DETECTING THE VIOLATION OF FACTORIAL INVARIANCE WITH AN UNKNOWN REFERENCE VARIABLE

A Dissertation

by

EUNJU JUNG

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

DOCTOR OF PHILOSOPHY

Chair of Committee,	Victor L. Willson
Co-Chair of Committee,	Myeongsun Yoon
Committee Members,	Robert J. Hall
	Oi-Man Kwok
Head of Department,	Victor L. Willson

August 2014

Major Subject: Educational Psychology

ABSTRACT

A widely used tool for testing measurement invariance is multi-group confirmatory factor analysis (MCFA). Identification of MCFA models is usually done by imposing invariance constraints on parameters of chosen reference variables (RV). If the chosen RVs were not actually invariant, one could draw invalid conclusions regarding the source of noninvariance. How can an invariant RV be selected accurately? To our knowledge, no method is yet available, yet two approaches have been suggested to detect non-invariant (or invariant) items without choosing specific RVs. One is the factor-ratio test (FR-T), and the other is the use of the largest modification index (Max-Mod). These two approaches have yet to be directly compared under the same conditions. To address unsolved problems in partial measurement invariance testing, two studies were conducted. The first aimed to identify a truly invariant RV using the smallest modification index. The second aimed to directly compare the performances of FR-T and the backward approach using the Max-Mod in correctly specifying the source of noninvariance. The second study also proposes a new method—the forward approach facilitated by the bias-corrected bootstrapping confidence intervals. The performances of the three methods was compared in terms of perfect recovery rates, model-level Type I error rates, and model-level Type II error rates. The results of the first study indicated that the Min-Mod successfully identify a truly invariant RV across all conditions. In the second study, overall, the backward approach also showed best performance under 99% confidence level ($\alpha = 0.01$) in both partial metric invariance (PMI) and partial scalar

invariance (PSI) conditions. The performance of the forward approach was comparable with that of the backward approach only in PMI conditions. The factor-ratio test had the poorest performance. Limitations and future directions are also discussed.

DEDICATION

To Jonghoon, Daniel, and Sophia who have gone through all the joys and agonies in the last decade of my life, my parents and parents-in-law for their love and sacrifice, my grandmother and grandfather for their love and all memories with them, and God for everything that he has done.

ACKNOWLEDGEMENTS

First of all, I would like to thank my committee members. I really appreciate my committee Co-Chair, Dr. Yoon's invaluable academic advice and sincere cares about me as a person. Thanks to her guidance, I could start to focus on very important topics in measurement invariance literature from the beginning of my doctoral study. I would like to thank my committee Chair, Dr. Willson for his discerning comments which allowed me to see the problems with different perspectives which I could have not come up with by myself. Dr. Hall, I cannot imagine my doctoral training without him. His intriguing introductory statistics class let me start the program of research, measurement, and statistics. Throughout the study at Texas A & M, his supports and caring comments have encouraged me to complete this process. Last but not least, I am very grateful to Dr. Kwok for his guidance and supports throughout the course of this research.

Thanks also go to Dr. Chen, who is such a delight to work with. I have been very productive while working with her. More importantly, she is a great person who is very enthusiastic to contribute to the society through her research. I would like to give special thanks to my friend Minjung who has been the best friend of mine from the first year of my study at Texas A & M. She always encourages me to complete this long journey. I also really appreciate Myunghee's advice for solving the error codes in SAS. Without her help, it might have taken more time for me to complete the dissertation study.

NOMENCLATURE

CFA	Confirmatory Factor Analysis
MCFA	Multi-group Confirmatory Factor Analysis
PFI	Partial Factorial Invariance
PMI	Partial Metric Invariance
PSI	Partial Scalar Invariance
ST-IM	Standardization Identification Method
RV	Reference Variable
RV-IM	Reference Variable Identification Method
VRV-IM	Variation of Reference Variable Identification Method
Min-Mod	Smallest Modification Index
Max-Mod	Largest Modification Index
BCBS-CI	Bias-corrected Bootstrapping Confidence Interval
FR-T	Factor-ratio Test
PR	Perfect Recovery Rate

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

Educational researchers are often interested in comparing educational outcomes across different demographic groups. In order to make meaningful comparisons in doing so one should establish what is known as measurement invariance (Vandenberg & Lance, 2000; Steenkamp & Baumgartner, 1998; Schmitt & Kuljanin, 2008). Measurement invariance broadly refers to the condition in which the measured variables are related to the construct(s) being measured in the same way across different groups of interest (Vandenberg & Lance, 2000). Measurement invariance can be examined using various statistical approaches such as multi-group confirmatory factor analysis (Vandenberg & Lance, 2000; Meade & Lautenschlager, 2004; Schmitt & Kuljanin, 2008), multiple-indicator multiple-cause modeling (Woods, 2009b; Kim, Yoon & Lee, 2012), and item response theory models (Stark, Chernyshenko, & Drasgow, 2006).

Of these, one that is widely used to test measurement invariance is the multi-group confirmatory factor analysis (MCFA). Using MCFA method, factorial invariance is examined as a special case of measurement invariance. One can specify different levels of factorial invariance models to test invariance of parameters of special interest including factor loadings, intercepts, and unique variance. Typically, four nested invariance models are tested hierarchically (Vandenberg & Lance, 2000): (1) configural (same factor model), (2) metric (invariance of factor loadings), (3) scalar (invariance of factor loadings and intercepts), and (4) strict invariance (invariance of loadings,

intercepts, and unique variances). Upon established configural invariance, we sequentially conduct tests of metric, scalar, and strict invariance. When any level of full invariance is rejected, the next possible step is examining the source of noninvariance (Cheung & Rensvold, 1999). Here we can pursue a partial metric invariance model in which some factor loadings are fixed to be invariant while the others are freely estimated (Byrne, Shavelson, & Muthen, 1989; Steenkamp; Schimitt & Kuljanin, 2008). We also investigate a partial scalar invariance model by allowing some intercepts to be different across groups. When studying partial factorial invariance (PFI) models, we must exercise caution in choosing the identification method. The reference variable identification method (RV-IM) is preferred to the standardization identification method (ST-IM) (Yoon & Millsap, 2007; Cheung & Rensvold, 1999) because ST-IM might distort the true status of measurement invariance in both full and partial invariance levels. Jung and Yoon (2012) showed the problem of ST-IM using empirical and generated data conditions in addition to mathematical illustrations. In the empirical data analyses, the conclusions drawn from RV-IM and ST-IM were not consistent. In the generated data examples, ST-IM leads to inaccurate results while RV-IM performed adequately. For example, when ST-IM was employed, the truly strictly invariant data were rejected at the metric invariance level given different factor variances across group. However, RV-IM supported strict invariance, which is the true status of the data. In addition, they mathematically illustrated how the truly invariant factor loadings or intercepts can be estimated not to be invariant with ST-IM. Yet RV-IM has also its weakness. If we choose a non-invariant RV, the true status of PFI model might be

compromised (Yoon & Millsap, 2007; Johnson, Meade, & DuVernet, 2009). However, information regarding which variable is invariant is not readily available in most studies, and one congruent theme for identifying an invariant RV is letting researchers rely on theories—which may not always exist (Millsap & Olivera-Aguilar, 2012). In addition, it is possible that the chosen RV is not invariant for the given samples although the RV is selected based on theories. Therefore, it is important to select an RV based on both theories (if available) and reliable empirical guidelines. To our knowledge, however, a methodological approach that correctly identifies an invariant RV has yet to be developed (Raykov, Marcoulides, & Li, 2012).

With the recognition of the problem of selecting a non-invariant RV, two approaches have been suggested for identifying non-invariant (or invariant) parameters without selecting a specific RV: (1) the factor-ratio test (FR-T) and (2) the sequential use of modification index (Mod) under a fully constrained invariance model. First, Cheung and Rensvold (1999) suggested the FR-T in which every set of an RV and an argument (i.e., a variable being tested for invariance) is tested. French and Finch's (2008) simulation study showed that the performance of the FR-T is promising with nominal false positive rates and high true positive rates across conditions except for complex model and/or high contamination conditions. Yet, the labor-intensive feature of the FR-T discourages researchers from using it. Even though Cheung and Lau (2011) recently simplified the FR-T using bias-corrected bootstrapping confidence intervals (BCBS-CIs), it has not yet been evaluated under known data conditions. In addition, the subsequent procedures that supplement the FR-T (e.g. triangular heuristics, step-wise

partitioning procedure, and list-and-delete method) do not provide a clear solution. Second, Yoon and Millsap (2007) suggested the sequential use of the largest modification index (Max-Mod) with a full metric invariance model as a baseline model. They found that using Max-Mod performs well under the conditions with low contaminations, large loading differences, and large samples. The sequence of the model specification using Max-Mod is backward—from the most restricted model to the least restricted model. As a result, it is prone to inflated type I error rates due to possible misspecifications in the baseline model (Kim & Yoon, 2011; Whittaker, 2012). Additionally, large sample size also inflates the size of modification index (Chou & Bentler, 1990). Although this method is relatively simple to be used, it has been noted that data driven model modification should be done with cautions (Yoon & Millsap, 2007).

In factorial invariance research, two problems remain to be solved. The first is to develop a method to correctly identify an invariant RV. The second is comparing performance of FR-T and the Max-Mod in partial invariance models. This dissertation aims to address both problems. In addition, I propose a new method less susceptible to respective weaknesses of the FR-T and the Max-Mod. The following section reviews the factorial invariance literature related to these problems and then presents this paper's purpose.

CHAPTER II

LITERATURE REVIEW

This chapter reviews the following important issues in measurement invariance literature: (1) definition of measurement invariance, (2) importance of measurement invariance, (2) factorial invariance, (3) evaluating factorial invariance, (4) partial factorial invariance, and (5) identification problems in factorial invariance testing as an unresolved issue. The research purposes of the two proposed studies are then introduced.

Definition of Measurement Invariance

Research often involves comparing scores from competitive conditions. Such conditions could include different socio-demographic groups, assessments in different languages, and scores from different time points. When comparing such scores, researchers usually utilize the same instrument. Doing so though is no guarantee that the differences in the observed scores are mainly the function of the different standings on the construct(s). To draw valid conclusions about differences in observed scores, we have to establish in advance measurement invariance.

In a seminal work, Mellenbergh (1989) formally defined measurement invariance as:

$$P(X|W, G) = P(X|W) \quad (1)$$

Here, P (.) denotes the probability function related to X, W, and G. X represents observed score (s) from an instrument. W indicates an individual's actual standing on the latent construct underlying the observed scores. G stands for group membership. In Equation 1, the conditional probability of attaining a specific observed score is unaffected by group membership after taking into account the variances explained by the latent construct. By contrast, measurement invariance is violated if the differences in the observed scores are not only the function of the latent construct but also unequal operation of the measurement. In other words, when measurement invariance does not hold, observed scores can be biased.

Importance of Measurement Invariance

When educational or psychological measurements are used for deciding admission or employment or assigning resources (e.g., services related to psychiatric disorders), measurement bias may lead to poor decisions. To avoid adverse consequences of measurement bias, researchers have made both preventive and remedial efforts in educational and psychological testing situations. Regarding preventive efforts, Zieky (2013) showed how to minimize measurement bias through fairness reviews. Fairness reviews concern identifying and excluding items at risk of measurement bias. Zeiky (2013) also noted that the fairness of a test should be ensured by conducting empirical measurement invariance testing when sufficient data are available after the test administration. Based on the invariance testing results, the item posing measurement bias can be removed or revised as remedial efforts.

Various simulation studies have looked at the impact of measurement bias so as to address different empirical situations. Millsap and Kwok (2004) demonstrated how partially violated measurement invariance could affect selecting people using simulated data. In addition, violating measurement invariance prevents the accurate tracing of the development of individuals on latent construct. For example, Wirth (2009) examined the effects of violation of measurement invariance on the parameter estimates of the latent growth model under simulated data conditions. He found that measurement bias distorted the growth trajectory in latent growth modeling. Recently, Whittaker (2013) demonstrated how non-invariant intercepts affected on the test for latent means under various multi-group confirmatory factor analysis (MCFA) models and multiple-indicator multiple-cause (MIMIC) models using a Monte Carlo study. She found that the Type I error rates for testing latent means increases as the degree of noninvariance in intercepts increases. Therefore, measurement invariance should be established, prior to making decisions or comparisons based on either observed scores or latent scores.

