333-343.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement, 35*(5), 339-361. doi: 10.1177/0146621611405984

Wu, J., & Kwok, O. (2012). Using SEM to analyze complex survey data: A comparison between design-based single-level and model-based multilevel approaches. *Structural Equation Modeling, 19*(1), 16-35. doi: 10.1080/10705511.2012.634703

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

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*(5), 737-757. doi: 10.1177/0013164404264853

Footnotes

[1] The first step for cross-level factorial invariance is to examine if the factor structure is configurally invariant across levels (i.e., the same number of factors and the same locations of zero and nonzero loadings). Once the factor structure is configurally invariant across levels, the invariance of factor loadings can be tested by comparing two models in which all factor loadings (except one for identification) are first relaxed for free estimation and then constrained equal across levels.

[2]Barendse, Oort, Werner, et al. (2012) introduced latent moderated structures (LMS; Klein & Moosbrugger, 2000) and random slope parameterization (RSP; Muthén & Asparouhov, 2003). The LMS allows the interaction effect between two latent variables on an observed variable. Because two latent variables are required for an interaction in LMS, for an observed covariate Barendse and colleagues created a latent variable on which the observed covariate loaded as a single indicator with a unit factor loading and near zero residual variance. In RSP, the interaction effect is estimated by specifying a random slope of the observed covariate which is in turn regressed on the latent factor.

[3]The correction method for positive SB LR demands considerable time and complexity in simulation (e.g., ensuring zero iteration in running a relaxed model with the baseline model parameter estimates) but the additional procedure is not expected to yield noticeably different results in the SB LR performance.

[4]The table of complete conditions (Table 1s) is available at http://www.coedu.usf.edu/main/faculty/ekim/supple_MLFI.html.

[5]Currently in the case of nonuniform ML MIMIC with numerical integration for the interaction between a factor and an observed variable Mplus does not produce sufficient

information for Oort adjustment which requires a baseline model chi-square statistic (see Oort, 1998, for the formula).

[6]http://www.coedu.usf.edu/main/faculty/ekim/supple_MLFI.html.

[7]Factor loadings and intercepts of all items but one are freely estimated and this free baseline model is subsequently compared to a model with the equality constraints on the factor loading and intercept of each item.

Table 1

True Positive (TP) and False Positive (FP) Rates of Multilevel Factor Mixture Model (Small ICC)

DIF size		**No (FP)**				**Small (TP)**								**Large (TP)**							
Location						**Factor loading**				**Intercept**				**Factor loading**				**Intercept**			
CN	**CS**	**SB**[a]	**AIC**	**BIC**	**SS**	**SB**[a]	**AIC**	**BIC**	**SS**	**SB**[a]	**AIC**	**BIC**	**SS**	**SB**[a]	**AIC**	**BIC**	**SS**	**SB**[a]	**AIC**	**BIC**	**SS**
60	**10**	.12	.13	.00	.01	.96	.49	.01	.09	.70	.95	.19	.78	.95	.98	.85	.97	.97	1.00	.99	1.00
	20	.13	.27	.00	.01	.95	.99	.05	.55	.69	1.00	.71	.95	.96	.99	.98	.99	.99	1.00	1.00	1.00
100	**10**	.10	.14	.00	.01	.98	.99	.04	.21	.88	.99	.40	.92	.97	.98	.96	.97	.97	1.00	1.00	1.00
	20	.10	.24	.01	.01	.98	.99	.13	.99	.91	1.00	.92	.99	.98	.99	.98	.99	.98	1.00	1.00	1.00
160	**10**	.08	.12	.00	.01	.99	.99	.07	.98	.95	1.00	.76	.99	.99	.99	.98	.99	.97	1.00	1.00	1.00
	20	.10	.26	.00	.00	.99	.99	.99	.99	.97	1.00	.99	1.00	.98	.99	.98	.98	.98	1.00	1.00	1.00

Note. DIF = differential item functioning or noninvariance, CN = number of clusters, CS = cluster size within a group, SB = Satorra-Bentler scaled likelihood ratio test, SS = sample-size adjusted BIC. [a]SB LR is computed only for the cases with positive Satorra-Bentler scaled likelihood ratio.

Table 2

True Positive and False Positive Rates of Multilevel MIMIC Model (Small ICC)

