June 2018

Covariates in Factor Mixture Modeling: Investigating Measurement Invariance across Unobserved Groups

Yan Wang
University of South Florida, Yan_Wang1@uml.edu

Covariates in Factor Mixture Modeling: Investigating Measurement Invariance across Unobserved Groups

by

Yan Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Curriculum and Instruction with a concentration in
Educational Measurement and Evaluation
Department of Educational and Psychological Studies
College of Education
University of South Florida

Major Professor: Eun Sook Kim, Ph.D.
John Ferron, Ph.D.
Robert Dedrick, Ph.D.
Tony Tan, Ed.D.
Stephen Stark, Ph.D.

Date of Approval:
May 16, 2018

Keywords: latent classes, measurement invariance, covariate effect, class enumeration

Acknowledgments

First, I would like to thank Dr. Kim for her mentoring and support during my doctoral journey, which has meant a lot to me. I especially appreciate how she has trained me to be an independent researcher. She has given me sufficient guidance so that I would not get lost, but she would also give me the freedom to explore the topic and find my own interest. She has set high standards for me, but always encourages me and supports me when I need her. I am also very grateful that she is someone I can trust one hundred percent and always gives me good advice when I share my personal concerns with her. I can never thank her enough for what she has done for me. She will be my mentor forever.

I would like to thank my committee members for their support and advice. Drs. Ferron and Dedrick always keep the office doors wide open and are willing to talk with me and provide good advice that would help me see the whole picture. Every time I walk out of their offices, I feel reassured and encouraged. I want to thank Dr. Ferron especially for his help with my job search, without which I could not have landed a job offer. I highly appreciate that Dr. Dedrick has helped me secure the assistantships for all five years including the summers. I enjoy working with Dr. Tan and Dr. Stark on research projects, which has helped me think about the substantive implications of my research.

My special thanks go to Dr. Kromrey and Dr. Chen. Although Dr. Kromrey has retired, the inspiration and motivation I have obtained through working with him will continue to guide me in my future career. Dr. Chen has always trusted me that I can achieve my goals and encouraged me when I feel not confident.

Last but not the least, I would like to thank my families and friends. Thank you, Mom and Dad, for encouraging me to study abroad and teaching me to be independent and considerate. Thank you, my husband, for always being proud of me and willing to comfort me when I need you. I want to thank all my friends who have listened to me patiently and cared for me genuinely.

Table of Contents

List of Tables

List of Figures

.

Abstract

Factor mixture modeling (FMM) has been increasingly used to investigate unobserved population heterogeneity. This Monte Carlo simulation study examined the issue of measurement invariance testing with FMM when there are covariate effects. Specifically, this study investigated the impact of excluding and misspecifying covariate effects on the class enumeration and measurement invariance testing with FMM. Data were generated based on three FMM models where the covariate had impact on the latent class membership only (population model 1), both the latent class membership and the factor (population model 2), and the latent class membership, the factor, and one item (population model 3). The number of latent classes was fixed at two. These two latent classes were distinguished by factor mean difference for conditions where measurement invariance held in the population, and by both factor mean difference and intercept differences across classes when measurement invariance did not hold in the population.

For each of the population models, different analysis models that excluded or misspecified covariate effects were fitted to data. Analyses consisted of two parts. First, for each analysis model, class enumeration rates were examined by comparing the fit of seven solutions: 1-class, 2-class configural, metric, and scalar, and 3-class configural, metric, and scalar. Second, assuming the correct solution was selected, the fit of analysis models with the same solution was compared to identify a best-fitting model. Results showed that completely excluding the covariate from the model (i.e., the unconditional model) would lead to under-extraction of latent classes, except when the class separation was large. Therefore, it is recommended to include covariate in FMM when the focus is to identify the number of latent classes and the level of invariance. Specifically, the covariate effect on the latent class membership can be included if there is no priori hypothesis regarding whether measurement invariance might hold or not. Then fit of this model can be compared with other models that included covariate effects in different ways

but with the same number of latent classes and the same level of invariance to identify a best-fitting model.

Chapter 1: Introduction

Over recent years, mixture modeling has been studied and used to investigate unobserved population heterogeneity, when the source of heterogeneity is not defined a priori (Lubke & Muthén, 2005; Tay, Newman, & Vermunt, 2011). Population heterogeneity is captured by latent classes. For example, in the field of psychopathology, there might be unobserved subtypes of the population that differ in anxiety sensitivity (Bernstein, Stickle, & Schmidt, 2013). In the educational context, students might be classified into latent classes of "masters" and "nonmasters", depending on how well they master the knowledge required in a cognitive test (Lubke & Muthén, 2005). Investigating unobserved heterogeneity might provide researchers with a more nuanced understanding of the population and design different interventions or curricula for different classes of individuals.

Increasing interest in the latent class approach to investigating population heterogeneity has originated from the downsides of the conventional approach that relies on manifest grouping variables, such as gender, race, ethnicity, and so on. These manifest grouping variables might be the surrogates for the true source of heterogeneity in how individuals respond to items (Cohen & Bolt, 2005). Even if the manifest grouping variable captures the population heterogeneity to some extent, homogeneity within the group is often assumed, which might not hold in reality. Samuelsen (2005) provided an example of within group heterogeneity about ethnicity. That is, the Hispanic group might come from different origins, such as Mexico, Cuba, Central or South America, and they might be of any race. Such heterogeneity within the group might indicate that, if using ethnicity as the manifest grouping variable, not all members of the Hispanic group will respond to items in the same way.

Due to these limitations of using manifest grouping variable to examine population heterogeneity, the latent class approach has been advocated (e.g., Samuelsen, 2005; Tay et al., 2011). It maximizes the differences across latent classes by identifying the underlying differences in how individuals respond to the items. Multiple covariates and covariate interactions could be incorporated as predictors of the latent class variable (Samuelsen, 2005). The substantive differences among latent classes can be further investigated in terms of their composition and other characterizations based on this set of covariates. Applications of this approach can be seen in many substantive studies on population heterogeneity using mixture modeling (e.g., Dimitrov, Al-Saud, & Alsadaawi, 2015; Dyer & Day, 2015).

Factor mixture modeling (FMM), one type of mixture modeling, can be used to investigate the unobserved population heterogeneity. FMM is a combination of confirmatory factor analysis (CFA) and latent class analysis (LCA), where the latent classes are distinguished by differences in measurement parameters and/or structural parameters across classes. FMM is analogous to the multiple group (MG) CFA and multiple-indicators multiple-causes (MIMIC) models with respect to the CFA part, and the major difference is that the grouping variable is latent in FMM. Similar to MG CFA and MIMIC models, measurement invariance testing can be conducted to examine whether items measure the factor equivalently across latent classes, prior to making any comparisons in the structural parameters (E. Kim, Joo, Lee, Wang, & Stark, 2016; Lubke & Muthén, 2005). This study aims to examine the performance of FMM in measurement invariance (MI) testing under a specific scenario, the presence of covariate effects.

The issue of covariates in FMM and mixture modeling in general has drawn the attention of many researchers. This might be attributable to the nature of the mixture modeling that once the classes are identified, further understanding of the latent classes is required, such as the characterization of classes, the cause of the class distinctiveness, and the potential consequences of being classified into a certain group. That is, the effect of covariates on the latent class membership and the effect of the latent class membership on the outcomes of interest are typically included in applications of FMM and other mixture models (e.g., Allan et al., 2014; Bernstein, Stickle, & Schmidt, 2013; Elhai, Naifeh, Forbes, Ractliffe, & Tamburrino, 2011). The exploration of covariate effects would help researchers better understand the connections between the latent class membership and some observed characteristics and behaviors of individuals. Recently, the possibility of some other covariate effects has been considered, such as the direct effects of the covariate on items that were used to identify classes in the latent class analysis (Asparouhov & Muthén, 2014; Nylund-Gibson & Masyn, 2016).

Methodological studies have investigated the impact of different approaches to specifying covariate effects in mixture modeling. For example, Lubke and Muthén (2007) suggested a one-step approach that includes the covariate when fitting the mixture models, where both the observed variables and the covariate are used to cluster observations into classes. This approach might improve the class enumeration and class assignment, as the incorporation of covariate might increase the class separation. The downside of this approach is that class enumeration and class assignment might change considerably when different covariates are included or different ways of specifying covariate effects are taken (Lubke & Muthén, 2005; Nylund-Gibson & Masyn, 2016). To prevent this change of classification, some researchers suggested excluding covariates in the latent class enumeration, i.e., specifying an unconditional mixture model, and including covariate effects in the subsequent steps. For example, based on the recently developed three-step procedure (Vermunt, 2010), latent class enumeration is done with the unconditional model (Step 1), and then observations are classified into one of the latent classes based on the most likely class membership (Step 2). Step 3 is to regress the class variable on covariates while taking into account the classification error. However, evidence from simulation studies has shown that when some covariate effects were ignored, bias would occur in class enumeration, class assignment, and the estimation of the covariate effect on the latent class membership (Asparouhov & Muthén, 2014; M. Kim, Vermunt, Bakk, Jaki, & Van Horn, 2016).

Although various ways to specifying covariate effects have been studied in the mixture modeling context, most simulation studies focused on LCA and class enumeration. Few studies have examined the issue of specifying covariate effects in the FMM context, specifically for measurement invariance (MI) testing. Therefore, this study aims to fill the methodological gap and investigates how the MI testing with FMM will be affected by different covariate inclusion strategies, which encompasses the exclusion of covariates and the misspecification of covariate effects. Data will be simulated under various conditions, including types of covariate effects (effect on the latent class membership, both the latent class membership and the factor, and the latent class membership, the factor, and the item), strengths of covariate effects (1 or 2 for effect on the latent class membership, .4 or .8 for effect on the factor and the item), magnitudes of measurement noninvariance (.4, .8, or 1.2 intercept difference), number of items with measurement noninvariance (1 or 2), mixing proportions (balanced or unbalanced), and sample sizes (500 or 2000). Through systematic evaluations of the covariate inclusion on MI testing with FMM, this simulation study aims to provide applied researchers guidelines and implications on whether the covariate should be included in the MI testing and if yes, how to include the covariate effects.

Chapter 2: Literature Review

The literature review part consists of three sections: measurement invariance (MI) testing in the confirmatory factor analysis (CFA) framework, factor mixture modeling (FMM) specification and MI testing with FMM, and the particular issue of covariate inclusion for MI testing with FMM.

Measurement Invariance (MI) Testing in CFA

Introduction of MI. When items are used to measure a latent construct (e.g., depression), MI testing is conducted to investigate if items measure the construct in the same way across subpopulations (e.g., males and females). The establishment of MI provides evidence for construct validity, while the violation of MI indicates potential measurement bias, which might lead to inaccurate interpretations of results. Specifically, MI refers to the fact that individuals who are at the same level of the construct/factor have the same probability to endorse an item, regardless of the subpopulations they belong to (Meredith, 1993). That is,

$$P(Y|\eta, G) = P(Y|\eta), \tag{1}$$

where Y is the observed item score, η is the factor score, and G indicates the group membership. Typically, MI testing is first conducted as an omnibus test where Equation 1 holds across all items and all groups. If MI is violated, item-level analysis can be conducted to identify items that function differently across groups (i.e., differential item functioning, or DIF items).

In the CFA framework, multiple group confirmatory factor analysis (MG CFA) has been commonly used to test MI. It is conducted by comparing a series of models with increasing restrictions of parameter invariance across groups. A typical MG CFA model can be expressed as:

$$\mathrm{Y_{ig}} = \nu_g + \Lambda_g \eta_{ig} + \varepsilon_{ig}, \tag{2}$$

where the response vector (Y_{ig}) of an individual i in group g (g = 1, 2, …, G) is a function of the intercept vector v_g, factor loading Λ_g, the vector of individual's factor scores η_{ig}, and the residual ε_{ig}. The configural invariance refers to the fact that the same factor structure (i.e., number of factors and the corresponding items) applies to all groups, but model parameters are freely estimated across groups (as expressed in Equation 2). The fit of the configural invariance model is checked first to ensure that the same factor structure does fit data of each group well. Building on the configural invariance model, the metric invariance model adds additional constraints on the equality of factor loadings (Λ) across groups, that is, $\Lambda_1 = \Lambda_2 = \Lambda_3 = \cdots = \Lambda_G$. Then the metric invariance model is compared with the configural invariance model. If the additional constraints on the equality of factor loadings do not worsen the model fit substantially, the metric invariance holds. In addition to both the factor structure and factor loadings, item intercepts (v) are constrained to be equal across groups ($v_1 = v_2 = v_3 = \cdots = v_G$) in a scalar invariance model. Likewise, scalar invariance can be established by no significant deterioration of model fit as compared with the metric invariance model. Note that on top of these restrictions, the equality of item residual variances can be included, which leads to the strict invariance model. This model is not included here because the establishment of scalar invariance has been considered as a prerequisite for factor mean comparisons, which is often the focus of substantive research (E. Kim, 2011; Meredith, 1993).

The fit of models with different levels of invariance can be evaluated by the chi-square goodness of fit and several model fit indices, such as the comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean squared residual (SRMR). Recommended cutoff values of those indices for good model fit are statistically nonsignificant chi-square ($p \geq .05$), CFI ≥ .95, RMSEA ≤ .06, and SRMR ≤ .08 (Hu & Bentler, 1999). Model comparisons are evaluated by the likelihood ratio test (LRT; or the chi-square difference test), and changes in model fit indices, such as the CFI and RMSEA. The LRT statistic is given by:

$$LR = -2(LL_0 - LL_1), \tag{3}$$

where LL_0 and LL_1 refers to the log-likelihood for the null and alternative models, respectively. The test statistic LR follows a chi-square distribution with degrees of freedom (*df*s) being the differences in the *df*s of the two models. The more constrained model (e.g., metric invariance versus configural invariance) is supported based on a nonsignificant LRT statistic ($p \geq .05$), ΔCFI ≤ .01, and ΔRMSEA ≤ .015 (Chen, 2008; Cheung & Rensvold, 2002).