Factorial Invariance

The most widely used empirical method for testing measurement invariance is confirmatory factor analysis (CFA). CFA is well known for its flexibility in testing measurement invariance. Using CFA to test measurement invariance is usually called "factorial invariance"—a special case of measurement invariance. In a CFA model, observed variables are linearly related to fewer latent variable(s) as in the following equation:

$$X = \tau + \Lambda\xi + \delta \quad (2)$$

Here, X is a $p \times 1$ vector of the observed scores, τ denotes a $p \times 1$ vector of intercepts, Λ represents a $p \times m$ factor loading matrix, ξ indicates an $m \times 1$ vector of the latent variables ($p > m$), and δ stands for a $p \times 1$ vector of unique factor scores. Equation 2 can be extended to a multi-group CFA (MCFA) model with the group indicator g as in Equation 3:

$$X_g = \tau_g + \Lambda_g\xi_g + \delta_g \quad (3)$$

Depending on the group membership, the parameters are allowed to vary in Equation 3. The assumption of uncorrelated latent variables and unique factor scores (i.e., *COV* (ξ,δ) =0)[1] leads to the following variance-covariance structure of X_g:

$$\Sigma_g = \Lambda_g\Phi_g\Lambda_g' + \Theta_g \quad (4)$$

In Equation 4, Σ_g represents a $p \times p$ variance-covariance matrix of observed scores (X_g) of the group g while Φ_g denotes an $m \times m$ variance-covariance matrix of latent factor scores (ξ_g). The final notation "Θ_g"indicates a $p \times p$ variance-covariance matrix of unique factor scores (δ_g). However, it is typically a diagonal matrix of the variance of

[1] COV (a, b) = Covariance between a and b

unique factor scores because unique factor scores are assumed to be uncorrelated one another (i.e. $COV(\delta_{ig}, \delta_{jg}) = 0$)[2].). The mean structure of Equation 3 can be written as below:

$$E(X_g) = \tau_g + \Lambda_g \kappa_g \quad (5)$$

Here, $E(X_g)$ refers to the mean vector of X_g while κ_g denotes the mean vector of ξ_g.The mean vector of δ_g does not appear in Equation 5 since the unique factor scores (δ_g) are, in the long run, expected to cancel out.

Equations 4 and 5 allow us to test every aspect of factorial invariance: configural invariance (i.e., equal form of the model), metric invariance (i.e., $\Lambda_g = \Lambda_{g'}$), scalar invariance (i.e., $\Lambda_g = \Lambda_{g'}$; $\tau_g = \tau_{g'}$), and strict invariance (i.e., $\Lambda_g = \Lambda_{g'}$; $\tau_g = \tau_{g'}$; $\Theta_g = \Theta_{g'}$). Given the condition of strict measurement invariance, the group indicator g for Λ, τ, and Θ can be eliminated in Equations 4 and 5 as shown below:

$$\Sigma_g = \Lambda \Phi_g \Lambda' + \Theta \quad (6)$$

$$E(X_g) = \tau + \Lambda \kappa_g \quad (7)$$

In Equation 6, the differences in the variance-covariance structure of observed scores originate from the differences in the variance-covariance structure of latent factor scores

[2] δ_{ig} and δ_{jg} stand for the unique factor score of i*th* and j*th* variables, respectively

when strict invariance is established. Based on Equation 7, we can infer that the differences in the observed mean scores can be accounted for by the difference in the means of the latent factor scores between groups. Equation 7 also implies that scalar invariance is a necessary condition for valid cross-group mean comparisons (Steenkamp & Baumgartner, 1998; Schmitt & Kuljanin, 2008).

A Closer Look at Each Level of Factorial Invariance

Vandenberg and Lance (2000) summarized the ideal steps for testing measurement invariance. Although they include testing the equality of the variance-covariance matrices across groups, empirical studies seldom test this too-strict condition. Researchers usually try to first fit a common factor model that works well for either group. Then, four levels of factorial invariance are tested sequentially: configural invariance, metric invariance, scalar invariance, and strict invariance.

Configural Invariance

Configural invariance refers to a MCFA model that stipulates the same number of factors and the same pattern of salient and non-salient factor loadings across groups (Vandenberg & Lance, 2000). Let's say that we have six variables that have two factors underlying them. The first three items are loaded on the first factor and the last three are loaded on the second factor. If the same structure holds across two groups, configural invariance is established as the following matrices:

$$\begin{bmatrix} \lambda_1^a & 0 \\ \lambda_2^a & 0 \\ \lambda_3^a & 0 \\ 0 & \lambda_4^a \\ 0 & \lambda_5^a \\ 0 & \lambda_6^a \end{bmatrix} = \begin{bmatrix} \lambda_1^b & 0 \\ \lambda_2^b & 0 \\ \lambda_3^b & 0 \\ 0 & \lambda_4^b \\ 0 & \lambda_5^b \\ 0 & \lambda_6^b \end{bmatrix} \quad (8)$$

Here, λ_i^a indicates the salient factor loading of the i^{th} item of Group *a* loaded on either the first or second factor. For Group *b*, the corresponding factor loading is expressed as λ_i^b, and the different group indicator *b* allows the value of the factor loading to differ from that of Group *a*. All zeros represent non-salient factor loadings. Establishing configural invariance is critical since it serves as the baseline model for the remaining higher-order factorial invariance models (Vandenberg & Lance, 2000).

Metric Invariance

After establishing configural invariance, we can test metric invariance by posing equality constraints on each set of corresponding factor loadings across groups. In the matrices, we can express metric invariance by removing the group indicators (i.e., *a* and *b* in Equation 8), as below:

$$\begin{bmatrix} \lambda_1 & 0 \\ \lambda_2 & 0 \\ \lambda_3 & 0 \\ 0 & \lambda_4 \\ 0 & \lambda_5 \\ 0 & \lambda_6 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ \lambda_2 & 0 \\ \lambda_3 & 0 \\ 0 & \lambda_4 \\ 0 & \lambda_5 \\ 0 & \lambda_6 \end{bmatrix} \quad (9)$$

In Equation 9, each set of factor loadings no longer varies across groups. Implied in its name, metric invariance refers to the comparability of the metrics or scale intervals across groups, according to Rock, Werts, and Flaugher (1978; as cited in Steenkamp & Baumgartner, 1998). In other words, when metric invariance holds, one unit change in the latent factor score results in the same unit change in the observed scores across groups. If metric invariance is not established, the strength of the relation between the latent construct to the observed variables is deemed to differ across groups (Vandenberg & Lance, 2000; Schmitt & Kuljanin, 2008).

Scalar Invariance

Under the condition of metric invariance, unless scalar invariance holds, observed scores can still be mathematically biased positively or negatively (Steenkamp & Baumgartner, 1998). In a scalar invariance model, more equality constraints on each set of corresponding intercepts are added to the metric invariance model, as shown below:

$$\begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \\ \tau_5 \\ \tau_6 \end{bmatrix} = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \\ \tau_5 \\ \tau_6 \end{bmatrix} \tag{10}$$

Here, τ_i denotes the intercept of the i^{th} variable. Scalar invariance is established across groups when every set of intercepts is equivalent. As implied in Equation 7, scalar

invariance is a critical condition for comparing cross-group means (Vandenberg & Lance, 2000; Schmitt & Kuljanin, 2008).

Strict Invariance

The most restricted form of factorial invariance is strict invariance (i.e., $\Lambda_g = \Lambda_{g'}$; $\tau_g = \tau_{g'}$; $\Theta_g = \Theta_{g'}$). Strict invariance tests are conducted by adding equality constraints on the corresponding unique factor variances on the scalar invariance model, as illustrated below:

$$\begin{bmatrix} \theta_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \theta_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \theta_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & \theta_6 \end{bmatrix} = \begin{bmatrix} \theta_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \theta_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \theta_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & \theta_6 \end{bmatrix} \quad (11)$$

Here, θ_i stands for the unique factor variance of the i^{th} variable. Strict invariance represents that, after taking into account the latent factor score(s), the uncertainty related to observed scores is equivalent across groups (Vandenberg & Lance, 2000; Schmitt & Kuljanin, 2008). Strict invariance is of course the preferred condition, yet in reality it is very difficult to be achieved (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000).

Identification Method

One model that relates observed variables to latent variables whose metric is unknown is confirmatory factor analysis (CFA) model. A CFA model cannot be

evaluated before identifying the metric of the latent variable(s). For a single-group CFA model, the scale of the latent variable can be constrained by fixing the factor variance to be a constant—usually 1. If the mean structure is considered, we can also fix the latent mean to 0 (or another constant). Second, we can assign the scale of the latent variable by fixing one of the factor loadings to a constant—usually 1. The metric of the chosen factor loading serves as the metric for the latent variable, and it is called a reference variable (RV). We can also assign the scale of the latent mean by constraining one of the intercepts to be 0 (or another constant). Conventionally, the same item serves as an RV for both variance-covariance structure and mean structure in a single analysis.

In a multi-group confirmatory factor analysis (MCFA), the identification method should be able to assign the unknown metric of latent variables as well as link the parameters of different groups. Therefore, in selecting our identification method for a MCFA, we need to be cautious. Measurement invariance literature puts forward three identification methods: (1) standardization identification method, (2) reference variable identification method, and (3) variation of the reference variable identification method.

Standardization Identification Method (ST-IM)

In an MCFA model, the model identification can be achieved by standardizing the structural parameters of each group, such as factor variances and factor means. Except for the case in which two (or more) groups are drawn from one population, it might be unreasonable to assume that the factor variances or means are identical across groups. If the factor variances or means are substantially different across groups, standardization will result in erroneous conclusions regarding the actual status of

factorial invariance (Cheung & Rensvold, 1999; Yoon & Millsap, 2007). For example, we are at risk of rejecting a truly invariant full metric invariance due to fallible standardization. For partial measurement invariance, the parameter estimates of factor loadings are also affected by the falsifying standardization, and as a result, the sources of noninvariance cannot be accurately detected. Those problems can affect the results of scalar invariance testing as well. In testing scalar invariance using ST-IM, factor means of both groups are fixed to zero. If the factor means are significantly different across groups, it is very likely to reject a model that is actually invariant. Therefore, standardizing factor variances or factor means should not be chosen to identify an MCFA model in testing factorial invariance (Cheung & Rensvold, 1999; Yoon & Millsap, 2007). When the default identification method is the ST-IM in the chosen latent variable modeling program, we need to override the default. For example, AMOS graphics uses the ST-IM to identify a scalar invariance model, and we need to change the identification method to specify a correct partial scalar invariance model.

Reference Variable Identification Method (RV-IM)

Using the RV-IM, an MCFA model is identified by fixing the parameters of a chosen variable to a specific value. For example, the factor loadings of the selected like-items are fixed to one to identify the model's variance-covariance structure. We identify the mean structure of the MCFA model by constraining to zero the intercepts of the chosen items. In the measurement invariance literature, RV-IM stands as the dominant identification method (Johnson et al., 2009). Nonetheless, it is hard to say that RV-IM is a safe method when the chosen reference variable (RV) is actually non-invariant.

Johnson et al. (2009) demonstrated how a non-invariant RV could distort the results of item-level measurement invariance tests. A good mathematical illustration is also available for the problem of choosing a non-invariant factor loading as an RV in detecting the source of noninvariance (Yoon & Millsap, 2007). However, researchers have yet to explain how to empirically choose a truly invariant RV (Raykov, Marcoulides, & Li, 2012). While many have emphasized the role of theory in choosing an invariant RV, in reality doing so is not always an option (Millsap and Olivera-Aguilar, 2012)

Variation of the Reference Variable Identification Method (VRV-IM)

As a variation of the RV-IM (VRV-IM), Reise, Widaman, and Pugh (1993) suggested another identification method. Using the VRV-IM, the variance-covariance structure of an MCFA model is identified by fixing to one the factor variances of a reference group while constraining as equal one set of corresponding factor loadings. The factor variance of the other group is freely estimated. To identify the mean structure, the factor mean of the reference group is fixed to zero with one set of equally constrained intercepts. For both cases, the equally constrained parameters need no specific value. There are two benefits to using the VRV-IM. First, we can be free of a priori invariance assumption for an RV when this identification method serves for a fully constrained factorial invariance model as in Yoon and Millsap (2007). Second, this method allows clear interpretations of the differences in the factor variances and factor means (Yoon & Millsap, 2007).

Identification Method in Empirical Studies

Since Rensvold and Cheung (1998; 2001) questioned the identification problems in testing factorial invariance, little attention has been paid to the identification issue in substantive areas. For example, the ST-IM can be frequently encountered in contemporary studies. We reviewed measurement invariance studies in 2000 through 2010 to identify the usage of each identification method. For keywords, we used "measurement invariance" or "measurement equivalence" and "confirmatory factor analysis." In the *PsychInfo* database, 232 studies were identified as eligible studies—testing measurement invariance under a MCFA. Among them, 29 studies (13%) used ST-IM while 59 (25%) offer no clue about the identification method. The remaining studies (62%) used RV-IM.

