TESTING MEASUREMENT INVARIANCE USING MIMIC: LIKELIHOOD RATIO TEST AND MODIFICATION INDICES WITH A CRITICAL VALUE ADJUSTMENT

A Dissertation

by

EUN SOOK KIM

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2011

Major Subject: Educational Psychology

Testing Measurement Invariance Using MIMIC: Likelihood Ratio Test and Modification Indices with a Critical Value Adjustment

TESTING MEASUREMENT INVARIANCE USING MIMIC: LIKELIHOOD RATIO TEST AND MODIFICATION INDICES WITH A CRITICAL VALUE ADJUSTMENT

A Dissertation

by

EUN SOOK KIM

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Victor L. Willson
	Myeongsun Yoon
Committee Members,	Bruce Thompson
	Oi-Man Kwok
	Michael Speed
Head of Department,	Victor L. Willson

August 2011

Major Subject: Educational Psychology

ABSTRACT

Testing Measurement Invariance Using MIMIC: Likelihood Ratio Test and Modification Indices with a Critical Value Adjustment. (August 2011)

Eun Sook Kim, B.S., Pusan National University;

M.S., Texas A&M University

Co-Chairs of Advisory Committee: Dr. Victor L. Willson
Dr. Myeongsun Yoon

Multiple-indicators multiple-causes (MIMIC) modeling is often employed for measurement invariance testing under the structural equation modeling framework. This Monte Carlo study explored the behaviors of MIMIC as a measurement invariance testing method in different research situations. First, the performance of MIMIC under the factor loading noninvariance conditions was investigated through model fit evaluations and likelihood ratio tests. This study demonstrated that the violation of factor loading invariance was not detected by any of the typically reported model fit indices. Consistently, the likelihood ratio tests for MIMIC models exhibited poor performance in identifying noninvariance in factor loadings. That is, MIMIC was insensitive to the presence of factor loading noninvariance, which implies that factor loading invariance should be examined through other measurement invariance testing techniques.

To control Type I error inflation in detecting the noninvariance of intercepts or thresholds, this simulation study with both continuous and categorical variables employed the likelihood ratio test with two critical value adjustment strategies, Oort

adjustment and Bonferroni correction. The simulation results showed that the likelihood ratio test with Oort adjustment not only controlled Type I error rates below the basal Type I error rates but also maintained high power across study conditions. However, it was observed that power to detect the noninvariant variables slightly attenuated with multiple (i.e., two) noninvariant variables in a model.

Given that the modification index is the chi-square difference after relaxing one parameter for estimation, this study investigated modification indices under four research scenarios based on a combination of the cutoffs of modification indices and the procedures of model modification: (a) the noniterative method (i.e., modification indices at the initial stage of model modification) using the conventional critical value, (b) the noniterative method using the Oort adjusted critical value, (c) the iterative procedure of model modification using the conventional critical value, and (d) the iterative procedure using the Oort adjustment. The iterative model search procedure using modification indices showed high performance in detecting noninvariant variables even without critical value adjustment, which indicates that iterative model search specification does not require critical value adjustment in identifying the noninvariance correctly. On the other hand, when the noniterative procedure was used, the Oort adjustment yielded adequate results.

DEDICATION

I dedicated this dissertation to my father who showed me love and wisdom through his short journey of life and to my mother who always encourages me to pursue my dream and to be an independent woman.

ACKNOWLEDGEMENTS

People barely accomplish anything for themselves. The accomplishment I made with this dissertation could not be possible without support and care of so many people around me throughout my doctoral life. Some of them directed and guided me to the world of research; some harnessed my knowledge and skills in the field of study through questions and discussions; others supported my study sometimes with sincere advice and sometimes with encouragement and cheers. Thanks to all these people, I could enjoy every minute here in Texas A&M University.

I learned the importance of scientifically-based research from my committee advisor, Dr. Willson. Dr. Willson guided me to the research connected with the issues and problems in the real world. The lessons I learned from him are invaluable not only for my dissertation but also throughout my career as an educational researcher. I thank my committee co-chair, Dr. Yoon for her rigorous review of my work and sincere comments. I also thank my committee members, Dr. Thompson, Dr. Kwok, and Dr. Speed who have inspired me with intriguing questions and thought-evoking instructions since the beginning of my graduate studies.

I would like to thank Dr. Hall for his care and concerns about me and my study. Without his technical support for the Educational Research and Evaluation Laboratory (EREL) my simulation studies could not be possible. I appreciate Dr. Hall's efforts and the department's support to keep the EREL the best work place with updated software programs and equipments for graduate students' research.

My special thanks go to my children. I appreciate their understanding and support whenever I went through difficulties in my study. On a Christmas Eve when I came back from the simulation work, I was touched with their present: all day long they completed house chores from laundry to dish wash, from living room to bath room to save Mom's time from house work. They knew that I wanted time for a Christmas present. After four years of doctoral studies, I realize with gratification that they outgrow me emotionally as well as physically.

I cannot express my gratitude and respect to my husband with any words. I could enjoy my study without stress and pressure because I knew that my husband would be proud of me as I am even when I fail to receive a doctoral degree. I appreciate his patience and endurance listening to all my complaints, disappointments, worries about my studies and his trust that I will achieve my goals.

Finally, I would like to say thanks to my friends. We shared knowledge and information along with laughter and tears encouraging each other for our studies. Thanks to all my family members in Korea who pray for me and make my dream come true.

TABLE OF CONTENTS

CHAPTER Page

LIST OF FIGURES

LIST OF TABLES

TABLE Page

CHAPTER I

INTRODUCTION: THE IMPORTANCE OF RESEARCH

With increasing attention to the measurement invariance across groups, testing measurement invariance has become a common practice before utilizing a measure in social science (Schmitt & Kuljanin, 2008). By definition, measurement invariance holds when people with identical ability in different groups have an identical probability to endorse a certain variable regardless of group membership (Mellenbergh, 1989). Although multiple group confirmatory factor analysis (CFA) is the preferred method for testing measurement invariance, multiple-indicators multiple-causes (MIMIC, Jöreskog & Goldberger, 1975) modeling is also often employed (Fleishman, Spector, & Altman, 2002; McCarthy, Pedersen, & D'Amico, 2009; Muthén, Kao, & Burstein, 1991; Rubio, Berg-Weger, Tebb, & Rauch, 2003).

MIMIC modeling allows the assessment of measurement invariance and latent mean difference across groups by incorporating grouping variables as covariates instead of testing separate models for each group as in multiple group CFA. Thus, MIMIC modeling can easily facilitate measurement invariance tests on multiple background variables and their interactions (e.g., gender, race, and gender by race) as well as on more than two groups per grouping variable of interest (e.g., four different race groups; Fleishman et al., 2002; Ainsworth, 2008). MIMIC modeling utilizes the total sample size

---

This dissertation follows the style of *Educational and Psychological Measurement.*

without splitting the data into groups (i.e., a single variance covariance matrix rather than a separate variance covariance matrix for each group), and may give more stable estimation when sample size is of concern. With the model flexibility and the sample size advantage, MIMIC modeling is prevalent in social science, especially for testing the equivalence of latent means across groups.

However, there are a couple of unsolved issues in MIMIC modeling which are the focal interests of this study. First, MIMIC modeling has an inherent limitation in testing partial invariance. That is to say, when measurement invariance is violated at some levels, MIMIC will not be able to locate the violations of measurement invariance (e.g., lack of invariance at factor loadings or intercepts) unlike multiple group CFA. Otherwise stated, a sequential procedure of measurement invariance testing from configural invariance (equivalence of factor structure across groups) to strict invariance (equivalence of factor structure, factor loading, intercept, and unique variance) cannot be conducted with MIMIC modeling. In addition, previous simulation studies on MIMIC as a measurement invariance test consistently reported high Type I error rates over the nominal level (e.g., Finch, 2005). Considering the prevalence of MIMIC in substantial fields of research for measurement invariance testing, this paper focused on the statistical approach to control the Type I error inflation, specifically when researchers employ the likelihood ratio test to detect noninvariant variables with MIMIC modeling.

When examining measurement invariance, researchers often rely on the likelihood ratio (LR) test. In the LR test, a model in which a variable or a set of parameters (e.g., factor loadings of all variables) are freely estimated is compared to a

baseline model with invariance constraints. The statistical significance between two rival models indicates the lack of invariance on the tested variable or the tested set of parameters (e.g., the violation of metric invariance or weak invariance). However, there are a couple of difficulties researchers may encounter in utilizing the likelihood ratio test. As a type of statistical significance testing, the likelihood ratio test highly depends on sample size. In large samples, a trivial chi-square difference can be detected as statistical significance. Another problem with a likelihood ratio test in invariance studies is that the chi-square difference between two models might not follow the chi-square distribution when the baseline model is misspecified. (Kim & Yoon, 2011; Oort, 1992, 1998; Stark, Chernyshenko, & Drasgow, 2006; Yuan & Bentler, 2004). The violation of distributional assumption, in turn, tends to lead to the inflation of Type I error in the likelihood ratio test.

In measurement invariance literature, Type I error and false positive are interchangeably used generally indicating the false detection of invariant variables as differential item function (DIF, or measurement noninvariance). The proportion of false positive cases across simulation replications is often defined as a Type I error rate or false positive rate. The Type I error inflation of the LR test with a misspecified baseline model can be a critical issue in MIMIC modeling because MIMIC modeling entails the LR test for measurement invariance testing. MIMIC with a grouping variable as a covariate assumes strict invariance, namely the equivalence of factor loadings, intercepts, and unique variances over groups (Thompson & Green, 2006). When the MIMIC model is utilized as a baseline model for measurement invariance testing and it

contains any noninvariant variable, Type I error inflation is expected because of the model misspecification (i.e., the violation of the strict invariance assumption). The details of the MIMIC model in measurement invariance testing will be discussed later.

A number of simulation studies on measurement invariance testing reported high Type I error rates when MIMIC modeling was used to detect noninvariant variables. Oort (1998) studied MIMIC with categorical indicators, and reported Type I error rates between .15 and .20. In Finch's (2005) study using MIMIC modeling, Type I error rates ranged from .08 to .22 (mean of .12) depending on the simulation conditions. Navas-Ara and Gomez-Benito (2002) reported Type I error rate of .36. In the study of Wang and colleagues (Wang, Shih, & Yang, 2009), MIMIC modeling showed the false positive rates as high as .48.

With the report of Type I error inflation, a body of literature contributed to explain and control for the Type I error inflation. Navas-Ara and Gomez-Benito (2002) utilized a scale purification method with categorical items and reported the improvement of Type I error rates (.07). The scale purification method is an iterative process in which biased items detected in the initial analysis are eliminated, and the bias detection procedure is repeated with unbiased items to identify remaining bias until no item is detected as noninvariance. Wang, Shih, and Yang (2009) applied scale purification procedures to MIMIC modeling and controlled for the Type I error below .10 for most study conditions although the Type I error rate still reached .24 with the 40% DIF contamination. In case of the LR test, Stark et al. (2006) suggested the Bonferroni correction of critical values calling attention to the chi-square statistic inflation in a

misspecified baseline model and subsequently Type I error rate elevation. On the other hand, Oort developed a formula to adjust the critical value to control the chi-square inflation of a misspecified baseline model. When the adjustment was applied to the iterative procedures using modification indices, the Type I error rates were reported under the nominal level.

Based on the previous findings, this study investigated the LR test in identifying noninvariant variables with two critical-value adjustment methods (i.e., Bonferroni correction and Oort adjustment) to control for the Type I error inflation. Following the reasoning that the chi-square difference in the LR test is not likely to follow the chi-square distribution when the baseline model is incorrect, this simulation study reevaluated the Oort adjustment in a variety of study conditions including both continuous and categorical data.

In addition, modification indices were utilized as an indicator of noninvariant variables in MIMIC modeling. A modification index (also called Lagrange multiplier) is the degree of expected chi-square change if a fixed or a constrained parameter is freely estimated (Brown, 2006) while constraining all other parameter estimates at the values obtained in the same analysis. When a model shows a poor fit, researchers often refer to modification indices to improve the model fit. In measurement invariance testing with a MIMIC model, modification indices may provide information about the noninvariant variables because the noninvariant variables under the assumption of invariance of the MIMIC model are likely to be the sources of model misfit.

For modification indices, prior Monte Carlo studies (Oort, 1998; Yoon & Millsap, 2007) demonstrated the superior performance of an iterative search procedure over the noniterative counterpart. In the noniterative procedure, all the parameters in the modification indices are relaxed for free estimation simultaneously. On the other hand, in the iterative search procedure, only a single parameter with the largest modification index is freely estimated. After the free estimation of the designated parameter, another parameter with the largest modification index is allowed to be estimated. This search process is repeated until there is no more modification index. This study, thus, compared iterative and noniterative search procedures of modification indices in combinations with different critical value adjustment methods (Bonferroni correction and Oort adjustment) to determine noninvariant variables in MIMIC modeling.

Incorporating the prior research, this study is expected to make unique contributions to the literature of measurement invariance and MIMIC modeling. First, the previous studies on MIMIC modeling as a method to detect measurement noninvariance did not include the source of noninvariance (i.e., factor loading noninvariance or intercept noninvariance) as a study condition. One plausible reason why the source of noninvariance is not of concern in MIMIC studies is that MIMIC modeling is typically used in testing the equivalence of intercepts in continuous data (or thresholds in categorical data) or in testing the latent factor mean difference across groups under the assumption of factor loading invariance. However, the performance of MIMIC modeling in the presence of factor loading noninvariance has not yet been explicitly investigated.

Second, this simulation study considered both categorical and continuous variables in identifying noninvariance using MIMIC. It appears that the research on measurement invariance testing for categorical and continuous data has developed and advanced in its own way. For categorical data, the research on measurement invariance has evolved typically in relation to IRT and focuses on the detection of biased items or DIF. On the other hand, the major interest in measurement invariance testing with continuous measures is on the establishment of the level of measurement invariance such as metric, scalar, and strict invariance. The two data types (i.e., continuous and categorical) were rarely studied together under the same interests of study.

Third, Oort (1998) studied the performance of MIMIC under the term restricted factor analysis (RFA) in the detection of DIF with categorical variables (either 2 or 7 response categories). However, due to the technical limitation that study failed to incorporate the weighted least squares (WLS) estimation which is appropriate for categorical measures. Because WLS is available as a default for categorical data in Mplus (Muthén & Muthén, 2008b), a study of the performance of MIMIC modeling with categorical variables can be done properly. This simulation study reevaluated the Oort adjustment with an appropriate data analytic method for categorical items (i.e., implementing threshold structures and using weighted least squares with robust mean and variance) in MIMIC modeling.

Fourth, the current study examined modification indices as indicators of measurement noninvariance and investigated the performance of modification indices of MIMIC modeling under various research situations such as different data type

(continuous, dichotomous, or polytomous) and different search procedures (iterative or noniterative) with different critical value adjustment methods (no adjustment, Bonferroni correction, or Oort adjustment).

In summary, this study proposed the following research questions:

1. How sensitive is MIMIC modeling to the violation of factor loading invariance? The MIMIC model assumes strict invariance or the invariance of factor loadings, intercepts, and residual variance when used with grouping variables as covariates. Given that the MIMIC model is typically employed for the latent mean comparison across groups and for the intercept invariance testing, the sensitivity of the MIMIC model to the violation of factor loading invariance is of question. This study examined the model fits, specifically chi-square goodness of fit, CFI, RMSEA, and SRMR (or WRMR for categorical data) when factor loading noninvariance existed in the MIMIC model. In addition, whether measurement invariance testing can detect the lack of invariance in factor loadings was examined.
2. How does MIMIC modeling behave when the invariance of intercepts over groups is violated? Which critical value adjustment method performs better with MIMIC modeling, Bonferroni correction or Oort adjustment? In the LR test using MIMIC, the chi-square inflation of a baseline model is expected when there is any noninvariant item in the model because the MIMIC assumes strict invariance over groups. This study investigated proper strategies to control for

high Type I error rates: Bonferroni correction and Oort adjustment on chi-square critical values.

3. Regarding modification indices, what is an optimal strategy in the detection of noninvariance: iterative or noniterative specification search? What is an optimal cutoff of modification indices? A modification index is, simply speaking, the chi-square difference when a parameter is relaxed for estimation (Brown, 2006). In detecting the violation of measurement invariance, the use of modification indices exceeding a certain chi-square values could replace the LR test. From this reasoning, modification indices were inspected as an alternative to the LR test. The critical value of modification indices was adjusted with the Oort correction and the model modification procedure with Oort-adjusted critical value was compared with the conventional model modification procedure. Along with the critical value adjustment, the iterative and noniterative model search procedures were compared as well.

CHAPTER II

LITERATURE REVIEW

The current review of literature in measurement invariance consists of four subsections: definition of measurement invariance and factorial invariance, measurement invariance with ordinal measures, levels of measurement invariance, and issues in measurement invariance studies. The last section includes the impacts of measurement noninvariance, partial measurement invariance, and the criteria of measurement invariance testing which all call for further studies. Followed by measurement invariance, the framework of MIMIC modeling as a measurement invariance testing method and modification indices were discussed.

Measurement Invariance

Measurement invariance has drawn attentions of researchers in social science since early studies (e.g., Meredith, 1993) and a systematic review (Vandenberg & Lance, 2000) on this topic. In a recent review of measurement invariance, Schmitt and Kuljanin (2008) reported increased interests and common practices of testing measurement invariance, which implies that social scientists has been cognizant of the importance of measurement invariance in the use of a measure.