Inclusion of covariates in MI testing. In addition to MG CFA, the multiple-indicators multiple-causes (MIMIC) model is another commonly used MI testing method within the typical CFA framework. Unlike MG CFA where the measurement model is built for each group and the invariance constraints can be imposed across groups, a single CFA model is fitted for the population and the group membership is included as a covariate in the MI testing. A generic MIMIC model for MI testing is shown in Figure 1. To test factor loading noninvariance (i.e., nonuniform DIF) of a particular item, the item is regressed on the interaction between the factor and the grouping variable ($\eta * G$). To test the intercept noninvariance (uniform DIF), the item is regressed on the grouping variable *G* (E. Kim, Yoon, & Lee, 2012; Woods & Grimm, 2011). Therefore using the two-group scenario as an illustration, the full model with both paths can be written as:

$$Y_{ij} = \lambda_j \eta_i + \beta_j G_i + \omega_j \eta_i * G_i + \varepsilon_{ij}, \tag{4}$$

$$\eta_i = \gamma G_i + \zeta_i, \tag{5}$$

where G_i is the dummy-coded grouping variable as a covariate in the MIMIC model (e.g., 0 = the reference group, 1 = the focal group); β_j denotes the effect of the covariate on the intercept of the *j*th item, i.e., intercept difference between two groups; and ω_j refers to the covariate effect on the factor loading of the *j*th item, that is, the factor loading difference between two groups.

MI testing (or DIF item detection) can be conducted as comparing metric and configural invariance models, and scalar and metric invariance models, respectively. The configural invariance model can be expressed as Equation 4 where differences in factor loadings and intercepts between groups are estimated. In the metric invariance model, the path coefficient ω_j is constrained to be zero, because factor loadings are equal across these groups. For the scalar invariance model, because both factor loadings and intercepts are equal across groups, both ω_j and β_j would be zero. If the scalar invariance can be established, factor means can be compared that γ in Equation 5 indicates the factor mean difference between groups. Similar to MG CFA, the configural, metric, and scalar invariance models can be compared sequentially based on the LRT and/or ΔCFI, ΔRMSEA. Another option available for the MIMIC modeling is the Wald z test that tests whether a single parameter is zero. The *p*-value of the test statistic is associated with each of the path coefficients, ω_j and β_j. A nonsignificant *p*-value indicates the absence of the covariate effect and thus the corresponding level of invariance holds.

A major advantage of the MIMIC modeling over MG CFA is the flexibility in accommodating covariates (E. Kim et al. 2011; Marsh, Tracey, & Craven, 2006). For example, covariates in the MIMIC modeling can be either categorical or continuous (e.g., age). In the MIMIC model, multiple covariates and/or the interactions among covariates can be included. In addition, for MG CFA sample sizes for each group need to be reasonable large. Testing DIG using a single group using MIMIC might not require the same number of cases.

Factor Mixture Modeling (FMM)

This section will present the specification of FMM first, followed by procedures of MI testing with FMM.

Specification of FMM. FMM (see Figure 2) is a combination of CFA and latent class analysis (LCA) (Lubke & Muthén, 2005). LCA allows us to model and capture the unobserved population heterogeneity in measurement and/or structural parameters through the latent class variable. The CFA part models the variability across individuals within each latent class, with the factor score quantifying the variability. The CFA part of the FMM can be extended from MG CFA where the observed group membership g is replaced by a latent class membership indicator *k*:

$$y_{ik} = \nu_k + \Lambda_k \eta_{ik} + \varepsilon_{ik}, \tag{6}$$

where *k* can take on values from 1, 2, …, *K*, where *K* is the number of latent classes. The subscript *k* is attached to parameters in Equation 6, indicating that parameters can vary across latent classes. Of note is that in MI testing, the measurement parameters (ν_k and Λ_k) can be constrained to indicate a certain level of invariance, which will be discussed shortly. The homogeneity within each class is assumed that the same measurement parameters, such as intercepts and factor loadings, applies to all individuals within the class. Residuals (ε_{ik}) are assumed to be normally distributed with a mean of zero and variance of Θ_k. It is also assumed that $\eta_{ik} \sim N(\alpha_k, \Phi_k)$. Thus the class-specific mean vectors and class-specific variance-covariance matrices can be expressed as:

$$\mu_k = \nu_k + \Lambda_k \alpha_k, \tag{7}$$

$$\Sigma_k = \Lambda_k \Phi_k \Lambda_k' + \Theta_k. \tag{8}$$

The structural part of the FMM model is achieved by regressing η_i on the latent class variable C_{ik}:

$$\eta_{ik} = AC_{ik} + \xi_{ik}. \tag{9}$$

C_{ik} is a multinomial variable with *K* – 1 categories: C_{ik} = 1 if individual *i* belongs to class *k*, and $C_{ik} = 0$ if individual belongs to a reference class. Similar to the MIMIC model, *A* is the vector containing the factor mean differences between the reference class and every other classes. Of note is that latent class membership is exclusive in a sense that individuals belong to one and only one latent class. Individuals are not classified into two or more classes. However, the probability of belonging to each of the latent classes is estimated through a multinomial regression model:

$$\ln\left[\frac{P(C_i = k|X_i)}{P(C_i = r|X_i)}\right] = \lambda_{ck} + \Gamma_{ck} X_i. \tag{10}$$

The left side of the equation denotes the log odds of the probability of belonging to a particular class *k* over that of belonging to a reference class *r*, given a vector of covariates X_i. The log odds can be predicted by covariates, with Γ_{ck} denoting the regression coefficients and λ_{ck} denoting intercepts. The subscript *c* indicates that the regression coefficients and intercepts are associated with the multinomial regression for the latent class variable, C_i. The subscript *k* indicates that both Γ_{ck} and λ_{ck} can vary across classes, that is, the relationship between covariates and the log odds of being assigned to a certain class can be class-specific. Note that the effect of covariates on the latent class membership is typically included in applications of FMM (e.g., Allan et al., 2014; Bernstein et al., 2013; Elhai et al., 2011). The rationale for this covariate inclusion is that researchers are often interested in understanding how to characterize classes and what causes the distinctiveness of classes. Individuals are assigned to a class based on their highest posterior probability of latent class membership. For the majority of the FMM applications, the number of classes is not known a priori and is determined by comparing FMMs with varying numbers of classes.

Testing MI across latent classes with FMM. Similar to MG CFA and MIMIC models, MI can be tested across subpopulations in FMM, to examine whether items measure the factor equivalently across latent classes, prior to making any comparisons in the structural parameters (E. Kim et al., 2016; Lubke & Muthén, 2005). That is, in the configural invariance model the same factor structure is fitted across latent classes but factor loadings and intercepts are free to vary, which is expressed by Equation 6. In the metric invariance model, factor loadings are equal across latent classes, suggesting $y_{ik} = \nu_k + \Lambda\eta_{ik} + \varepsilon_{ik}$. An additional constraint of intercept equality across latent classes is imposed in the scalar invariance model, $y_{ik} = \nu + \Lambda\eta_{ik} + \varepsilon_{ik}$. Similar to MG CFA, LRT, ΔCFI, and ΔRMSEA can be used in model comparisons. If a certain level of invariance is violated, item-level DIF analysis can be conducted based on the LRT.

However, the MI testing procedures described above might not be applicable, when the optimal number of classes is not known a priori. The task of determining the optimal number of classes will be entangled with the MI testing. In other words, should we determine the number of classes first and then test MI across classes? Or should we investigate both simultaneously? Both approaches have been adopted in the methodological and substantive literature (e.g., Clark et al., 2013; E. Kim et al., 2016; E. Kim, Cao, Wang, & Nguyen, 2017; Lubke & Neale, 2008; Tay et al., 2011). For example, both the E. Kim et al. (2016) and Tay et al. (2011) studies determined the number of classes first and then test MI across the classes. When fitting FMMs with varying numbers of classes, the measurement (loadings, intercepts/thresholds) and structural parameters (factor means, factor variances/covariances) were class-specific. In other words, the configural invariance model was specified in determining the number of classes. Once the optimal number of classes was determined, MI was tested across the classes.

By contrast, Clark et al. (2013) fitted a series of FMM models to determine the number of classes and the level of invariance simultaneously. They took a more exploratory approach to FMM that FMM models fitted depended upon the LCA and EFA analyses that were conducted first. Because results from LCA and EFA showed that the three-class model and the one-factor model fitted data well, respectively, FMM models constructed included one- to three-class one-factor model. Lubke and Neale (2008) took the exploratory approach as well, but they examined the number of classes, the level of invariance, together with the number of factors simultaneously. The approach Lubke and Muthén (2005) took in their demonstration was more confirmatory FMM where the number of factors was fixed at two, and only determining the number of classes and testing MI was conducted simultaneously. Similarly, E. Kim et al. (2017) also suggested this simultaneous approach which compared models included 1-class, 2-class configural, 2-class metric, 2-class scalar, 3-class configural, 3-class metric, 3-class scalar, and so on.

Note that no simulation study has been conducted to compare the sequential and the simultaneous approaches to MI testing with FMM. Intuitively, the simultaneous approach might have the advantage that model comparisons are more comprehensive, including possible combinations of the number of latent classes and the level of invariance, which might lead to more accurate results in class enumeration and MI testing. However, involving more model comparisons might be time-intensive. In comparison, the sequential approach does not involve many model comparisons. It might meet the needs of applied researchers better if typology and MI are separate interests. However, for the sequential approach, results of MI testing might be impacted greatly by the class enumeration. If class enumeration is incorrect, MI testing would be problematic.

Depending on the approach to testing the number of latent classes and testing MI, model comparisons will be different. If the sequential approach is taken, models with different numbers of classes can be compared with bootstrap LRT (BLRT), Lo-Mendell-Rubin (LMR) test, ad hoc adjusted

LMR (aLMR) and information criteria (IC), which will be discussed below. Then to investigate the level of MI that holds, LRT, ΔCFI, and ΔRMSEA can be used. If the simultaneous approach is taken, BLRT, LMR, aLMR and IC can be used in model comparisons.

Class enumeration. Likelihood-based tests and IC have been used to compare the fit of FMMs with different numbers of classes. Although the regular LRT test (see Equation 4) has been commonly used in model comparisons in structural equation modeling (SEM), it cannot be applied to the context of FMM or mixture modeling in general (Lubke & Muthén, 2005; McLachlan & Peel, 2000). That is, when testing the *k*-class model (null hypothesis) against the model with *k*+1 classes (alternative hypothesis), the *k*-class model can be specified by one of the class proportions being zero. Thus, due to the problems of the true parameter under the null hypothesis being on the space boundary (proportions ranging from 0 to 1) and parameter non-identification, the regularity conditions do not hold. In this case, the LRT statistic does not follow the asymptotic chi-square distribution.

Alternative LRT tests have been developed and applied in model comparison in the mixture modeling context (including FMM), such as LMR (Lo, Mendell, & Rubin, 2001), the aLMR (Lo et al. 2001), and BLRT (McLachlan & Peel, 2000). The LMR test adopts an approximation of the sampling distribution of LR and compares models with *k* and *k*-1 classes, favoring the *k* class model if rejecting the null hypothesis (*p*-value below the nominal alpha level). The aLMR incorporates a correction term to the test statistic to improve the accuracy of LMR. The BLRT generates an empirical sampling distribution for LR by drawing many bootstrap samples and computing LR for the null and alternative models. This empirical sampling distribution is then used to estimate *p*-value for the observed LR statistic.

In addition to these likelihood-based tests, ICs are also commonly used in model comparisons, including Akaike information criterion (AIC; Akaike, 1974), consistent AIC (cAIC; Bozdogan, 1987), Bayesian information criterion (BIC; Schwarz, 1978), and sample size adjusted BIC (saBIC; Sclove, 1987). These ICs are based on the log-likelihood of the model and penalizes the model for the number of free parameters, sample size, or both. Smaller values of ICs indicate a better model fit. Entropy has also been used in model selection as a classification-based measure (E. Kim et al., 2016; Tein, Coxe, & Cham, 2013; Ramaswamy, DeSarbo, Reibstein, & Robinson, 1993). The classification is evaluated at the individual level and entropy indexes the aggregated classification accuracy. Values of entropy are bounded by 0 and 1. Higher entropy indicates that the posterior probabilities are very distinctive across classes, classes are well separated, and classification of individuals is more accurate. Therefore, a model with a higher value of entropy is more desirable.

Regarding the performance of various measures for model selection, there has been inconsistency among the methodological literature primarily due to different model specifications and simulation conditions. For example, BLRT was shown to have greater power than other likelihood-based tests across all sample sizes (Nylund, Asparouhov, & Muthén, 2007). E. Kim et al. (2016) found that BLRT, as well as AIC, outperformed other model selection methods when class separation was small or sample size was small. However, two concerns have been raised about BLRT: the long execution time and possible model nonconvergence (E. Kim et al., 2016; Nylund et al., 2007). Overextraction of classes was observed for AIC, especially when sample size was larger (Cho & Cohen, 2010; Henson, Reise, & Kim, 2007; Lukočienė & Vermunt, 2010; Nylund et al., 2016). BIC and SaBIC have been suggested for model selection, particularly with relatively large sample sizes (E. Kim et al., 2016; Nylund et al., 2007; Tein et al., 2013). Entropy has been shown to be an unreliable measure for model selection (Tein et al., 2013).

The Inclusion of Covariates in MI Testing with FMM

A major issue particularly to mixture modeling that perplexes the MI testing in FMM is the inclusion of covariates. The next section will be devoted to discussing this issue in detail. Specifically, different ways to specify covariate effects in FMM will be described, followed by discussions on the impact of misspecifying covariate effects in FMM and mixture modeling in general. This section will conclude with a review of studies that examined the issue of covariates in MI testing with FMM.