			Factor loading DIF								**Intercept DIF**									
			False positive				**True positive**				**False positive**					**True positive**				
DIF	**CN**	**CS**	**SB**[a]	**AIC**	**BIC**	**SS**	**SB**[a]	**AIC**	**BIC**	**SS**	**SB**[a]	**AIC**	**BIC**	**SS**	**Oort**[b]	**SB**[a]	**AIC**	**BIC**	**SS**	**Oort**[b]
Small	**60**	**10**	.14	.02	.00	.01	.90	.99	.77	.99	.08	.21	.01	.09	*.02*	.79	1.00	.87	.98	.90
		20	.13	.04	.00	.01	.90	.99	.99	.99	.08	.38	.06	.18	*.06*	.83	1.00	.97	.99	.95
	100	**10**	.16	.03	.01	.01	.97	.98	.98	.98	.09	.28	.02	.10	*.02*	.87	1.00	.99	1.00	.99
		20	.17	.07	.01	.01	.96	.99	.99	.99	.10	.48	.08	.21	*.04*	.90	1.00	1.00	1.00	1.00
	160	**10**	.19	.04	.00	.01	.98	.99	.99	.99	.13	.38	.03	.13	*.01*	.90	1.00	1.00	1.00	1.00
		20	.21	.11	.01	.01	.98	.99	.99	.99	.15	.58	.12	.27	*.04*	.94	1.00	1.00	1.00	1.00
Large	**60**	**10**	.14	.03	.00	.01	.91	1.00	1.00	1.00	.14	.40	.06	.23	*.01*	.93	1.00	1.00	1.00	*1.00*
		20	.14	.09	.01	.02	.91	1.00	1.00	1.00	.16	.59	.18	.37	*.03*	.95	1.00	1.00	1.00	*1.00*
	100	**10**	.25	.06	.01	.01	.94	1.00	.99	1.00	.20	.55	.10	.31	*.01*	.93	1.00	1.00	1.00	*1.00*
		20	.22	.17	.01	.03	.93	1.00	1.00	1.00	.24	.73	.29	.50	*.01*	.95	1.00	1.00	1.00	*1.00*
	160	**10**	.29	.10	.01	.02	.97	1.00	1.00	1.00	.29	.70	.19	.42	*.00*	.96	1.00	.99	.99	*1.00*

20	.33	.33	.01	.05	.97	1.00	1.00	1.00	.36	.85	.42	.62	*.00*	.96	1.00	1.00	1.00	*1.00*

Note. DIF = differential item functioning or noninvariance, CN = number of clusters, CS = cluster size within a group, SB = Satorra-Bentler scaled likelihood ratio test, SS = sample-size adjusted BIC. [a]SB LR is computed only for the cases with positive Satorra-Bentler scaled likelihood ratio. [b]Multilevel MIMIC model for uniform noninvariance (i.e., without the interaction term between a factor and an observed grouping variable) was conducted.

Table 3

True Positive and False Positive Rates of Multilevel Factor Mixture Model (ML FMM) and Design-Based Multigroup CFA (MG CFA) When Population Within and Between Factor Structures Are Different

			Complete invariance (FP)								**Factor loading noninvariance (TP)**							
			ML FMM				**MG CFA**				**ML FMM**				**MG CFA**			
ICC	**CN**	**CS**	**SB**[a]	**AIC**	**BIC**	**SS**	**SB**[a]	**AIC**	**BIC**	**SS**	**SB**[a]	**AIC**	**BIC**	**SS**	**SB**[a]	**AIC**	**BIC**	**SS**
Small	**60**	**10**	.06	.06	.00	.00	.07	.07	.00	.01	1.00	.99	.00	.08	1.00	.92	.00	.01
		20	.06	.04	.00	.00	.06	.06	.00	.00	1.00	1.00	.01	1.00	1.00	1.00	.00	.98
		50	.05	.05	.00	.00	.06	.05	.00	.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Med	**60**	**10**	.07	.06	.00	.00	.10	.09	.00	.03	1.00	.98	.00	.01	1.00	.06	.00	.00
		20	.09	.08	.00	.00	.11	.09	.00	.02	1.00	1.00	.00	1.00	1.00	1.00	.00	.06
		50	.07	.06	.00	.00	.10	.08	.00	.00	1.00	1.00	1.00	1.00	1.00	1.00	.58	1.00
Large	**60**	**10**	.05	.04	.00	.00	.15	.12	.00	.03	1.00	.95	.00	.01	.99	.01	.00	.00
		20	.07	.07	.00	.00	.12	.12	.00	.02	1.00	1.00	.00	1.00	1.00	.38	.00	.00
		50	.04	.03	.00	.00	.13	.10	.00	.00	1.00	1.00	1.00	1.00	1.00	1.00	.01	.36

Note. ICC = intraclass correlation, CN = number of clusters, CS = cluster size within a group, SB = Satorra-Bentler scaled likelihood ratio test, SS = sample-size adjusted BIC. [a]SB LR is computed only for the cases with positive Satorra-Bentler scaled likelihood ratio.

Table 4

Bias and Relative Bias of Factor Loading Noninvariance for Multilevel Factor Mixture (ML FMM) Model and Design-Based Multigroup CFA (MG CFA) When Population Within and Between Factor Structures Are Different

			ML FMM		MG CFA	
ICC	**CN**	**CS**	**Bias**	**Rel. bias**	**Bias**	**Rel. bias**
Small	**60**	**10**	-.021	-.086	-.015	-.061
		20	-.013	-.052	-.016	-.063
		50	-.006	-.024	-.015	-.061
Med	**60**	**10**	-.027	-.109	-.034	-.138
		20	-.014	-.058	-.034	-.134
		50	-.006	-.026	-.033	-.133
Large	**60**	**10**	-.031	-.123	-.067	-.266
		20	-.016	-.065	-.066	-.263
		50	-.007	-.027	-.065	-.260

Note. ICC = intraclass correlation, CN = number of clusters, CS = cluster size within a group, Rel. bias = relative bias. Bias and relative bias of intercept noninvariance are all zero.