Measurement invariance holds when a measure utilized under different conditions yields the same observed scores for people who have identical attributes being measured (Drasgow, 1987; Meade & Bauer, 2007; Schmitt and Kuljanin, 2008). Measurement invariance testing is commonly practiced in several measurement

conditions (Meade & Bauer, 2007; Vandenberg & Lance, 2000): measurement invariance (a) across subgroups of a population, (b) over longitudinal changes, and (c) over different mediums of measurement. Typical subgroups of a population include ethnicity, gender, and age. With the increase in cross-cultural studies, measurement invariance over different ethnic and cultural groups is often of interest (e.g., Blake, Kim, & Lease, 2011; Riordan & Vandenberg, 1994). For studies conducted over time with repeated measures, the invariance of a measure across different time points can be questioned. There are also some study conditions in which the methods of measurement are not consistent. For example, a test developed in a language is translated and utilized in a different language. In other cases, the scores of a test presented in different formats (e.g., pencil-and-paper or online; Meade, Michels, & Lautenschlager, 2007) are compared. In all these study conditions, measurement invariance is of great concern.

Although measurement invariance is an issue under diverse study conditions, in this study the scope of measurement invariance is limited to group comparisons. However, all defined equations in the following can be also applied to other research settings such as measurement invariance in longitudinal models.

Measurement invariance is mathematically equivalent to the fact that the conditional probability to attain an observed score given ability is independent of group membership (Mellenbergh, 1989; Meredith & Millsap, 1992; Yoon & Millsap, 2007).

$$P(X \mid \xi, G) = P(X \mid \xi), \tag{2.1}$$

where X is the observed score, ξ is the latent construct, and G denotes group membership. That is, measurement invariance holds when persons of identical ability (or attributes) on a construct have the same probability distribution of observed scores regardless of group membership.

Factorial Invariance

A measurement invariance study is commonly conducted with linear confirmatory factor analysis (CFA). Measurement invariance in a factor model is called factorial invariance (Meredith, 1993; Widaman & Reise, 1997; Yoon, 2008). Measurement invariance is a broad term encompassing linear and nonlinear relationship between observed variables and latent factors taking into account the whole score distribution (Yoon, 2008). On the other hand, factorial invariance, a special case of measurement invariance, is expressed within a linear factor model with mean and covariance structures. Under the CFA framework, factorial invariance is defined as the equivalence of parameters specified in the model across groups. Therefore, depending on the parameters in testing of invariance, different levels of factorial invariance can be determined. The levels of factorial invariance will be discussed later.

Assuming a single unidimensional factor, we can specify the relationship between the common factor scores (ξ_i) and the continuous observed scores (X_{ij}) in ordinary CFA as follows:

$$X_{ij} = \tau_j + \lambda_j \xi_j + \delta_{ij}, \qquad (2.2)$$

where X_{ij} is an observed score of an individual *i* on a variable *j*; τ_j and λ_j are an intercept and a factor loading of a variable *j*, respectively; ξ_i is a common factor score of an individual *i*, and δ_{ij} is a unique factor score. A measurement model of a single factor with six observed variables is illustrated in Figure 1. Equation 2.2 expands to specify the relationships between multiple common factors and multiple observed variables in a matrix format as follows:

$$X = \tau + \Lambda_x \xi + \delta, \tag{2.3}$$

where X is a vector of observed variables, τ is a vector of intercepts, Λ_x is matrix of factor loadings, ξ is a vector of common factors, and δ is a vector of unique variables (Kaplan, 2009). Under the assumption that the common factors and unique factors are uncorrelated, that is, $\text{cov}(\xi,\delta) = 0$, the covariance structure is defined as:

$$\Sigma = \Lambda_X \Phi \Lambda_X' + \Theta_\delta, \tag{2.4}$$

where Σ is a population covariance matrix, Φ is a variance covariance matrix for factors, and Θ_δ is a variance covariance matrix for the unique factors. Further assuming that the expected value of a unique variable is zero, $E(\delta) = 0$, and the expected value of a common factor is defined as $E(\xi) = \kappa$, the mean structure of a general factor model is derived from Equation 2.3 as follows:

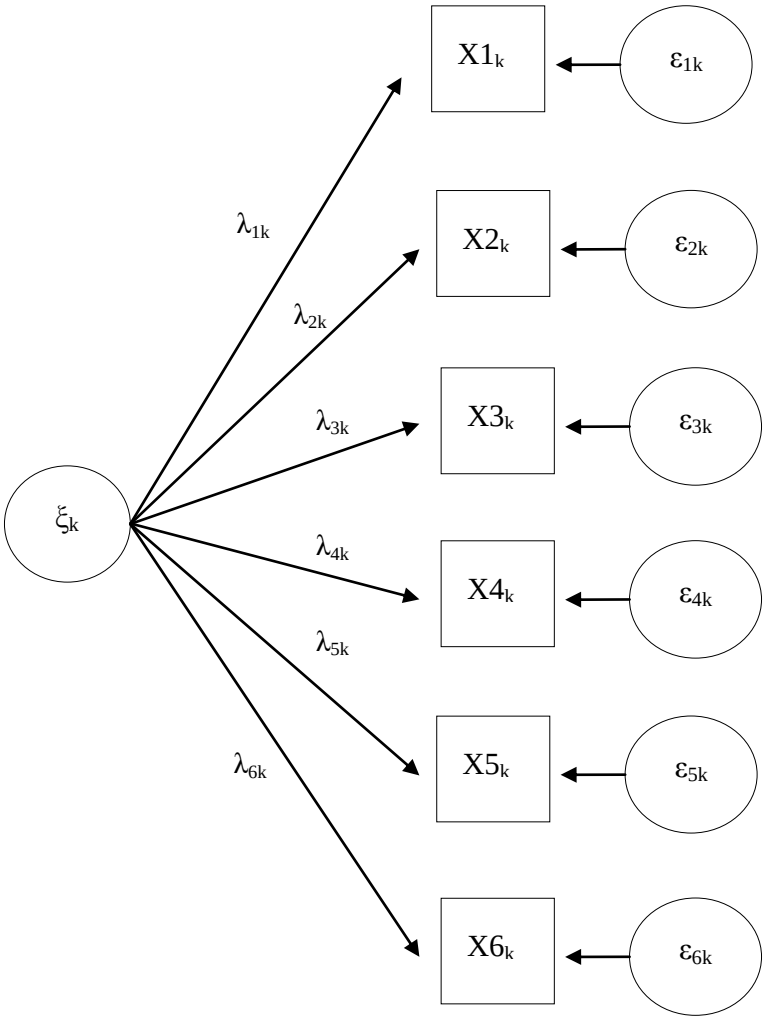

Figure 1. Multiple-group CFA. *k* denotes group membership.

$$E(X) = \tau + \Lambda_X \kappa, \tag{2.5}$$

where κ is a vector of factor means.

The general structural equation models defined in Equation 2.3 through Equation 2.5 are expanded to multiple group CFA model by incorporating a group indicator.

$$\begin{aligned} X_g &= \tau_g + \Lambda_{xg}\xi_g + \delta_g, \\ \Sigma_g &= \Lambda_{Xg}\Phi_g\Lambda_{Xg}'+\Theta_{\delta g}, \\ E(X_g) &= \tau_g + \Lambda_{Xg}\kappa_g, \end{aligned} \tag{2.6}$$

where a subscript *g* is a group indicator (*g* = 1, 2, …, G) and others are as defined above. Under the assumption of normal distribution of observed scores (X_{ij}), factorial invariance holds if the conditional mean and variance covariance of observed scores given factor scores are independent of group membership (*g*). To put it another way, the equivalence of a set of parameters in the former equations across groups indicates factorial invariance. Thus, factorial invariance can be tested with a series of null hypotheses which denote identical parameters over groups.

$$\begin{aligned} H_{\Sigma 0} &: \Sigma_1 = \Sigma_2 = ... = \Sigma_G. \\ H_{\Lambda 0} &: \Lambda_1 = \Lambda_2 = ... = \Lambda_G. \\ H_{\tau 0} &: \tau_1 = \tau_2 = ... = \tau_G. \end{aligned} \tag{2.7}$$

$$H_{\Theta_\delta 0} : \Theta_{\delta 1} = \Theta_{\delta 2} = ... = \Theta_{\delta G}.$$

$$H_{\Phi 0} : \Phi_1 = \Phi_2 = ... = \Phi_G.$$

$$H_{\kappa 0} : \kappa_1 = \kappa_2 = ... = \kappa_G.$$

Precisely speaking, factorial invariance testing includes the first four null hypotheses which test the invariance of variance covariance matrices of observed variables, factor loadings, intercepts, and unique variance of observed variables, respectively. The last two null hypotheses test the equalities of factor variance covariance and factor means over groups, respectively.

Measurement Invariance with Ordinal Measures

So far, the discussions on the factorial invariance are limited to continuous observed variables which assume multivariate normal distributions. However, the items of a measure are often discrete such as dichotomous or polytomous. Although ordered categorical variables with more than five categories are often treated as continuous variables in practice, the ordinary linear factor model is not appropriate for ordered categorical variables because the normality assumption and the linear relationship between the observed variables and the latent factors of the ordinary factor model do not hold for ordinal measures unless the distribution has been validated through methods such as those developed by Thurstone (1928) and Likert (1932).

To incorporate ordered categorical variables in the factor model, latent response variates are employed. The latent response variates are assumed to be continuous and multivariate normally distributed although the corresponding manifested variables are

discrete and non-normal. Then, the factor model specifies the linear relationship between the latent response variates and the factor scores. A one-factor model with continuous latent response variates, X_{ij}^{*} that underlie the observed scores, X_{ij} is

$$X_{ij}^{*} = \tau_j + \lambda_j \xi_{ij} + \delta_{ij}, \tag{2.8}$$

where all terms are as defined earlier. The variance covariance matrix of the latent response variates is termed tetrachoric variance-covariance for dichotomous variables and polychoric variance-covariance for polytomous variables. These latent correlation matrices are derived from the variance covariance matrices of observed variables and utilized in the factor analysis of ordered categorical variables.

The relationship between the latent response variates and the observed variables is modeled with a threshold structure.

$$\begin{aligned} X_{\mathrm{ij}} &= 0, \text{ if } \nu_{j0} < X_{ij}^{*} \le \nu_{j1}, \\ X_{\mathrm{ij}} &= 1, \text{ if } \nu_{j1} < X_{ij}^{*} \le \nu_{j2}, \\ &\cdots \\ X_{\mathrm{ij}} &= c, \text{ if } \nu_{jc} < X_{ij}^{*} \le \nu_{j(c+1)}, \end{aligned} \tag{2.9}$$

where c indicates the C ordered-categorical responses of the jth item ($c = 0, 1, \ldots, C - 1$), ν_{jc} is a threshold of the c category response ($\nu_{j0} = -\infty; \nu_{j(c+1)} = \infty$), and other terms

are as defined above (Wirth & Edwards, 2007). That is, the observed categorical responses are determined by a set of thresholds in relation to the latent response variates. The number of thresholds of a variable equals the number of categories of the variable minus one. For example, binary variables have a single threshold. If the threshold of a dichotomous variable is -1.0, then any latent variate score below -1.0 is manifested as 0. On the other hand, the latent variate score of -1.0 or above corresponds to the observed response score of 1. By embedding a threshold structure in a model, multiple group CFA for categorical variables includes the invariance of a threshold structure. Thus, the null hypothesis of equal thresholds over groups ($H_{\nu_c 0} : \nu_{c1} = \nu_{c2} = ... = \nu_{cG}$) is tested for ordinal measures instead of the equivalence of intercepts which is tested for continuous variables. The relationship between the latent response variates and the manifested variables is illustrated in Figure 2.

Identification. In the estimation of unknown parameters, the number of available pieces of information is critical since the model is not determined or identified when the available information is short of the number of parameters to be estimated. Simply speaking, we cannot derive more than what we have. For example, when a mean structure is entered in a model, the number of pieces of information we can utilize is the number of observed means. For example, with six observed variables, six observed means are used for model estimation. However, the number of parameters to estimate exceeds the number of available pieces of information because the number of parameters

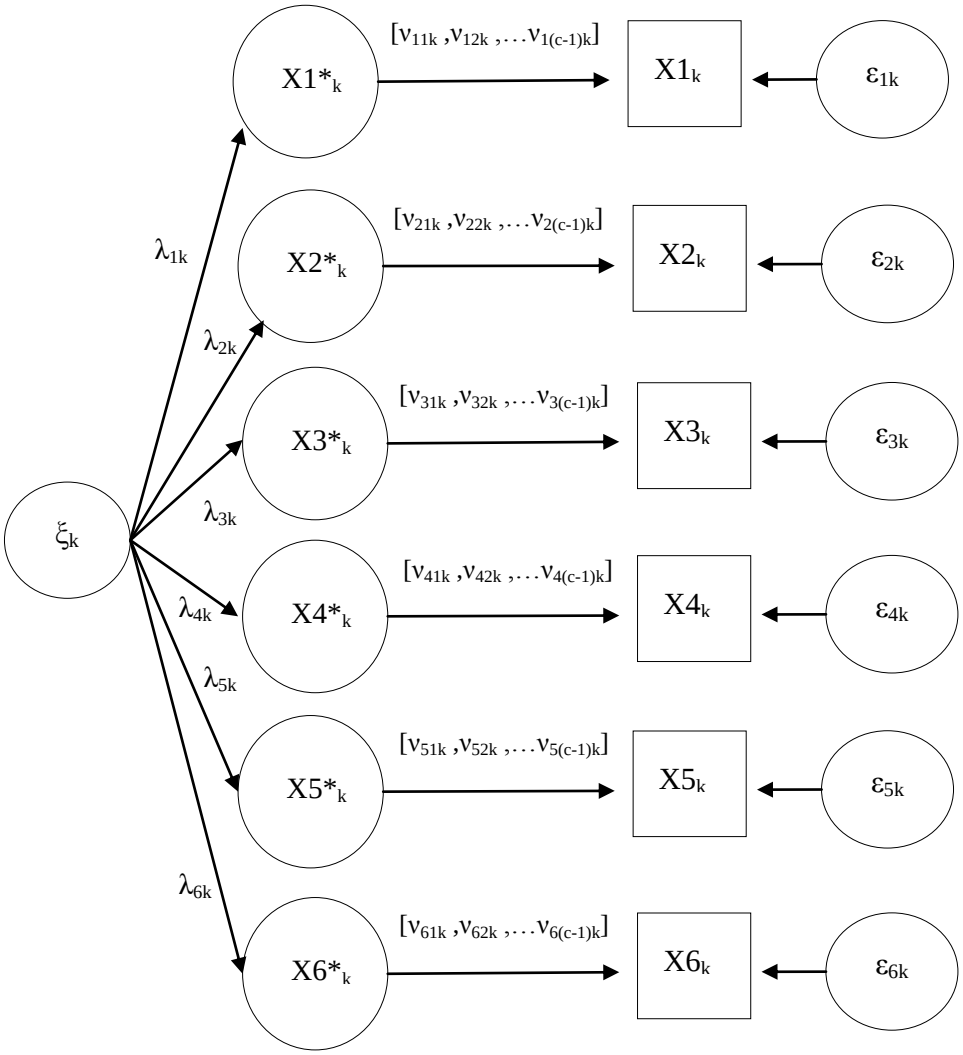

Figure 2. Multiple-group categorical CFA. *k* denotes group membership; *c* is the number of categories in each item.

to estimate includes factor means in addition to intercepts (or means of observed variables). That is, some constraints on the parameter estimates are necessary for identification. For multiple group analysis with continuous observed variables, all factor means of one group (reference group, hereafter) are typically fixed at zero, and the corresponding parameters of other groups (focal groups, hereafter) are freely estimated (Brown, 2006). Intercepts of one variable are set to be equal for all groups.

The same identification issues emerge when a threshold structure is entailed in the model. Because the distribution of latent response variates is analyzed in relation to a threshold structure for ordinal measures, the mean (μ^*) and covariance (Σ^*) structure of latent response variates along with the threshold structure (ν) should be identified. Millsap and Yun-Tein (2004) extensively discussed the identification of multiple group CFA model with ordinal measures in use of either Mplus (Muthén & Muthén, 2008a) or LISREL (Jöreskog & Sörbom, 2006). In summary, for general factor models allowing the loadings on more than one factor with polytomous variables, the following constraints should be imposed to identify the threshold structure (Millsap & Yun-Tein, 2004; Yoon, 2008): (a) mean and variance of latent variates are constrained at 0 and 1 for reference group and (b) two of thresholds of each variable are restricted equal across groups. To further identify the factor structure, (a) the factor means of reference group is fixed at zero, (b) the intercepts of all groups are constrained at zero, and (c) constraints on factor loadings are imposed to render uniqueness of Λ_G.

Parameterization. Because latent factors and latent response variates do not consist of any real scores, scaling these latent variables (i.e., parameterization) is another

issue in a factor model with ordinal measures. Before the discussion of parameterizations in multiple group CFA with ordinal measures, it should be noted that the parameterization of latent variables are undertaken in the process of identification simultaneously (e.g., fixing factor means of a group to zero renders the scales to the factors). Kamata and Bauer (2008) classified the parameterization of multiple group categorical CFA models into four categories. They considered two types of scaling choices for latent response variates and latent factors, respectively. In scaling of latent response variates, (a) the variances of latent response variates are fixed to unity (termed marginal parameterization or delta parameterization), or (b) the unique variances are constrained to unity (termed conditional or theta parameterization). In scaling of common factors, (a) the scale of a reference indicator is chosen, or (b) the variance of a common factor is fixed at 1. It should be noted that the choice of parameterization is up to research questions, and the choice of different parameterizations does not influence model structure and model fit, but the interpretation of parameters will be different. For factorial invariance testing, the equivalence of unique factors across groups is often of interest. In this case, the choice of theta parameterization (fixing the unique factors of reference group at unity and freely estimating those parameters of other groups) is optimal (Millsap & Yun-Tein, 2004).