However, neither has much effort been made to select an invariant RV when using RV-IM. Johnson et al. (2009) reviewed studies testing, from 2005 and 2007, factorial invariance. They found that among the 153 eligible studies only 17 heeded the problem of non-invariant RVs.

In sum, it is obvious that in testing factorial invariance the substantive studies have seldom attended to the problem of identification methods. Nevertheless, we cannot overstate the importance of choosing an appropriate identification method; in testing both full and partial measurement invariance, the RV-M is preferred to the ST-IM (Cheung & Rensvold; 1999). When using the RV-IM, we need to select an invariant RV to correctly detect non-invariant items (Johnson et al., 2009). However, it is widely

accepted that there is no existing method to correctly identify an invariant RV (Raykov, Marcoulides, & Li, 2012).

Detecting Noninvariance without a Specific Reference Variable

Factor-ratio Test

Cheung and Rensvold (1998) suggested the factor-ratio test (FR-T) as the first method without a specific reference variable (RV). When using FR-T, every variable in a factor model serves as an RV while one of the remaining variables is tested for invariance. Once full metric or scalar invariance is rejected, we can employ FR-T. If more than two factors in the model are being tested, the researcher conducts the omnibus test of factor loadings separately for each factor. Once a factor is identified as the source of noninvariance, the FR-T test is administered to identify the non-invariant items. The following illustrates how to conduct FR-T with four variables under one factor:

1. When the first variable serves as an RV, we test three sets of combinations: (1, 2), (1, 3), and (1, 4).
2. When the second variable is an RV, we need to test only two pairs of variables (2, 3) and (2, 4) because the fit statistics of the pair (2, 1) is exactly same with that of (1, 2) the first step.
3. Similarly, we need to test only one pair, (3, 4), when the third variable is selected as an RV.

The total number of FR-Ts necessary for the case above is six. Generally, we need $\frac{p\,(p-1)}{2}$ FR-Ts to test every pair of an RV and an argument with p variables under one factor. In an FR-T, each model is compared to the unconstrained model (i.e., the model with one equal constraint for identification purposes) using chi-square difference testing.

A significant chi-square difference fit statistic indicates that the tested set may have non-invariant item(s).

French and Finch (2008) tested the performance of the factor-ratio test (FR-T) using simulated data. They found that the FR-T maintained nominal false positive rates while having good power across various conditions. Despite the FR-T's value being recognized, researchers have realized that with more indicators and factors, the process is labor-intensive. Recently, Cheung and Lau (2011) simplified FR-T by incorporating bias-corrected bootstrap confidence intervals (BCBS-CIs), a process still waiting to be investigated under known data conditions.

However, the FR-T informs only whether the tested pair is significant or not. Hence, a subsequent procedure should be conducted to distinguish the invariant variable from the non-invariant ones. Up to now, researchers have put forward three approaches: (1) trianglular heuristic (Cheung & Rensvold, 1998), (2) list-and-delete method (Rensvold & Cheung, 2001; Cheung & Lau, 2011), and (3) stepwise partitioning procedures (Rensvold & Cheung, 2001; French & Finch, 2008). In some data conditions, however, all three methods result in more than one invariant variable set in some data conditions (e.g., Cheung & Rensvold, 1999; Rensvold & Cheung, 2001; French & Finch, 2008; Cheung & Lau, 2011). When multiple invariant sets are identified, researchers should rely once again on theory (Rensvold & Cheung, 2001; Cheung & Lau, 2011). Such ambiguity might dishearten researchers from conducting the FR-T and its subsequent procedures.

Backward Approach Using the Largest Modification Index

The size of a modification index (Mod) indicates the amount of expected drop in the chi-square statistics (df =1) when one parameter constraint in the given model is released. Even though the literature has been cautious about making this kind of post-hoc model modification, it is appropriate in some cases to employ the sequential use of the largest modification index (Max-Mod; Yoon & Millsap, 2007). Yoon and Millsap (2007) utilized the Max-Mod of each factor loading to detect non-invariant items under a full metric invariance model. The chosen identification method was the variation of the reference variable identification method (VRV-IM). As mentioned above, this method allows us to avoid the a priori assumption that the chosen reference variable is invariant. If there is any Mod which exceeds a pre-specified significance level, the equality constraints having the Max-Mod were relaxed until no more significant Mod was left. The method performed well under the conditions of low contamination, large sample, a large loading difference condition, and mixed pattern of noninvariance condition. In some conditions, however, there were present high Type I error rates.

Forward Approach Using BCBS-CIs

The baseline model of the forward approach is the configural invariance model which does not have any invariant constraints on either factor loadings or intercepts except for the constraints for the reference variable. Therefore, the forward approach is less susceptible to the inflated Type I error rate due to the misspecification as of the baseline model of the backward approach. In the study of comparing the forward and backward approaches to detect measurement bias of categorical data, Khalid (2011)

found that the forward approach performed better than the backward approach in terms of the Type I error rates, yet he pointed out that the forward approach is only useful when we can correctly specify a truly invariant reference variable (RV).

In the newly proposed forward approach, the reference variable (RV) will be a truly invariant parameter. Then we will use BCBS-CIs for testing invariance instead of conducting separate likelihood ratio test for each tested variable, combined with the "MODEL CONSTRAINT" command in M*Plus*7.0. Here, we will only examine the BCBS-CI of each tested variable except for the reference variable by adopting the idea of Cheung and Lau (2011). For example, the BCBS-CIs of X2 through X6 will be examined when we choose X1 as an RV. Each confidence interval represents the confidence interval of the difference between corresponding parameters (e.g., $\lambda_2^a - \lambda_2^b$). If the BCBS-CI does not include zero, we will conclude that the tested parameter is not invariant across groups.

We expect this method to perform better in terms of both simplicity and accuracy. First, this new method needs only one data analysis to make all possible comparisons. Second, this method is expected not to have the problem of the inflated Type I error rates due to the misspecification in the baseline model.

Purpose of the Studies

Study I

The purpose of the study is to evaluate the performance of the smallest modification index (Min-Mod) in identifying a truly invariant reference variable (RV) in fully constrained factorial invariance models (e.g., full metric/ factor loading invariance

model, intercept invariance model, or scalar invariance model). This method is very similar to the "all others as anchors" (AOAA) identification method in the item response theory literature (Woods, 2009a; Mead & Wright, 2012), though using Min-Mod is a much simpler process.

Study II

The second study aims to evaluate the performances of (a) the forward approach, (b) the backward approach, and (c) the factor-ratio test in terms of perfect recovery rates, model-level Type I error rates, model-levelType II error rates, item-level power, and item-level Type I error rates. For both of the forward approach and factor-ratio test, we utilized the bias-corrected bootstrapping confidence intervals in detecting non-invariant item parameters (forward approach) or non-invariant pair of reference and argument (factor-ratio test). We referred to both 95% and 99% BCBS-CIs for the forward approach and factor-ratio test while, for the backward approach, we employed two significant modification index (Mod) values with one degrees of freedom , Mod = 3.841 at $\alpha = 0.05$ and Mode = 6.635 at $\alpha = 0.01$.

CHAPTER III

METHOD

Simulation Conditions

Study I

Using the Monte Carlo feature of M*Plus*7.0 (Muthen & Muthen, 1998-2012), data were generated considering both fixed and manipulated design elements. First, the fixed conditions were the number of factors and number of variables (one factor under six variables) and the proportion of noninvariance variables (33% across all conditions). Next, the manipulated conditions were: (1) location of noninvariance (either factor loadings or intercepts and both factor loadings or intercepts), (2) size/pattern of noninvariance (small- , large-, mixed-size-, and nonuniform-difference), and (3) sample size (N = 100, 250, 500, and 1000). The number of replications was set to 1000 for each condition. Table 1 summarizes the data conditions of the current study according to the manipulated design factors.

Location of Noninvariance

Many studies have pointed out that a necessary condition for cross-group latent mean comparison is establishing scalar invariance in which factor loadings and intercepts are invariant across groups (Vandenberg & Lance, 2000; Steenkamp & Baumgartner, 1998). A great deal of literature has discussed metric invariance (i.e., factor loading invariance) to evaluate the performance of a multi-group confirmatory factor analysis (MCFA) models in testing measurement invariance (Meade &

Lautenschlager, 2004; French & Finch, 2006; Meade & Bauer, 2007). Nonetheless, only recently have researchers begun to include scalar invariance in their simulation studies (Kim & Yoon, 2011; Kim, Yoon, & Lee, 2012). Neither study examined scalar invariance to evaluate the performance of the factor-ratio test (French & Finch, 2008) or the use of modification indices (Yoon & Millsap, 2007). In this dissertation, the location of noninvariance was chosen as one important design condition. We created four different scenarios: (1) partial metric invariance (PMI) in which noninvariance was only in factor loadings ($\lambda_{x2}^{g1} \neq \lambda_{x2}^{g2}$; $\lambda_{x4}^{g1} \neq \lambda_{x4}^{g2}$), (2) partial scalar invariance (PSI) in which noninvariance was only in intercepts ($\tau_{x2}^{g1} \neq \tau_{x2}^{g2}$; $\tau_{x4}^{g1} \neq \tau_{x4}^{g2}$), (3) partial metric and scalar invariance of the same variables (PMSI-S) in which noninvariance was in both factor loadings and intercepts in the same variables ($\lambda_{x2}^{g1} \neq \lambda_{x2}^{g2}$; $\lambda_{x4}^{g1} \neq \lambda_{x4}^{g2}$; $\tau_{x2}^{g1} \neq \tau_{x2}^{g2}$; $\tau_{x4}^{g1} \neq \tau_{x4}^{g2}$), and (4) partial metric and scalar invariance of different variables (PMSI-D) in which noninvariance was in both factor loadings and intercepts in different variables ($\lambda_{x2}^{g1} \neq \lambda_{x2}^{g2}$; $\lambda_{x4}^{g1} \neq \lambda_{x4}^{g2}$; $\tau_{x1}^{g1} \neq \tau_{x1}^{g2}$; $\tau_{x3}^{g1} \neq \tau_{x3}^{g2}$).

Size/Pattern of Noninvariance

In the literature an important data condition has also been the degree of differences between groups (Yoon & Millsap, 2007). Previous studies showed that power is higher when the difference is larger (Meade & Lautenschlager, 2004; French & Finch, 2006; Yoon & Millsap, 2007). In addition, the power increased when non-invariant variables had a nonuniform pattern, which means that some loadings are higher in the reference group while some loadings are higher in the focal group (Meade and Lautenschlager,2004; Yoon and Millsap, 2007). Therefore, we generated four levels of

the size/pattern of noninvariance depending on the magnitude differences and the direction of differences. In the small-difference condition, we manipulated two parameters to be noninvariant across groups with the same degree of small difference in the same direction. For example, the second and fourth factor loadings of Group 1 were 0.2 higher than those of Group 2 in the small-difference partial metric invariance (PMI) condition. In the small-difference partial scalar invariance (PSI) condition, the second and fourth intercepts of Group 1 was 0.3 lower than that of Group 2. In the large-difference conditions, the two noninvariant parameters had a large difference in the same direction. In the large-difference PMI condition, the second and fourth noninvariant factor loadings of Group 1 were 0.4 higher than those of Group 2 while in large-difference PSI condition the second and fourth intercepts of Group 1 were 0.6 lower than those of Group 2. Additionally, we simulated two types of mixed pattern conditions which are plausible in real settings: mixed-size-difference condition and nonuniform-difference condition. In the mixed-size-difference condition, two noninvariant parameters had varying degree of differences in the same direction. For example, in the mixed-size-difference PMI condition, the second and fourth factor loadings of Group 1 were 0.3 and 0.5 higher, respectively, than those of Group 2. In the nonuniform-difference condition, two noninvariant parameters were in different directions. For instance, the second factor loading of Group 1 was 0.3 higher than that of Group 2 while the fourth factor loading of Group 1 was 0.3 lower than that of Group2.