Specification of FMM with covariate(s). Possible covariate effects in FMM are illustrated in Figure 3. Path 1 (see Equation 10), denoting the covariate effect on the latent class membership, was most frequently examined in substantive studies. Path 2 ($\gamma_{\eta k}$) refers to the covariate effect on the factor mean, similar to the MIMIC model. This indicates factor mean differences across the levels of the covariate within each latent class. The factor mean can thus be expressed as:

$$\eta_{ik} = AC_{ik} + \gamma_{\eta k}X_i + \zeta_{ik}. \tag{11}$$

Note that the covariate effect $\gamma_{\eta k}$ can be class-invariant or class-specific. If it is class-invariant, the effect of covariate on the factor mean is the same across all latent classes. If it is class-specific, the covariate impacts the factor mean differently across classes. In other words, there is an interaction between the covariate and the latent class.

Paths 3 and 4 represent covariate effects on the measurement model. Similar to the logic of the MIMIC model, the covariate can have a direct effect on the observed variable (Path 3), indicating that some of the within-class variation in that observed variable can be explained by the covariate. The covariate can have an effect on the factor loadings as well (Path 4), creating the interaction between the latent factor and the covariate. That is,

$$y_{ik} = \nu_k + \Lambda_k\eta_{ik} + \beta_k X_i + \omega_k\eta_i * X_i + \varepsilon_{ik}. \tag{12}$$

Note that β_k and ω_k denote the covariate effect on the intercepts and factor loadings, respectively. These two path coefficients indicate measurement noninvariance with respect to the covariate within latent classes. For example, if the covariate is gender and β_k and ω_k are substantial across two latent classes, it suggests that within each latent class, males and females respond differently to the item. Note that β_k and ω_k can be class-invariant or class-specific. If class-invariant, the direction and magnitude of DIF are the same across latent classes. If class-specific, item might favor one group in one latent class, but not the other in the other latent class; or it favors one group to different extents across latent classes.

De Ayala, Kim, Stapleton, and Dayton (2002) demonstrated this possibility of class-specific DIF. Class 1 was dominated by Black examinees (67.56%), and the majority of Class 2 were White examinees (80.97%). DIF analysis (using the Mantel-Haenszel method) within each latent class showed that 22 out of 50 items displayed DIF against Black examinees in both classes, while some items were biased against Black examinees in Class 1, but not in Class 2. Tay et al. (2011) also illustrated that gender DIF was only shown in one class which was dominated by individuals with relatively more work experience. Gender DIF was not displayed for the other class dominated by those with limited work experience.

Please note that although the covariate effect can be present with respect to the residual variance of observed variables, it is not considered in this study, because scalar invariance (equality in factor loadings and intercepts) is sufficient for comparisons of structural parameters. Additionally, although including distal outcomes of the latent class membership can be of interest (e.g., Lanza, Tan, & Bray, 2013), this is beyond the scope of the study and thus will not be illustrated here.

The impact of misspecified covariate effects. The inclusion of covariates has been a long-standing issue in the general framework of mixture modeling, not particularly for MI testing with FMM. The impact of covariate effect misspecification on the class enumeration and parameter estimates of the mixture modeling has been investigated. A summary of the relevant literature is provided in Table 1.

Nylund-Gibson and Masyn (2016) focused on the impact of misspecifying covariate effects on the class enumeration in latent class analysis through a simulation study. They observed over-extraction of latent classes when the covariate effect was misspecified. That is, three classes were extracted rather than two, when only the covariate effect on the latent class membership was specified, while the true effects included an additional effect on the item (indicating uniform DIF within latent classes). Such over-extraction was more severe with larger sample size (1000 versus 500), and unbalanced class sizes (80-20 versus 50-50 split). BIC performed remarkably better than LMR and BLRT and only BIC correctly detected two latent classes with sample size of 500. More severe over-extraction was evidenced when the misspecification of covariate effects was more substantial, such as only specifying the covariate effect on latent class membership with true covariate effects on latent class membership, item, and the path from the factor to the item. Note that this last path indicates nonuniform DIF within latent classes. Even BIC failed to detect the correct number of classes when the true covariate effect on the item was ignored and the effect on latent class membership was estimated instead. However, the unconditional model omitting the covariates (as Step 1 in the three-step procedure) could still extract the correct number of classes, regardless of the true relationship between the covariates and latent class model parameters in the population model. One exception was when there was more severe violations of the local independence assumption underlying LCA, i.e., two out of five observed variables had direct relationships with the covariates. Therefore, authors suggested the three-step procedure based on results of their simulation study.

However, Asparouhov and Muthén (2014) found that the three-step procedure would be contaminated if the direct covariate effects on items were ignored in the latent class analysis. Specifically, ignoring the direct effects on the item tended to overestimate the relationship between the covariate and the latent class membership. Overestimation became more substantial as the number of omitted direct effects on items increased and/or the entropy became smaller. Therefore the largest absolute bias (.76) was observed when five covariate effects on items (out of 10 items) were ignored and the entropy was .6. When the direct covariate effects on items were modeled, the three-step procedure performed much better, but the relationship between the covariate and the latent class membership would still be overestimated when the entropy was .6. They argued that the three-step procedure has the flaw that the latent class membership could not be well measured because the classification model excludes the covariate. This can be problematic because if there are direct covariate effects on two or more items, these items would be correlated through the covariate, while LCA assumes items are independent given the latent class membership. In this case, the formation of latent classes and the accuracy of classification might be affected, and the prediction model might have bias in the estimation of covariate effect on the latent class membership. Even if the direct effects on items were included, the classification might still be incorrect, though to a much smaller degree than ignoring the direct effects. This is because the classification model does not include the effect of covariate on the latent class membership, and such effect can be absorbed by the covariate effects on items, which might lead to inaccurate estimation of the parameters in the classification model. In comparison, the one-step approach performed very well in estimating the relationship between the covariate and the latent class membership, with the covariate effects on items and the latent class membership correctly modeled.

Consistent with the findings of the Asparouhov and Muthén (2014) study, Lubke and Muthén (2007) also observed the benefit of including covariate when estimating the latent class membership. That is, when the covariate had a medium to large effect on the class membership, including such covariate effect would improve the correct class assignment and the coverage of factor mean differences even if with small class separation. Aligned with this approach, Curran, Cole, Bauer, Hussong, and Gottfredson (2016) also demonstrated in their simulation study that including covariates in a single-factor measurement model would improve the accuracy and precision of factor score estimation. Although it is still questionable whether this finding in the common factor analysis framework can be confirmed in mixture models, the potential impact of including background covariates on the model parameter estimates was evidenced. Interestingly, they found that factor scores were estimated accurately when only the covariate effect on the factor was specified, while the correct specification should be covariate effects on both the factor and observed variables.

The impact of omitting covariate effects was also examined in regression mixture models by M. Kim et al. (2016). They studied the efficacy of the one-step and the three-step approaches when the covariate effect on the outcome variable was ignored and only the regression of the outcome variable on a predictor was modeled. They noticed that when ignoring this covariate effect, substantial bias in the class enumeration occurred in the one-step approach where the covariate was included to predict the latent class membership; the class enumeration remained accurate in the three-step approach. However, ignoring the same covariate effect in Step 1 of the three-step approach, the covariate effect on the class membership was not estimated adequately. One exception occurred when the covariate effect on the outcome was zero and the covariate had zero correlation with the predictor. In the one-step approach, even if the covariate effect on the outcome was modeled, the three-step approach showed biased estimates of the covariate effect on the latent class membership most of the time. This limited ability of the three-step approach with the presence of the covariate effect on the outcome variable seemed to confirm what Asparouhov and Muthén (2014) speculated in their simulation study. That is, the exclusion of the covariate effect on the latent class membership in the classification model might lead to biased estimates of the model and inaccurate class enumeration.

The inclusion of covariates in MI testing with FMM. Different approaches with respect to including covariates were also found in the methodological literature on MI testing and DIF (differential item functioning) detection across latent classes (see Table 2 for a summary of the literature). Distinctions among these approaches were even more nuanced than those discussed above (e.g., LCA), because the model complexity under FMM increased with a CFA measurement model. Considerations of the covariate effects were not limited to items, but also the factor measured by items. For instance, Lubke and Muthén (2005) demonstrated using factor mixture modeling to investigate measurement invariance across classes with two covariates included at the onset. Both the factor and the class variable were regressed on covariates, thus estimating the variations in the factor scores and between-class variations due to the covariate effects. An increasing number of classes was specified, with varying levels of class-invariant restrictions imposed on the intercepts of observed variables.

Maij-de Meij et al. (2010), on the other hand, only included the covariate effect on the latent class variable in the mixture Rasch model to detect DIF. In their simulation study, they generated two latent classes with equal and unequal class sizes and 1/3 of the items exhibited DIF across latent classes. When sample size was relatively small (1000), the one-class model was supported across all simulation conditions based on BIC. When sample size was medium (5000), the two-class model was supported over the one-class model if correlations between the covariate and the latent class variable were high (over .6). Regarding the large sample size (25000), the two-class model was selected with equal class sizes, but exceptions occurred that the one-class model was shown to provide best fit when the unequal class size was coupled with low correlations between the covariate and the latent class variable. They also observed that when the covariate had zero correlation with the latent class variable, excluding the covariate in the mixture Rasch model would yield more accurate results in DIF detection, with the number of false negatives much lower as compared with the approach including the covariate. However, when the covariate had nonzero (even if as low as .2) correlation with the latent class variable, including the covariate effect on the latent class variable improved DIF item detection.

Tay et al. (2011) proposed and demonstrated an integrated mixture measurement IRT model with covariates (or MM-IRT-C as used in their paper) to examine items displaying DIF across latent classes, as well as observed DIF (i.e., DIF with respect to the covariate). Authors suggested that determining the optimal number of classes can be conducted first without covariates. Then testing covariate effects and MI across latent classes can be employed in this order: covariate effect on the factor, covariate effect on the latent class variable, nonuniform DIF across classes, uniform DIF across classes, class-specific uniform and nonuniform DIF within each latent class, and factor mean comparisons across latent classes if MI was established across classes. Nonsignificant covariate effects were set to zero and parameters for items without DIF (either latent or observed) were constrained to be equal.

Significance of This Study

Although simulation studies have been conducted to examine different approaches to covariate inclusion in the mixture modeling context, few studies have examined this issue systematically in FMM, particularly for the MI testing. Most simulation studies on this topic have considered LCA and focused on the class enumeration and the estimation of the relationship between the covariate and the latent class membership (e.g., Asparouhov & Muthén, 2014; Lubke & Muthén, 2007; Nylund-Gibson & Masyn, 2016). The Lubke and Muthén (2005) and Tay et al. (2011) studies focused on this issue of including covariates in MI testing, but both were demonstrations. Maij-de Meij et al. (2010) is the only relevant simulation study that investigated the issues of covariates in the testing of MI across latent classes in the mixture modeling framework. But their focus was on the mixture Rasch model. In addition, they only considered the covariate effect on the latent class membership. Other covariate effects, such as those on the factor, item, and the path from the factor to the item, remain unexplored in MI testing with FMM.

Taken together, it is not clear how the MI testing with FMM would perform under different ways of including the covariate, such as excluding or misspecifying the covariate effects. Therefore, this study aims to fill this gap by conducting a Monte Carlo simulation study to investigate the issue of including covariates in MI testing under FMM. Additionally, this study will also examine results in terms of class enumeration, because it is of common interest in mixture modeling. More specifically, this study aims to address the following research questions:

1. What is the impact of excluding and misspecifying covariate effects on the class enumeration of factor mixture modeling (FMM)?
2. What is the impact of excluding and misspecifying covariate effects on the measurement invariance testing in FMM?

Table 1. A Summary of Literature on the Inclusion of Covariates in Mixture Modeling

Study	Model	Inclusion of Covariate Effects	Outcomes of Interest
Nylund-Gibson & Masyn, 2016	LCA	Omit covariate(s) when determining the number of classes, except for substantial covariate effects on items	Class enumeration
Asparouhov & Muthén, 2014	LCA	When substantial covariate effects on items are found, those effects should be included in the LCA model	Bias and coverage of the covariate effect on the latent class variable
Lubke & Muthén, 2007	FMM	Include the covariate effect on the latent class variable	Class assignment, coverage of factor mean differences and intercept differences
Kim, Vermunt, Bakk, Jaki, & Van horn, 2016	Regression mixture	Omit covariate(s) when determining the number of classes, but include covariate(s) when estimating model parameters	Class enumeration, parameter estimates

Table 2. A Summary of Literature on the Inclusion of Covariates in Measurement Invariance Testing with Factor Mixture Modeling

Study	Study Type	Model	Inclusion of Covariate Effects
Lubke & Muthén, 2005	Demonstration	FMM	Covariate effects on the latent class variable and the factor
Maij-de Meij, Kelderman, & Van der Flier, 2010	Simulation	Mixture Rasch model	Covariate effects on the latent class variable
Tay, Newman, & Vermunt, 2011	Demonstration	Mixture 2-parameter logistic (2PL) model	Test covariate effects on the factor and latent class variable prior to MI testing; remove if nonsignificant

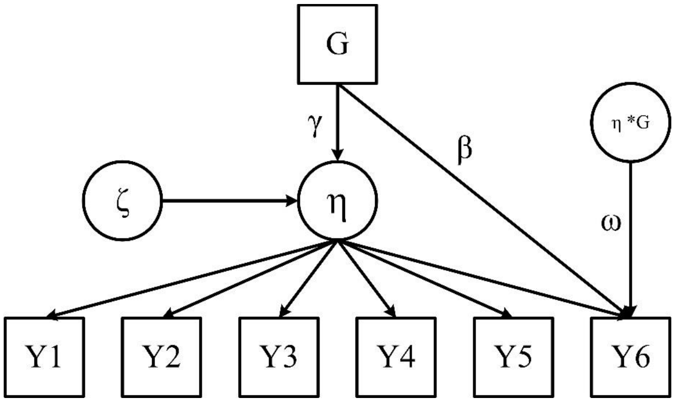


Figure 1. A generic MIMIC model for MI testing.