Sequence of Measurement Invariance Testing

As introduced earlier, depending on which set of parameters are tested for group equality, different levels of factorial invariance are established. Concerning the sequence of measurement invariance testing, two suggestions are commonly made. First, the full

invariance of all levels is not easily attainable in reality, but also it is not necessary in practice. Second, the order and the choice of different levels of invariance testing mostly depend on research questions and interest (Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000; Widaman & Reise, 1997). These points will be discussed in detail.

1. The invariance of variance covariance matrices of observed variables ($\Sigma_G = \Sigma$)

   The elements of variance covariance matrices of observed variables are compared over groups. This level of invariance indicates the full invariance of a factor model over groups and no further invariance testing is required. However, the invariance of variance covariance matrices is hardly achieved in reality (Schmitt & Kuljanin, 2008). In the systematic review of measurement invariance, Schmitt and Kuljanin (2008) reported that virtually no study conducted the invariance testing of variance covariance matrices since Vandenberg and Lance (2000). Thus, in some literature of measurement invariance this level of invariance is not even included in discussions.

2. The configural invariance

   Widaman and Reise (1997) distinguished this type of invariance from others in that the configural invariance is nonmetric. The configuration of a model over groups is of interest. Thus, the pattern of a factor model is tested for invariance with values of parameters varying over groups. The parameters are allowed to vary except the minimal constraints for identification. The configural invariance model is served as a baseline for subsequent measurement invariance tests. If the configural invariance is violated, the subsequent factorial invariance testing

cannot be done. The violation of configural invariance indicates either that the proposed model may not be correct, showing a poor fit although factor structures are equivalent across groups, or that there is lack of configural invariance across groups. With poor fit of a configural invariance model, it is recommended to fit the model in different ways (e.g., with a total sample or with a different factor structure over groups) to find the source of poor fit.

3. The metric invariance ($\Lambda_G = \Lambda$)

   The metric invariance is also called weak invariance (Meredith, 1993). Before Meredith emphasized the testing of scalar invariance, the metric invariance was mostly tested in factorial invariance testing. The establishment of metric invariance is important because the nonequality of factor loadings suggests that the relationship between the observed variables and the common factors are different across groups. In other words, one unit change in a common factor results in different unit changes in an outcome depending on group membership. The metric invariance is also a prerequisite to identify factor variance covariances over groups (Widaman & Reise, 1997). Once the metric invariance is established, rescaling of latent variables does not affect the relationship among latent variables (i.e., variance covariances of common factors). To compare latent group means, the metric invariance should be attained along with the scalar invariance.

4. The scalar invariance ($\tau_G = \tau$)

The scalar invariance is also called strong invariance (Meredith, 1993). The scalar invariance refers to the equivalence of intercepts across groups. As pointed out earlier, for latent mean comparisons over groups, metric and scalar invariance are necessary. Since Meredith called for the inclusion of mean structure in measurement invariance testing and the establishment of scalar invariance in group mean comparisons, scalar invariance testing becomes a common practice of invariance testing whereas scalar invariance testing was the least conducted in the report by Vandenberg and Lance (2000). The scalar invariance was discussed least because location parameters or intercepts were considered arbitrary and sample specific.

5. The strict invariance ($\Theta_{\delta G} = \Theta_{\delta}$)

   The strict invariance is defined as the homogeneous unique variances over groups. The necessity to establish strict invariance is controversial. It is said that with metric and scalar invariance, the difference in latent means can be appraised. The invariant unique variance over groups is not necessary (Widaman & Reise, 1997). In addition, it is difficult to achieve strict invariance in reality. However, others argued that the unique variances are related to factor loadings and the unique variance should be achieved as well. One misconception related to the strict invariance is that the equivalence of unique variances over groups is considered as identical reliability over groups. However, the strict invariance can be analogous to the reliability equivalence only when the factor variances are invariant over groups (Vandenberg & Lance, 2000).

6. The equivalence of factor variances ($\Phi_G = \Phi$)

   The structural part of a model which includes factor variances and means can be tested for equivalence across groups. Other than the first two steps (the equivalence of variance covariance matrices and the configural invariance), researchers agreed that the equivalence of measurement model (configuration, factor loadings, intercepts) should be established before the invariance of structural part (factor variance covariance, factor mean).

7. The equivalence of factor means ($\kappa_G = \kappa$)

   The equivalence of factor means is viable only when metric and scalar invariance hold. The increasing interest in the invariance of factor means was observed in the review of measurement invariance literature (Schmitt & Kuljanin, 2008). Comparing latent means between groups has merits over comparing observed group means because measurement errors are taken into account in latent factor means. On the other hand, the comparison of observed group means such as ANOVA assumes perfect reliability of measurement which is not likely to be achieved in reality. Although the equality of factor variances and means is discussed in the literature of measurement invariance, the purpose of testing the equivalence of factor variances and means is to examine group differences rather than to establish measurement invariance

Evaluations of Measurement Invariance Testing

Measurement invariance under the SEM framework is typically tested through the likelihood ratio test between a baseline model and sequentially constrained models.

The likelihood ratio test is also called chi-square difference test. The chi-square difference ($\Delta\chi^2$) follows the chi-square distribution with the degrees of freedom difference (Δdf) if the original models meet the assumptions to apply the chi-square goodness-of-fit tests (e.g., multivariate normality). The literature of measurement invariance has led to the conclusion that the chi-square difference test can be too sensitive because it is heavily dependent on the size of a sample. Alternatively, the changes in overall model fit indices were investigated to assess measurement invariance in the likelihood ratio test. Since Cheung and Rensvold (2002) initiated the evaluation of model fit indices as alternatives to chi-square difference testing, a number of simulation studies (Chen, 2007; Fan & Sivo, 2009; Meade, Johnson, & Braddy, 2008) were conducted to derive the cutoffs of the Δ alternative fit indices (ΔAFI) and to examine their performance in detecting the lack of measurement invariance (e.g., power, sensitivity, and generalizability).

However, the findings of simulation studies were not consistent and the suitability of ΔAFI as criteria of measurement invariance testing seemed not promising although ΔAFI was less sensitive to sample size than chi-square difference testing. For example, the cutoff values derived from the percentile of the invariant distributions were not consistent across studies. Cheung and Rensvold recommended that researchers assess CFI changes with the cutoff of .02. However, Meade et al. found the cutoff of .002 for ΔCFI. Meade et al pointed out that many ΔAFIs provided redundant information on measurement invariance. For example, ΔTLI(Tucker-Lewis index), ΔIFI (incremental fit index), and $\Delta\hat{\gamma}$ did not produce additional information over ΔCFI. Because the cutoffs

of ΔCFI and ΔMc (McDonald's Noncentrality Index, 1989) were uniformly applicable across different research conditions, it was recommended to report ΔCFI and ΔMc over other relative fit indices. According to Fan and Sivo (2009), ΔMc showed better performance than other ΔAFIs in detecting the factor mean difference over groups. However, because the performance of ΔAFI in latent group mean difference testing relies highly on model size (e.g., number of factors, number of indicators per factor, etc.) as well as sample size, ΔAFI appears not to be a valid choice to test latent group means.

Partial Invariance

The most salient change in measurement invariance studies since Vandenberg and Lance (2000) was the increasing utilization of partial invariance (Schmitt and Kuljanin, 2008). Partial invariance is one of the current controversial issues in measurement invariance studies because it is argued that only full invariance should be considered for the utilization of a measure. Schmitt and Kuljanin reported that approximately 50% of the reviewed studies conducted partial invariance. Since full invariance is difficult to attain in practice, researchers alternatively consult partial invariance.

In the practice of evaluating partial invariance, it was reported that not many researchers relied upon theoretical considerations and even did not practice post hoc interpretations when they allowed the difference across groups in a set of parameters. Instead, their decision of partial invariance heavily relied upon modification indices and other statistical results. Despite the increasing practice of partial invariance, there has not been much study of the suitability of partial invariance. Millsap and Kwok (2004)

studied the impact of partial measurement invariance on the selection of members and demonstrated that the sensitivity in the selection of focal group members decreased as the degree of partial invariance increased. However, there is a call for further studies on selection bias under partial invariance in different simulation settings such as with categorical variables (Chen, 2008; Millsap & Kwok, 2004). Because the suitability of a partial invariance model over full invariance is not well established yet, the researchers who utilize partial invariance should be cautious in the resulting interpretation and should incorporate theoretical explanations on their decisions of partial invariance.

Importance of Measurement Invariance

Measurement invariance is understood as the negation of bias or as the lack of bias (Borsboom, 2006). Hence, the importance of measurement invariance has been discussed with respect to test bias. According to Borsboom, measurement invariance seems always important in a selection context. For example, suppose a depression measure is biased against males, specifically, males' factor loading is lower than females'. Then, with one unit change in the true depression trait we expect smaller unit change in observed scores for males than for females. In this case, the measure of depression is more sensitive to females' depression but less to males' depression. This measure is not likely to identify males with depression as well as females with depression. If this measure is used to select a patient for appropriate treatment of depression, males in the same depression level is less likely to be selected for treatment.

However, when the measure is utilized for research purposes, measurement invariance is important, but the degree of bias matters (Borsboom, 2006; Meredith &

Teresi, 2006). Borsboom suggested the shift of question from ‘is a test biased’ to ‘does the amount of bias matter?’ Measurement invariance testing allows the statistical assessment on test bias. However, measurement invariance testing is also heavily dependent on statistical significance testing. With a substantive sample size, even very trivial differences can be detected as bias in a measure. Borsboom suggested that measurement invariance should be interpreted in context. For example, the same amount of bias can be very critical in a medical test related to human life, but can be acceptable if the effect size of group mean difference is huge. Beyond statistical significance, therefore, practical significance which focuses on the actual magnitude of group difference (e.g., effect sizes) or clinical significance which evaluates the degree to meet diagnostic criteria after intervention (Thompson, 2006) is important in making decisions of measurement noninvariance.

Multiple Indicators Multiple Causes Modeling

Multiple indicators multiple causes (MIMIC) modeling, in general, allows causal indicators of factors as well as effect indicators. For measurement invariance testing, the MIMIC model includes grouping variables (X_i) in the model as causal indicators (Kaplan, 2009; Kline, 2005; Thompson & Green, 2006). For the grouping variables, different coding schemes (e.g., dummy coding or contrast coding) can be chosen with respect to research purposes. Because group membership is indicated as a predictor in the model, MIMIC modeling does not need a subscript of a group indicator (g) in the equations as multiple group CFA does (see Equation 2.6). The observed score of individual, i on variable, j is related to the latent factor score, η_i as follows:

$$Y_{ij} = \lambda_j \eta_i + \varepsilon_{ij}\,. \tag{2.10}$$

Since a covariate (in this study, a dummy-coded grouping variable) explains the latent factor, η we can further model the latent factor in relation to the covariate:

$$\eta_i = \gamma X_i + \zeta_i\,, \tag{2.11}$$

where X_i denotes a dummy variable indicating group membership, γ is the path coefficient of the grouping variable on the latent factor, and ζ_i is the disturbance of the latent factor (see Figure 3 without a dotted line). Because the expected value of the disturbance of the latent factor equals zero, the expected value of the latent factor is expressed as

$$E(\eta_i) = \gamma X_i\,. \tag{2.12}$$

Therefore, for the reference group ($X_i = 0$) the expected value of the factor scores is zero whereas the focal group ($X_i = 1$) has the expected value of γ. In other words, with a dummy-coded grouping variable (X_i) γ represents the group difference in latent factor means (Hancock, 2001; Hancock, Lawrence, & Nevitt, 2000; Thompson & Green, 2006). That is, the latent factor mean of the focal group ($X_i = 1$) is γ unit higher (or lower) than that of the reference group ($X_i = 0$).

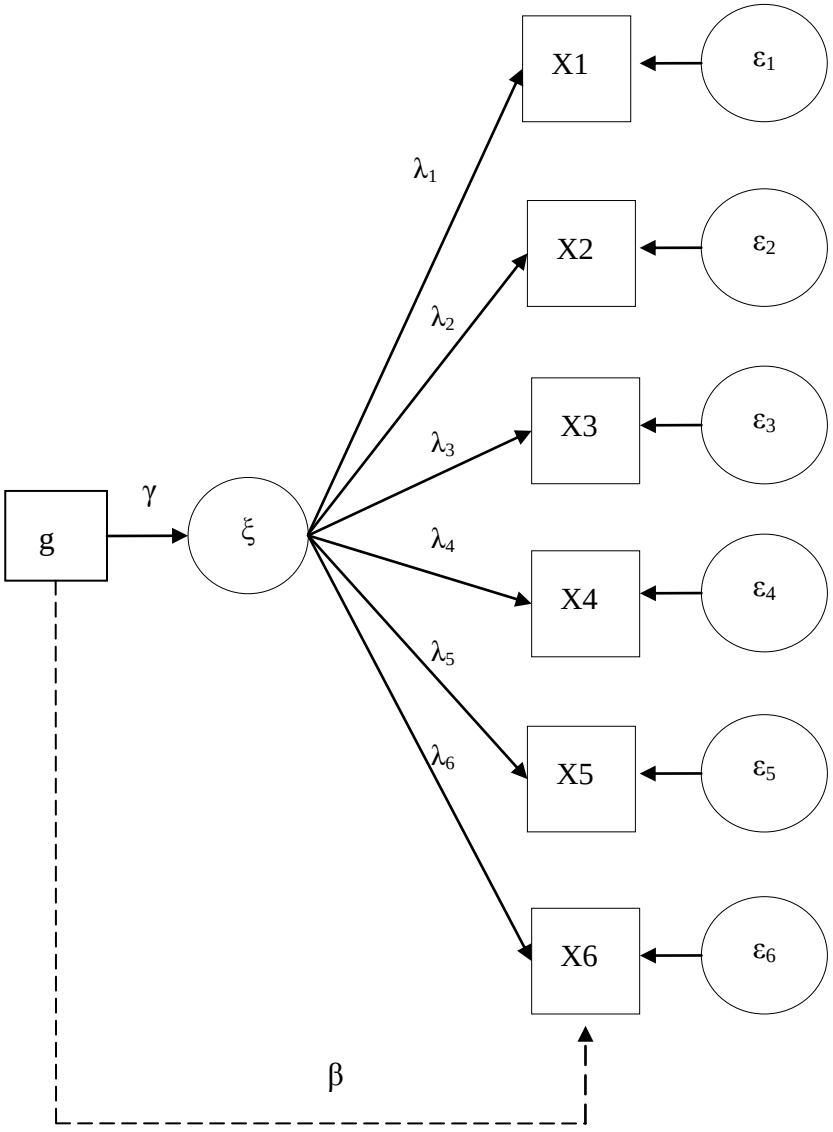

Figure 3. .MIMIC modeling with a grouping variable as a covariate for continuous data. *g* denotes a grouping variable.

Thus, statistical significance testing on γ (H_0: γ = 0) directly tests the statistical significance of latent group mean difference.

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

$$\begin{aligned} Y_{ij}^* &= \lambda_j \eta_i + \varepsilon_{ij}, \\ \eta_i &= \gamma X_i + \zeta_i. \end{aligned} \tag{2.13}$$

The relationship between the observed categorical responses and the latent response variates was explained earlier with the threshold structure and will not be repeated here. To test measurement invariance and the latent group mean difference, a grouping variable X_i is introduced as a causal indicator of the latent factor η as in the continuous model (see Equation 2.11). The MIMIC model which incorporates the threshold structure with latent response variates is illustrated in Figure 4.