Table 1. Simulation Conditions

	Group 1	Group 2			
		Small-difference	Large-difference	Mixed-size-difference	Nonuniform-difference
		33%	33%	33%	33%
Factor Loading					
λ_{x1}	0.7	0.7	0.7	0.7	0.7
λ_{x2}	0.7	**0.5**	**0.3**	**0.4**	**0.4**
λ_{x3}	0.7	0.7	0.7	0.7	0.7
λ_{x4}	0.7	**0.5**	**0.3**	**0.2**	**1.0**
λ_{x5}	0.7	0.7	0.7	0.7	0.7
λ_{x6}	0.7	0.7	0.7	0.7	0.7
Intercept					
τ_{x1}	0.1	0.1	0.1	0.1	0.1
τ_{x2}	0.1	**0.4**	**0.7**	**0.5**	**0.5**
τ_{x3}	0.1	0.1	0.1	0.1	0.1
τ_{x4}	0.1	**0.4**	**0.7**	**0.7**	**-0.3**
τ_{x5}	0.1	0.1	0.1	0.1	0.1
τ_{x6}	0.1	0.1	0.1	0.1	0.1
Unique variances					
ε_{x1} - ε_{x6}	0.3		0.3		
Factor variance					
φ	1		1.3		
Factor Mean					
κ	0		0.5		

Note. All conditions presented in the table have four levels of sample size (N =100, 250, 500, and 1000 per group). ; In Study I, λ_{x2}, λ_{x4}, τ_{x1}, and τ_{x3} were noninvariant when the location of noninvariance existed in both factor loadings and intercepts of different items; In Study II, we also created baseline conditions in which every set of factor loadings, intercepts, and unique variances were invariant across groups.

Sample Size

We simulated four sample size conditions: N = 100, 250, 500, and 1000 per group. This was done to include a wide range of sample sizes in real settings. Only balanced sample sizes between the two groups were considered.

In sum, we manipulated four locations of noninvariance, four size/patterns of noninvariance, and four sample sizes. Thus, the resulting number of simulated conditions was 4 * 4 * 4 = 64 in Study I.

Study II

In Study II, we only tested two partial factorial invariance conditions among the four levels of locations of noninvariance in Study I: (1) noninvariance only in factor loadings ($\lambda_{x2}^{g1} \neq \lambda_{x2}^{g2}$; $\lambda_{x4}^{g1} \neq \lambda_{x4}^{g2}$) and (2) noninvariance only in intercepts ($\tau_{x2}^{g1} \neq \tau_{x2}^{g2}$; $\tau_{x4}^{g1} \neq \tau_{x4}^{g2}$). We also simulated the baseline condition in which all parameters except for the structural parameters (i.e., factor variance and factor means) are equal across groups in order to evaluate basal Type I error rates. All the other conditions (i.e., size/pattern of noninvariance, sample sizes, and number of replications) were same with Study I. Therefore, the total number of conditions simulated in Study II was 40 (2 locations of noninvariance * 4 size/pattern of noninvariance * 4 sample sizes + 2 baseline condition * 4 sample sizes).

Data Analysis Procedure

Study I

The value of a modification index indicates the amount of decrement in the chi-square statistics when the indicated constrained parameter is released with all other

parameters fixed. Thus, we hypothesized that the smallest modification index (Min-Mod) would represent the smallest difference within the set of parameters to be constrained invariant. For example, we fixed all sets of corresponding factor loadings to be equal across groups to test metric invariance. We considered that the factor-loading set having the Min-Mod had the smallest difference across groups among the all sets of equally constrained factor loadings. We identified the models using the variation of the reference variable identification method (VRV-IM), which does not require any invariant assumption under a full factorial invariance model. We score the result that the Min-Mod performed accurately when the factor loading or intercept indicated by the Min-Mod belonged to the group of truly invariant parameters.

There are two possible invariance models which can serve the process of searching an RV. For example, to search for an invariant factor loading set, we can refer to the Min-Mod under a metric invariance model in which every set of factor loadings is fixed to be invariant, or a scalar invariance model in which every set of factor loadings and intercepts is constrained to be equal across groups. To select an invariant set of intercepts, we can refer to the Min-Mod under a scalar invariance model or an intercept invariance model in which only every set of intercepts is equally constrained over groups.

Study II

Presented here first is the detailed analytic procedure of each of the three methods: the forward approach using the bias-corrected bootstrapping confidence intervals (BCBS-CIs), the backward approach using the largest modification index (Max-Mod), and the factor-ratio test. The process of the factor-ratio test was also

simplified by the BCBS-CIs. Second, we described how to evaluate the performance of each approach and defined perfect recovery rate, model-level Type I error rate, model-level Type II error rate, item-level power, and item-level Type I error rate. Third, we explained how to analyze the effects of chosen approach and manipulated design factors on the perfect recovery rate, model-level Type I error rate, and model-level Type II error rate.

Analytic Procedure

First, for the forward approach using BCBS-CIs the baseline model was either a configural invariance model in testing partial metric invariance or a metric invariance model in testing partial scalar invariance. The parameter with the smallest modification index served as an RV for this method from the Study I. We also retrieved the BCBS-CIs of the differences of the tested parameter pairs and referred to both 95% and 99% confidence intervals. If the confidence interval did not include "zero," the parameter was categorized as noninvariant. The M*Plus* syntaxes for the proposed forward method are provided in Appendix C and D for testing factor loadings and intercepts, respectively.

Second, the baseline model for the backward approach using the Max-Mod was either a full metric invariance model (i.e., all factor loadings are equally constrained) or full scalar invariance model (i.e., all factor loadings and all intercepts are equally constrained). Both models were identified using the variation of the reference variable identification method (VRV-IM). The VRV-IM also requires one set of equally constrained parameters (i.e., reference variable) without a specific value for the identification of a configural invariance model. We did not need to choose, though, a

specific variable for the identification purpose because the baseline model was a full metric or scalar invariance model. As in Yoon and Millsap (2007), we sequentially relaxed the equality constraint of the parameters with the largest modification index (Max-Mod) until there were left no more significant modification indices for either factor loadings or intercepts (Mod = 3.841 at α=0.05; Yoon & Millsap, 2007). To make more comprehensive comparisons, we also investigated the performance of the backward approach at the alpha level of 0.01 (Mod=6.635) which has not yet been examined.

Third, to conduct the factor-ratio test we used the bias-corrected bootstrapping confidence intervals (BCBS-CIs) as Cheung and Lau (2011) suggested. The baseline model for this analysis was a configural invariance model, and the tested parameters were the cross-group difference for the ratio of a reference variable (RV) and an argument. The equality of each set of corresponding parameters was tested using a "MODEL CONSTRAINT." We retrieved a bias-corrected bootstrapping confidence interval by asking "CINTERVAL(BCBOOTSTRAP)" under the "OUTPUT" command. The M*Plus* syntaxes used in testing factor loadings and intercepts are available in Appendixes A and B, respectively. As with the forward approach, we considered both 95% and 99% confidential intervals. If the confidence interval of a pair of an RV and an argument did not include zero, the pair was considered to be noninvariant.

Outcome Variables

The performances of the forward approach using BCBS-CIs, the backward approach using the Max-Mod, and the factor-ratio test were primarily evaluated in terms

of perfect recovery rates, model-level Type I error rates, and model-level Type II error rates. Additionally, we also looked at the item-level power and Type I error rate.

First, the most interesting outcome of the study ought to be how well each method perfectly recovered the true partial factorial invariance (PFI) model because perfect recovery rate can be deemed to be the most rigorous forms of power without any errors. To achieve a perfect recovery, all noninvariant items should be detected as such while no invariant items get so detected. The perfect recovery rate was calculated within the 1000 replications of each condition, by the size/pattern of noninvariance, and by the sample size with respect to the location of noninvariance.

Second, we also examined model-level Type I and II errors to see the sources of inaccuracy in specifying true PFI models. In this study, model-level Type I error was defined as detecting any invariant parameter as noninvariant in the finally recovered PFI model. Similarly, model-level Type II error was defined as any failure to detect noninvariant parameters as such in the finally specified PFI model. Both model-level Type I and II error rates were calculated within the 1000 replication of each PFI condition by the size/pattern of noninvariance, by the sample size.

Finally, we also examined item-level power and Type I error rate which are often reported in simulation studies. Item-level power is defined as the proportion of the detected noninvariant items (pairs) over the number of tested noninvariant items (pairs) *1000 replications. Similarly, item-level Type I error rates were calculated by dividing the total occurrence of Type I errors by the number of tested invariant items (pairs)* 1000 replications.

Effects of Design Factors

We conducted the analysis of variance (ANOVA) to determine the effects of the methods (e.g., forward approach, backward approach, and factor-ratio test), the size/pattern of noninvariance (e.g., small-, large-, mixed-size-, and nonuniform-difference), and the sample size (e.g., N = 100, 250, 500, and 1000) on the perfect recovery rate, model-level Type I error rate, and model-level Type II error rate with respect to the location of noninvariance (factor loadings or intercepts). In the ANOVA model, we also added all two- and three- way interaction effects among the main factors. The variance in each outcome variable (i.e., perfect recovery rate, Type I error rate, and Type II error rate) was partitioned with accordance to the three main factors and their interactions. The chosen effect size was eta-squared ($\eta 2$) which is the proportion of the variance explained by each factor in the total variance in the perfect recovery rates, model-level Type I error rates, or model-level Type II error rates of each approach.

CHAPTER IV

RESULTS

Study I. Searching for a Truly Invariant Reference Variable Using the Smallest Modification Index

The purpose of this study was to evaluate the performance of the smallest modification index (Min-Mod) within the limited set of equally constrained parameters (e.g., factor loadings or intercepts) in the model. As a value of a modification index indicates the amount of decrease in the chi-square statistic (with one degrees of freedom), the Min-Mod within the set of invariant constraints was assumed to indicate the parameter having the smallest difference between groups. Thus, we selected the factor loading set with the Min-Mod as a reference variable (RV) in either a fully constrained metric invariance model or scalar invariance model while we selected the intercept with the Min-Mod as an RV in either a fully constrained scalar invariance model or a fully constrained intercept invariance model. When the chosen factor loading or intercept belongs to the truly invariant parameter group, it was coded as "accurate." The mean accuracy (%) was reported across 1000 replications within each partial factorial invariance (PFI) condition by the sample size. We simulated four possible PFI scenarios: (1) noninvariance only in factor loadings (partial metric invariance: PMI), (2) noninvariance only in intercepts (partial scalar invariance: PSI), (3) noninvariance in both factor loadings and intercepts of the same items (partial metric & scalar invariance of the same items: PMSI-S), and (4) noninvariance in both factor loadings and intercepts

of different items (partial metric & scalar invariance of different items: PMSI-D), and thus, the results are presented by each PFI scenario. In each scenario, 16 different PFI conditions were simulated depending on the size/pattern of noninvariance (small-, large-, mixed-size-, and nonuniform-difference) and the sample sizes (N = 100, 250, 500, and 1000).

Partial Metric Invariance Scenario

In the partial metric invariance scenario, two factor loadings were manipulated to be noninvariant across groups. Here, our focus was to identify a truly invariant factor loading, and we investigated the modification index of factor loadings under the two fully constrained invariant models in which every set of factor loadings was equally constrained: (1) metric invariance model and (2) scalar invariance model. As presented in Table 2, the accuracy rates of identifying a truly invariant factor loading were almost perfect across all PMI conditions except in the small-difference with N =100, regardless of the type of fully constrained model (i.e., metric invariance model or scalar invariance model). Although the accuracy rates were not perfect for all conditions with N = 100, the error rates were negligible or very low. The highest error rates were found when a small difference was combined with small sample size, however, the error rates were still very low under both metric invariance model and scalar invariance model ($\leq$ 7.3% across 1000 replications).

Table 2. Accuracy of Identifying a Truly Invariant Factor Loading under PMI Scenario

Condition	N	Metric Invariance Model	Scalar Invariance Model
Small-difference	100	92.7	93.0
	250	99.8	99.5
	500	100.0	100.0
	1000	100.0	100.0
Large-difference	100	100.0	100.0
	250	100.0	100.0
	500	100.0	100.0
	1000	100.0	100.0
Mixed-size-difference	100	99.4	99.4
	250	100.0	100.0
	500	100.0	100.0
	1000	100.0	100.0
Nonuniform-difference	100	99.9	99.9
	250	100.0	100.0
	500	100.0	100.0
	1000	100.0	100.0

Note. N = sample size per group; The baseline accuracy rate is 66.7% given four invariant factor loadings among the six variables.