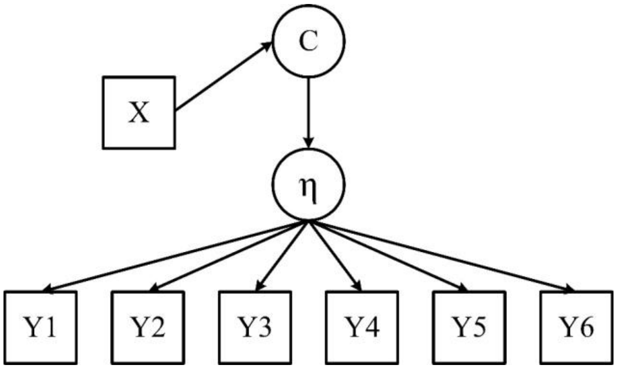


Figure 2. An example of the factor mixture modeling (FMM) model.

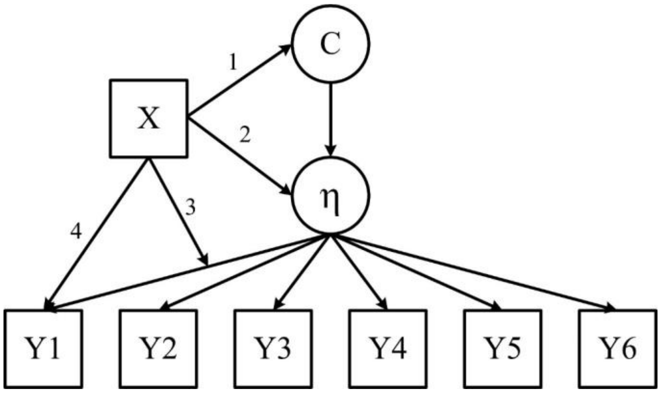


Figure 3. A generic model of FMM with covariate effects.

Chapter 3: Methodology

Data Generation

Data were generated based on the factor mixture model with covariate effects. Depending on the locations of the covariate effects, three population models were considered, which will be described in detail shortly. Covariate was assumed to be continuous and normally distributed, with mean of zero and variance of one. The CFA model consisted of a single factor and six continuous items. Items were normally distributed with mean of zero and variance of one. Residual variance for each item was specified in data generation and the values depended upon the factor loadings used, that is, for each item the sum of squared loading and residual variance should be one. Factor loadings ranged from .5 to .8 across items and were assumed to be equal across latent classes. The number of latent classes was fixed at two. Classes were separated based on factor mean difference for conditions with measurement invariance and based on both factor mean difference and intercept difference for conditions with measurement noninvariance. Factor means were simulated to be 0 and .5 (corresponds to .5 effect size) for two latent classes, respectively. The rationale for fixing some simulation factors in the simulation design was to ensure that the scale of the study is manageable and results are interpretable. However, it is acknowledged that it might be of future research interest to vary some factors, such as the differences in factor loadings or the number of latent classes. Data were generated and analyzed using *M*plus 7.1 (Muthén & Muthén, 1998-2014). Two hundred replications were simulated for each condition. The default robust maximum likelihood estimation was employed.

Simulation Design Factors

Manipulated factors included three population models, each representing one type of covariate effect, level of DIF magnitude (0, .4, .8, and 1.2), number of DIF items (1 and 2), strength of covariate effects (1 and 2 for the effect on the latent class membership; .4 and .8 for the effect on the factor; and .4

and .8 for the effect on the item depending on the population model), sample size (500 and 2000), and mixing proportions (balanced 50-50, and unbalanced 30-70). Because population models involved different covariate effects and thus different simulation factors, Table 3 summarizes simulation factors by population model.

Population models. Three population models considered are displayed in Figure 4. Figure 4(a) represents the population model when only the covariate effect on the latent class variable is present. Figure 4(b) adds an additional effect from the covariate on the factor, suggesting factor mean differences with respect to the covariate within each latent class. Figure 4(c) refers to the case when the covariate has a direct effect on one of the items (Y2), indicating uniform DIF with respect to the covariate within each latent class.

To facilitate the understanding of each population model, consider a hypothetical scenario where researchers identified two latent classes based on whether students mastered the knowledge required to solve problems in a math test or not (masters and nonmasters). They also collected data on a covariate, students' attitude towards math. Population model (a) indicated that students' attitude towards math predicted their being in the masters or nonmasters class. For example, students with more positive attitude towards math might be more likely to belong to the masters class. Model (b) implied that in addition to the impact on the latent class membership, students' attitude towards math also explained the variability in the test scores within each latent class. On top of these two covariate effects, students' attitude towards math also affected students' response to item Y2 within each latent class in model (c). In other words, for both masters and nonmasters, Y2 had DIF in terms of the covariate. For instance, students with a more positive attitude towards math tend to get higher responses on Y2 than those with less positive attitude even though their math abilities are the same. This is true for both masters and nonmasters. Note that this DIF was related with the students' attitude towards math, not with the latent classes.

It was expected as stronger covariate effects were ignored or misspecified, class enumeration would be less accurate. Specifically, if covariate effects were ignored, more latent classes might be selected to absorb the effects. For models (a) and (b) where there is no direct effect on item(s), MI testing might remain less affected than model (c), when the covariate effects were ignored.

Magnitude of DIF. For the majority of the simulation studies on DIF, the magnitude of uniform DIF ranged from .3 to 1.5 (i.e., DIF effect sizes; e.g., Jackman, 2012; E. Kim et al., 2017; Maij-de Meij et al., 2010). Therefore, three levels of uniform DIF magnitude were selected: .4, .8, and 1.2, representing small, medium, and large DIF, respectively. That is, across population models, item intercepts in one class were fixed at zero, while intercepts in the other class were .4, .8, or 1.2 for the DIF items. It is important to highlight that this DIF was between classes, not the DIF in terms of the covariate (i.e., the direct effect of X on Y2). For null conditions, equal intercepts across latent classes were generated. It was expected that as intercept differences across classes became larger, class separation would be greater and more accurate class enumeration might be observed, under the exclusion or misspecification of covariate effects. In addition, when covariate effects were ignored or misspecified, the correct level of MI might be better detected, as the intercept differences increased.

Number of DIF items. For population models 4(a) and 4(b), one or two items out of six were DIF items (i.e., DIF between latent classes), which corresponds to about 17% and 33% DIF contamination. Overall, compared with the one DIF item scenario, it was anticipated that with two DIF items, class separation would be greater. Therefore, class enumeration and scale-level MI testing should yield more accurate results, under the covariate exclusion or misspecification. For population model 4(c), only one item was simulated to have DIF. In addition, DIF across latent classes occurred for Y4 and DIF within latent classes occurred for Y2.

Strength of covariate effects. Across all population models, the effect of the covariate on the latent class variable varied at two levels, 1 and 2, which is consistent with previous simulation studies on mixture modeling with covariates (Lubke & Muthén, 2007; Maij-de Meij et al., 2010; Nylund-Gibson & Masyn, 2016). These levels of effects correspond to odds ratios of 2.72, and 7.39. The covariate effects on the factor and item(s) both varied at .4 and .8 (Nylund-Gibson & Masyn, 2016). Overall, excluding or misspecifying a larger covariate effect was expected to yield worse class enumeration. When a larger direct covariate effect on item(s) was ignored or misspecified, MI testing across latent classes could be affected to a greater extent.

Sample sizes. The sample sizes varied at 500 and 2000, representing small to medium sample sizes observed in applied and simulation studies on mixture modeling (e.g., Asparahov & Muthén, 2014; Bernstein et al., 2013; Lubke & Neale, 2008; Tein et al., 2013). Larger sample size was expected to yield higher class enumeration rates for the correct model in MI testing.

Mixing proportions. Balanced (50-50) and unbalanced (30-70) proportions were considered, which is consistent with simulation studies on mixture modeling (e.g., De Ayala et al., 2002; Park & Yu, 2016). With covariate exclusion/misspecification, class enumeration and MI testing might be worse under unbalanced proportions, as compared with the balanced proportions.

Fitted Models and Simulation Outcomes of Interest

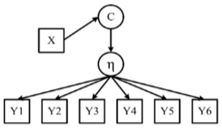
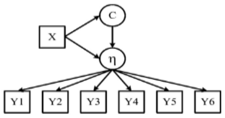
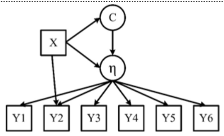
For data generated based on each of the population models shown in Figure 4, four models were fitted. Figure 5 lists these models: (a) an unconditional model excluding the covariate; (b) covariate is included in the model but only the effect on the latent class membership is modeled; (c) both covariate effects on the latent class membership and the factor are included; and (d) model estimates covariate effects on the latent class membership, the factor, and all the items. Covariate effects on all the items were included, because in applied studies, researchers were uncertain for which item the effect occurs (Nylund-Gibson & Masyn, 2016). This fitted model thus mimicked the approach applied researchers might take in examining covariate effects.

For each of the fitted models, the number of classes and the level of MI were tested simultaneously. Specifically, models with an increasing number of classes were constructed simultaneously while testing measurement invariance. More specifically, seven models were compared: 1-class, 2-class configural, 2-class metric, 2-class scalar, 3-class configural, 3-class metric, and 3-class scalar. This study adopted this simultaneous approach, due to its advantage in conducting more comprehensive model comparisons over the sequential approach. Although it might be ideal to compare across likelihood-based tests and ICs, only AIC, BIC, and SaBIC were used for model comparisons.

Although overall BLRT performed well, the execution time would be very long considering the number of conditions and the number of fitted models to run. Thus it was not considered in this study.

For each fitted model, class enumeration results were summarized and the proportions of replications selecting each class were reported. For conditions where there was no DIF across latent classes, the 2-class scalar model was expected to be selected. For conditions that had intercept noninvariance, the 2-class metric model was expected to be selected. In addition to analyzing class enumeration results within each analysis model, the model fit across analysis models was examined as well. Specifically, the fit of the true model was compared across the four analysis models, to see which analysis model fitted the data better. Non-convergence and inadmissible solutions (e.g., negative residual variances) were checked and reported. Note that only converged models with proper solutions were compared for model selection.

Table 3. Simulation Factors by Population Model

Population model	Manipulated factors		Number of conditions
	Covariate effect on class (1, 2) Number of DIF items (1, 2)	Level of DIF magnitude (0, .4, .8, 1.2) Sample size (500, 2000) Mixing proportions (50-50, 30-70)	2*2*3*2*2 + 2*2*2 = 56
	Covariate effect on class (1, 2) Covariate effect on factor (.4, .8) Number of DIF items (1, 2)		2*2*2*3*2*2 + 2*2*2*2 = 112
	Covariate effect on class (1, 2) Covariate effect on factor (.4, .8) Covariate effect on item (.4, .8)		2*2*2*3*2*2 + 2*2*2*2*2 = 128

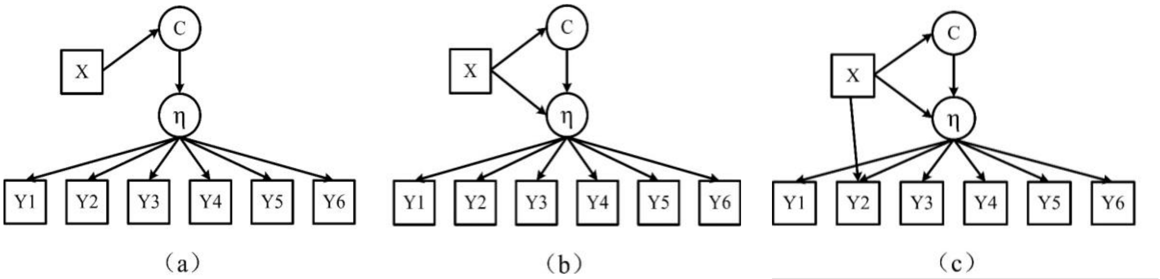


Figure 4. Population models used for data generation.

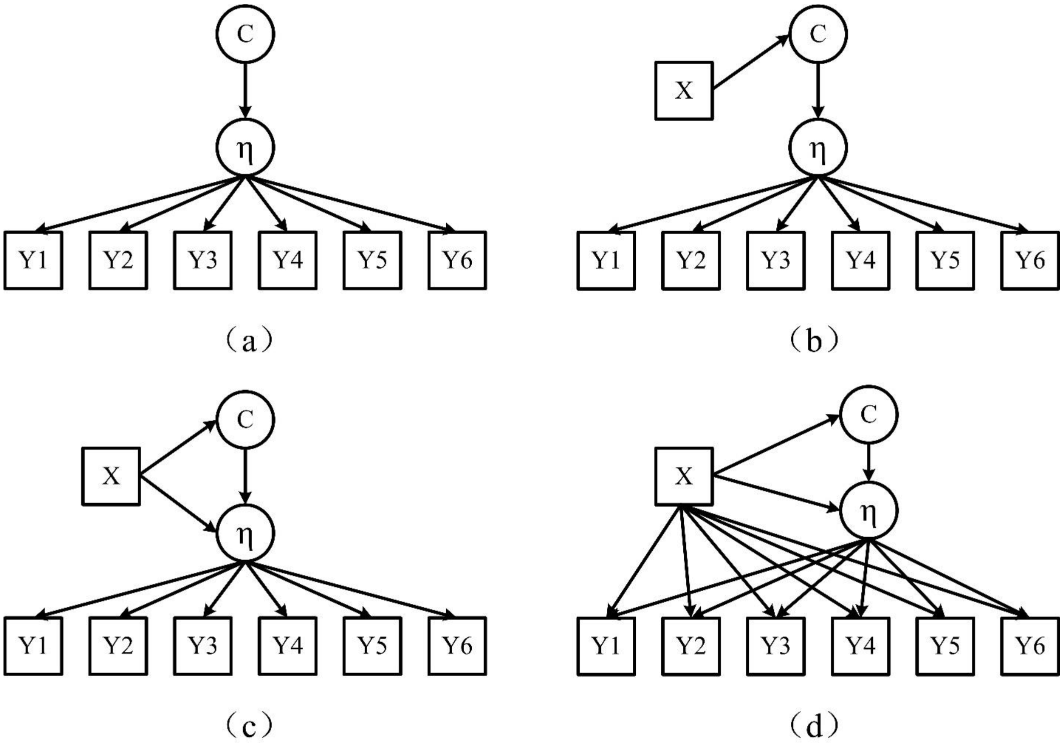


Figure 5. Fitted models in class enumeration of factor mixture modeling.