Measurement Invariance Testing with MIMIC Modeling

To identify noninvariant variables, the direct path from the grouping variable to the observed variable is created in the model (see Figures 3 and 4 with a dotted line for continuous and categorical variables, respectively). The model with the direct path from the grouping variable to the measured variable can be rewritten as:

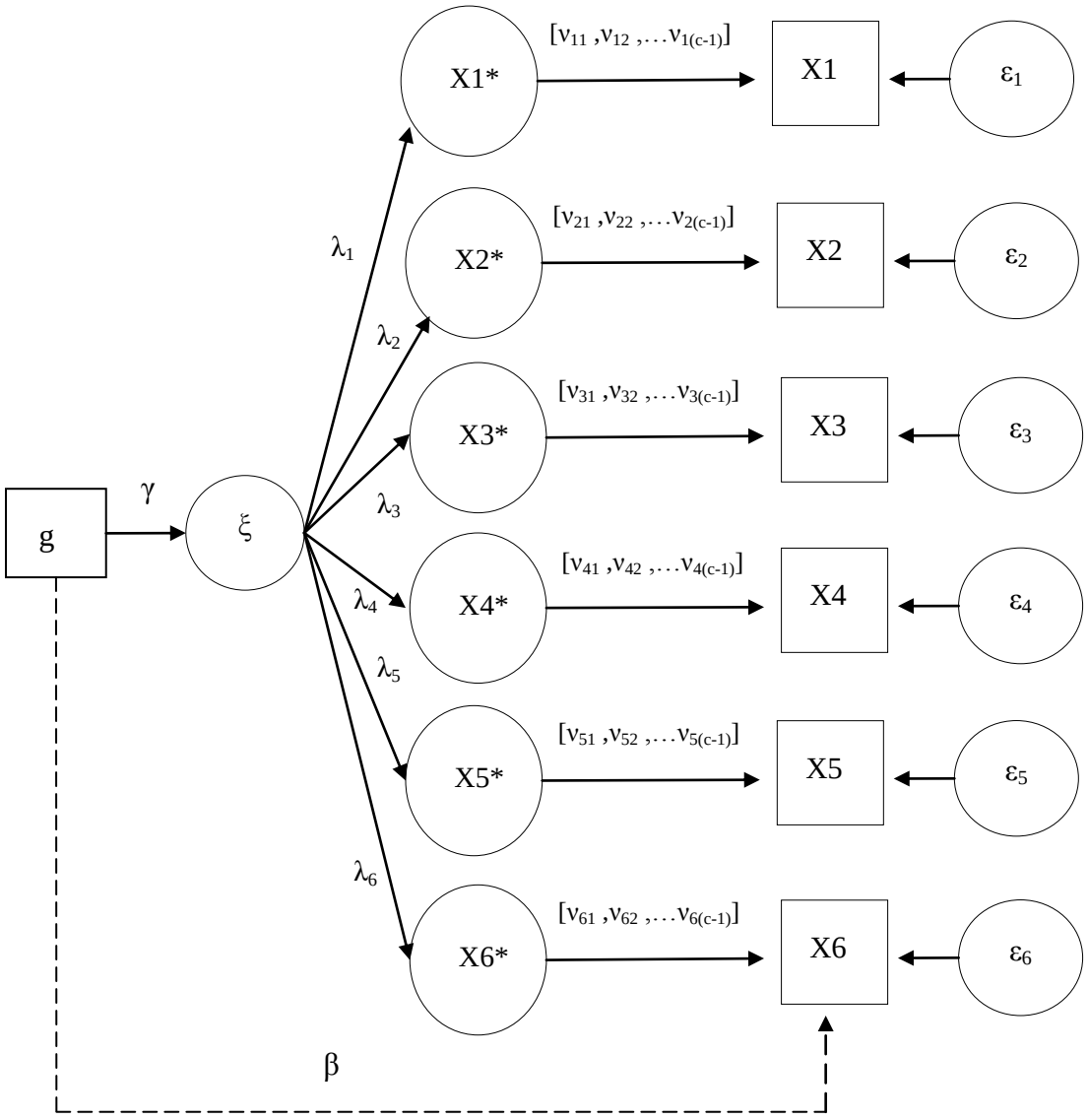

Figure 4. MIMIC modeling with a grouping variable as a covariate for categorical data. *g* denotes a grouping variable; *c* is the number of response categories of an item.

$$Y_{ij} = \lambda_j \eta_i + \beta_j X_i + \varepsilon_{ij},$$
$$\eta_i = \gamma X_i + \zeta_i, \quad (2.14)$$

where β_j is a path coefficient of the grouping variable in relation to the *j*th observed variable (or the *j*th latent variate for ordinal measures) (Finch, 2005; Kaplan, 2009). The β coefficient represents the group effect on an observed variable controlling for the effect of the latent factor. Thus, this model allows the statistical significance test on the group difference of the measured variable (β_j) as well as on the group difference of the latent means (γ). Otherwise stated, the statistically significant β coefficient indicates the violation of measurement invariance across groups (*noninvariance* of the intercept) of *j*th variable. Of note is the fact that the path coefficient, β tests group invariance controlling for the latent factor effects ($\lambda_j \eta_i$) assuming the factor loading is invariant across groups (i.e., estimating only one set of factor loadings across groups, not for each group) (Woods, 2009). Considering that this assumption of factor loading invariance is often violated in practice, the equivalence of factor loadings should be tested and not simply be assumed. Hence, when there is lack of invariance in factor loadings, the behaviors of MIMIC modeling was inspected in this study. Furthermore, whether β can reflect the lack of invariance in factor loadings was examined.

Type I Error Inflation in MIMIC Modeling

MIMIC modeling which incorporates grouping variables as covariates of latent factors assumes strict invariance (i.e., the equivalence of factor loadings, intercepts, and unique variances across groups) because a single model is constructed regardless of the

number of groups in the model. The construction of a single model over groups reduces model complexity, but at the same time decreases model flexibility by imposing a single model to all groups in comparison. Therefore, when the groups of interest do not share an identical model, the over-restriction of MIMIC modeling possibly leads to statistically invalid results. For example, when the assumption of strict invariance is violated, Type I error inflation is likely to occur in measurement invariance testing. As explained earlier, measurement invariance is tested in MIMIC modeling through the LR test between a baseline model (a MIMIC model without a direct path from a grouping variable to an observed variable, a MIMIC model without the dotted line in Figures 3 and 4) and a model with the group effect on an observed variable freely estimated (a MIMIC model with a direct path from a grouping variable to an observed variable, a MIMIC model with the dotted line in Figures 3 and 4). When the baseline model is contaminated with noninvariant variables, the chi-square statistic of the baseline model is expected to inflate, showing a poor fit due to model misspecification. Subsequently, the chi-square difference between the misspecified baseline model and the tested model is likely to elevate, which may lead to Type I error inflation. Thus, Type I error inflation has been an issue in the research on the measurement invariance testing with MIMIC modeling.

Yuan and Bentler (2004) conducted an extensive study on the chi-square difference test when the baseline model was misspecified. Suggesting that the misspecified baseline model likely misleads the conclusions of chi-square difference

tests, they illustrated the Type I error inflation in the factor loading invariance testing with a heuristic example.

To control Type I error inflation in measurement invariance testing, researchers introduced several different statistical methods which will be explained one by one in the following.

Scale purification. Navas-Ara and Gomez-Benito (2002) utilized scale purification with categorical items to compare six different DIF detection techniques (e.g., Mantel-Haenszel, logistic regression, restricted factor analysis, IRT-based indices). In the scale purification procedure, noninvariant items detected in the initial DIF analysis are excluded in a new DIF analysis. That is to say, a new DIF analysis is sequentially conducted with only invariant variables of the previous analysis to identify any remaining noninvariant variables. In Navas-Ara and Gomez-Benito's study, scale purification improved false positive rates (or Type I error rates) as well as false negative rates (failing to detect noninvariant items) and power (detecting biased items) for all six DIF detection techniques. Applied to IRT-based indices, scale purification made the greatest improvement in terms of power, false positive and false negative rates. Wang, Shih, and Yang (2009) employed the scale purification to MIMIC modeling of measurement invariance testing and reported improved Type I error rates.

Forward procedure in the LR test. In the LR test, Stark et al. (2006) recommended researcher use the forward testing procedure rather than the backward procedure. In the forward procedure, the least constrained model in which all parameters except minimal constraints for identification are allowed to vary across groups serves as

a baseline model. Then, a more restricted model with the equality constraints of one or more parameters across groups is compared with the freely estimated baseline model. On the contrary, the backward procedure takes the fully-constrained model as a baseline model, and conducts the LR test with a less constrained (i.e., one or more parameter relaxed) model. When the model includes one or more noninvariant variables, the fully-constrained baseline model in the backward procedure violates the assumption of invariance across groups, which leads to the misfit of chi-square difference to chi-square distribution given degrees of freedom and possibly increases Type I error. Thus, Stark et al. recommended the forward procedure which does not impose any equality constraints in the baseline model.

Bonferroni correction. Stark et al. (2006) also suggested the use of the Bonferroni correction of critical values as a solution of Type I error inflation. Bonferroni correction is commonly used in consideration of the experimentwise (or familywise) Type I error (Thompson, 2006). Experimentwise Type I error is the probability to reject one of the true null hypotheses when two or more null hypotheses are tested in a study. Bonferroni correction divides the significance level (i.e., α) by the number of null hypotheses analyzed in a study. Stark and colleagues adjusted the conventional critical *p* value divided by the number of items in the model. The Type I error rates were still high when the Bonferroni correction was applied in the backward LR tests. However, in the free-baseline forward procedure, the Type I error rates were near zero throughout their simulation conditions. French & Finch (2008) also applied the Bonferroni correction in

their simulation study on measurement invariance testing, specifically, locating invariant reference variables in multiple group CFA.

Oort adjustment. On the other hand, Oort (1992, 1998) suggested a formula to adjust the critical value to control the chi-square inflation of a misspecified baseline model. The Oort adjustment takes into consideration the chi-square value and degrees of freedom of the baseline model in the likelihood ratio test:

$$K' = \left(\frac{\chi^2_0}{K + df_0 - 1}\right) * K \qquad (2.15)$$

where K' is the adjusted critical value, K is the original critical value chosen from the chi-square distribution given the degree-of-freedom difference between models, χ^2_0 is the chi-square value of a baseline model, and df_0 is the corresponding degree of freedom. Given degree of freedom (df_0), the critical value (K') becomes larger as the degree to which the baseline chi-square (χ^2_0) is inflated. When the degree of freedom is one, the adjusted critical value equals the baseline chi-square. This equation can be utilized when the tested chi-square difference in the LR test is *not* assumed to follow the chi-square distribution. Oort (1998) applied this correction formula in the measurement invariance testing with restricted factor analysis. When the adjustment was applied to the iterative procedures using modification indices, the Type I error rates were reported under the nominal level.

Benjamini-Hochberg procedure. Woods (2009) utilized the Benjamini-Hochberg (1995) procedure to control Type I error and compared the MIMIC model to two-group IRT in the detection of noninvariance. Benjamini and Hochberg developed a Type I error control procedure to prevent over rejection of true null hypothesis in multiple tests and to maintain reasonable power because Bonferroni correction is often adverse to power and criticized to be too conservative. In the Benjamin-Hochberg procedure, the *p* values of all hypotheses are sorted in an ascending order and the hypotheses meeting the following requirement are all rejected.

$$p_{(i)} \leq \frac{\alpha}{m} i \qquad (2.16)$$

where $p_{(i)}$ is the *p* value of the *i*th hypothesis, α is the unadjusted critical *p* value, *m* is the number of hypotheses tested, and *i* is the largest index which satisfies the above requirement ($i = 1, 2, \ldots, m$). For example, suppose the observed *p* values of five null hypotheses are .0010, .0131, .0200, .0420, and .0650. Because the third *p* value is the largest *p* value which meets the aforementioned requirement when α is set at .05 and *m* equals 5 (that is, .0200 is smaller than .05 × 3 ÷ 5 = .03 whereas .0420 is larger than .05 × 4 ÷ 5 = .04), the first three null hypotheses will be rejected for statistical significance. On the other hand, with the Bonferroni-corrected critical *p* value (i.e., .05/5 = .01), only the first hypothesis achieves statistical significance. In measurement invariance testing, Woods showed that the Benjamini-Hochberg procedure adequately controlled the Type I error rates.

Modification Indices

The modification index (Sörbom, 1989) is also referred to the Lagrange multiplier (LM) test (Brown, 2006). The LM test is utilized to test the validity of the constraints placed on the model (Kaplan, 2009). The LM test follows the chi-square distribution with the degrees of freedom of difference between more constrained and less constrained models. Researchers frequently resort to the modification indices to improve the fit of a given model by removing a restriction placed in the model.

Because a misspecified equality constraint of a parameter over groups plausibly causes model misfit, modification indices are expected to provide information about the measurement noninvariance across groups. In other words, the parameter of a variable with the modification index over a certain cutoff can be considered noninvariant across groups, and the variable can be identified as a noninvariant variable. Moreover, modification indices are given by request in many SEM software programs. Researchers can painlessly obtain modification indices with a simple command or by clicking a checkbox of modification indices. For these reasons, modification indices are employed to locate the sources of noninvariance (specifically, which parameter of which variable).

However, a couple of questions arise in the utilization of the modification indices. Woods (2009) articulated two problems of modification indices in measurement invariance testing. The modification indices could lead to erroneous conclusions because (a) the cutoff of the modification index is not established (e.g., how large is large; Muthén, 1988) and (b) the information of modification index is pertinent only when a parameter is relaxed one at a time. In addition, because each variable is tested for

invariance assuming the invariance of all other variables, the Type I error inflation can be of issue when noninvariant variables are in the model. MacCallum (1986) demonstrated that model revisions through specification searches with modification indices did not find the true model in a simulation study. On the other hand, a reasonable performance of modification indices was reported, especially when the iterative model search strategy was applied in model modification (Oort, 1998; Yoon & Millsap, 2007).

Atheoretical decisions implied by modification indices to improve a model fit requires a caution (Brown, 2006; Kaplan, 2009; Kline, 2005). The problems of atheoretical model modification include overfitting and capitalization of chance in a sample (Brown, 2006; Kline, 2005). The respecified model may include unnecessary parameters due to sampling error. Data-driven model specification may hinder the generalization of the model to the population of the study. Therefore, it is recommended that the decision on model respecification be based on prior research or theory.

CHAPTER III

METHODOLOGY

This study is intended to explore the behaviors of MIMIC modeling in detecting noninvariant variables with continuous and categorical measures. First, MIMIC modeling allows researchers to test the equivalence of intercepts over groups assuming the invariance of factor loadings over groups. Given that the weak invariance or the invariance of factor loadings is not always met, the sensitivity of MIMIC modeling to the violation of the weak invariance is worthy of investigation. Second, the overall performance of MIMIC modeling as a measurement invariance testing technique (testing the equivalence of intercepts over groups) is the primary interest of this study. Considering that the previous studies on MIMIC modeling for measurement invariance testing reported high Type I error rates, this study employed and compared two critical value adjustment strategies to control Type I error inflation. Finally, modification indices were considered as an indication of measurement noninvariance in measurement invariance testing. Instead of likelihood ratio tests, modification indices were employed to detect the noninvariant variables and the subsequent performance was investigated. For these research questions, Monte Carlo study was conducted under various simulation conditions.

Simulation Conditions and Data Generation

Simulation conditions includes data type (continuous or categorical), for categorical data, number of categories (dichotomous or polytomous), source of

noninvariance (factor loading or intercept for continuous data and threshold for categorical data), degree of noninvariance or effect size (small or large), number of noninvariant variables (zero, one or two of six), and sample size (200, 400, 1000, or 2000).

Type of data. Three types of data were generated through Mplus 5.2 (Muthén & Muthén, 2008a): continuous, dichotomous and polytomous variables. Six variables (*X*1 – *X*6) were created under a single factor following unidimensionality. The polytomous variables have five responses which are assumed to take ordered-categorical values.

Sample size. Two balanced groups with sample size 100, 200, 500, and 1000 each were examined in this study. Although in many research settings two groups may be disproportionate (e.g., 90% Caucasian and 10% African American), studies in which two group sizes are roughly equal are not uncommon (e.g., boys and girls, primary school students and secondary school students, etc.). Woods (2009) studied an optimal sample size for MIMIC modeling in the detection of DIF and found that the focal group sample size smaller than 100 (e.g., 25, 50, or 100) yielded very low power to detect the DIF items. Thus, this study included the minimum sample size as low as 100 per group.

Number of noninvariant variables. Concerning the number of noninvariant items, two conditions of noninvariance contamination were simulated: only one noninvariant variable (about 17% contamination) and two noninvariant variables (about 33% contamination). The noninvariance contamination is less than 50% because it is more likely that the majority of variables are invariant across groups, and a small portion of variables may exhibit noninvariance. *X*5 was simulated as a noninvariant variable, and

*X*2 was added as a noninvariant item for the two noninvariant-variable conditions. This study also included the condition in which all six variables were invariant across groups to establish basal Type I error rates.

Source of noninvariance. The location of noninvariance varied at either factor loadings or intercepts/thresholds. In the previous studies on MIMIC modeling in measurement invariance testing, the source of noninvariance was not considered as a simulation condition. This study focused on examining the behaviors of MIMIC with different sources of noninvariance in the model. Different research questions were addressed depending on the source of noninvariance. For the noninvariance in the factor loadings over groups, the sensitivity of MIMIC modeling to the factor loading noninvariance was questioned because MIMIC modeling does not test the factor loading equivalence over groups explicitly but assumes the invariance of factor loadings. For the noninvariance in the intercepts over groups, the performance of MIMIC modeling to detect the noninvariant variables was investigated with different critical value adjustments in the likelihood ratio tests.

Magnitude of noninvariance. The magnitude of noninvariance was manipulated with small and large difference. For the factor loadings of the focal group, .2 and .4 were subtracted from the factor loadings of the reference group for small and large effect size, respectively. In terms of intercept or threshold noninvariance, approximately .3 for small difference and .6 for large difference were added to the intercepts or thresholds of the reference group. Accordingly, in case of the two noninvariant-variable conditions, the noninvariance was uniform in favor of the reference group.

The factor mean and variance of the reference group were 0 and 1, respectively. The corresponding parameters of the focal group were assigned as 0.5 and 1.3, respectively. The simulated factor mean difference between groups is not presumably related to the behaviors of MIMIC in detecting noninvariant variables (Stark et al., 2006). The residual variances were homogeneous across groups as 0.3. The parameters of intercepts were specified for continuous variables whereas a set of thresholds were specified for dichotomous and polytomous variables. Dichotomous variables have a single threshold with two response categories whereas polytomous variables in this study take five response categories yielding four thresholds. The parameter values used for the generation of both continuous and categorical data are presented in Table 1. Under each condition 500 replications were generated.