Partial Scalar Invariance Scenario

In the partial scalar invariance scenario, two intercepts were manipulated to be noninvariant across groups. In order to identify a truly invariant intercept we examined the modification index of intercepts under the two full invariance models in which every set of intercepts was fixed to be equal across groups: (1) intercept invariance model and (2) scalar invariance model. In Table 3, the accuracy rates of identifying a truly invariant intercepts using the Min-Mod were presented according to the invariance model. Similarly in the partial metric invariance scenario, the Min-Mod identified a truly invariant intercept with very high accuracy rates. Under both invariance models, the accuracy rates in the small-difference condition with N =100 and in the mixed-size

difference conditions with N = 100, 250, and 500 were lower than the other conditions. The accuracy rates of those conditions were higher under a scalar invariance model than those under an intercept invariance model. For instance, the lowest accuracy rate was found in the mixed-size-difference conditions with N =100 under an intercept invariance model (92.5%), but the accuracy rate increased under a scalar invariance model (96.7%). It seemed that examining the Min-Mod of intercepts under a scalar invariance model yielded to higher accuracy rates than under an intercept invariance model when noninvariance existed only in intercepts. But, the maximum difference of the accuracy rates between the models was only 4.2%.

Table 3. Accuracy of Identifying a Truly Invariant Intercept under PSI Scenario

Condition	N	Intercept Invariance Model	Scalar Invariance Model
Small-difference	100	96.5	97.1
	250	99.9	99.8
	500	100.0	100.0
	1000	100.0	100.0
Large-difference	100	99.7	100.0
	250	100.0	100.0
	500	100.0	100.0
	1000	100.0	100.0
Mixed-size-difference	100	92.5	96.7
	250	96.2	99.3
	500	97.5	99.8
	1000	99.5	99.9
Nonuniform-difference	100	100.0	100.0
	250	100.0	100.0
	500	100.0	100.0
	1000	100.0	100.0

Note. N = sample size per group; the baseline accuracy rate is 66.7% given four invariant factor loadings among the six variables.

Partial Scalar and Metric Invariance of the Same Items Scenario

In the partial scalar and metric invariance of the same variable scenario (PMSI-S), two factor loadings and two intercepts of the same items were manipulated to be noninvariant across groups. In selecting an invariant factor loading, we examined the modification indices of factor loadings under either a full metric invariance model or a full scalar invariance model. The factor loading with the Min-Mod was selected as a reference variable (RV). To search a truly invariant intercept, we looked at the modification indices of intercepts under either a full intercept invariance model or a full scalar invariance model, and selected the intercept with the Min-Mod as an RV.

Accuracy of Identifying a Truly Invariant Factor Loading

Under both full metric and full scalar invariance models, the Min-Mod worked perfectly well to identify a truly invariant factor loading except for some conditions with small-difference or small sample size (see Table 4). For example, the accuracy rates were lower in all conditions with N = 100 and the small-difference condition with N =250. In those conditions, the Min-Mod had higher accuracy rates under a metric invariance model than under a scalar invariance model with the maximum difference of 10.6% in the small-difference condition with N =100.

Accuracy of Identifying a Truly Invariant Intercept

The Min-Mod performed very accurately across all conditions except for the small-difference conditions with N =100 regardless of the invariance models in which the Min-Mod of intercepts were examined. Even in the small-difference conditions with N =100, the accuracy rates were 98.7 % and 97.0%, respectively, under a full intercept

invariance model and under a full scalar invariance model. Although the accuracy rates in that condition were slightly higher under the intercept invariance model than the scalar invariance model, the difference was negligible (1.7%).

Table 4. Accuracy of Identifying a Truly Invariant Factor Loading or Intercept under PMSI-S Scenario

		Factor Loading		Intercept	
Condition	N	Metric Invariance Model	Scalar Invariance Model	Intercept Invariance Model	Scalar Invariance Model
Small-difference	100	92.7	81.1	98.7	97.0
	250	99.8	92.2	100.0	99.7
	500	100.0	98.8	100.0	100.0
	1000	100.0	99.9	100.0	100.0
Large-difference	100	100.0	94.4	100.0	100.0
	250	100.0	99.7	100.0	100.0
	500	100.0	100.0	100.0	100.0
	1000	100.0	100.0	100.0	100.0
Mixed-size-difference	100	99.4	94.3	99.9	99.8
	250	100.0	99.0	100.0	100.0
	500	100.0	100.0	100.0	100.0
	1000	100.0	100.0	100.0	100.0
Nonuniform-difference	100	99.8	98.3	100.0	100.0
	250	100.0	100.0	100.0	100.0
	500	100.0	100.0	100.0	100.0
	1000	100.0	100.0	100.0	100.0

Note. N = sample size per group; The baseline accuracy rate is 66.7% given four invariant factor loadings among the six variables.

Overall, the Min-Mod almost perfectly identify either a truly invariance factor loading or a truly invariant intercept except for some conditions with small sample size and small difference in the PMSI-S scenario. It seems that a truly invariant factor

loading was better identified by the Min-Mod under a metric invariance model than under a scalar invariance model in the conditions with lower accuracy rates. In selecting a truly invariant intercept, the accuracy rates were almost perfect no matter which invariance model was used except for the small-difference with N =100. Although the accuracy rate was slightly higher in the intercept invariance model than in the scalar invariance model, the pattern was not clear as in searching for an invariant factor loading.

Partial Scalar and Metric Invariance of the Different Items Scenario

In the partial scalar and metric invariance of different variables scenario (PMSI-D), two factor loadings and two intercepts were manipulated to be noninvariant in different variables across groups. That is, the factor loadings of the second and fourth variable were set to be noninvariant while the intercepts of the first and third variables were chosen to be noninvariant. Similar in the PMSI-S scenario, the Min-Mod of factor loadings were observed under both metric invariance model and scalar invariance model. In selecting a truly invariant intercept, we referred to the Min-Mod under either a full intercept invariance model or a full scalar invariance model.

Accuracy of Identifying a Truly Invariant Factor Loading

The Min-Mod correctly identified a truly invariant factor loading except in the small-difference condition with N = 100 regardless of the invariance models (see Table 5). In the small-difference condition with N =100, the Min-Mod worked better under a scalar invariance model than under a metric invariance model.

Accuracy of Identifying a Truly Invariant Intercept

In the PMSI-D scenario, we could observe that the pattern of the accuracy rates was not inconsistent depending on the size/ pattern of noninvariance and the sample size under different invariance model. In the small-difference conditions, the accuracy rates were lower with N =100 than with the other sample sizes. In that condition, the accuracy rates were higher under a scalar invariance model than under intercept invariance model. In the large-difference condition, the accuracy rates were slightly lower with N = 100 than the other sample size conditions. Similarly in the small-difference with N =100 condition, the Min-Mod performed better under a scalar invariance model than an intercept invariance model. In the mixed-size-difference condition, the accuracy rate was higher under a scalar invariance model than the intercept invariance model. However, in the remaining conditions with large sample sizes, higher accuracy rates observed under an intercept invariance model than a scalar invariance model. In the nonuniform-difference condition, the Min-Mod achieved perfect accuracies across all sample sizes.

In the PMSI-D scenario, the pattern of accuracy rates was not very clear although the Min-Mod maintained very high accuracy rates across all conditions with some exceptions. To search for a truly invariant factor loading, the Min-Mod achieved almost perfect accuracy rates across all conditions except for one condition (i.e., small-difference condition with N =100) under both metric and scalar invariance models. However, the performance of the Min-Mod was not consistent in selecting a truly invariant intercept. If we can consider that larger sample sizes are more important than N

= 100, choosing an intercept under an intercept invariance model might be a better option.

Table 5. Accuracy of Identifying a Truly Invariant Factor Loading and Intercept under PMSI-D Scenario

		Factor Loading		Intercept	
Condition	N	Metric Invariance Model	Scalar Invariance Model	Intercept Invariance Model	Scalar Invariance Model
Small-difference	100	92.7	96.3	91.7	95.4
	250	99.8	100.0	99.0	99.8
	500	100.0	100.0	99.8	100.0
	1000	100.0	100.0	100.0	100.0
Large-difference	100	100.0	100.0	97.1	98.9
	250	100.0	100.0	100.0	99.9
	500	100.0	100.0	100.0	100.0
	1000	100.0	100.0	100.0	100.0
Mixed-size-difference	100	99.4	100.0	83.5	88.9
	250	100.0	100.0	92.3	88.2
	500	100.0	100.0	97.8	88.4
	1000	100.0	100.0	99.9	92.4
Nonuniform-difference	100	99.8	99.8	100.0	100.0
	250	100.0	100.0	100.0	100.0
	500	100.0	100.0	100.0	100.0
	1000	100.0	100.0	100.0	100.0

Note. N = sample size per group; The baseline accuracy rate is 66.7% given four invariant factor loadings among the six variables.

Study II. Specifying a True Partial Invariance Model

In Study II, we compared the performances of the forward approach using the BCBS-CIs (BCBS-CIs), the backward approach using the largest modification index (Max-Mod), and the factor-ratio test (FR) in specifying either a true partial metric invariance (PMI) or a partial scalar invariance (PSI) model. The performance of each method was evaluated in two ways. First, we examined the basal Type I error rates of each approach under the conditions without any type of measurement noninvariance to determine the adequacy of each approach. Then, under either PMI or PSI conditions, we examined the perfect recovery rates, model-level Type I error rates, and model-level Type II error rates across 1000 replications in the finally specified partial factorial invariance model. In Study II, perfect recovery rate is defined as the percentage of specifying the original partial factorial invariance (PFI) model across the 1000 replications within each PFI condition by sample size. Model-level Type I error was the proportion of any occurrence of Type I error in the finally specified PFI model across 1000 replications. Similarly, model-level Type II error was also calculated by dividing the number of PFI models with any Type II error by 1000. Additionally, we examined the item-level power and the Type I error rates. The item-level power was calculated by dividing the number of items (or pairs) which were to be detected as noninvariant by the number of truly noninvariant items (or pairs) tested * 1000. The item-level Type I error rates were attained by dividing the number of items which were detected as noninvariant by the number of truly invariant items/pairs * 1000. Although the main interest of the study was the perfect recovery rates, the model-level Type I and Type II error rates, and

item-level power and Type II error rate were also important to informing about the source of errors in the perfect recovery rates.

Simulation Baseline Check

With respect to the location of the measurement parameter (i.e., factor loadings and intercept), the basal Type I error rate was examined for each approach with the data condition in which all factor loadings and intercepts were invariant over groups. According to Bradley's (1978) formula ($\alpha \pm \frac{\alpha}{2}$), an acceptable Type I error rate is between 0.025 and 0.75 or between 0.005 and 0.015 when the chosen α is 0.05 or 0.01, respectively. As the width of BCBS-CIs and cut-off values of the modification index was set for each tested item (for the forward and backward approaches) or pair (for the factor-ratio test), we needed to consider the total number of tested items and pairs in each approach to calculate the model-level nominal Type I error rates in both full factorial invariance conditions and partial factorial invariance conditions.

Forward Approach

Using the forward approach, we conducted a total of five tests given six variables since one of the variables served as a reference variable. When all measurement parameters are invariant, the nominal model-level Type I error rate equal to five times the chosen item-level significance or confidence level. Therefore, the nominal model-level Type I error rate is 0.05 and 0.25 with 99% and 95% confidence levels, respectively. Based on Bradley's generous criteria, an acceptable model-level Type I error rate is between 0.025 and 0.075 (99% confidence level) or between 0.125 and

0.375 (95% confidence level) if the given data do not have noninvariance in the tested parameters.

As shown in Table 6, the model-level Type I error rates of the forward approach were far below the nominal Type I error rates in testing factor loadings. However, in testing intercepts, we observed much higher model-level Type I error rates, and the Type I error rates were inflated as the sample size grew. Even though the model-level Type I errors for intercepts were within Bradley's criteria with a sample size equal to or less than 500, it exceeded the criteria given a sample size of 1000 with both 99% and 95% BCBS-CIs.

Backward Approach

Using the backward approach, it is not very clear when the model specification stops. However, the model-level α level can be calculated by multiplying the item-level significance level by six given six variables in the model. Thus, the nominal model-level Type I error rate is 0.06 at $\alpha = 0.01$ while the nominal model-level Type I error rate is 0.30 at $\alpha = 0.05$. Based on Bradley's formula, an acceptable model-level Type I error rate can be found between 0.03 and 0.09 ($\alpha = 0.01$) or between 0.15 and 0.45 ($\alpha = 0.05$).

Table 6 presents the basal Type I error rates of the backward approach in either testing factor loadings or intercepts when all variables are invariant. Although our expectation was that the backward approach would have higher Type I error rates as sample size grew, it showed homogeneous level of Type I error rates across the sample sizes regardless of the tested parameters (factor loadings or intercepts). Across all conditions, the basal Type I error rates were beyond the nominal model-level Type I

error rate. In testing factor loadings, it had higher model-level basal Type I error rates, however, in testing intercepts, it showed lower error rates than the forward approach.