Chapter 4: Results

Non-Convergence and Inadmissible Solutions Check

Convergence was examined for all fitted models (including 1-class, 2-class configural, metric, and scalar, and 3-class configural, metric, and scalar models) under each simulation condition. The proportion of converged replications out of 200 replications ranged from .74 to 1.00 across all fitted models and conditions for Population Model 1 (PM1), .67 to 1.00 for Population Model 2 (PM2), and .76 to 1.00 for Population Model 3 (PM3). One exception occurred that all replications of the 1-class model did not converge for Analysis Model 4 (AM4) across population models. Also, only 103 out of 200 replications converged for the 3-class metric model under one condition (sample size 2000, balanced proportions, 1 DIF item, .4 DIF magnitude, covariate effect on the latent class variable = 2, covariate effect on the factor = .8, and covariate effect on the item = .8). Overall the 3-class models, especially the 3-class configural model, had lower convergence rates compared with other models. It can be observed that the convergence rates for AM1 tended to be slightly lower than other analysis models. The presence of inadmissible solutions for the true model was checked when the true model was supported. The rates for inadmissible solutions were near zero across simulation conditions.

Class Enumeration and Measurement Invariance Testing When Measurement Invariance Held in the Population

When measurement invariance held, factor mean difference of .5 between two latent classes was simulated. Overall AIC did not perform very well. The proportion of replications that selected the correct model (2-class scalar) ranged between .01 and .19 across conditions and analysis models for PM1, .00 to .43 for PM2, and .00 to .09 for PM3. Instead of 2-class scalar, AIC tended to select a more complex model (i.e., either 2-class configural, 2-class metric, 3-class configural or 3-class metric). BIC and saBIC performed much better than AIC. Eta-squared analyses were conducted to investigate the impact of simulation factors and their two-way interaction on the proportion of replications that selected the 2-class scalar model for BIC and saBIC. Results of eta-squared analyses were summarized in Table 5 by the population model.

For PM1 where the covariate had impact only on the class membership in the population, analysis model (η^2= .99) was shown to have significant impact on the MI testing when using BIC in model comparisons. When saBIC was used in model comparisons, analysis model (η^2= .70), the interaction between analysis model and sample size (η^2= .17), and sample size (η^2= .11) had significant impact on the MI testing. As can be seen from Table 6, overall A2 and A4 yielded better results in terms of identifying the correct model, where A2 matched the population model used to generate data (i.e., covariate effect on the latent class) and A4 had covariate effects on the latent class, factor, and all items. Taking a closer look at the results, when the covariate was ignored (A1, unconditional model), almost all replications selected 1-class model. Results for A3 (i.e., covariate effect on the latent class and factor) were very comparable to those under A1, that is, 1-class model was selected. saBIC tended to be more sample size dependent than BIC. In other words, when sample size was small (i.e., 500), BIC was more reliable than saBIC. For example, when using AM2, 83% of the replications correctly detected the 2-class scalar model for BIC, but saBIC only had 39% of replications detecting the correct model.

For PM2 where the covariate had impact on both the latent class membership and the factor, analysis model (η^2= .76), the interaction between analysis model (η^2= .17), and the covariate effect on the factor (η^2= .06), and the covariate effect on the factor significantly affected the MI testing when BIC was used in the model comparisons. Analysis model (η^2= .47), the interaction between analysis model and the covariate effect on the factor (η^2= .24), the interaction between analysis model and sample size (η^2= .13), and the covariate effect on the factor (η^2= .07) had significant impact on the MI testing when saBIC was used in the model comparisons. As shown in Table 6, AM4 worked very well in identifying the correct model, regardless of the covariate effect on the factor. But AM2 performed well only when the covariate effect on the factor was moderate (.4). When a stronger covariate effect (.8) was ignored, 3-class scalar model was supported instead of 2-class scalar. AM1 and AM3 supported the 1-class model. saBIC did not perform as well as BIC when sample size was 500 and AM4 was used.

For PM3 where the covariate had impact on the latent class membership, the factor, and the item, analysis model (η^2= .99) had significant impact on the MI testing when BIC was used in model comparison. When saBIC was used in model comparisons, analysis model (η^2= .79) and the interaction between analysis model and sample size (η^2= .16) had significant impact on the MI testing. As can be seen in Table 6, only AM4 performed well in detecting the correct model. BIC was more reliable than saBIC when sample size was 500. The other three analysis models could not well detect the 2-class scalar model. Specifically, AM1, AM2 and AM3 had near zero proportion of replications selecting the true model across conditions and information criteria. Instead, 1-class model was supported under AM1; 3-class metric model was supported under AM2; 2-class metric and 3-class metric models were supported under AM3.

Comparing the Fit of Analysis Models When Measurement Invariance Held in the Population

Assuming the correct model (i.e., 2-class scalar) was selected, the fit of analysis models that had different covariate effects was compared. The purpose was to examine which approach to including the covariate effect had the best fit for each population model. Results of the model comparisons were presented in Tables 7, 8, and 9 for PM1, PM2, and PM3, respectively. For PM1, overall AM1, AM2, and AM3 could all be selected as the best-fitting model. Compared with AIC and saBIC, BIC tended to have more replications selecting AM2, which matched PM1. For PM2, AM2 and AM4 were shown to have better model fit than A1 and A3. Specifically, AM2 had the best model fit when the covariate effect on the factor was .4 and AM4 had the best model fit when that effect was .8. In other words, when the covariate effect on the factor was not very strong, ignoring that effect yielded the best model fit compared with other ways of modeling the covariate effects. However, when the effect on the factor was strong, including that effect led to the best model fit. For PM3, it was apparent that AM4 almost always yielded the best fit.

Class Enumeration and Measurement Invariance Testing When Measurement Invariance Did Not Hold in the Population

Similar to the results under the establishment of measurement invariance, AIC did not perform very well in detecting the correct model (i.e., the 2-class metric model). Overall the proportions of replications that detected the correct model were low, ranging from .04 to .26 for AM1, .08 to .28 for AM2, .06 to .19 for AM3, and .03 to .12 for AM4. The 1-class model was almost never selected across analysis models. For other models, there was no consistent pattern which model was preferred over others. BIC and saBIC performed much better than AIC. Results of eta-squared analyses on the class enumeration rates were summarized in Table 10 by the population model. Eta-squared analyses showed that for PM1 where the covariate had impact only on the latent class membership, DIF magnitude (η^2= .28), analysis model (η^2= .22), number of DIF items (η^2= .12), and sample size (η^2= .08) had significant impact on the MI testing results when using BIC in model comparisons. Similarly, these four design factors had significant impact on MI testing (η^2= .27, .23, .14, and .08, respectively) when saBIC was used in model comparisons. The proportion of replications that selected the correct model (i.e., 2-class metric invariance model) was presented in Table 11 based on the simulation factors that showed significant impact. AM2 performed the best among all analysis models, followed by AM3, AM4, and AM1. Overall across all analysis models, the proportions of replications that selected the correct model were higher as the DIF magnitude increased, the number of DIF items increased, and the sample size increased. For example, the class enumeration rates increased from .24 to .54 and .73 for saBIC as the DIF magnitude increased from .4 to .8 and 1.2, respectively, when there was 1 DIF item and sample size was 500. The class enumeration rates were .54 and .82 for 1 and 2 DIF items, respectively, when the DIF magnitude was .8 and sample size was 500. Overall saBIC performed better than BIC. For AM2, saBIC almost always supported the correct model when the DIF magnitude was above .4 and sample size was 2000. When there were 2 DIF items, the DIF magnitude was larger (.8 and 1.2), AM3, AM4, and AM1 could also detect the 2-class metric model.

For PM2 where the covariate had impact on both the latent class membership and the factor, eta-squared analyses showed that DIF magnitude (η^2= .27), the number of DIF items (η^2= .14), the interaction between analysis model and the covariate effect on the factor (η^2= .08), and the analysis model (η^2= .08) had significant impact on the class enumeration for BIC. The same simulation factors impacted the class enumeration rates for saBIC substantially, but eta-squared values were different (please refer to Table 10 for the values). As shown in Table 12, AM3 performed well in identifying the correct model when the DIF magnitude was .8 and 1.2. Note that AM3 matched the data generation model where there were covariate effects on the latent class variable and the factor. As the DIF magnitude increased, the class enumeration rates increased substantially across all analysis models. AM2 (i.e., ignoring the covariate effect on the factor) worked well in detecting the 2-class metric model when the covariate effect on the factor was .4. When the effect was .8, ignoring that effect would lead to an over-extraction of latent classes, that is, 3-class metric instead of 2-class metric model. AM4, which was the most complex model with covariate effects on all items, supported the 2-class metric model only when DIF magnitude was .8 or 1.2 with 2 DIF items; it supported the 2-class scalar model instead for other conditions. A1 only worked well with 1.2 DIF magnitude and 2 DIF items. But for most conditions, it tended to select the 1-class model.

For PM3 where the covariate had impact on the latent class membership, the factor, and the item, simulation factors that had significant impact on the class enumeration included: analysis model (η^2= .25), the interaction between analysis model and the covariate effect on the item (η^2= .11), and the interaction between analysis model and sample size (η^2= .06) for BIC. For saBIC, the interaction between analysis model and the covariate effect on the item (η^2= .23), analysis model (η^2= .14), the interaction between analysis model and DIF magnitude (η^2= .11), the interaction between analysis model and covariate effect on the factor (η^2= .07), the interaction between analysis model and sample size (η^2= .07), and the DIF magnitude (η^2= .07) all affected the class enumeration. Surprisingly, overall AM3 performed better than AM4 in detecting the 2-class metric model, as seen in Table 13. In other words, it was better to ignore the covariate effect on the item if the effect was present than including covariate effects on all items. But this was the case only when the covariate effect on the item was .4. When the effect was .8, ignoring that effect tended to result in an additional latent class (i.e., 3-class metric instead of 2-class metric). AM4 only worked well when sample size was 2000 and the DIF magnitude was 1.2. AM2 (i.e., ignoring both covariate effects on the factor and the item) only worked well when these two covariate effects were .4 and sample size was 500; for other conditions, an over-extraction of latent classes (i.e., 3-class metric) was observed. A1 tended to support the 2-class scalar model rather than the 2-class metric model.

Comparing the Fit of Analysis Models When Measurement Invariance Did Not Hold in the Population

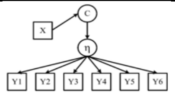
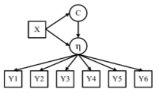
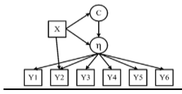
For PM1, BIC and saBIC showed that overall AM2 had the best model fit among all analysis models (see Table 14). For PM1, more replications selected AM2 as the best-fitting model as the DIF magnitude, sample size, number of DIF items, and the covariate effect on the latent class variable increased. Other analysis models were much less likely to be selected as the best-fitting model. As expected, for PM2, AM3, which matched PM2, showed the best model fit (see Table 15). For PM3, AM4, which modeled the covariate effect on all items, had the best model fit (see Table 16).

Table 4. Non-Convergence and Inadmissible Solutions Check

	Proportion of converged replications
Population model 1	
Analysis model 1	.74 ~ 1.00
Analysis model 2	.78 ~ 1.00
Analysis model 3	.86 ~ 1.00
Analysis model 4	.84 ~ 1.00[a]
Population model 2	
Analysis model 1	.67 ~ 1.00
Analysis model 2	.87 ~ 1.00
Analysis model 3	.87 ~ 1.00
Analysis model 4	.83 ~ 1.00[a]
Population model 3	
Analysis model 1	.76 ~ 1.00
Analysis model 2	.88 ~ 1.00
Analysis model 3	.88 ~ 1.00
Analysis model 4	.84 ~ 1.00[a]

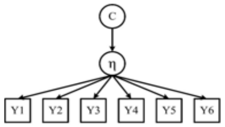
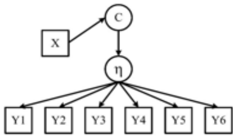
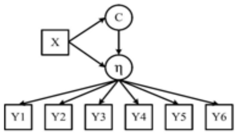
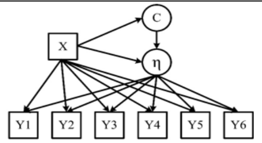
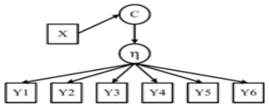
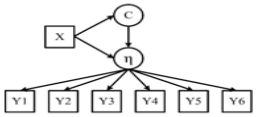
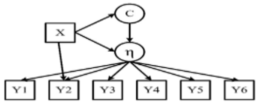
Note. [a]The range of convergence rates did not include the 1-class model, because all replications of the 1-class model did not converge.

Table 5. Results of Eta-Squared Analyses by Population Model When Measurement Invariance Held in the Population

Population model	BIC		saBIC	
	Simulation factor	η^2	Simulation factor	η^2
	Analysis model	.99	Analysis model	.70
			Analysis model*sample size	.17
			Sample size	.11
	Analysis model	.76	Analysis model	.47
	Analysis model*covariate effect on factor	.17	Analysis model*covariate effect on factor	.24
	Covariate effect on factor	.06	Analysis mode*sample size	.13
			Covariate effect on factor	.07
	Analysis model	.99	Analysis model	.79
			Analysis model*sample size	.16

Note. BIC = Bayesian information criterion, saBIC = sample size adjusted BIC.