The parameter values, the magnitude of DIF, sample size, and number of replications were selected with the reference to the previous simulation studies on the similar research conditions (e.g., Meade & Lautenschlager, 2004; Muthén & Asparouhov, 2002; Stark et al., 2006; Yoon, 2008). The simulation conditions of prior studies on measurement invariance and MIMIC modeling were reviewed and summarized in Appendices A-D.

Data Analysis

Model identification and estimation. For the identification of the MIMIC model with a grouping variable as a covariate, factor variance was fixed at 1 instead of constraining one of the factor loadings at 1. This identification strategy allows freely estimating the factor loadings of all observed variables and testing all variables for

Table 1

Design of Monte Carlo Study

Data type	Group	Item	Small DIF					Large DIF				
			λ	ν_1	ν_2	ν_3	ν_4	λ	ν_1	ν_2	ν_3	ν_4
Dichotomous/ Continuous	Reference group	*X1*	.9	-0.15				.9	-0.15			
		X2	**.7**	**0.25**				**.7**	**0.25**			
		X3	.6	0.15				.6	0.15			
		X4	.8	-0.25				.8	-0.25			
		X5	**.7**	**-0.10**				**.7**	**-0.10**			
		X6	.6	0.10				.6	0.10			
	Focal group	*X1*	.9	-0.15				.9	-0.15			
		X2	**.5**	**0.58**				**.3**	**0.82**			
		X3	.6	0.15				.6	0.15			
		X4	.8	-0.25				.8	-0.25			
		X5	**.5**	**0.20**				**.3**	**0.50**			
		X6	.6	0.10				.6	0.10			
Polytomous	Reference group	*X1*	.9	-0.05	0.35	0.75	1.05	.9	-0.05	0.35	0.75	1.05
		X2	**.7**	**-0.80**	**-0.40**	**0.00**	**0.40**	**.7**	**-0.80**	**-0.40**	**0.00**	**0.40**
		X3	.6	-0.55	-0.05	0.45	0.85	.6	-0.55	-0.05	0.45	0.85
		X4	.8	0.05	0.50	0.85	1.15	.8	0.05	0.50	0.85	1.15
		X5	**.7**	**-0.50**	**-0.10**	**0.25**	**0.65**	**.7**	**-0.50**	**-0.10**	**0.25**	**0.65**
		X6	.6	0.15	0.40	0.70	1.25	.6	0.15	0.40	0.70	1.25
	Focal group	*X1*	.9	-0.05	0.35	0.75	1.05	.9	-0.05	0.35	0.75	1.05
		X2	**.5**	**-0.55**	**-0.10**	**0.30**	**0.75**	**.3**	**-0.24**	**0.25**	**0.60**	**0.98**
		X3	.6	-0.55	-0.05	0.45	0.85	.6	-0.55	-0.05	0.45	0.85
		X4	.8	0.05	0.50	0.85	1.15	.8	0.05	0.50	0.85	1.15
		X5	**.5**	**-0.20**	**0.20**	**0.55**	**0.95**	**.3**	**0.10**	**0.50**	**0.85**	**1.25**
		X6	.6	0.15	0.40	0.70	1.25	.6	0.15	0.40	0.70	1.25

Note. The parameters of DIF items are written in bold.

invariance. For estimation, ML (maximum likelihood) for continuous data and WLSMV (weighted least square with robust mean and variance) with theta parameterization for categorical data were utilized. Both estimations are the defaults of the M*plus* program for continuous and categorical data, respectively.

Model evaluation. When the equivalence of factor loadings over groups is violated, the MIMIC model is expected to exhibit a poor fit due to model misspecification. To evaluate a model fit under the violation of the factor loading invariance assumption, a set of model fit statistics were examined. In the current study, model fit indices including a chi-square fit statistic and alternative fit indices (AFI) were inspected when a proportion of factor loadings were simulated to be noninvariant. The following alternative fit indices were analyzed: (a) the weighted root mean square residual (WRMR) for categorical items or the standardized root mean square residual (SRMR) for continuous items; (b) comparative fit index (CFI); and (c) the root mean square error of approximation (RMSEA). Recommended cutoff values of these AFI measures for a good model fit are CFI $\geq$ .95, RMSEA $\leq$ .05, and SRMR $\leq$ .08, and WRMR $\leq$ 1.0 (Hu & Bentler, 1999; Yu, 2002) in addition to statistically non-significant chi-square ($p \geq .05$). For each fit statistic, the value out of the given range can be considered as a flag of model misspecification due to noninvariant factor loadings. The proportion of cases in which the model was correctly flagged as a poor fit was computed in addition to the mean of each fit index across 500 replications.

Likelihood ratio test in MIMIC modeling. Instead of statistical significance testing on β coefficients in Equation 2.14, this study employed the likelihood ratio test to

detect the noninvariant variables. A likelihood ratio test is conducted with two nested models in the attempt to obtain a better fit model with more parsimony. In this study, the MIMIC model with a direct path from the grouping indicator to each variable (augmented model with β, that is, with the dotted path in Figures 3 and 4) is compared to the model constraining the corresponding path parameter (β) at zero assuming invariance (baseline model, a model without the dotted path in Figures 3 and 4). The statistical significance of the chi-square difference given degree of freedom between two models (in this study, $df = 1$) indicates the direct effect of group membership on the tested variable in favor of the augmented model. In other words, the tested variable is considered *non*invariant over groups. The statistical significance testing on the β coefficient yields identical results as the likelihood ratio test. However, in this study, the LR test was intentionally selected for two reasons: (a) to apply the Oort adjustment to the critical values which requires the baseline model chi-square and degrees of freedom, and (b) to make a connection to modification indices which is tantamount to the LR test with one degree of freedom.

As speculated earlier, when a baseline model is misspecified, the chi-square fit statistic of the baseline model is expected to inflate reflecting the misspecification (Oort, 1998; Stark et al, 2006; Yuan & Benter, 2004). Subsequently, the chi-square difference between baseline and augmented models is likely to increase, which may lead to the rejection of the null hypothesis when there is no difference between competing models. The MIMIC model is inherently a full invariance model assuming measurement invariance of all parameters of all observed variables across groups because the groups

share a single model. This full-invariance-assumed MIMIC model serves as a baseline model in the likelihood ratio test. However, when the model contains noninvariant variables (i.e., the model is misspecified) as simulated in this study, the chi-square inflation and consecutive Type I error increase possibly occur. To correct the inflated Type I error rates, two critical value adjustment methods, Oort adjustment and Bonferroni correction, were employed in this simulation study.

Oort adjustment to control Type I error. In this study, the chi-square difference between baseline and augmented models is likely not to conform to the chi-square distribution because of the baseline model misspecification. Thus, the likelihood ratio tests were conducted with the Oort adjusted chi-square critical values. The Oort adjustment was compared to the Bonferroni correction. Stark et al. (2006) suggested Bonferroni correction to lower the inflated Type I error rates in the LR tests. For the Bonferroni correction, critical *p* value .008 (= .05 / 6) was adopted because six likelihood ratio tests were performed for each replication.

Modification indices. The final concern of this study is modification indices which are similar to the chi-square difference values with one degree of freedom in the likelihood ratio test. In the utilization of modification indices, one critical decision researchers should make is to set the cutoff values of modification indices. For example, in M*plus* the default cutoff of modification index is 10. Instead, the conventional chi-square value at the .05 level for one degree of freedom ($\chi^2[1] = 3.84$) can be utilized as a cutoff. In this study, taking into account the Type I error rate inflation, Oort-adjusted chi-square critical values were selected and compared to the conventional cutoff of

modification index (3.84). In case of Oort correction each replication adopts a different cutoff value depending on the degree of adjustment. The variable with a modification index greater than the Oort-adjusted critical value was considered as noninvariance.

This study examined two different strategies of modification indices: (a) noniterative method in which modification indices in the baseline model are examined for noninvariance without subsequent model modification, and (b) iterative method in which models are sequentially modified according to the modification indices. The first method corresponds to the initial stage of iterative modification procedures. The baseline model assumes full invariance of all variables. Because one or two variables were simulated noninvariant, it is expected that the noninvariant variables will have modification indices over a given cutoff indicating lack of invariance. In the iterative procedure, model modification was sequentially conducted by relaxing one noninvariant variable with the largest modification index at a time until all modification indices were below a certain criterion or the model turns into a good fit ($p \geq .05$ for a chi-square goodness-of-fit statistic). In the first stage, all variables were constrained equal across groups and the modification indices were examined. Then, only *one* variable with the maximum modification index was selected as a noninvariant variable and relaxed for free estimation in the subsequent model. This process was repeated as long as any noninvariance was detected with the modification index over the cutoff under a poor model fit.

It is recommended to use the iterative model modification procedure because the existence of any noninvariant variable in the model is likely to distort the correct model

fit (Oort, 1998; Wang et al., 2009; Yoon & Millsap, 2007). In Yoon and Millsap's simulation study on the multiple group CFA using continuous variables, the sequential model modification procedure yielded better results than the noniterative method. Oort (1998) showed the outperformance of the iterative procedure in the ordinal measures using ordinary linear MIMIC. This study embraced data types (continuous or categorical) as a design variable, and examined the modification indices of the MIMIC model with either continuous or ordered-categorical variables. To sum up, this study included four different strategies of modification indices in combination of the cutoffs of modification indices and the procedures of model modification: (a) the noniterative method (i.e., modification indices at the initial stage of model modification) using the conventional critical value, (b) the noniterative method using the Oort adjusted critical value, (c) the iterative procedure of model modification using the conventional critical value, and (d) the iterative procedure using the Oort adjustment.

CHAPTER IV

RESULTS

Simulation Baseline Check

The basal Type I error rate was established and the adequacy of simulation was checked under the conditions of measurement invariance over groups. In this study, Type I error refers to the false detection of an invariant variable as noninvariance. When all six variables were invariant across groups, the Type I error rate was measured at the critical *p* value of .05 (Table 2). Because the model was specified correctly without noninvariance, Type I error rates should not considerably exceed the sampling error rate, which was set at .05. Therefore, Type I error rates were expected around .05 regardless of simulation conditions.

In most study conditions, Type I error rate was about .05. According to Bradley (1978), the acceptable range of Type I error rates is computed with a formula, $\alpha \pm 1/2\alpha$. When α is .05, the Type I error rates between .025 and .075 are considered reasonable.

Table 2

Basal Type I Error Rates

Sample size	Continuous	Categorical	
		Dichotomous	Polytomous
200	.050	.058	.048
400	.044	.057	.059
1000	.049	.057	.061
2000	.053	.081	.081

Note. $\alpha = .05$

The Type I error rates of most study conditions fell within the Bradley's range. In general, the Type I error rates of categorical variables were slightly higher than those of continuous variables. In the condition of large sample size (2000) of categorical variables the Type I error rate was over the predetermined critical *p* value ($\alpha = .05$). However, previous simulation studies on measurement invariance showed similar basal Type I error ranges. In Finch's (2005) simulation with MIMIC modeling, Type I error rates were .050 and .064 for sample size 100 and 500 per group, respectively. Stark and colleagues reported basal Type I error rates of .05 and .09 for sample size 500 and 1000, respectively in case of dichotomous data and .04 and .08 in case of polytomous data with multiple group CFA. The Type I error rates in Wang and colleagues' (2009) study on MIMIC modeling ranged between .01 and .14. Overall, the baseline check in the present study showed that the false detection occurred merely due to sampling error or by chance and data were adequately simulated.

Factor Loading Noninvariance

M**odel fit evaluation.** When MIMIC model is used to test factorial invariance, strict invariance is assumed as a baseline. A violation of this assumption, specifically, the noninvariance in factor loadings across groups, should lead to misfit of the MIMIC model. With the expectation of a poor fit of MIMIC models due to the violation of invariant factor loadings across groups, the model fit indices were examined.

Overall, the simulation study showed good model fits across all simulation conditions (namely, data type, sample size, and degree of noninvariance). With the chi-square fit statistic, statistical significance tests were conducted at the significance level

.05 ($\alpha = .05$) to evaluate model fit. That is, if $p \geq .05$, the model fit was considered adequate. On average, chi-square *p* values were considerably higher than the significance level failing to reject the null hypothesis of a good model fit. The mean of chi-square *p* values of 500 replications ranged from .11 to .50 across all simulation conditions. Only 3 of 24 simulation conditions (12%) yielded the average chi-square *p* value less than .30.

When one or more factor loadings were noninvariant in MIMIC, the proportions of cases correctly flagged as a poor fit were near .05 with the range of .04 and .09 regardless of sample size and magnitude of DIF for dichotomous variables (see Table 3). The proportions of correct detections of measurement noninvariance in factor loadings through chi-square did not diverge from the significance level ($\alpha = .05$), which indicates that MIMIC modeling did not detect the violation of factor loading invariance over the chance rate for dichotomous variables.

For both polytomous and continuous variables, the proportions of the cases showing a poor model fit increased as sample size and degree of noninvariance increased (Table 3). For example, with a large DIF of polytomous cases, the proportion to detect the model misspecification correctly was .14, .25, and .59 for sample size 400, 1000, and 2000, respectively. However, for all other scenarios of polytomous variables, the misspecified MIMIC model assuming factor loading invariance were more likely to exhibit good fits: the proportions of poor fit cases ranged .04 through .19 depending on the simulation conditions. The continuous variable conditions exhibited similar results to the polytomous variables as presented in Table 3.

Table 3

The Power of Model Fit Indices: Factor Loading Noninvariance

DIF	Sample size	χ^2 (*p*)	CFI	RMSEA	SRMR
			Dichotomous variables		
Small	200	0.06	0.00	0.11	0.00
	400	0.05	0.00	0.01	0.00
	1000	0.05	0.00	0.00	0.00
	2000	0.08	0.00	0.00	0.00
Large	200	0.05	0.00	0.11	0.00
	400	0.04	0.00	0.01	0.00
	1000	0.05	0.00	0.00	0.00
	2000	0.09	0.00	0.00	0.00
			Polytomous variables		
Small	200	0.04	0.00	0.10	0.00
	400	0.08	0.00	0.03	0.00
	1000	0.08	0.00	0.00	0.00
	2000	0.19	0.00	0.00	0.00
Large	200	0.06	0.00	0.14	0.00
	400	0.14	0.00	0.07	0.00
	1000	0.25	0.00	0.00	0.00
	2000	0.59	0.00	0.00	0.00
			Continuous variables		
Small	100	0.07	0.00	0.12	0.00
	200	0.06	0.00	0.01	0.00
	500	0.07	0.00	0.00	0.00
	1000	0.11	0.00	0.00	0.00
Large	100	0.08	0.00	0.15	0.00
	200	0.11	0.00	0.03	0.00
	500	0.20	0.00	0.00	0.00
	1000	0.47	0.00	0.00	0.00

Note. Power is defined as the proportion of the cases in which the model with noninvariant variables showed a poor fit. CFI = comparative fit index, RMSEA = root mean squared error of approximation, SRMR = standardized root mean squared residual. The cutoff values of a poor model fit are chi-square *p*-values < .05, CFI < .95, RMSEA < .05, and SRMR < .08.

Table 4

The Mean of Model Fit Indices: Factor Loading Noninvariance

DIF	Sample size	χ^2 (*p*)	CFI	RMSEA	WRMR
			Dichotomous variables		
Small	200	0.49	1.00	0.02	0.54
	400	0.48	1.00	0.01	0.53
	1000	0.48	1.00	0.01	0.54
	2000	0.44	1.00	0.01	0.55
Large	200	0.49	1.00	0.02	0.54
	400	0.48	1.00	0.01	0.54
	1000	0.47	1.00	0.01	0.54
	2000	0.42	1.00	0.01	0.56
			Polytomous variables		
Small	200	0.50	1.00	0.02	0.37
	400	0.46	1.00	0.01	0.38
	1000	0.41	1.00	0.01	0.39
	2000	0.31	1.00	0.01	0.42
Large	200	0.46	1.00	0.02	0.38
	400	0.39	1.00	0.02	0.40
	1000	0.26	1.00	0.02	0.44
	2000	0.11	1.00	0.02	0.51
			Continuous variables		
Small	100	0.47	1.00	0.02	0.02
	200	0.46	1.00	0.01	0.01
	500	0.46	1.00	0.01	0.01
	1000	0.37	1.00	0.01	0.01
Large	100	0.44	1.00	0.02	0.02
	200	0.39	1.00	0.02	0.02
	500	0.31	1.00	0.01	0.01
	1000	0.16	1.00	0.02	0.01

In addition to the statistical significance of chi-square fit statistic, alternative fit indices (i.e., CFI, RMSEA, and SRMR for continuous data and WRMR for categorical data) were evaluated. The mean of the alternative fit indices are presented in Table 4 (see the last three columns). Consistent to chi-square fit statistics, all AFIs, in general, supported good model fits of MIMIC with noninvariant factor loadings. For continuous variables, (a) the mean of CFI was 1.00 irrespective of sample size and DIF magnitude; (b) the mean of RMSEA was below .02; (c) the mean of SRMR was below .02 across simulation conditions. The average WRMR of dichotomous variables was about .55 for all simulation conditions. To sum up, MIMIC modeling, in general, failed to detect the violations of factor loading invariance through model evaluations.

Measurement invariance testing. Concerning factor loading noninvariance, this study questioned whether the factorial invariance test using MIMIC (i.e., estimating the direct effect of group membership on each observed variable) can detect the violation of equivalent factor loading assumption. To address this question, likelihood ratio tests were conducted as explained in the method section, and power and Type I error were examined. Two methods of critical value adjustment (Bonferroni and Oort) were applied in the LR tests.