Factor-ratio Test

In testing factorial invariance using the factor-ratio test, the number of tests to be conducted is 15 given six variables. Thus, the nominal model-level Type I error rate is 15* the chosen α level (0.15 with 99% confidence level; 0.45 with 95% confidence level). If we employ Bradley's criteria, an acceptable model-level Type I error rate lies between 0.075 and 0.225 (99% confidence level) or between 0.225 and 0.675 (95% confidence level).

Table 6 presents the basal model-level Type I error rates of the factor-ratio test in testing either factor loadings or intercepts given an invariant model condition. In testing factor loadings, the factor-ratio test maintained basal Type I error rates around the nominal level. Similar with the forward approach, its Type I error rates were inflated in testing intercepts, and the inflation increased given a larger sample size. However, the basal model-level Type I error rates were within Bradley's criteria.

In sum, only the backward approach showed acceptable Type I error rates in testing both factor loadings and intercepts. Although the forward approach maintained the lowest basal Type I error rates in testing factor loadings, unexpectedly, it had much higher error rates in testing intercepts. The factor-ratio test presented similar pattern with the forward approach since both approaches used BCBS-CIs. However, the model-level Type I error rates of the factor-ratio test were too high under both 99% and 95% confidence levels since the number of tests to be conducted increases given a larger

number of variables. To avoid too high model-level Type I error rates, the factor-ratio test may need critical value adjustment which is allowed in LISREL not in MPlus.

Table 6. Type I Error Rates in the Baseline Conditions

	N	Forward		Backward		Factor-ratio Test	
		99%	95%	99%	95%	99%	95%
Factor Loading	100	0.001	0.032	0.059	0.244	0.117	0.344
	250	0.003	0.029	0.062	0.249	0.119	0.390
	500	0.001	0.023	0.050	0.241	0.097	0.373
	1000	0.002	0.033	0.060	0.256	0.115	0.406
Intercept	100	0.078	0.246	0.049	0.240	0.130	0.386
	250	0.082	0.266	0.058	0.289	0.155	0.437
	500	0.138	0.321	0.072	0.275	0.202	0.486
	1000	0.159	0.406	0.061	0.260	0.269	0.580

Partial Metric Invariance Conditions

Forward Approach

In all partial metric invariance conditions (PMI) conditions, we had two noninvariant factor loadings among the six variables under one factor. Table 7 shows the performance of the forward approach in terms of perfect recovery rates, model-level Type I error rates, model-level Type II error rates, item-level power, and item-level Type I error rates by the width of bias-corrected bootstrapping confidence intervals (BCBS-CIs) in specifying the true PMI model.

The forward approach maintained very high perfect recovery rates across the except for all PMI conditions with the small sample size (N =100) and the small-difference condition with N=250. Particularly, in those conditions, 95% BCBS-CIs

showed higher perfect recovery rates than those with 99% BCBS-CIs. For the remaining conditions, 99% BCBS-CIs outperformed 95% BCBS-CIs.

The nominal model-level Type I error rate is .03 and .15 given 99% and 95% BCBS-CIs, respectively, because there are three invariant factor loadings to be tested under the forward approach. The forward approach showed very low model-level and item-level Type I error rates below the nominal level regardless of the width of the BCBS-CIs. Model-level Type II error rates were found to be higher as the sample size was smaller using 99% BCBS-CIs. The same pattern was presented using 95% BCBS-CIs although model-level Type II error rates slightly lower than using 99% BCBS-CIs. Particularly, the highest model-level Type I error rates were observed in the small-difference condition with N =100 under both CIs. For both BCBS-CIs, the item-level power remained perfect (1.00) except for all small sample size conditions (N=100) and small-difference condition with N = 250 and 500. Item-level power also showed the consistent pattern with the model-level Type II error rates. In such conditions, 95% BCBS-CIs could detect more noninvariant items than 99% BCBS-CIs.

In sum, given that the forward approach exhibited very low Type I error rates across all conditions, the main source of the low perfect recovery rates appeared to be related to the high Type II error rates (low power) in the conditions with small sample size or small sample size combined with small-difference of noninvariance. However, we can expect that we can successfully differentiate noninvariant variables from invariant variables using the forward approach given a substantially large sample size ($N \geq 250$) using 99% CIs.

Table 7. Performance of the Forward Approach in PMI Conditions

Condition	N	99% BCBS-CIs					95% BCBS-CIs				
		Perfect recovery	Type I Error	Type II Error	Power/ # of items	Type I error/ # of items	Perfect recovery	Type I error	Type II error	Power/ # of items	Type I error/ # of items
Small-difference	100	0.116	0.003	0.884	0.558	0.001	0.290	0.025	0.703	0.649	0.008
	250	0.495	0.009	0.501	0.750	0.003	0.722	0.055	0.243	0.879	0.018
	500	0.913	0.011	0.078	0.961	0.004	0.927	0.055	0.020	0.990	0.018
	1000	0.988	0.011	0.001	1.000	0.004	0.938	0.062	0.000	1.000	0.021
Large-difference	100	0.837	0.018	0.151	0.925	0.006	0.912	0.056	0.034	0.983	0.019
	250	0.986	0.014	0.000	1.000	0.005	0.927	0.073	0.000	1.000	0.024
	500	0.990	0.010	0.000	1.000	0.003	0.943	0.057	0.000	1.000	0.019
	1000	0.988	0.012	0.000	1.000	0.004	0.934	0.066	0.000	1.000	0.022
Mixed-size-difference	100	0.412	0.011	0.585	0.708	0.004	0.647	0.047	0.322	0.839	0.016
	250	0.947	0.016	0.038	0.981	0.005	0.925	0.069	0.007	0.997	0.023
	500	0.986	0.014	0.000	1.000	0.005	0.944	0.056	0.000	1.000	0.019
	1000	0.986	0.014	0.000	1.000	0.005	0.936	0.064	0.000	1.000	0.021
Nonuniform-difference	100	0.766	0.000	0.234	0.883	0.000	0.932	0.006	0.062	0.969	0.002
	250	1.000	0.000	0.000	1.000	0.000	0.991	0.009	0.000	1.000	0.003
	500	0.999	0.001	0.000	1.000	0.000	0.981	0.019	0.000	1.000	0.006
	1000	0.992	0.008	0.000	1.000	0.003	0.960	0.040	0.000	1.000	0.013

Note. Perfect recovery rates and item-level power greater than or equal to 0.90 were underlined.

Backward Approach Using the Max-Mod

The performance of the backward approach in which the factor loading indicated by the largest modification index (Max-Mod) was sequentially relaxed until no more significant modification index (Mod) remained in the model. To decide the size of the significant Mod, we employed two criteria: Mod = 3.841 $\alpha = .05$ and 6.635 at $\alpha = .01$. The values of modification index are corresponding to the χ^2 statistic value with $df = 1$ at the chose alpha levels. Depending on the significance values for the modification index, the backward approach showed very different performances. That is, it yielded more promising results for $\alpha = 0.01$ than $\alpha = 0.05$. Table 8 shows the performance of the backward approach in specify a PMI model.

In correctly recovering the original PMI model with two noninvariant factor loading, the backward approach showed very high perfect recovery rates (greater than .95) using the modification index value (6.635) at $\alpha = .01$ except for the conditions with small difference and small sample size. However, it could not achieve perfect recovery rates above .85 using the cut-off value at $\alpha = .05$ (Mod = 3.841). Although it showed higher perfect recovery rates at $\alpha = .05$ than at $\alpha = .01$ in the small-difference conditions with N =100 and 250 and in the nonuniform-difference condition with N=100, none of them was very high (below .80).

Given four invariant factor loading among the six variables, the model-level nominal Type I error rate is .04 at $\alpha = .01$ while it is .20 at .05. Based on Bradley's formula, acceptable Type I error rates are between .02 and .06 ($\alpha = .01$) or between .10 and .30 ($\alpha = .05$). Although Type I error rates were below Bradley's criteria across all

PMI conditions, Type I error rates at α = .05 were too high while diminishing the perfect recovery rates. However, the Type I error rates decreased substantially when we employed the significant value at α = .01 (Mod = 6.635). Similar to the forward approach, the backward approach showed high Type II error rates all conditions with N =100 and the small-difference condition with N = 250 at α = .01. In the same conditions, we observed smaller Type II error rates given α = .05. However, the model-level Type II error rates were very low or zero in the remaining conditions regardless of the significance values. Item-level power also exhibited a consistent pattern; lower power with lower sample size and smaller difference of noninvariance.

In sum, the backward approach maintained very low Type I error rates when we applied more conservative α level (=.01), and thus, it showed very high perfect recovery rates. In the previous study on the backward approach used only the modification index value at α = .05, and it reported high Type I error rates and low perfect recovery rates except for some ideal conditions with large difference and large sample sizes (Yoon & Millsap, 2007). However, the results of the current study indicate that adjusting the critical value with more conservative alpha level improved the perfect recovery rates in the simulated PMI condition by critically reducing the model-level Type I error rates with maintaining similar high powers with the powers at α = .05.

Table 8. Performance of the Backward Approach in PMI Conditions

Condition	N	Mod = 6.635 (α = 0.01)					Mod = 3.841 (α = 0.05)				
		Perfect recovery	Type I error	Type II error	Power/ # of items	Type I error/ # of items	Perfect recovery	Type I error	Type II error	Power/ # of items	Type I error/ # of items
Small-difference	100	0.128	0.092	0.856	0.293	0.026	0.324	0.236	0.533	0.532	0.073
	250	0.714	0.074	0.236	0.833	0.021	0.754	0.189	0.066	0.931	0.056
	500	0.965	0.029	0.006	0.995	0.008	0.850	0.150	0.000	0.999	0.039
	1000	0.962	0.038	0.000	1.000	0.010	0.825	0.175	0.000	1.000	0.047
Large-difference	100	0.921	0.042	0.038	0.975	0.011	0.832	0.164	0.004	0.996	0.044
	250	0.965	0.035	0.000	1.000	0.009	0.822	0.178	0.000	1.000	0.046
	500	0.969	0.031	0.000	1.000	0.008	0.843	0.157	0.000	1.000	0.040
	1000	0.957	0.043	0.000	1.000	0.011	0.810	0.190	0.000	1.000	0.051
Mixed-size-difference	100	0.779	0.047	0.174	0.905	0.012	0.777	0.173	0.050	0.965	0.046
	250	0.962	0.035	0.003	0.998	0.009	0.831	0.169	0.000	1.000	0.044
	500	0.973	0.027	0.000	1.000	0.007	0.847	0.153	0.000	1.000	0.039
	1000	0.959	0.041	0.000	1.000	0.011	0.810	0.190	0.000	1.000	0.051
Nonuniform-difference	100	0.586	0.056	0.359	0.795	0.014	0.702	0.179	0.119	0.906	0.047
	250	0.959	0.036	0.005	0.997	0.009	0.832	0.168	0.000	0.999	0.044
	500	0.969	0.031	0.000	1.000	0.008	0.849	0.151	0.000	1.000	0.038
	1000	0.965	0.035	0.000	1.000	0.009	0.818	0.182	0.000	1.000	0.049

Note. Perfect recovery rates and item-level power greater than or equal to 0.90 were underlined.

Factor-ratio Test

When using the factor-ratio test, the ratios of 15 factor-loading pairs were tested using either 99%% or 95% bias-corrected bootstrapping confidence intervals (BCBS-CIs). Among the tested pairs, each of the nine noninvariant factor-loading pairs was expected to have a non-zero value in its BCBS-CIs while each of the six invariant pairs was expected to include zero in its BCBS-CIs. The performance of the factor-ratio test is shown in Table 9.

Interestingly, the perfect recovery rates of the factor-ratio test were extremely low in the small- and large-difference conditions across all sample sizes regardless of the width of BCBS-CIs. Among those conditions, the highest perfect recovery rate (PR = .047) was found in the large-difference condition with N = 250 using 95% BCBS-CIs. That is only 47 cases out of 1000 replications were perfectly recovered when the focal group's noninvariant factor loadings had the same magnitude of difference in the same direction. However, the factor-ratio test performed much better in the mixed-size-difference condition and nonuniform-difference condition, especially, referring to 99% BCBS-CIs. With 95% BCBS-CIs, the factor-ratio test could achieve the highest perfect recovery rates (PR = .785) in the nonuniform-difference condition with N = 1000.