Table 6. Class Enumeration and Measurement Invariance (MI) Testing When MI Held in the Population

			Analysis Model							
										
Population Model	Covariate Effect on Factor	Sample Size	BIC	saBIC	BIC	saBIC	BIC	saBIC	BIC	saBIC
		500	.01	.14	**.83**	.39	.00	.03	**.92**	.32
		2000	.00	.05	**.93**	**.84**	.00	.00	**.98**	**.84**
	.4	500	.00	.15	**.95**	**.72**	.00	.03	**.94**	.33
		2000	.00	.03	**.92**	**.74**	.00	.05	**.99**	**.85**
	.8	500	.00	.15	.12	.00	.00	.02	**.94**	.38
		2000	.00	.04	.00	.00	.01	.10	**.98**	**.84**
		500	.00	.00	.00	.00	.08	.00	**.94**	.33
		2000	.00	.01	.00	.00	.00	.00	**.99**	**.87**

Note. BIC = Bayesian information criterion, saBIC = sample size adjusted BIC.

Table 7. Comparing Analysis Models When Measurement Invariance Held in the Population under Population Model 1

Sample size	Mixing proportion	Covariate effect on class	AIC_a1	AIC_a2	AIC_a3	AIC_a4	BIC_a1	BIC_a2	BIC_a3	BIC_a4	saBIC_a1	saBIC_a2	saBIC_a3	saBIC_a4
500	50-50	1	.37	.37	.25	.02	**.53**	.39	.09	.00	.39	.39	.23	.00
		2	.25	.34	.38	.04	.45	.40	.15	.01	.30	.38	.32	.02
	30-70	1	.37	.39	.22	.03	**.58**	.37	.06	.00	.41	.42	.17	.01
		2	.33	.30	.35	.03	**.53**	.33	.14	.00	.37	.31	.32	.01
2000	50-50	1	.18	.27	.49	.07	.33	.47	.20	.00	.24	.40	.36	.01
		2	.02	.24	**.69**	.07	.06	**.66**	.28	.00	.04	.45	**.51**	.01
	30-70	1	.23	.27	.45	.06	.44	.39	.17	.00	.33	.36	.32	.00
		2	.05	.20	**.69**	.06	.15	**.56**	.30	.00	.08	.41	**.51**	.01

Note. The suffix of a1, a2, a3, and a4 for AIC, BIC, and saBIC indicates that analysis models 1, 2, 3, and 4 (see Figure 5a, 5b, 5c, and 5d), respectively.

Table 8. Comparing Analysis Models When Measurement Invariance Held in the Population under Population Model 2

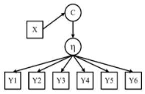
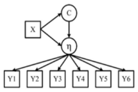
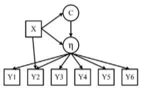
Sample size	Mixing proportion	Covariate effect on class	Covariate effect on factor	AIC_a1	AIC_a2	AIC_a3	AIC_a4	BIC_a1	BIC_a2	BIC_a3	BIC_a4	saBIC_a1	saBIC_a2	saBIC_a3	saBIC_a4
500	50-50	1	4	.07	.40	.00	**.53**	.14	**.85**	.00	.02	.08	**.71**	.00	.22
			8	.00	.02	.00	**.99**	.00	.02	.00	**.99**	.00	.02	.00	**.99**
		2	4	.12	.39	.00	**.51**	.27	**.71**	.00	.03	.20	**.66**	.00	.15
			8	.00	.01	.00	**1.00**	.00	.01	.00	**1.00**	.00	.01	.00	**1.00**
	30-70	1	4	.07	.37	.00	**.57**	.13	**.87**	.00	.01	.08	**.67**	.00	.26
			8	.00	.02	.00	**.99**	.00	.02	.00	**.99**	.00	.02	.00	**.99**
		2	4	.13	.40	.00	.49	.22	**.77**	.00	.02	.16	**.66**	.00	.19
			8	.00	.01	.00	**.99**	.00	.01	.00	**.99**	.00	.01	.00	**.99**
2000	50-50	1	4	.00	.06	.00	**.95**	.00	**.99**	.00	.01	.00	**.58**	.00	.43
			8	.00	.01	.00	**1.00**	.00	.01	.00	**1.00**	.00	.01	.00	**1.00**
		2	4	.00	.06	.00	**.94**	.00	**.99**	.00	.02	.00	**.66**	.00	.34
			8	.00	.01	.00	**.99**	.00	.01	.00	**.99**	.00	.01	.00	**.99**
	30-70	1	4	.00	.05	.00	**.96**	.00	**.97**	.00	.04	.00	**.53**	.00	.47
			8	.00	.02	.00	**.99**	.00	.02	.00	**.99**	.00	.02	.00	**.99**
		2	4	.00	.26	.00	**.74**	.00	**1.00**	.00	.00	.00	**1.00**	.00	.01
			8	.00	.02	.00	**.99**	.00	.02	.00	**.99**	.00	.02	.00	**.99**

Note. The suffix of a1, a2, a3, and a4 for AIC, BIC, and saBIC indicates that analysis models 1, 2, 3, and 4 (see Figure 5a, 5b, 5c, and 5d), respectively.

Table 9. Comparing Analysis Models When Measurement Invariance Held in the Population under Population Model 3

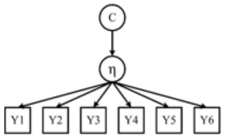
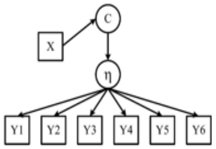
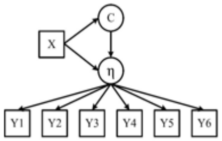
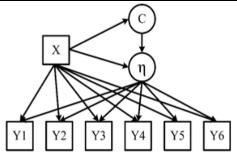
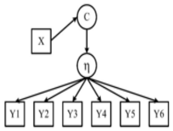
Sample size	Mixing proportion	Covariate effect on class	Covariate effect on factor	Covariate effect on item	AIC_a1	AIC_a2	AIC_a3	AIC_a4	BIC_a1	BIC_a2	BIC_a3	BIC_a4	saBIC_a1	saBIC_a2	saBIC_a3	saBIC_a4
500	50-50	1	4	4	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**
				8	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**
			8	4	.00	.00	.01	**.99**	.00	.00	.37	**.63**	.00	.00	.04	**.96**
				8	.00	.00	.01	**1.00**	.00	.00	.01	**1.00**	.00	.00	.01	**1.00**
		2	4	4	.00	.00	.01	**1.00**	.00	.00	.01	**1.00**	.00	.00	.01	**1.00**
				8	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**
			8	4	.00	.00	.02	**.99**	.00	.00	.35	**.65**	.00	.00	.05	**.95**
				8	.00	.00	.01	**1.00**	.00	.00	.01	**1.00**	.00	.00	.01	**1.00**
	30-70	1	4	4	.00	.00	.01	**1.00**	.00	.00	.01	**1.00**	.00	.00	.01	**1.00**
				8	.00	.00	.01	**1.00**	.00	.00	.01	**1.00**	.00	.00	.01	**1.00**
			8	4	.00	.00	.01	**.99**	.00	.00	.39	**.61**	.00	.00	.06	**.95**
				8	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**
		2	4	4	.00	.00	.01	**1.00**	.01	.00	.00	**1.00**	.00	.00	.01	**1.00**
				8	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**
			8	4	.00	.00	.01	**.99**	.00	.00	.35	**.66**	.00	.00	.04	**.96**
				8	.00	.00	.01	**.99**	.00	.00	.01	**.99**	.00	.00	.01	**.99**
2000	50-50	1	4	4	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**
				8	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**
			8	4	.00	.00	.03	**.98**	.00	.00	.03	**.98**	.00	.00	.03	**.98**
				8	.00	.00	.01	**1.00**	.00	.00	.01	**1.00**	.00	.00	.01	**1.00**
		2	4	4	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**
				8	.00	.00	.01	**1.00**	.00	.00	.01	**1.00**	.00	.00	.01	**1.00**
			8	4	.00	.00	.01	**.99**	.00	.00	.01	**.99**	.00	.00	.01	**.99**
				8	.00	.00	.01	**.99**	.00	.00	.01	**.99**	.00	.00	.01	**.99**
	30-70	1	4	4	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**
				8	.00	.00	.01	**.99**	.00	.00	.01	**.99**	.00	.00	.01	**.99**
			8	4	.00	.00	.02	**.99**	.00	.00	.02	**.99**	.00	.00	.02	**.99**
				8	.00	.00	.01	**1.00**	.00	.00	.01	**1.00**	.00	.00	.01	**1.00**
		2	4	4	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**
				8	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00	.00	.00	**1.00**
			8	4	.00	.00	.02	**.99**	.00	.00	.02	**.99**	.00	.00	.02	**.99**
				8	.00	.00	.02	**.98**	.00	.00	.02	**.98**	.00	.00	.02	**.98**

Table 10. Results of Eta-Squared Analyses by Population Model When Measurement Invariance Did Not Hold in the Population

Population model	BIC		saBIC	
	Simulation factor	η^2	Simulation factor	η^2
	DIF magnitude	.28	DIF magnitude	.27
	Analysis model	.22	Analysis model	.23
	Number of DIF items	.12	Number of DIF items	.14
	Sample size	.08	Sample size	.08
	DIF magnitude	.27	DIF magnitude	.20
	Number of DIF items	.14	Analysis model*covariate effect on factor	.15
	Analysis model*covariate effect on factor	.08	Analysis model	.12
	Analysis model	.08	Number of DIF items	.11
	Analysis model	.25	Analysis model*covariate effect on item	.23
	Analysis model*covariate effect on item	.11	Analysis model	.14
	Analysis model*sample size	.06	Analysis model*DIF magnitude	.11
			Analysis model*covariate effect on factor	.07
			Analysis model*sample size	.07
			DIF magnitude	.07

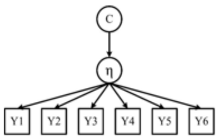
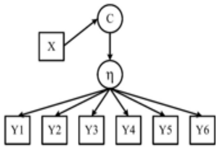
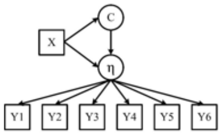
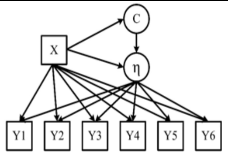
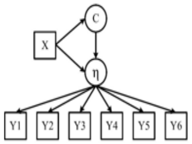
Note. BIC = Bayesian information criterion, saBIC = sample size adjusted BIC.

Table 11. Class Enumeration and Measurement Invariance (MI) Testing When MI Did Not Hold in the Population under Population Model 1

				Analysis Model							
											
Population Model	DIF Magnitude	Number of DIF Items	Sample Size	BIC	saBIC	BIC	saBIC	BIC	saBIC	BIC	saBIC
	.4	1	500	.00	.02	.11	.24	.00	.06	.02	.15
			2000	.00	.00	.29	**.60**	.00	.07	.01	.05
		2	500	.00	.02	.18	.40	.00	.13	.02	.15
			2000	.00	.01	**.70**	**.92**	.05	**.51**	.01	.13
	.8	1	500	.00	.02	.36	**.54**	.00	.32	.03	.15
			2000	.00	.00	**.92**	**.98**	.46	**.78**	.03	.18
		2	500	.00	.16	**.82**	**.82**	.19	**.71**	.14	.39
			2000	.07	**.88**	**1.00**	**1.00**	**.99**	**1.00**	**.95**	**.97**
	1.2	1	500	.00	.03	**.69**	**.73**	.13	**.63**	.03	.21
			2000	.00	.00	**1.00**	**1.00**	**.84**	**.99**	.34	**.60**
		2	500	.36	**.84**	**1.00**	**.87**	**.93**	**.83**	**.96**	**.74**
			2000	**1.00**	**.98**	**1.00**	**.99**	**1.00**	**.99**	**.99**	**.97**

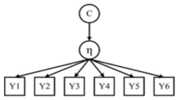
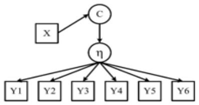
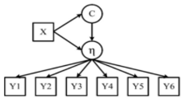
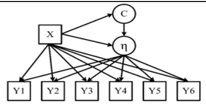
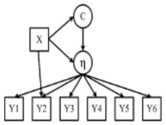
Note. BIC = Bayesian information criterion, saBIC = sample size adjusted BIC.

Table 12. Class Enumeration and Measurement Invariance (MI) Testing When MI Did Not Hold in the Population under Population Model 2

				Analysis Model							
											
Population Model	DIF Magnitude	Number of DIF Items	Covariate Effect on Factor	BIC	saBIC	BIC	saBIC	BIC	saBIC	BIC	saBIC
	.4	1	.4	.00	.02	.14	.41	.00	.09	.01	.10
			.8	.00	.02	.00	.00	.00	.06	.01	.11
		2	.4	.00	.01	.46	**.69**	.10	.44	.02	.15
			.8	.00	.02	.02	.00	.08	.47	.01	.15
	.8	1	.4	.00	.02	**.65**	**.80**	.25	**.65**	.03	.18
			.8	.00	.01	.01	.00	.23	**.64**	.02	.23
		2	.4	.08	**.56**	**.94**	**.88**	**.71**	**.88**	**.56**	**.69**
			.8	.07	**.57**	.32	.01	**.76**	**.90**	**.55**	**.72**
	1.2	1	.4	.00	.02	**.86**	**.86**	**.60**	**.84**	.20	.43
			.8	.00	.02	.08	.00	**.54**	**.86**	.25	**.52**
		2	.4	**.75**	**.90**	**.98**	**.61**	**1.00**	**.90**	**.96**	**.85**
			.8	**.76**	**.92**	.34	.08	**1.00**	**.91**	**.95**	**.87**

Note. BIC = Bayesian information criterion, saBIC = sample size adjusted BIC.