For dichotomous data, power was below .10 regardless of sample size, magnitude of noninvariance, and critical value adjustment strategies (see Table 5). That is, power was not different from the basal Type I error rates. When Bonferroni correction was utilized, the power rates were almost zero (.01 ~ .02). For polytomous and continuous data, sample size and degree of noninvariance made a positive impact on

Table 5

The Power and Type I Error Rates of the LR Tests: Factor Loading Noninvariance

DIF	Sample size	No adjustment		Bonferroni		Oort	
		Power	*Type I error*	Power	*Type I error*	Power	*Type I error*
				Dichotomous variables			
Small	200	.06	*.06*	.02	*.01*	.07	*.08*
	400	.03	*.06*	.01	*.01*	.05	*.07*
	1000	.05	*.06*	.01	*.01*	.07	*.08*
	2000	.08	*.08*	.01	*.01*	.08	*.09*
Large	200	.05	*.05*	.01	*.01*	.07	*.08*
	400	.06	*.06*	.01	*.01*	.07	*.08*
	1000	.06	*.06*	.01	*.00*	.08	*.08*
	2000	.10	*.08*	.02	*.01*	.09	*.09*
				Polytomous variables			
Small	200	.07	.05	.02	*.01*	.09	*.08*
	400	.11	*.06*	.03	*.01*	.12	*.07*
	1000	.17	*.07*	.05	*.01*	.18	*.08*
	2000	.37	*.09*	.15	*.02*	.33	*.08*
Large	200	.15	.05	.04	*.01*	.15	*.07*
	400	.26	*.07*	.09	*.01*	.25	*.07*
	1000	.52	*.09*	.26	*.01*	.48	*.06*
	2000	.82	*.12*	.61	*.03*	.78	*.05*
				Continuous variables			
Small	200	.08	*.05*	.02	*.01*	.10	*.06*
	400	.11	*.05*	.03	*.01*	.14	*.07*
	1000	.20	*.05*	.05	*.01*	.23	*.06*
	2000	.36	*.06*	.15	.01	.35	*.06*
Large	200	.16	*.05*	.05	*.01*	.19	*.06*
	400	.27	*.05*	.10	*.01*	.30	*.06*
	1000	.58	*.06*	.33	*.01*	.60	*.04*
	2000	.87	*.07*	.67	*.02*	.88	*.02*

Note. Bonferroni means Bonferroni correction on the critical values; Oort means Oort correction on the critical values. The Type I error rates were italicized.

power. Thus, power reached over .80 in the conditions of large DIF and large sample size (n = 2000 in Table 5). However, in most conditions, power was considerably lower although Type I error was reasonably controlled below .10 across conditions (Table 5). Compared with the conventional critical p value (α = .05), neither Bonferroni nor Oort correction improved the performance of MIMIC in detecting the factor loading noninvariance. Bonferroni correction slightly degraded Type I error rates while dropping power considerably. In general, Oort correction worsened Type I error rates in most conditions of dichotomous variables. Interestingly, when Oort adjustment was applied for continuous variables, the power rates exceeded the counterparts of no adjustment conditions (e.g., for no adjustment, the power rate in the condition of large DIF with sample size 1000 was .58; for Oort adjustment, .60). Because the critical value is tailored according to the baseline chi-square in Oort adjustment, the power rate could improve over the no adjustment conditions whereas Bonferroni correction uniformly lessens the power, taking more conservative critical p values. Although Oort adjustment improved the power rates under certain simulation conditions, it should be noted that in factor loading noninvariance scenarios critical value adjustment appeared not to be required because Type I error was reasonable under no adjustment across all conditions.

Intercept/Threshold Noninvariance

No critical value adjustment. The statistical behaviors of the MIMIC model in detecting intercept (or threshold) noninvariance were assessed with two types of summary statistics: power and Type I error rates as defined earlier. Before we applied certain types of critical value adjustment to control Type I error, no adjustment

conditions were investigated first. Table 6 presented power and Type I error of continuous and categorical data, respectively when the conventional alpha (.05) was employed without correction. When a single noninvariant variable existed for continuous variables, the power rate was simply 1.00 regardless of sample size and degree of noninvariance. Even a small size difference in intercept between groups was detected almost all the time. In case of categorical variables except small DIF and small sample size ($n \leq 200$ for polytomous and $n \leq 400$ for dichotomous), the noninvariant variable was detected with nearly 100% power.

When two of six variables were noninvariant over groups, two types of proportions under power were reported. When each noninvariant variable was tested for invariance and detected as DIF, the proportion of detected cases over 500 replications was defined as the power rate. The average power rate of two noninvariant variables is presented at the top of each cell (Tables 7 and 8). The value at the bottom of each cell represents the proportion of replications in which both variables were detected as noninvariance. Similar to one-DIF conditions presented above, power was substantial (near 1.00 or above) in almost all simulation scenarios. Higher power was observed in continuous data than polytomous data; polytomous data displayed higher power than dichotomous data.

As reported in previous studies, substantial Type I error elevations were observed throughout conditions. The Type I error rate inflated as sample size and degree of noninvariance increased. In combination of large sample size and large degree of noninvariance, the inflation of Type I error rate was the most serious (see Tables 6, 7,

Table 6

The Power and Type I Error Rates of the LR Tests: Intercept/Threshold Noninvariance in One Noninvariant Variable

		No Adjustment		Bonferroni		Oort	
DIF	Sample size	Power	*Type I Error*	Power	*Type I Error*	Power	*Type I Error*
				Dichotomous variables			
Small	200	.63	*.08*	.35	*.01*	.62	*.06*
	400	.87	*.10*	.65	*.02*	.85	*.04*
	1000	1.00	*.17*	.99	*.05*	.99	*.02*
	2000	1.00	*.30*	1.00	*.12*	1.00	*.00*
large	200	1.00	*.14*	.95	*.04*	.99	*.03*
	400	1.00	*.22*	1.00	*.08*	1.00	*.01*
	1000	1.00	*.48*	1.00	*.23*	1.00	*.00*
	2000	1.00	*.75*	1.00	*.52*	1.00	*.00*
				Polytomous variables			
Small	200	.82	*.09*	.63	*.02*	.84	*.05*
	400	.99	*.14*	.94	*.04*	.98	*.03*
	1000	1.00	*.29*	1.00	*.11*	1.00	*.01*
	2000	1.00	*.48*	1.00	*.26*	1.00	*.00*
large	200	1.00	*.20*	1.00	*.06*	1.00	*.02*
	400	1.00	*.37*	1.00	*.16*	1.00	*.00*
	1000	1.00	*.70*	1.00	*.47*	1.00	*.00*
	2000	1.00	*.92*	1.00	*.79*	1.00	*.00*
				Continuous variables			
Small	200	.93	*.10*	.80	*.03*	.91	*.04*
	400	1.00	*.16*	.99	*.05*	1.00	*.02*
	1000	1.00	*.33*	1.00	*.14*	1.00	*.00*
	2000	1.00	*.56*	1.00	*.32*	1.00	*.00*
large	200	1.00	*.24*	1.00	*.08*	1.00	*.01*
	400	1.00	*.42*	1.00	*.21*	1.00	*.00*
	1000	1.00	*.77*	1.00	*.56*	1.00	*.00*
	2000	1.00	*.94*	1.00	*.83*	1.00	*.00*

Note. Bonferroni means Bonferroni correction on the critical values; Oort means Oort correction on the critical values. The Type I error rates were italicized.

Table 7

The Power and Type I Error Rates of the LR Tests: Threshold Noninvariance in Two Noninvariant Variables of Categorical Data

		No Adjustment		Bonferroni		Oort	
DIF	Sample size	Power	*Type I Error*	Power	*Type I Error*	Power	*Type I Error*
				Dichotomous variables			
Small	200	.48	*.15*	.22	*.05*	.37	*.09*
		.19		.04		.05	
	400	.73	*.27*	.47	*.08*	.50	*.07*
		.52		.19		.15	
	1000	.98	*.54*	.93	*.29*	.74	*.06*
		.97		.86		.48	
	2000	1.00	*.83*	1.00	*.62*	.90	*.02*
		1.00		1.00		.81	
Large	200	.94	*.40*	.79	*.17*	.67	*.07*
		.88		.61		.35	
	400	1.00	*.68*	.98	*.41*	.81	*.03*
		.99		.97		.62	
	1000	1.00	*.97*	1.00	*.87*	.93	*.01*
		1.00		1.00		.86	
	2000	1.00	*1.00*	1.00	*1.00*	.99	*.00*
		1.00		1.00		.98	
				Polytomous variables			
Small	200	.67	*.22*	.40	*.07*	.48	*.09*
		.41		.13		.11	
	400	.92	*.39*	.73	*.18*	.65	*.07*
		.85		.50		.32	
	1000	1.00	*.74*	1.00	*.51*	.85	*.05*
		1.00		.99		.71	
	2000	1.00	*.96*	1.00	*.85*	.96	*.02*
		1.00		1.00		.93	
Large	200	1.00	*.65*	.98	*.39*	.83	*.05*
		.99		.96		.66	
	400	1.00	*.91*	1.00	*.75*	.95	*.03*
		1.00		1.00		.89	
	1000	1.00	*1.00*	1.00	*1.00*	1.00	*.01*
		1.00		1.00		1.00	
	2000	1.00	*1.00*	1.00	*1.00*	1.00	*.00*
		1.00		1.00		1.00	

Note. The Type I error rates were italicized. The second value of power denotes the proportion of the cases across 500 replications in which both noninvariant variables were detected simultaneously.

and 8). It appeared that the Type I error elevation became worse as the number of noninvariant variables increased from one to two. Although MIMIC showed reasonable power to detect noninvariance, high Type I error rates deteriorated the performance of MIMIC as a valid tool for measurement invariance testing. As articulated earlier, a statistical consideration to control for high Type I error rates appears to be necessary.

Table 8

The Power and Type I Error Rates of the LR Tests: Intercept Noninvariance in Two Noninvariant Variables of Continuous Data

		No adjustment		Bonferroni		Oort	
DIF	Sample size	Power	*Type I error*	Power	*Type I error*	Power	*Type I error*
Small	200	.81	*.30*	.58	*.13*	.58	*.08*
		.64		.29		.23	
	400	.98	*.54*	.92	*.31*	.76	*.08*
		.96		.84		.53	
	1000	1.00	*.87*	1.00	*.71*	.91	*.04*
		1.00		1.00		.82	
	2000	1.00	*.98*	1.00	*.92*	.98	*.03*
		1.00		1.00		.95	
large	200	1.00	*.74*	.99	*.52*	.86	*.08*
		1.00		.98		.72	
	400	1.00	*.93*	1.00	*.81*	.98	*.05*
		1.00		1.00		.95	
	1000	1.00	*1.00*	1.00	*1.00*	1.00	*.02*
		1.00		1.00		1.00	
	2000	1.00	*1.00*	1.00	*1.00*	1.00	*.01*
		1.00		1.00		1.00	

Note. Bonferroni means Bonferroni correction on the critical values; Oort means Oort correction on the critical values. The Type I error rates were italicized. The second value of power denotes the proportion of the cases across 500 replications in which both noninvariant variables were detected simultaneously.

Bonferroni correction. When the critical *p* value were adjusted with Bonferroni correction (that is, $\alpha = .008$), the Type I error rates showed slight improvement (e.g., see Tables 6, 7, and 8). However, the Type I error rates remained substantial across most large noninvariance, large sample, and large contamination conditions regardless of data type (e.g., in the condition of large DIF and sample size 1000 with 2 noninvariant variables, Type I error was 1.00). More seriously, Bonferroni correction diminished the power rates as it improved the Type I error rates. For example, for the small DIF, sample size 400, small contamination condition with dichotomous variables, the power rate to detect a noninvariant variable decreased to .65 from .87 while the Type I error rate became smaller than expected or desired (from .10 to .02, see Table 6). Adopting more conservative critical *p* value resulted in the loss of power to identify the noninvariance properly. Overall, the findings of Bonferroni correction indicated that simply taking more conservative critical *p* value was not an optimal solution to correct the inflated Type I error rates.

Oort adjustment. When Oort-adjusted critical chi-square values were applied, Type I error rates were controlled around the basal rates (between .00 and .09). Most conditions of small contamination (namely, a single noninvariant variable) showed near zero Type I error rates (Tables 6, 7, and 8). That is, MIMIC appeared to classify the invariant variables as invariant almost all the time when utilized with Oort adjustment. At the same time, the power rates remained almost same as before-adjustment when the model had a single noninvariant variable (Table 6). Whereas Bonferroni correction

reduced the power rates, Oort adjustment maintained the power to detect the noninvariance, simultaneously controlling for the Type I error rates.

On the other hand, the MIMIC model with more than one noninvariant variable displayed loss of power with Oort correction. Across small DIF conditions, power was considerably lower with the Oort adjustment than no adjustment (Tables 7 and 8). As observed in Bonferroni correction, the decrease of power was substantial in detecting both noninvariant variables simultaneously (the second type of the reported power rates at the bottom). When the variables were dichotomous in sample size 1000, the power to detect both small DIF variables was .97 without adjustment but .48 with Oort adjustment (Table 7). However, for polytomous and continuous variables the power loss was less obvious or did not occur with a large sample size. With a sufficient sample size such as 400 or more the large intercept/threshold difference between groups was reasonably detected. In case of continuous variables the power was over 95% with large DIF and sample size over 400. Overall, unless sample size and DIF were small, the Oort correction maintained acceptable power. Considering the high power rates across simulation conditions and the Type I error rates below the basal rates, MIMIC with the Oort adjustment can be a choice of measurement invariance testing. The power degradation of Oort correction with two noninvariant variables will be explained in the discussion section.

Modification Indices

The performance of modification indices in detecting noninvariance in intercepts was explored using the data with two-noninvariant variables. Tables 9 and 10 show

power and Type I error for four different search strategies of modification indices to identify the lack of invariance in intercepts/thresholds across group: noniterative model modification method with the unadjusted chi-square critical value, noniterative method with the Oort-adjusted critical value, iterative method with the unadjusted critical value, and iterative method with Oort adjustment. First, for the noniterative method (i.e., the modification indices of the baseline model assuming strict invariance), the overall behaviors of the modification indices were very similar to those of the LR test (Table 9). Without adjustment in the cutoff of modification indices, Type I error was substantial across simulation conditions. Overall, due to the extremely high Type I error, the modification indices over the unadjusted cutoff ($\chi^2[1] = 3.84$) did not provide any useful information on the noninvariant variables in the model although power was considerably high. However, when the Oort adjustment was applied to the cutoff of modification indices, Type I error rates were controlled below the basal rates with reasonable power if one of the conditions was met: (a) sample size over 400, (b) large size of DIF, or (c) low contamination (one DIF variable). As observed in the LR tests, the power to detect both noninvariant variables was noticeably lower when sample size was small. However, modification indices with Oort-adjusted cutoff values, in general, performed adequately to detect large DIF even at the initial stage of model modification when sample size was substantial (i.e., 1000 or more).

When the iterative procedure was implemented, the performance of modification indices improved substantially, especially with the conventional cutoff (Table 10). Type I error rates were near or below the basal error rates throughout simulation conditions.