For the factor-ratio test, the model-level nominal Type I error rate is nine times the chosen alpha level (.09 with 99% BCBS-CIs; .45 with 95% BCBS-CIs). Using 99% BCBS-CIs, the model-level Type I error rates were very low (below the nominal level) across all PMI conditions. However, using 95% BCBS-CIs, the model-level Type I error rates were too high although all of them were below the nominal level (.45). Type II

error rates were very high in the small- and large-difference conditions across all sample sizes. However, the item-level power rates were high with much lower Type II error rates in the same conditions. It seemed that in those conditions, the ratio of the two noninvariant factor loadings with the same difference in the same direction were not likely to be detected. Except for that pair, the other noninvariant pairs appeared to be detected well.

In sum, the factor-ratio test had strikingly low perfect recovery rates in the small- and large-difference conditions in which the two noninvariant factor loadings had the same difference in the same direction. In both levels of BCBS-CIs it could not detect that pair no matter which sample size was given. However, the factor-ratio test could successfully detect noninvariant pairs in the mixed-size-difference and nonuniform-difference with larger sample sizes (N = 500 and 100) using 99% BCBS-CIs. But, 95% BCBS-CIs were not effective to perfectly recover the original PMI model in the mixed-size-and nonuniform-difference conditions due to the high model-level Type I error rates. Overall, the results from the current study showed that the factor-ratio test might not be an optimal method to specify the original PMI model, particularly, when the size and direction of noninvariance was same for the noninvariant factor loadings.

Table 9. Performance of the Factor-ratio Test in PMI Conditions

		99% BCBS-CIs					95% BCBS-CIs				
Condition	N	Perfect recovery	Type I error	Type II error	Power/ # of items	Type I error/ # of items	Perfect recovery	Type I	Type II	Power/ # of items	Type I error/ # of items
Small-difference	100	0.000	0.052	1.000	0.231	0.011	0.000	0.198	1.000	0.427	0.048
	250	0.000	0.055	1.000	0.642	0.012	0.005	0.207	0.994	0.790	0.052
	500	0.004	0.041	0.996	0.868	0.008	0.032	0.208	0.962	0.890	0.048
	1000	0.007	0.060	0.992	0.890	0.014	0.050	0.223	0.940	0.896	0.058
Large-difference	100	0.006	0.055	0.994	0.831	0.012	0.026	0.203	0.970	0.880	0.050
	250	0.011	0.057	0.989	0.890	0.012	0.047	0.219	0.948	0.895	0.054
	500	0.010	0.045	0.988	0.890	0.009	0.041	0.213	0.946	0.895	0.048
	1000	0.011	0.067	0.988	0.890	0.015	0.041	0.225	0.948	0.895	0.058
Mixed-size-difference	100	0.092	0.057	0.905	0.766	0.011	0.338	0.201	0.606	0.897	0.049
	250	0.833	0.060	0.118	0.985	0.013	0.768	0.210	0.027	0.997	0.053
	500	0.959	0.040	0.001	1.000	0.009	0.784	0.216	0.000	1.000	0.050
	1000	0.937	0.063	0.000	1.000	0.015	0.773	0.227	0.000	1.000	0.060
Nonuniform-difference	100	0.041	0.061	0.959	0.604	0.012	0.202	0.200	0.776	0.793	0.048
	250	0.615	0.057	0.373	0.941	0.000	0.720	0.210	0.135	0.982	0.053
	500	0.946	0.042	0.015	0.998	0.008	0.783	0.215	0.004	1.000	0.050
	1000	0.938	0.062	0.000	1.000	0.014	0.785	0.215	0.000	1.000	0.058

Note. Perfect recovery rates and item-level power greater than or equal to 0.90 were underlined.

Comparisons of the Performances of the Three Approaches: PMI Conditions

Perfect Recovery Rates

In order to clearly compare the performance of the three methods, we located their perfect recovery rates on a graph based on both criteria: 99% and 95% BCBS-CIs for the forward approach and factor-ratio test or the critical value of modification index at $\alpha = .01$ and at $\alpha = .05$ (Mod = 3.841 and Mod = 6.635 with $df = 1$, respectively) for the backward approach. However, we will use the terms 99% confidence level (or $\alpha = .01$) and 95% confidence level (or $\alpha = .05$) in order to refer to the chosen confidence (or significance) level when comparing the three approaches on the same page.

Figure 1 shows the perfect recovery rates of the three approaches by the sample size. The factor-ratio test had strikingly low perfect recovery rates across all sample size conditions regardless of the chosen criteria. The factor-ratio test exhibited extremely low perfect recovery rates across all sample sizes regardless of the confidence (significance) levels. Under 99% confidence level (or $\alpha = .01$), both the forward and backward approaches showed very low perfect recovery rates with N =100. Although the perfect recovery rates got higher with N = 200, they were still low for both approaches. However, both approach presented very high perfect recovery rates with larger sample sizes ($N \geq 500$). The backward approach had higher perfect recovery rate than the forward approach with N = 500 whereas the forward approach showed better performance than the backward approach with N = 1000. Under 95% confidence level (or $\alpha = .05$), both forward and backward approach had very low perfect recovery rates (PR = .290 and .324, respectively) with N =100. With N = 250, the perfect recovery

rates increase for both approach but seemed not to be acceptable. With N ≥ 500, the forward approach showed higher perfect recovery rates than the backward approach. Overall, both forward and backward approach worked well given larger sample sizes (N ≥ 500) when using 99% confidence level (or α = .01). However, none of the three approaches could detect noninvariance successfully when the size of noninvariance was combined with the small sample size (N = 100) regardless of which criteria was used.

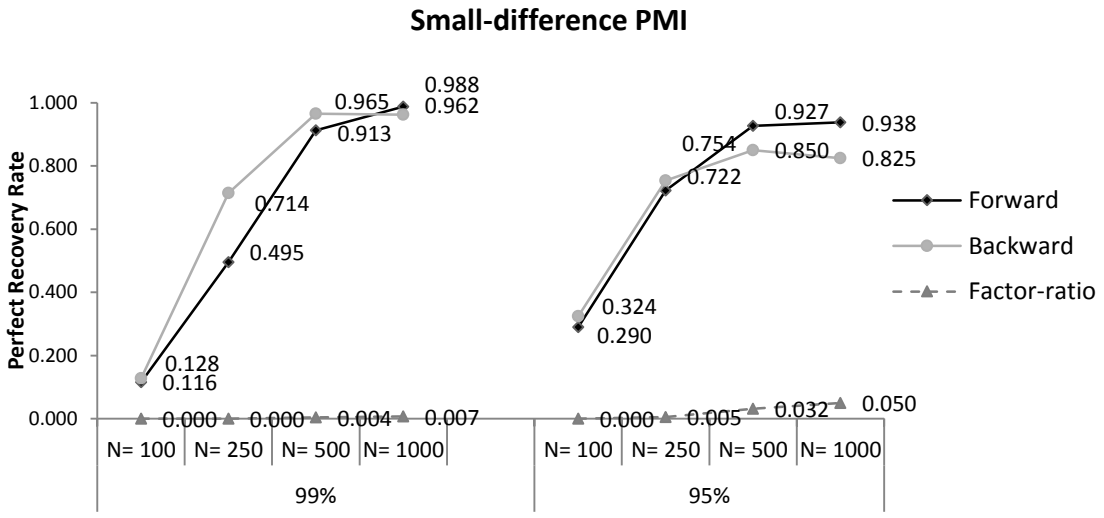

Figure 1. Perfect Recovery Rates in the Small-difference PMI Conditions

Figure 2 shows the perfect recovery rates of the three approaches in the large-difference PMI conditions. At 99% confidence level (or α = .01), the forward approach achieved very high perfect recovery rates (PR ≥ .986) with N ≥ 250 while having relatively lower perfect recovery rate (PR = .837) with N =100. The backward approach also performed very well across all sample sizes at α = .01. With larger samples (N ≥

250), the forward approach outperformed the backward approach whereas the backward approach had higher perfect recovery rate with N =100. Across all sample size conditions, the factor-ratio test exhibited extremely low perfect recovery rates as in the small-difference PMI conditions (PR ≤ .011). When we referred to the results under 95% confidence level (or α = .05), the factor-ratio still showed very poor performance (PR ≤ .047). With the same criterion, the forward approach maintained high perfect recovery rates (PR ≥ .912) across all sample sizes while the perfect recovery rates of the backward approach got to be approximately 10% (or more) lower than those under 95% confidence level (or α = .05). Overall, the forward and backward approach presented very promising perfect recovery rates with more conservative criterion when there are two noninvariant factor loadings with same magnitude of large difference in the same direction.

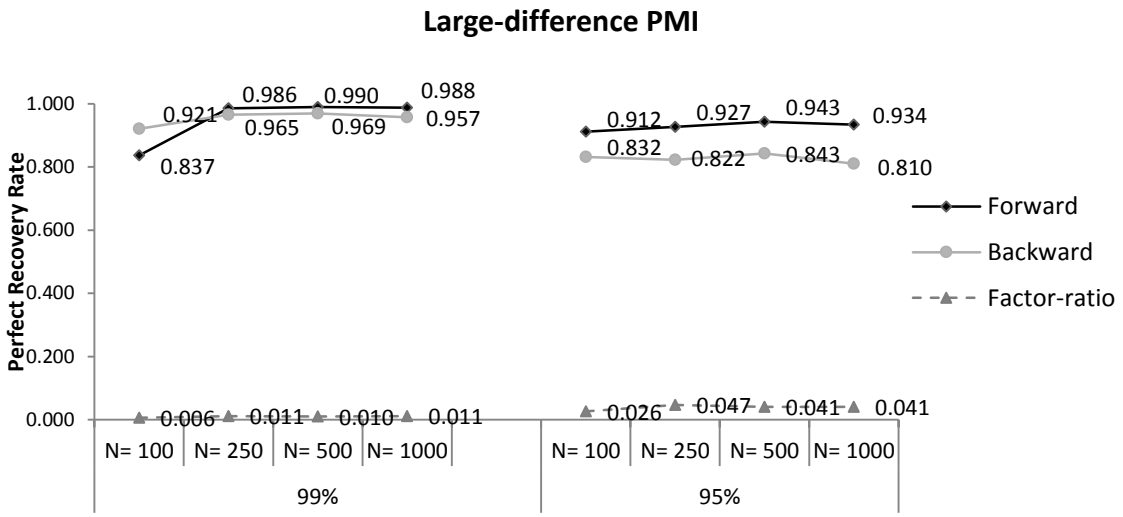

Figure 2. Perfect Recovery Rates in the Large Difference PMI Conditions

Figure 3 shows the perfect recovery rates of the three approaches in the mixed-size-difference PMI conditions. Under more conservative criterion (99% confidence level or $\alpha = .01$), the backward approach had the highest perfect recovery rates (PR = .779) with N = 100. With the same sample size, the forward approach showed much lower perfect recovery rate (PR = .412) than that of the backward approach while the factor-ratio test presented extremely low perfect recovery rate (PR = .092). With N = 250, the backward approach performed best (PR = .962) while the forward approach also showed a high perfect recovery rates (PR = .947). The perfect recovery rates of the factor-ratio test (PR = .833) also improved with N = 250, but it is still much lower than those of the other methods. With larger sample sizes ($N \geq 500$), all three approaches presented very high perfect recovery rates while the forward approach performed best. When we using the less conservative criterion (95% confidence level or $\alpha = .01$), the forward approach maintained the highest perfect recovery rates with $N \geq 250$.The backward approach outperformed the others with N = 100. The perfect recovery rates of the backward approach decreased more at $\alpha = .05$ than at $\alpha = .01$. Among the three approaches, the factor-ratio test performed worst under 95% confidence level as under 99% confidence level.