Table 13. Class Enumeration and Measurement Invariance (MI) Testing When MI Did Not Hold in the Population under Population Model 3

					Analysis Model							
												
Population Model	Covariate Effect on Item	Sample Size	Covariate Effect on Factor	DIF Magnitude	BIC	saBIC	BIC	saBIC	BIC	saBIC	BIC	saBIC
	.4	500	.4	.4	.00	.03	**.99**	.41	**.91**	**.63**	.02	.18
				.8	.00	.03	**.99**	.35	**.98**	**.58**	.02	.18
				1.2	.00	.11	**.99**	.25	**1.00**	.49	.06	.30
			.8	.4	.00	.01	.07	.00	.08	**.72**	.03	.17
				.8	.00	.02	.13	.00	.39	**.76**	.02	.22
				1.2	.00	.04	.14	.00	**.80**	**.76**	.08	.33
		2000	.4	.4	.00	.00	.17	.00	**.98**	.22	.01	.05
				.8	.00	.04	.08	.00	**.91**	.07	.08	.33
				1.2	.04	**.58**	.02	.00	**.71**	.01	**.61**	**.80**
			.8	.4	.00	.00	.00	.00	**1.00**	**.99**	.01	.06
				.8	.00	.00	.00	.00	**1.00**	**.96**	.14	.46
				1.2	.00	.14	.00	.00	**.75**	**.67**	**.54**	**.76**
	.8	500	.4	.4	.00	.04	.00	.00	.00	.00	.02	.15
				.8	.00	.28	.00	.00	.00	.00	.02	.17
				1.2	.13	**.67**	.00	.00	.00	.00	.05	.30
			.8	.4	.00	.03	.00	.00	**1.00**	.33	.02	.18
				.8	.00	.12	.00	.00	**1.00**	.34	.02	.18
				1.2	.01	**.49**	.00	.00	**1.00**	.29	.09	.31
		2000	.4	.4	.00	.07	.00	.00	.00	.00	.01	.07
				.8	.48	**.88**	.00	.00	.00	.00	.08	.34
				1.2	**.98**	**.99**	.00	.00	.00	.00	**.62**	**.80**
			.8	.4	**.80**	.01	.00	.00	**.66**	.02	.02	.07
				.8	.12	**.69**	.00	.00	.47	.01	.14	.47
				1.2	**.89**	**.98**	.00	.00	.20	.00	**.77**	**.89**

Note. BIC = Bayesian information criterion, saBIC = sample size adjusted BIC.

Table 14. Comparing Analysis Models When Measurement Invariance Did Not Hold in the Population under Population Model 1

DIF magnitude	Sample size	Number of DIF items	Covariate effect on class	AIC_a1	AIC_a2	AIC_a3	AIC_a4	BIC_a1	BIC_a2	BIC_a3	BIC_a4	ssBIC_a1	ssBIC_a2	ssBIC_a3	ssBIC_a4
4	500	1	1	.26	.29	.32	.14	.53	.30	.17	.00	.35	.32	.29	.04
			2	.15	.23	.39	.24	.31	.43	.26	.01	.19	.31	.41	.10
		2	1	.17	.26	.37	.20	.37	.39	.24	.00	.21	.34	.39	.07
			2	.04	.30	.35	.31	.12	**.62**	.25	.01	.06	.48	.37	.10
	2000	1	1	.10	.23	.39	.29	.25	**.54**	.21	.00	.15	.45	.38	.03
			2	.00	.43	.24	.33	.00	**.90**	.09	.00	.00	**.76**	.21	.03
		2	1	.01	.36	.33	.31	.03	**.84**	.13	.00	.02	**.70**	.27	.02
			2	.00	**.66**	.17	.17	.00	**.99**	.02	.00	.00	**.93**	.07	.01
8	500	1	1	.16	.26	.30	.29	.35	.46	.19	.00	.24	.37	.30	.10
			2	.03	.36	.25	.37	.08	**.76**	.16	.01	.03	**.54**	.27	.16
		2	1	.02	**.51**	.31	.17	.05	**.80**	.15	.00	.02	**.66**	.26	.06
			2	.00	**.67**	.20	.13	.00	**.93**	.07	.01	.00	**.82**	.15	.04
	2000	1	1	.01	.41	.21	.38	.01	**.93**	.06	.00	.01	**.81**	.17	.02
			2	.00	**.71**	.20	.10	.00	**.98**	.02	.00	.00	**.91**	.09	.00
		2	1	.00	**.79**	.17	.05	.00	**.98**	.02	.00	.00	**.95**	.05	.00
			2	.00	**.81**	.16	.03	.00	**.99**	.02	.00	.00	**.96**	.04	.00
12	500	1	1	.06	.33	.27	.34	.16	**.70**	.14	.01	.09	**.50**	.29	.13
			2	.00	**.53**	.16	.32	.00	**.92**	.07	.01	.00	**.73**	.16	.11
		2	1	.00	**.77**	.20	.04	.00	**.97**	.03	.00	.00	**.87**	.13	.00
			2	.00	**.79**	.18	.03	.00	**.97**	.03	.00	.00	**.88**	.12	.00
	2000	1	1	.00	**.71**	.18	.11	.00	**.97**	.03	.00	.00	**.92**	.07	.01
			2	.00	**.81**	.19	.01	.00	**.98**	.02	.00	.00	**.94**	.06	.00
		2	1	.00	**.81**	.17	.02	.00	**.99**	.01	.00	.00	**.97**	.03	.00
			2	.00	**.83**	.15	.02	.00	**.99**	.02	.00	.00	**.95**	.05	.00

Note. The suffix of a1, a2, a3, and a4 for AIC, BIC, and saBIC indicates that analysis models 1, 2, 3, and 4 (see Figure 5a, 5b, 5c, and 5d), respectively.

Table 15. Comparing Analysis Models When Measurement Invariance Did Not Hold in the Population under Population Model 2

Covariate effect on factor	Sample size	DIF magnitude	Number of DIF items	AIC_a1	AIC_a2	AIC_a3	AIC_a4	BIC_a1	BIC_a2	BIC_a3	BIC_a4	ssBIC_a1	ssBIC_a2	ssBIC_a3	ssBIC_a4
4	500	4	1	.02	.05	**.62**	.31	.09	.19	**.71**	.01	.04	.09	**.74**	.14
			2	.02	.11	**.52**	.36	.08	.26	**.65**	.01	.03	.15	**.68**	.13
		8	1	.00	.03	**.53**	.44	.02	.23	**.74**	.01	.00	.08	**.72**	.20
			2	.00	.03	**.77**	.20	.00	.19	**.80**	.01	.00	.07	**.87**	.06
		12	1	.00	.01	**.60**	.39	.01	.15	**.83**	.02	.00	.04	**.78**	.18
			2	.00	.01	**.95**	.05	.00	.02	**.99**	.00	.00	.01	**.99**	.01
	2000	4	1	.00	.01	**.56**	.44	.00	.03	**.96**	.00	.00	.01	**.94**	.05
			2	.00	.00	**.68**	.32	.00	.04	**.96**	.00	.00	.02	**.96**	.03
		8	1	.00	.00	**.71**	.29	.00	.01	**.99**	.00	.00	.00	**.97**	.03
			2	.00	.00	**.95**	.05	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00
		12	1	.00	.00	**.91**	.09	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00
			2	.00	.00	**.97**	.04	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00
8	500	4	1	.00	.00	**.70**	.31	.00	.00	**.99**	.01	.00	.00	**.87**	.13
			2	.00	.00	**.59**	.41	.00	.00	**.99**	.02	.00	.00	**.84**	.17
		8	1	.00	.00	**.53**	.47	.00	.00	**.98**	.02	.00	.00	**.77**	.23
			2	.00	.00	**.80**	.20	.00	.00	**1.00**	.00	.00	.00	**.93**	.07
		12	1	.00	.00	**.65**	.35	.00	.00	**.99**	.01	.00	.00	**.83**	.17
			2	.00	.00	**.96**	.04	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00
	2000	4	1	.00	.00	**.65**	.35	.00	.00	**.99**	.01	.00	.00	**.97**	.03
			2	.00	.00	**.77**	.24	.00	.00	**1.00**	.00	.00	.00	**.99**	.02
		8	1	.00	.00	**.76**	.24	.00	.00	**1.00**	.00	.00	.00	**.98**	.02
			2	.00	.00	**.94**	.06	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00
		12	1	.00	.00	**.93**	.07	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00
			2	.00	.00	**.95**	.05	.00	.00	**1.00**	.00	.00	.00	**1.00**	.00

Note. The suffix of a1, a2, a3, and a4 for AIC, BIC, and saBIC indicates that analysis models 1, 2, 3, and 4 (see Figure 5a, 5b, 5c, and 5d), respectively.

Table 16. Comparing Analysis Models When Measurement Invariance Did Not Hold in the Population under Population Model 3

Covariate effect on item	Covariate effect on factor	Sample size	AIC _a1	AIC _a2	AIC _a3	AIC _a4	BIC _a1	BIC _a2	BIC _a3	BIC _a4	ssBIC _a1	ssBIC _a2	ssBIC _a3	ssBIC _a4
4	4	500	.00	.00	.02	**.98**	.00	.03	.38	**.59**	.00	.00	.05	**.95**
		2000	.00	.00	.01	**.99**	.00	.00	.01	**.99**	.00	.00	.01	**.99**
	8	500	.00	.00	.17	**.83**	.00	.00	**.91**	.09	.00	.00	.37	**.63**
		2000	.00	.00	.06	**.94**	.00	.00	**.66**	.34	.00	.00	.24	**.76**
8	4	500	.00	.00	.01	**.99**	.00	.00	.01	**.99**	.00	.00	.01	**.99**
		2000	.00	.00	.01	**.99**	.00	.00	.01	**.99**	.00	.00	.01	**.99**
	8	500	.00	.00	.02	**.98**	.00	.00	.31	**.69**	.00	.00	.04	**.96**
		2000	.00	.00	.00	**1.00**	.00	.00	.01	**1.00**	.00	.00	.00	**1.00**

Note. The suffix of a1, a2, a3, and a4 for AIC, BIC, and saBIC indicates that analysis models 1, 2, 3, and 4 (see Figure 5a, 5b, 5c, and 5d), respectively.

Chapter 5: Discussion

This simulation study examined the impact of covariate effect inclusion on measurement invariance testing with factor mixture modeling. Different covariate effects were simulated, including the covariate effect on the latent class variable, the factor, and the item. Then different analysis models were fitted where the covariate effects were misspecified and class enumeration was examined to see how the misspecification affected the class enumeration. Key findings will be summarized and discussed first, followed by implications of this study on applied and methodological research in the future.

When measurement invariance (MI) held, two latent classes were distinguished by factor mean difference. Overall models that included the covariate effect on the latent class membership only (i.e., analysis model 2) and the covariate effects on the latent class membership, factor, and all items (i.e., analysis model 4) performed well. Specifically, both models could well identify the correct model (i.e., 2-class scalar in this study), if there was only covariate effect on the latent class membership or the covariate effect on the factor was weak (.4 in this study). Under these circumstances, the simpler model including only the covariate effect on the latent class membership had better fit than the more complex model including direct covariate effects on all items. However, only the more complex model identified the 2-class scalar model when the covariate effect on the factor was strong (i.e., .8 in this study) or there was direct covariate effect on the item. Under these circumstances, the more complex model fitted data better than the simpler model. However, in reality, it is unknown what the underlying population model is, whether the covariate effect on the factor is strong, or whether there are direct covariate effects on the items. Therefore, we recommend the most complex model in the class enumeration process when MI holds. Once the number of latent classes and the level of invariance are identified, a simpler model (i.e., covariate effect on the latent class membership only) with these factors fixed (e.g., 2-class scalar) can be fitted and compared to the same solution under the more complex model. Then the model that has better model fit can be selected based on information criteria and parameter estimates can be examined and interpreted. However, it should be noted that due to the complexity of this analysis model, it might not be a practical solution when there are multiple covariates (Nylund-Gibson & Masyn, 2016). It might be challenging to estimate model parameters when all the included covariates have direct effects on all the items.

When MI held, the unconditional model tended to select the 1-class model, which might result from the low class separation. That is, classes were separated only by the factor mean difference so both the class separation and classification accuracy could be low. Therefore, information criteria failed to identify the 2-class scalar model. The poor performance of the unconditional model compared with other model that included covariate effects showed that including covariate effects might help improve class separation and thus the class enumeration in MI testing (Lubke & Muthén, 2007; Maij-de Meij et al., 2010). However, the inclusion of covariate effects does not necessarily guarantee accurate results for MI testing. Specifically, the analysis model including both covariate effects on the latent class membership and the factor also supported 1-class model. This might be because although the covariate effect on the factor was simulated within each latent class and the factor mean difference was simulated between latent classes, the covariate path to the factor in the fitted model captured the variability in the factor mean as a whole. Therefore, the 1-class model was supported instead of the 2-class scalar model. Other possible explanations have been ruled out by results of a few additional simulation conditions. That is, larger factor mean difference (1.5 rather than the original .5) and the covariate effect on the latent class variable being zero were simulated separately. Similar results were found with these additional simulations that adding the covariate effect on the factor would lead to the 1-class solution. That is, even though the factor mean difference between classes got larger or the covariate had no relationship with the latent class membership, the difference in the factor mean would still be captured by the covariate effect. Overall, when MI held, BIC was more reliable than saBIC. The performance of saBIC was sample-dependent.