Table 9

The Power and Type I Error Rates of Modification Indices: Noniterative Procedure

		Dichotomous				Polytomous				Continuous			
		No adjustment		Oort		No adjustment		Oort		No adjustment		Oort	
DIF	*n*	Power	*Type I error*	Power	*Type I error*	Power	*Type I error*	Power	*Type I error*	Power	*Type I error*	Power	*Type I error*
Small	200	.36	*.28*	.24	*.10*	.63	*.51*	.52	*.16*	.97	*.82*	.92	*.28*
		.11		.01		.32		.05		.64		.20	
	400	.74	*.64*	.65	*.18*	.94	*.87*	.91	*.21*	1.00	*.98*	1.00	*.24*
		.45		.12		.81		.26		.96		.52	
	1000	.99	*.98*	1.00	*.23*	1.00	*1.00*	1.00	*.13*	1.00	*1.00*	1.00	*.13*
		.96		.49		1.00		.70		1.00		.81	
	2000	1.00	*1.00*	1.00	*.11*	1.00	*1.00*	1.00	*.06*	1.00	*1.00*	1.00	*.08*
		1.00		.82		1.00		.94		1.00		.95	
Large	200	.96	*.87*	.93	*.18*	1.00	*.99*	1.00	*.10*	1.00	*1.00*	1.00	*.16*
		.85		.25		.99		.50		1.00		.66	
	400	1.00	*1.00*	1.00	*.08*	1.00	*1.00*	1.00	*.05*	1.00	*1.00*	1.00	*.09*
		1.00		.61		1.00		.85		1.00		.92	
	1000	1.00	*1.00*	1.00	*.04*	1.00	*1.00*	1.00	*.00*	1.00	*1.00*	1.00	*.02*
		1.00		.91		1.00		1.00		1.00		1.00	
	2000	1.00	*1.00*	1.00	*.00*	1.00	*1.00*	1.00	*.00*	1.00	*1.00*	1.00	*.00*
		1.00		1.00		1.00		1.00		1.00		1.00	

Table 10

The Power and Type I Error Rates of Modification Indices: Iterative Procedure

		Dichotomous				Polytomous				Continuous			
		No adjustment		Oort		No adjustment		Oort		No adjustment		Oort	
DIF	*n*	Power	*Type I error*	Power	*Type I error*	Power	*Type I error*	Power	*Type I error*	Power	*Type I error*	Power	*Type I error*
Small	200	.29	*.11*	.22	*.07*	.55	*.10*	.49	*.08*	.77	*.11*	.73	*.10*
		.06		.03		.15		.11		.36		.32	
	400	.68	*.11*	.62	*.09*	.90	*.09*	.87	*.07*	.99	*.05*	.98	*.05*
		.25		.22		.59		.56		.83		.83	
	1000	.98	*.05*	.98	*.04*	1.00	*.03*	1.00	*.02*	1.00	*.02*	1.00	*.01*
		.87		.87		1.00		1.00		1.00		1.00	
	2000	1.00	*.03*	1.00	*.02*	1.00	*.03*	1.00	*.02*	1.00	*.02*	1.00	*.01*
		1.00		1.00		1.00		1.00		1.00		1.00	
Large	200	.93	*.10*	.90	*.08*	1.00	*.03*	1.00	*.03*	.99	*.03*	.99	*.02*
		.59		.56		.96		.96		.97		.97	
	400	.99	*.04*	.99	*.04*	1.00	*.04*	1.00	*.01*	1.00	*.02*	1.00	*.01*
		.95		.95		1.00		1.00		1.00		1.00	
	1000	1.00	*.02*	1.00	*.02*	1.00	*.03*	1.00	*.02*	1.00	*.02*	1.00	*.01*
		1.00		1.00		1.00		1.00		1.00		1.00	
	2000	1.00	*.03*	1.00	*.02*	1.00	*.02*	1.00	*.01*	1.00	*.02*	1.00	*.01*
		1.00		1.00		1.00		1.00		1.00		1.00	

The choice of critical value adjustments did not make any difference in the power and Type I error of modification indices. Interestingly, the behaviors of unadjusted and adjusted modification indices were almost identical within a few hundredth differences. Hence, when researchers utilize the sequential process of model modification, Type I error appears to be no concern irrespective of whether they apply critical value adjustment to modification indices or not. However, the power to properly detect the small DIF required large sample size, especially for dichotomous variables (e.g., sample size of 1000 for small noninvariance with dichotomous data). When the magnitude of noninvariance was large, modification indices detected the noninvariant variables almost all the time even with a small sample such as 200. When sample size and DIF were small, the noninvariance in the continuous data was better detected followed by polytomous data and dichotomous data.

In comparing the Oort-adjusted likelihood ratio test and the iterative method of modification indices, the likelihood ratio tests yielded slightly lower power due to the Oort adjustment effect than the iterative modification indices whereas the Oort-adjusted likelihood ratio test outperformed the iterative procedure of modification indices in terms of Type I error although the difference was very small. The more striking difference was observed in the power to detect both noninvariant items. The iterative model search procedure using the modification indices exhibited better results in detecting noninvariance in two variables. The relatively lower power in the likelihood ratio tests with Oort adjustment in detecting multiple DIF variables will be discussed later.

Simulation Design Factors

With respect to intercept/threshold noninvariance of MIMIC modeling, the major simulation factors explaining the variation of the power rates were determined through ANOVA (Analysis of Variance). The variance of the power rates was partitioned according to the simulation design factors of the study when the likelihood ratio tests were employed for measurement invariance testing (Table 11). For this analysis only the simulation condition which yielded desirable results in terms of power and Type I error was selected: one DIF condition with Oort adjustment.

When the likelihood ratio tests were conducted to detect the intercept noninvariance in MIMIC modeling, sample size and the magnitude of DIF appeared to play a role in determining the power for noninvariance. Sample size, magnitude of noninvariance, and their interaction explained about 66% of the variance of power, which indicates (a) that the large magnitude of noninvariance with a large sample is well detected, (b) that sample size does not matter to a large extent when the DIF size is large,

Table 11

The Proportion of Variance Explained by the Simulation Design Factors: Likelihood Ratio Tests of the Intercept Noninvariance (One DIF with Oort Adjustment)

Simulation design factors	η^2
Data type	8.58
Magnitude of noninvariance (DIF)	15.40
Sample size	26.03
Data type*DIF	7.80
Data type*Sample size	9.49
DIF*Sample size	24.39
Data type*DIF*Sample size	8.31

Note. η^2 = the proportion of variance explained by each variable (%).

and (c) that a large sample size is required to detect a small size of noninvariance. Making interpretations more complicated, the types of data and the interactions with other design factors were also related to power. For example, the contributions of sample size and DIF size appeared less compelling for continuous data than for dichotomous data because the power of continuous data was simply 100% regardless of sample size and DIF size.

For the iterative procedure of modification indices with no critical value adjustment, the proportion of variance explained by each design factor is presented in Table 12. The power rates were highly attributable to sample size, magnitude of noninvariance, and their interaction as observed in the likelihood ratio tests. The effects of sample size and DIF were predominant in detecting both noninvariant variables with modification indices explaining approximately 85% of the variance. On the other hand, the same interaction effects of data type and other design factors were detected as in the likelihood ratio tests but the effects were less salient with two noninvariant variables.

Table 12

The Proportion of Variance Explained by the Simulation Design Factors: Modification Indices of the Intercept Noninvariance with the Iterative Procedure (Two DIFs with No Adjustment)

Simulation design factors	η^2
Data type	6.29
Magnitude of noninvariance (DIF)	21.87
Sample size	42.65
Data type*DIF	1.06
Data type*Sample size	4.27
DIF*Sample size	20.42
Data type*DIF*Sample size	3.44

Note. η^2 = variance explained by each variable (%).

The effects of data type in terms of power were further investigated through the inspection of power rate changes across different data types. The same conditions selected for ANOVA were used for this analysis. Because the continuous data showed the highest power, the power reduction was computed by subtracting the power rates of either dichotomous or polytomous data from the corresponding power rates of continuous data. Then, the difference was divided by the power rates of continuous data. As presented in Tables 13 and 14, when the magnitude of noninvariance was large, the types of data (dichotomous, polytomous, or continuous) did not make a difference with respect to power. The power reduction rates between continuous data and polytomous data or between continuous data and dichotomous data were .00 in most sample size conditions. That is, irrespective of data type, the noninvariance was detected almost all the time. However, when DIF and sample size were small, the power rates of dichotomous data were noticeably lower than those of continuous data, especially with

Table 13

Likelihood Ratio Test Power Reduction Rates across Data Types for Intercept Noninvariance in One Noninvariant Variable

DIF	Sample size	Con – Poly	Con – Dich
Small	200	.08	.32
	400	.02	.15
	1000	.00	.01
	2000	.00	.00
Large	200	.00	.01
	400	.00	.00
	1000	.00	.00
	2000	.00	.00

Note. Con – Poly = $(power_{continuous} - power_{polytomous})/ power_{continuous}$, Con – Dich = $(power_{continuous} - power_{dichotomous})/ power_{continuous}$.

two noninvariant variables. For example, when the magnitude of noninvariance was small with sample size of 400, the power of dichotomous data was 70% lower than that of continuous data. In comparison of dichotomous and polytomous data, the degree of power reduction rates were greater in the dichotomous data with small DIF and small sample size (Table 14).

Table 14

Modification Index Power Reduction Rates across Data Types for Intercept Noninvariance in Two Noninvariant Variables

DIF	Sample size	Con – Poly	Con – Dich
Small	200	.58	.83
	400	.29	.70
	1000	.00	.13
	2000	.00	.00
Large	200	.01	.39
	400	.00	.05
	1000	.00	.00
	2000	.00	.00

Note. Con – Poly = $(power_{continuous} - power_{polytomous})/ power_{continuous}$, Con – Dich = $(power_{continuous} - power_{dichotomous})/ power_{continuous}$.

CHAPTER V

CONCLUSIONS

Discussions

Likelihood ratio tests. This study explored the behaviors of MIMIC modeling in testing measurement invariance under a variety of research conditions. When factor loading was noninvariant over groups, the primary interest of this study was the sensitivity of MIMIC modeling to the presence of factor loading noninvariance through the model fit evaluations and measurement invariance testing. For the noninvariance in the intercepts of continuous variables or in the thresholds of categorical variables, the present study focused on the power and Type I error rates of MIMIC modeling with different critical value adjustment strategies using either likelihood ratio tests or modification indices.

This study observed the insensitivity of model fit indices to the violation of factor loading invariance assumption of the MIMIC model. This finding implies that good model fit of a MIMIC model may not guarantee the equivalence of factor loadings over groups. All examined model fit indices including chi-square *p*, CFI, RMSEA, and SRMR/WRMR consistently showed that MIMIC models generally failed to identify the factor loading noninvariance exhibiting good fits almost all the time. None of the fit indices evaluated in this study properly detected the factor loading noninvariance of MIMIC modeling. The proportions detected as a poor fit were not different from the predetermined significance level (that is, .05) when dichotomous data were analyzed.

With the good fit of a MIMIC model, factor loadings may not be invariant across groups as illustrated in this simulation study. To use a MIMIC model for measurement invariance testing or for the comparison of the latent means across groups, the violation of factor loading invariance should be tested with other measurement invariance testing techniques (e.g., multiple group CFA).

One of limitations of MIMIC is that it does not allow a partial invariance model. That is, researchers cannot establish the levels of measurement invariance such as configural, metric, and scalar invariance using MIMIC. Of note is the fact that the proposed measurement invariance testing method of MIMIC in this study purports to assess the intercept or threshold noninvariance. The direct effect of group membership on the observed variables (β_j in Figures 3 and 4 or Equation 2.14) tests the noninvariance of intercept /threshold holding the latent factor effects ($\lambda_j\eta_i$ in Equation 2.14) constant. However, if the controlled latent factor effects ($\lambda_j\eta_i$) are not invariant over groups due to the noninvariance of factor loadings, this violation of the invariant factor loading assumption may be reflected in the process of testing the intercept/threshold invariance. From this reasoning, this paper included the noninvariant factor loadings in the simulation conditions of the LR tests. Only when both sample size and degree of noninvariance were large, the noninvariance of a factor loading was adequately detected. However, in most conditions power was not high with a minimum of .03 (small DIF of dichotomous variables with sample size 400) and maximum of .87 (large DIF of polytomous variables with sample size 2000). Especially for dichotomous data, the observed power remained the basal Type I error rates. Therefore, MIMIC, in general,

appears not to be an optimal model to test factor loading equivalence. Again, other measurement invariance testing techniques are recommended to detect factor loading noninvariance.

With respect to intercept/threshold noninvariance, the major factors influencing power included sample size and degree of noninvariance. Sample size and degree of noninvariance were positively related to power. However, the degree of noninvariance exhibited interaction with sample size. That is to say, sample size matters if the degree of noninvariance is small. For example, when noninvariance was large, power reached 100% throughout different sample size conditions. However, small difference across groups was less detectable if sample size was small. When both effect size and sample size were small for dichotomous variables, the detection rate of a single noninvariant variable was about .60; power was below .05 with two noninvariant variables. On the other hand, for polytomous and continuous variables with small effect size and small sample size, MIMIC modeling was still tenable as a detection method. The LR tests identified the noninvariance of polytomous or continuous data with great precision across all conditions irrespective of the degree of noninvariance and sample size.

In addition, the power to detect the violation of measurement invariance depended on number of noninvariant variables (or degree of contamination) and type of data (dichotomous, polytomous, or continuous). As the number of noninvariant variables increased from 1 to 2 (for contamination rate, from 17% to 33%), power was slightly attenuated. As to data type, continuous data on average displayed the highest power, followed by polytomous data and dichotomous data. The overall behaviors of

polytomous variables are worthy of further comments. The polytomous variables behaved more like continuous variables than dichotomous variables across all scenarios. Given that the number of categories of polytomous variables in this simulation was five, polytomous variables may be more similar to continuous than dichotomous conditions. This finding implies that the common practice in measurement and psychometric research (that is, polytomous data with more than five response categories are often treated as continuous) appears to be reasonable, at least in measurement invariance research.

Previous simulation studies consistently reported high Type I error rates in detecting noninvariant variables either with multiple group CFA or with MIMIC modeling (Finch, 2005; Navas-Ara & Gomez-Benito, 2002; Oort, 1998; Wang et al., 2009). So did this simulation study before the Oort adjustment was applied to the critical chi-square values. The Type I error inflation in MIMIC (also in multiple group CFA) is explained by the misspecification of the baseline model in the likelihood ratio test. When the model includes any noninvariant variable and is analyzed with the assumption of invariance across groups, which is inherently done in MIMIC, the model fit of this model is poor due to the misspecification. The chi-square fit statistic reflects this poor model fit. As sample size and the degree of noninvariance increase, the chi-square statistic to capture the misspecification becomes more sensitive and tends to yield high values. Due to this heightened chi-square baseline statistic, the likelihood ratio test between an augmented (or relaxed) model and this misspecified baseline model with inflated chi-square fit statistic would produce an inflated chi-square difference value that

leads to the rejection of the null hypothesis more frequently than it should. This was demonstrated in the no-adjustment conditions of this study.

Bonferroni correction is one option of critical value adjustment. In this study, each replication went through six likelihood ratio tests, which might inflate Type I error. However, the major cause of Type I error inflation appeared beyond the experimentwise Type I error inflation. As explained in the previous section, the inflation more likely originated from the baseline model misspecification with the oversensitivity of chi-square distribution on the model misfit when sample size is substantial. Therefore, Bonferroni correction will not be an appropriate remedy in this case although it could lower Type I error. As the results showed, after the Bonferroni adjustment, the Type I error rates were still considerably higher than acceptable throughout all conditions. Another concern of Bonferroni correction is in the power reduction. Adopting more conservative critical *p* value reduced the power to identify the noninvariance. However, this power shrinkage was less obvious when sample size and the degree of noninvariance were large. To sum up, it can be concluded that Bonferroni correction is not an appropriate method to suppress the inflated Type I error rates when Type I error is expected due to the baseline model misspecification.

Oort correction does not merely lower the critical value but takes into account the magnitude of baseline chi-square value given degree of freedom (see Equation 2.15). Therefore, instead of evaluating the model fit with one fixed critical value such as 3.84 (χ^2 with one degree of freedom at α = .05), the critical chi-square value was tailored for each model depending on the degree of inflation of chi-square. The results of simulation

showed that the Oort adjustment worked remarkably well when the inflation was severe. Because Oort adjustment tailored the critical chi-square value for each baseline model, it did not lessen the power to detect the noninvariance much. When there was only one noninvariant variable, MIMIC modeling with Oort adjustment performed decently even with small sample size. The LR test detected the large degree of noninvariance all the time even with sample size as small as 200 while the Type I error rate remained near zero.

However, two adverse situations to Oort adjustment were observed in this study: (a) Oort correction did not make an improvement in controlling Type I error when power was low, and (b) Oort correction attenuated power when more than one DIF variables existed. Because Oort correction was developed to adjust the inflated chi-square of the baseline model, when the inflation of chi-square was severe, the Oort correction showed high performance: the Type I error rate of the large sample and large DIF condition dropped from 1.00 to .01 after the Oort correction was applied for continuous variables. However, when power was low (approximately below .50; e.g., most of factor loading noninvariance conditions), and accordingly the chi-square inflation was not likely to occur, Oort correction did not improve Type I error much. That is, it is not necessary to correct the critical values in these situations.

Although the overall performance of Oort-adjusted LR tests in the two-noninvariant-variable conditions was reasonable, the deterioration of power was notable compared to the one-noninvariant-variable cases. Given that the LR test relaxed only one variable at a time, even when the less restricted model correctly specified one of the DIF

items, another noninvariant variable existed in the model. That is, when the baseline model had two noninvariant variables, the less restricted model in which one of DIF was relaxed for inequality over groups still had another noninvariant variable. Subsequently, both models (baseline model with two DIF variables and augmented model with one DIF variable) in the LR test are incorrectly specified. As explained by Yuan and Bentler (2004), the LR tests between a misspecified baseline model and a misspecified unconstrained model did not yield the same power in detecting DIF items as the LR tests did with only one DIF item.

This understanding about the power degradation with more than one noninvariance calls for the iterative procedure of the LR test. Because the Type I error rates were considerably low throughout conditions with Oort adjustment, the detected variable in the first LR test was more likely to be one of the DIF variables. Hence, if this detected item is free to be estimated across groups and if this unconstrained model is utilized as a baseline for the following LR test, the same high performance of MIMIC is expected as observed in the one-noninvariance conditions. It should be noted that adequate performance of the iterative LR tests is expected when the detected variables are likely to be noninvariant variables (i.e., when the Type I error rate is low). Therefore, the statistical approach to control for the Type I error rates (e.g., Oort adjustment) to identify a noninvariant variable with accuracy will play a critical role in the LR tests using MIMIC modeling.

Modification indices. The simulation results of this study showed adequate performance of MIMIC when used for likelihood ratio tests with Oort adjustment to

detect the intercept/threshold noninvariance. One important application of the findings is in the utilization of modification indices with the Oort-adjusted cutoff values. Modification indices are easily obtainable with one command in most SEM statistical packages. Modification index is the chi-square difference computed by relaxing one fixed parameter to make an improvement of the model fit. One approach to utilize modification indices for measurement invariance testing is to find the noninvariant variables by requesting modification indices over a certain cutoff such as chi-square value of one degree of freedom at the full invariance model. Another usage of modification indices is to modify the model by relaxing the parameter with the highest modification index sequentially until the model shows a good fit or until modification index meets a certain criterion. I applied the findings of the LR test to modification indices by setting the Oort-adjusted chi-square critical value as the cutoff of modification indices. It should be of note that the cutoff should be adjusted at each step of modification because Oort adjustment depends on the chi-square statistic and degrees of freedom of a baseline model.

In terms of modification indices, the iterative procedures either with or without cutoff adjustment outperformed the noniterative procedures. In case of the iterative procedure to search noninvariance through a series of model modification, critical value adjustment did not make a noticeable difference. In fact, the results were almost identical between adjusted and unadjusted iterative methods. Therefore, if the iterative procedure is employed, the critical value adjustment appeared not to be necessary. The iterative procedures showed high performance in controlling the Type I error rates. Consistent

with the previous studies, Type I error appeared to be of no concern in the iterative procedures. The power rates remained near 100% with a sufficient size of sample (1000 for small effect size; 400 for large effect size). Moreover, the iterative procedure detected the small-size noninvariance better than the noniterative method. Considering both power and Type I error, a modification index in the iterative model search procedures can be an accurate indicator of a noninvariant variable. This finding implies that the utilization of modification indices may replace the likelihood ratio tests for the purpose of the identification of noninvariance.

When the noniterative method is utilized, the critical value adjustment using Oort formula seemed necessary. As observed in the LR tests, the noniterative procedure without critical value adjustment was not viable due to extremely high Type I error rates. On the other hand, for the noniterative method with Oort adjusted cutoffs even the initial information of modification indices showed great precision in detecting the noninvariant variables with large sample size, over 1000 (e.g., power = 1.00, Type I error = .00). If sample size is substantial, then all existing noninvariant variables, especially with large DIF can be detected simultaneously even in the first stage of model modification if the cutoff values are adjusted in consideration of baseline model misfit. However, I caution the use of modification indices at the initial stage of model modification because under the situations of higher contamination (more noninvariant variables) the performance of modification indices at the initial stage has not been determined yet. Considering that the Oort-adjusted critical value needs to be reestimated depending on the baseline model, whether a single Oort adjustment based on the full invariance model will work for a

large number of noninvariant variables is still in question. This indicates a call for further research on the noniterative method with more noninvariant variables: whether the performance of the noniterative method with the cutoff adjustment is consistently comparable with the iterative method when more than two noninvariant variables are present.

In this simulation study, MIMIC modeling utilized for measurement invariance testing showed the highest performance with continuous variables followed by polytomous and dichotomous variables, respectively. Moreover, the behavior of polytomous variables in terms of power and Type I error in detecting noninvariant variables is closer to that of continuous data rather than dichotomous data. Given that the polytomous variables in this study have five response categories, the behavior of polytomous variables with less than five categories needs further investigation to explore the relations between categorical and continuous data in measurement invariance testing.

Conclusions

Measurement invariance testing is important to establish the validity of a measure across subpopulations of interest. The detection of noninvariant variables is necessary to improve test quality and, furthermore, to understand the meaning of noninvariance over groups.

The findings of this study were three fold. First, when noninvariance existed in factor loading only, the MIMIC model did not detect the noninvariance. The model fit indices of MIMIC were insensitive to the unequal factor loadings over groups showing good fits. Thus, the MIMIC model should be used only when the metric invariance is

achieved. Second, for the LR tests using MIMIC with a direct path from a grouping variable to an observed variable, if the baseline model was contaminated with noninvariant variables, the Type I error rates were inclined to inflate. In this case, Oort adjustment on critical values controlled the high Type I error rates reasonably while keeping the power rates high. Third, the process of LR tests to detect the noninvariant items implied the utilization of modification indices. When utilizing modification indices, an iterative procedure is recommended, especially when the adjustment of the cutoff is not applied. With large sample size, even at the initial stage of model modification, the large degree of noninvariance was well detected when the cutoff of modification index was properly adjusted. However, without knowledge of the magnitude of noninvariance in reality, iterative procedure should be the choice of method.

REFERENCES

Ainsworth, A. T. (2008). Dimensionality and invariance: Assessing differential item functioning using bifactor multiple indicator multiple cause models. *Dissertation Abstracts International, 68*(9), 6383B-6494B.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289-300.

Blake, J. J., Kim, E. S., & Lease, A. M. (2011). Exploring the incremental validity of nonverbal social aggression: The utility of peer nominations. *Merrill Palmer Quarterly, 57*, 293-318.

Borsboom, D. (2006). When does measurement invariance matter? Commentary. *Medical Care, 44*(11), S176-S181.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152.

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

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464-504.

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.

Cheung, G. W. (2008). Testing equivalence in the structure, means, and variances of higher-order constructs with structural equation modeling. *Organizational Research Methods, 11*(3), 593-613.

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.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*(1), 19-29.

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

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*(4), 278-295.

Fleishman, J., Spector, W., & Altman, B. (2002). Impact of differential item functioning on age and gender differences in functional disability. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 57*(5), S275-S284.

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling: A Multidisciplinary Journal, 15*(1), 96-113.

Gelin, M. N. (2005). Type I error rates of the DIF MIMIC approach using Jöreskog's covariance matrix with ML and WLS estimation. *Dissertation Abstracts International, 66*(2), 1214B-1328B.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, *66*(3), 373-388.

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2000). Type I error and power of latent mean methods and MANOVA in factorially invariant and noninvariant latent variable systems. *Structural Equation Modeling: A Multidisciplinary Journal, 7*(4), 534-556.

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

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

Jöreskog, K.G., & Sörbom, D. (2006). LISREL 8.8 for Windows [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal, 15*, 136-153.

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

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*, 1-17.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: The Guildford Press.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*(140), 55.

MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100*(1), 107-120.

McCarthy, D. M., Pedersen, S. L., & D'Amico, E. J. (2009). Analysis of item response and differential item functioning of alcohol expectancies in middle school youths. *Psychological Assessment, 21*(3), 444-449.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6*(1), 97-103.

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 611-635.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568-592.

Meade, A. W., & Kroustalis, C. M. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods*, *9*(3), 369-403.

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

Meade, A. W., Michels, L. C., & Lautenschlager, G. J. (2007). Are internet and paper-and-pencil personality tests truly comparable? An experimental design measurement invariance study? *Organizational Research Methods, 10*(2), 322-345.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*(2), 127-143.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-43.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*(2), 289-311.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*(11), S69-S77.

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

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*(3), 479.

Muthén, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 213-238). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Muthén, B. O., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus Web Notes: No. 4.* Retrieved July 5, 2008, from http://www.statmodel.com/download/webnotes/CatMGLong.pdf.

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

Muthén, B. O., & Muthén, L. K. (2008a). Mplus (Version 5.2) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., & Muthén, L. K. (2008b). *Mplus User's Guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.

Navas-Ara, M. J., & Gomez-Benito, J. (2002). Effects of ability scale purification on the identification of dif. *European Journal of Psychological Assessment, 18*(1), 9-15.

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika, 6*(2), 150-166.

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

Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management, 20*(3), 643-671.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., & Rauch, S. M. (2003). Validating a measure across groups: The use of MIMIC models in scale development. *Journal of Social Service Research, 29*(3), 53-67.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*(4), 210-222.

Shih, C.-L., & Wang, W.-C. (2009). Differential item functioning detection using the multiple indicators multiple causes method with a pure short anchor. *Applied Psychological Measurement*, *33*(3), 184-199.

Sörbom, D. (1989). Model modification. *Psychometrika, 54*(3), 371-84.

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.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach.* New York: The Guilford Press.

Thompson, M. S., & Green, S. B. (2006). Evaluating between-group differences in latent variable means. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation*

modeling: A second course (pp. 119-169). Greenwich, CT: Information Age Publishing.

Thurstone, L. L. (1928). Attitudes can be measured. *The American Journal of Sociology, 33*(4), 529-554.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-69.

Wang, W.-C., Shih, C.-L., & Yang, C.-C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational & Psychological Measurement, 69*(5), 713-731.

Wanichtanom, R. (2001). *Methods of detecting differential item functioning: A comparison of item response theory and confirmatory factor analysis*. Unpublished doctoral dissertation, Old Dominion University, Norfolk.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Wilse, J. T., & Goodman, J. T. (2008). Comparison of multiple-indicators, multiple-causes- and item response theory-based analyses of subgroup differences. *Educational and Psychological Measurement, 68*(4), 587-602.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*(1), 58-79.

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

Yang, Y. (2008). *Partial invariance in loadings and intercepts: Their interplays and implications for latent mean comparisons*. Unpublished doctoral dissertation, University of Nebraska, Lincoln.

Yoon, M. (2008). Statistical power in testing factorial invariance with ordinal measures. *Dissertation Abstracts International, 68*(11), 7705B-7854B.

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

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

Yuan, K. H., & Bentler, P. M. (2004). On chi-square difference and z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement, 64*(5), 737-757.

APPENDIX A

Summary of Simulation Conditions in the Literature about Measurement Invariance

Authors	Year	Sample size (Group size)	Items	Factors	Rep	Program used
Fan & Sivo	2009	20* number of items = 80~240	4/8/12	2	1000	LISREL 8.51
		40* number of items = 240~720	6/12/18	3		
		60* number of items = 480~1440	8/16/24	4		
Cheung	2008	400 (200)	32	4 [a]	200	
				4 [b]		
French & Finch	2008	500/1000(250/500)	6	1	1000	LISREL
			6	2		PRELIS
			12			
Meade et al.	2008	100/200/400/800/1600/6400	16	4	500	LISREL
			32	4		PRELIS
			8	2		
			16	2		
Yang[c]	2008	500 (250)	6	1	1800	Mplus
			12			
Yoon[c]	2008	200/400/1000/2000	6		500	Mplus 4.21
		(100/200/500/1000)	12			
Chen	2007[1]	300/500/1000(150/250/500)	8	1	500	Mplus 3.01
			12	2		
	2007[2]	300(150:150/200:100/240:60)	8	1	500	
		500(250:250/333:167/400:100)	12	2		
		1000(500:500/666:334/800:200)				

Summary of Simulation Conditions in the Literature about Measurement Invariance (Continued)

Authors	Year	Sample size (Group size)	Item	Factor	Rep	Program used
Meade & Bauer	2007	200/400/800 (100/200/400)	20	6	500	LISREL
			20	3		PRELIS
Yoon & Millsap	2007	400 (200)	6		100	LISREL
		1000 (500)	12			
Meade & Kroustalis	2006[1]	500 ~ 10000 by 500 increment	32	2	300	
	2006[2]	100/200/300/400/500	32	2	300	
	2006[3]	100/200/300/400/500/1000/10000	32	2	100	
	2006[4]	100/200/300/400/500	32	2	100	
	2006[5]	100/200/300/400/500	32	2		
Millsap & Kwok	2004[1]		4	1		
	2004[2]		4/8/12/16			
Cheung & Rensvold	2002	300/600 (150/300)	3/factor	2	1000	AMOS 3.6
			4/factor	3		
			5/factor			
Wanichtanom[c]	2001	2000 (1000)	50	1	25	SAS
						MULTILOG
						SAS Proc Calis

Note. Item = Number of items, Factor = Number of factors, Rep = Number of replications.
[a] 4 factors with 2 second-order factors in direct effect, [b] 4 factors with 2 second-order factors in correlation, [c] dissertation, [1] study 1, [2] study 2, [3] study 3, [4] study 4, [5] study 5,

APPENDIX B

Summary of DIF Conditions in the Literature about Measurement Invariance

Authors	Year	DIF location	DIF size	% contamination (Number of DIF)
Fan & Sivo	2009	mean	0/.1/.2/.3/.4/.5/.8	
Cheung	2008	a	0.6	(2 of 32)
		b	2	
French & Finch	2008	a	0.25	0/17/34
Meade et al.	2008	a	0~0.4 by .02	25
		b	0~0.3 by .02	(4 of 16/ 8 of 32)
Yang	2008	a	0.1/0.2	33/66
		b	0.1/0.5	
Yoon	2008	a	0.08/0.16/0.24[a]	33
		b	10/20/30[b]	
		both		
Chen	2007			25/50/75/100
Meade & Bauer	2007	a	0.2	
Yoon & Millsap	2007	a	0.1/0.2/0.3	33/67
Meade & Kroustalis	2006[1]	a	0.654	4 items per factor
	2006[2]	a	0.02 ~ 0.40 by .02	4 items per factor[c]
	2006[3]	b	0.4	4 items per factor
	2006[4]	b	0.05 ~0 .40 by .05	4 items per factor[c]
Millsap & Kwok	2004	a	0.2	0/25/50/75/100
		b		
Wanichtanom	2001	a	0.5[d]	18 (9 of 50)
		b	0.2[d]	
		both		

Note. a = factor loading or discriminant parameter, b = intercept or difficulty parameter, mean = latent factor mean

[a] size of SRMR (standardized root mean squared residual); [b] % reduction in probability for each response variate; [c] 4 items per factor, 1 item per parcel, 2 items in 4 parcels, or all items, [d] size of Raju area difference, [1] study 1, [2] study 2, [3] study 3, [4] study 4.

APPENDIX C

Summary of Simulation Conditions in the Literature about MIMIC Modeling

Authors	Year	Sample size (group size)	Item	Factor	Rep	Res	Program used
Shih & Wang	2009	1000/2000/3000 (500/1000/1500)	20/30/40		100		MATLAB Mplus
Wang et al.	2009		50		100		MATLAB Mplus
Woods	2009[1]	(500:25/100/200/400) (1000:25/50/100/200/400)	6/12/24		100	2 5	Mplus IRTLRDIF
	2009[2]	(500:50/100/200/400) (1000:50/100/200/400)	6/12/24				MULTILOG
Ainsworth[a]	2008	900 (500/400)	20	4[b]	100	2	TESTFACT
Wilse & Goodman	2008	2000 (1000:1000) 2000(1800:200)			100		R PARSCALE Mplus
Finch	2005	600/1000 (500:100/500:500)	20/50		500		Mplus
Gelin[a]	2005	400/1000/2000 (200/500/1000) (100:900/200:800/300:700/400:600)	10/20		1000	4	LISREL PRELIS
Navas-Ara & Gomez-Benito	2002	1000	25	1	3	2	BMDP PASCAL LISREL
Hancock et al.	2000	200/400/800/1600 (100/200/400/800) (80:120/160:240/320:480/640:960) (50:150/100:300/200:600/400:1200)	6	2	1000		EQS 5.7 GAUSS

Summary of Simulation conditions in the Literature about MIMIC Modeling (Continued)

Authors	Year	Sample size (group size)	Item	Factor	Rep	Res	Program used
Oort	1998	200/2000 (100/1000)	40	1	10	2	LISREL
		400 (200:200/300:100)	40	1	3	7	PASCAL

Note. Item = Number of items, Factor = Number of factors, Rep = Number of replications, Res = Number of response categories.

[a] dissertation, [b] 4 factors with 1 general factor, [c] dissertation, [1] study 1, [2] study 2, [3] study 3, [4] study 4, [5] study 5.

APPENDIX D

Summary of DIF Conditions in the Literature about MIMIC Modeling

Authors	Year	DIF location	DIF size	% contamination (Number of DIF)	Estimation
Shih & Wang	2009			0/10/20/30/40	
Wang et al.	2009	b	0.6	8/20/28/40	WLS
Woods	2009[1]	both	0.3~0.7		MLR
		b			
	2009[2]	b	0.9[a]		
Wilse & Goodman	2008				WLSMV ML
Finch	2005		0/0.6	0/15	WLS
Gelin	2005		0	0	ML/WLS
Navas-Ara & Gomez-Benito	2002	b	0.75	40 (10)	WLS
Hancock et al.	2000	a	0.2/0.4	(4 of 6)	ML
Oort	1998[1]	b	0.5[b]	25 (10 of 40)	ML
			0.8[b]		
	1998[2]	b	0.2		
		a	0.5		
		both	0.8		

Note. a = factor loading or discriminant parameter, b = intercept or difficulty parameter, WLS = weighted least squares, WLSMV = WLS with robust mean and variance, ML = maximum likelihood.

[a] normal distribution of mean 0.9 and *SD* 0.4, [b] *SD* of item scores, [1] study 1, [2] study 2.

VITA

Name: Eun Sook Kim

Address: Department of Educational Psychology
Texas A&M University
4225 TAMU
College Station TX 77843-4225

Email Address: lyriceye@gmail.com

Education: B.A., English Education, Pusan National University, 1993
M.S., Educational Psychology, Texas A&M University, 2007
Ph.D., Educational Psychology, Texas A&M University, 2011

Publications:

Kim, E. S., Kwok, O., & Yoon, M. (in press). Testing factorial invariance in multilevel data: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal*.

Thoemmes, F., & Kim, E. S. (in press). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research.*

Blake, J. J., Kim, E. S., & Lease, A. M. (in press). Exploring the incremental validity of nonverbal aggression: The utility of peer nominations. *Merrill-Palmer Quarterly*.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: Comparison of multiple- group categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal, 18*, 1-17.

Blake, J. J., Kim, E. S., McCormick, A. S., & Hayes, D. (2011). The dimensionality of relational victimization: A preliminary investigation. *School Psychology Quarterly, 26*(1), 56-69.

Kim, E. S., & Willson, V. L. (2010). Evaluating pretest effects in pre-post studies. Educational *and Psychological Measurement, 70*(5), 744-759.

Willson, V. L., & Kim, E. S. (2010). Pretest sensitization. In N. Salkind (Ed.), *Encyclopedia of Research Design: Vol. 2* (pp. 1091-1094). Thousands Oaks, CA: Sage Publications.