When MI did not hold, overall the choice of analysis models depended upon the population model and class separation. When there were covariate effects on the latent class variable only (population model 1) or both the latent class variable and the factor (population model 2), all analysis models including the unconditional model performed well if the class separation was large; otherwise, the analysis model that matched the population model performed the best. For population model 1, in addition to the analysis model that matched the population model, the slightly over-specified model including covariate effects on both the latent class membership and the factor also performed well, when the class separation was large. For population model 2, the analysis model that ignored the covariate effect on the factor performed well when the omitted covariate effect was not strong. When the covariate effect on the factor was strong, the analysis model that matched the population model (i.e., covariate effects on both the latent class membership and the factor) performed the best. When there were covariate effects on the latent class membership, the factor, and the item (population model 3), the model that ignored the covariate effect on the item performed very well when the omitted effect was not strong. When the effect was strong, none of the models performed satisfactorily; however, with large sample size and large separation, the unconditional model was acceptable.

Overall, when there was measurement noninvariance, the unconditional model should not be recommended, because it tended to select the 1-class model regardless of the population model. This is similar to the finding under measurement invariance conditions and consistent with previous findings in the literature (Lubke & Muthén, 2007; Maij-de Meij et al., 2010). That is, including covariates in the class enumeration could improve the class separation and thus covariates should be included in the factor mixture model when testing measurement invariance across latent classes. Note that the poor performance of the unconditional model was not observed in previous simulation studies using latent class analysis model (Nylund-Gibson & Masyn, 2016) and regression mixture model (M. Kim et al., 2016). Both studies found that the unconditional model performed well in terms of class enumeration. This discrepancy in the performance of the unconditional model might occur due to several differences between this study and the other two studies mentioned above. First, this study focused on the factor mixture modeling, which is a combination of confirmatory factor analysis and latent class analysis. Therefore, the model could be considered as more complex than the latent class analysis and regression mixture model. It might be more difficult to distinguish latent classes because simulated differences across classes might be absorbed by other model parameters. If this happens, including covariates could improve the class separation and classification accuracy, which would further help the class enumeration. Second, class enumeration is defined differently for this study and the other two studies. This study examined the number of latent classes and the level of invariance simultaneously in the class enumeration process, while the other two studies only considered the number of latent classes. Overall, it would be beneficial to include covariate effects into the factor mixture model in class enumeration.

Specifically, it seems that including covariate effects on both the latent class membership and the factor yielded desirable results consistently across population models, when there was measurement noninvariance. However, the model should be interpreted with caution because when the class separation was not large, that is, smaller DIF magnitude coupled with fewer DIF items, the model tended to select the 1-class model. Only including the covariate effect on the latent class membership might lead to over-extraction of latent classes if there were other covariate effects but were omitted. In other words, additional latent classes emerged due to the omitted effects. Nevertheless, if the direct covariate effects on all the items were modeled in the most complex analysis model, the 2-class scalar model was selected instead of the 2-class metric model. This might be because the covariate effects on items absorbed the intercept noninvariance.

Although including covariate effects on the latent class variable and the factor seemed to work well in identifying the correct model under measurement noninvariance, this approach did not have the best model fit across simulation conditions. That is, with intercept noninvariance, the analysis model that matched the population model yielded the best fit to the data, as compared with other analysis models. Therefore, instead of recommending a single model that includes covariate effects on the latent class variable and the factor, we suggest that applied researchers can use this model as a starting point in the class enumeration process and compare that model to other analysis models. That is, first, identify the number of latent classes and the level of invariance using the model that includes covariate effects on the latent class variable and the factor. Second, fit other analysis models that modeled the covariate effects differently with the number of latent classes and the level of invariance identified at the first step. The fitted analysis models include the model that has the covariate effect on the latent class membership only, and the most complex analysis models that includes covariate effects on the latent class membership, the factor, and all items. Third, compare the fit of the analysis models with the model used at the first step and choose an analysis model that yields the best fit.

It is important to note that the recommendation provided above regarding the inclusion of covariate effects in testing measurement invariance are assuming measurement noninvariance. In other words, the recommendation can be taken if applied researchers hypothesize measurement noninvariance based on substantive theory or previous research. If the hypothesis is wrong and measurement invariance actually holds, following the recommendation would lead to biased results. That is, although including covariates on the latent class variable and the factor is recommended in testing measurement invariance across latent classes assuming measurement noninvariance, this way of modeling covariate effects would not work in testing factor mean differences across classes when measurement invariance actually held. In other words, if the only difference across classes was in the factor mean, including covariate effects on both the latent class variable and the factor would not lead to the identification of the scalar invariance model. Instead, the 1-class model would be supported, because the factor mean difference would be absorbed by the covariate effect on the factor, as discussed earlier. In this case, including covariate effects on both the latent class membership and the factor was not a good option. If applied researchers had no hypothesis about whether measurement invariance holds or not, including the covariate effect only on the latent class membership seems to be a reasonable approach. This approach would lead to satisfactory class enumeration results except when the covariate effect on the factor was strong or there were direct effects on the items. Then the fit of this model can be compared with other analysis models, including the model that has the covariate effects on both the latent class membership and the factor, and the model that includes the covariate effects on the latent class membership, the factor, and all the items. The number of latent classes and the level of invariance for these analysis models are fixed to be the same as the solution identified by the model with covariate effect on the latent class membership only.

Some additional words of caution need to be pointed out in interpreting or generalizing results and recommendations. First, this study focused on the impact of excluding/misspecifying covariate effects on the class enumeration for MI testing with factor mixture modeling. Future research could further examine the classification accuracy and parameter estimates once the correct solution and the best-fitting analysis model is identified. Second, only measurement noninvariance in intercepts was considered in this study and factor loadings were constrained to be equal across latent classes. It would be interesting to examine the performance of the analysis models under loading noninvariance only or both loading and intercept noninvariance. Third, this study focused on detecting measurement noninvariance across latent classes (i.e., latent DIF), but results showed that the presence of observed DIF related with the covariate could distort the results for testing latent DIF. That is, when there was a strong direct covariate effect on the item (i.e., observed DIF), all analysis models fitted in this study failed to detect the correct level of measurement invariance across latent classes. Future methodological research can further examine how to conduct measurement invariance testing across latent classes with the presence of observed DIF and what approaches or model building process should be used to identify latent and observed DIF. Note that Tay et al. (2011) proposed a procedure to test latent and observed DIF in the item response theory framework. Masyn (2017) proposed a stepwise MIMIC approach to testing observed nonuniform and uniform DIF related with a covariate in latent class analysis. Nevertheless, simulation studies are needed in the future to investigate the performance of the approach and other possible approaches.

In summary, this study shows that covariates should be included in the factor mixture modeling when the focus is to identify the number of latent classes and the level of invariance. It is not a good option to exclude the covariate effects because this approach would lead to the 1-class solution. Instead, when testing measurement invariance, the covariate effect on the latent class membership can be included if there is no priori hypothesis regarding whether measurement invariance might hold or not. If measurement invariance might not hold based on substantive theory or prior research, the covariate effects on the latent class membership and the factor can be included to identify a solution. In addition, it is good to know that larger sample size and larger class separation would help the class enumeration.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. doi:10.1109/TAC.1974.1100705

Allan, N. P., Raines, A. M., Capron, D. W., Norr, A. M., Zvolensky, M. J., & Schmidt, N. B. (2014). Identification of anxiety sensitivity classes and clinical cut-scores in a sample of adult smokers: Results from a factor mixture model. *Journal of Anxiety Disorders, 28*(7), 696-703. doi: 10.1016/j.janxdis.2014.07.006

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(3), 329-341. doi: 10.1080/10705511.2014.915181

Bernstein, A., Stickle, T. R., & Schmidt, N. B. (2013). Factor mixture model of anxiety sensitivity and anxiety psychopathology vulnerability. *Journal of Affective Disorders*, *149*(1-3), 406-417. doi: 10.1016/j.jad.2012.11.024

Bozdogan, H. (1987). Model selection and Akaike’s information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370. doi:10.1007/BF02294361

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005-1018. doi: 10.1037/a0013193

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233-255. doi: 10.1207/S15328007SEM0902_5

Cho, S., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics*, *35*(3), 336–370. doi: 10.3102/1076998609353111

Clark, S. L., Muthén, B. O., Kaprio, J., D'Onofrio, B. M., Viken, R., & Rose, R. J. (2013). Models and strategies for factor mixture analysis: An example concerning the structural underlying psychological disorders. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(4), 681-703. doi: 10.1080/10705511.2013.824786

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis for differential item functioning. *Journal of Educational Measurement, 42*(2), 133-148. doi: 10.1111/j.1745-3984.2005.00007

Curran, P. J., Cole, V., Bauer, D. J., Hussong, A. M., & Gottfredson, N. (2016). Improving factor score estimation through the use of observed background characteristics. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(6), 827–844. doi: 10.1080/10705511.2016.1220839

De Ayala, R. J., Kim, S.-H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing, 2*(3-4), 243-276. doi: 10.1080/15305058.2002.9669495

Dimitrov, D. M., Al-Saud, F. A. A.-M., & Alsadaawi, A. S. (2015). Investigating population heterogeneity and interaction effects of covariates: The case of a large-scale assessment for teacher licensure in Saudi Arabia. *Journal of Psychoeducational Assessment, 33*(7), 674-686. doi: 10.1177/0734282914562121

Dyer, W. J., & Day, R. D. (2015). Investigating family shared realities with factor mixture modeling. *Journal of Marriage and Family, 77*(1), 191-208. doi: 10.1111/jomf.12158

Elhai, J. D., Naifeh, J. A., Forbes, D., Ractliffe, K. C., & Tamburrino, M. (2011). Heterogeneity in clinical presentations of posttraumatic stress disorder among medical patients: Testing factor structure variation using factor mixture modeling. *Journal of Traumatic Stress, 24*(4), 435-443. doi: 10.1002/jts.20653

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: A Multidisciplinary Journal, 14*(2), 202–226. doi:10.1080/10705510701293478

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

Jackman, M. G.-A. (2012). *A Monte Carlo investigation of the performance of factor mixture modeling in the detection of differential item functioning* (Doctoral dissertation). Retrieved from ProQuest dissertation & these: A&I. (3480453)

Kim, E. S. (2011). *Testing measurement invariance using MIMIC: Likelihood ratio test and modification indices with a critical value adjustment*. (Doctoral dissertation). Retrieved from ProQuest dissertation & these: A&I. (3486109)

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-21. doi: 10.1080/10705511.2017.1304822

Kim, E. S., Joo, S.-H., Lee, P., Wang, Y., & Stark, S. (2016). Measurement invariance testing across between-level latent classes using multilevel factor mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(6), 870-887. doi: 10.1080/10705511.2016.1196108

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(2), 212–228. doi: 10.1080/10705511.2011.557337

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

Kim, M., Vermunt, J., Bakk, Z., Jaki, T., & Van Horn, M. L. (2016). Modeling predictors of latent classes in regression mixture models. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(4), 601-614. doi: 10.1080/10705511.2016.1158655

Lanza, S. T., Tan, X., & Bray, B. C. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(1), 1–26. doi: 10.1080/10705511.2013.742377

Li, M., & Harring, J. R. (2016). Investigating approaches to estimating covariate effects in growth mixture modeling: A simulation study. *Educational and Psychological Measurement*, 1-26. doi: 10.1177/0013164416653789

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*(3), 767–778. doi:10.1093/biomet/88.3.767

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

Lubke, G., & Muthén, B. O. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(1), 26-47. doi: 10.1080/10705510709336735

Lubke, G., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research, 41*(4), 499-532. doi: 10.1207/s15327906mbr4104_4

Lukočienė, O., & Vermunt, J. K. (2010). Determining the number of components in mixture models for hierarchical data. In A. Fink, L. Berthold, W. Seidel, & A. Ultsch (Eds.), Advances in data analysis, data handling and business intelligence (pp. 241–249). Berlin, Germany: Springer. doi:10.1007/978-3-642-01044-6_22

Maij-de Meij, A. M., Kelderman, H., & Van der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research, 45*(6), 975-999. doi: 10.1080/00273171.2010.533047

Marsh, H. W., Tracey, D. K., & Craven, R. G. (2006). Multidimensional self-concept structure for preadolescents with mild intellectual disabilities. *Educational and Psychological Measurement, 66*(5), 795-818. doi: 10.1177/0013164405285910Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. doi:10.1007/BF02294825

Masyn, K. E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(2), 180-197. doi: 10.1080/10705511.2016.1254049

McLachlan, G., & Peel, D. (2000). Finite mixture models. Hoboken, NJ: Wiley.

Muthén, L. K., & Muthén, B. O. (1998-2004). *M*plus User's guide. Third Edition. Los Angeles, CA: Muthén & Muthén.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 535–569. doi:10.1080/10705510701575396

Nylund-Gibson, K., & Masyn, K. E. (2016). Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(6), 782-797. doi: 10.1080/10705511.2016.1221313

Park, J., & Yu, H.-T. (2016). The impact of ignoring the level of nesting structure in nonparametric multilevel latent class models. *Educational and Psychological Measurement, 76*(5), 824-847. doi: 10.1177/0013164415618240

Ramaswamy, V., Desarbo, W. S., Reibstein, D. J., & Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science, 12(*1), 103–124. doi:10.1287/mksc.12.1.103

Samuelsen, K. M. (2005). Examining differential item functioning from a latent mixture perspective. In G. R. Hancock & K. M. Samuelson (Eds.), Advances in latent variable mixture modeling (pp. 177–197). Charlotte, NC: Information Age.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. doi:10.1214/aos/1176344136

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*(3), 333–343. doi:10.1007/BF02294360

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*(6), 1292–1306. doi: 10.1037/0021-9010.91.6.1292

Tay, L., Newman, D. A., & Vermunt, J. K. (2011). Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organizational Research Methods, 14*(1), 147-176. doi: 10.1177/1094428110366037

Tein, J., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(4), 640–657. doi:10.1080/10705511.2013.824781

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis, 18*(4), 450-469. doi: 10.1093/pan/mpq025

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