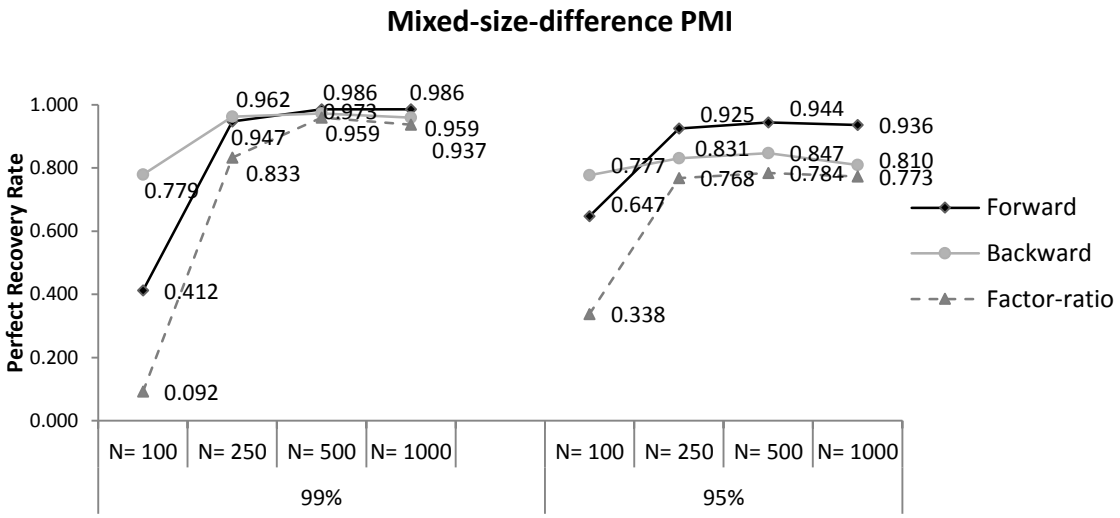

Figure 3. Perfect Recovery Rates in the Mixed-size-difference PMI Conditions

Figure 4 shows the perfect recovery rates of the three approaches in the nonuniform-difference PMI conditions. The forward approach outperformed the other two methods in all sample sizes at 99% confidence level. It showed very high perfect recovery rates (PR ≥ .992) with N ≥ 250 while it had much lower perfect recovery rates (PR = .766) with N =100. The backward approach also showed similar pattern in the perfect recovery rates with the forward approach. Yet, it had lower perfect recovery rates across all sample sizes. The factor-ratio test exhibited the lowest perfect recovery rates than the other two, but its perfect recovery rates got to be very high (≥ .938) with N ≥ 500. With the less conservative criterion (95% confidence level; $\alpha = .05$), the forward approach performed best with very high perfect recovery rates (Range: .932 – .991) across all sample sizes. The next well-performing approach was the backward approach, however, its perfect recovery rates were much lower (Range: .702 – .849) than those of

the forward approach. Here, the factor-ratio showed the lowest perfect recovery rates (Range: .202 – .785).

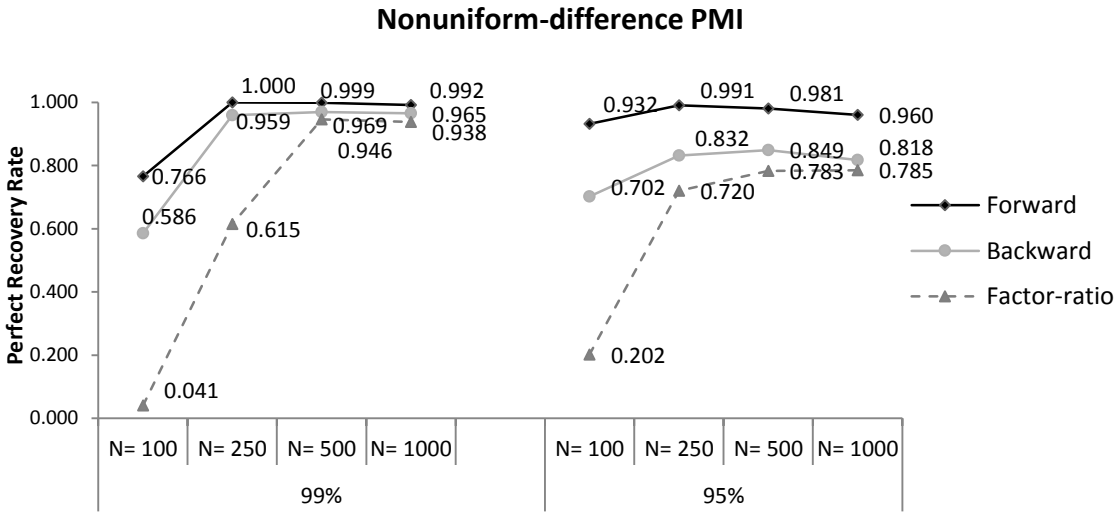

Figure 4. Perfect Recovery Rates in the Nonuniform-difference PMI Conditions

Model-level Type I Error Rate

The model-level Type I error rate in Study II was defined as erroneously detecting any truly invariant factor loading as noninvariant in the finally specified PMI model. To compare the model-level Type I error rates of the three approaches, we located them on the same graph based on the selected confidence (significance) levels, the size/patterns of noninvariance, and the sample sizes (see Figure 5). Under both alpha levels, the forward approach showed the lowest model-level Type I error rates across all PMI conditions. The backward approach had the next lowest model-level Type I error rates while the factor-ratio test presented the highest model-level Type I error rates in

most conditions. Generally, all approaches had higher Type I error rates with less conservative alpha level and larger sample sizes. In sum, the forward approach outperformed the other two approaches in terms of the model-level Type I error rates across all PMI conditions.

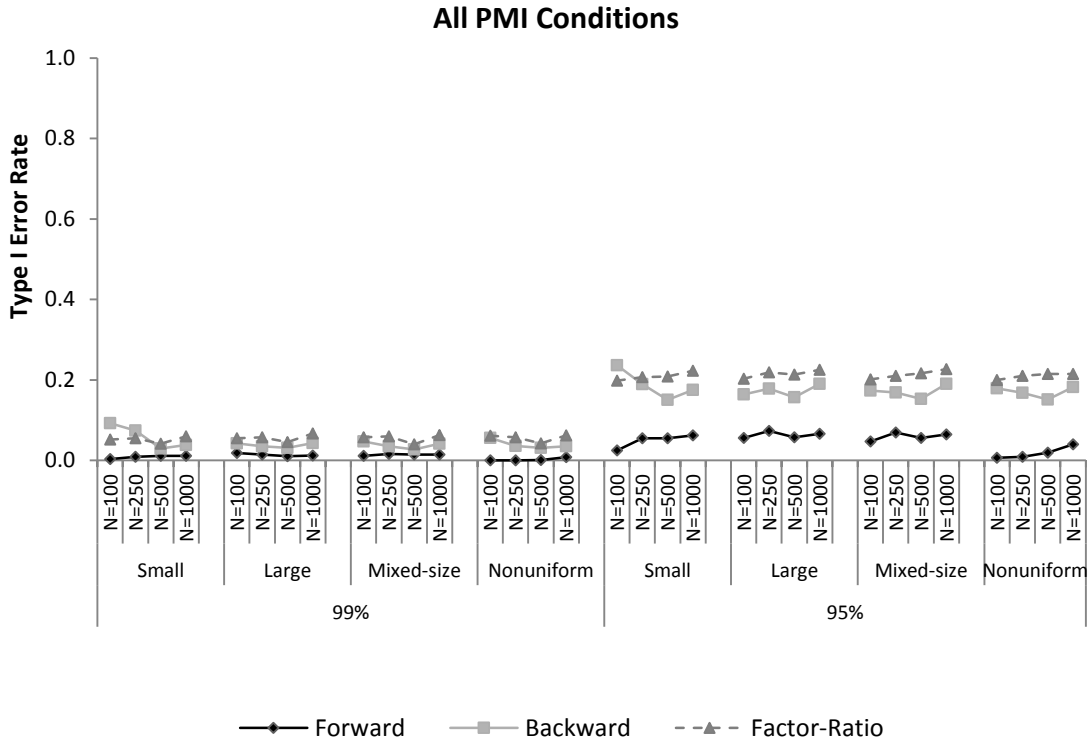

Figure 5. Model-level Type I Error Rates across PMI Conditions

Model-level Type II Error Rate

In Study II, model-level Type II error rate refers to any failure to detect truly noninvariant factor loadings in the finally specified PMI model. Figure 6 shows the Type II errors of the three approaches based on the confidence (significance) level, the size/

patterns of noninvariance, and the sample sizes. Both forward and backward approaches maintained very low model-level Type II error rates except for some conditions with small sample sizes and small difference under both confidence (significance) levels. The factor-ratio test presented extremely high Type II error rates in both small- and large-difference conditions while it still had higher Type II error rates in the remaining conditions that the other two approaches. To summarize, the backward approach performed best in terms of Type II error rates across all PMI conditions.

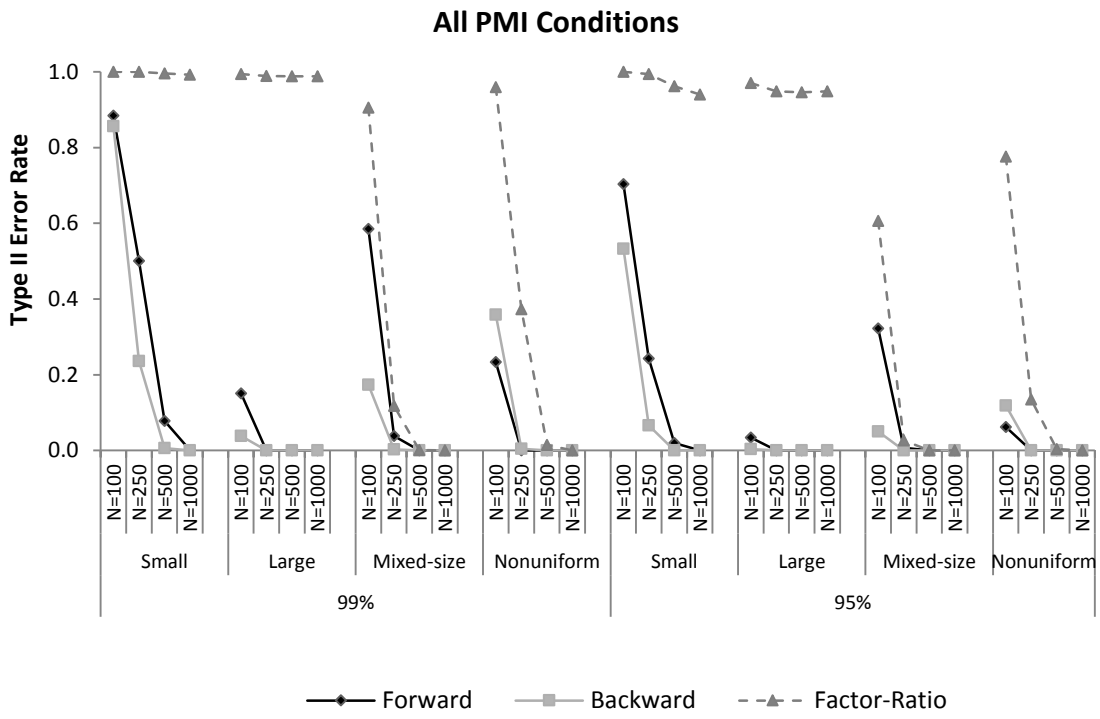

Figure 6. Model-level Type II Error Rates across PMI Conditions

Partial Scalar Invariance Conditions

Across all partial scalar invariance (PSI) conditions, we had two noninvariant intercepts among the six variables under one factor. As in the partial metric invariance conditions, we evaluated the performances of the forward approach, backward approach, and factor-ratio test in terms of perfect recovery rates, model-level Type I error rates, model-level Type II error rates, item-level power, and item-level Type I error rates in specifying the original PSI model. In addition, all results were presented based on both 99% and 95% confidence (significance) levels.

Forward Approach

The performance of the forward approach is presented in Table 10. The perfect recovery rates of the forward approach were lower in the PSI conditions compared to those in the PMI conditions. One major difference was observed in the conditions with larger sample size (≥ 500). In the large-, mixed-size-, and nonuniform-difference conditions, the highest perfect recovery rates were found with N = 250 while the perfect recovery rates were lower given larger samples. Additionally, the perfect recovery rates in all conditions with N = 100 were slightly higher than those in the PMI conditions. Among the all PSI conditions, the maximum perfect recovery rate was found in the nonuniform condition with N = 250 while the minimum perfect recovery rate was found in the small-difference conditions with N =100.

The nominal model-level Type I error rate is .03 and .15 given 99% and 95% BCBS-CIs, respectively, because there are three invariant intercepts to be tested using the forward approach. Under Bradley's criteria an acceptable model-level Type I error

rate is between .015 and .045 (99% confidence level) or between .075 and .225 (95% confidence level). However, the forward approach showed very high model-level Type I error rates across all sample size conditions as sample size grew as we saw the similar pattern in the basal Type I error rates in specifying intercepts. Only 50% of the conditions achieved acceptable model-level Type I error rates based on the Bradley's criteria (≤ .045) under both 99% and 95% levels. Particularly, the forward approach presented the lowest Type I error rates in the nonuniform-difference condition. We also observed inflated item-level Type I error rates, especially with larger sample sizes (≥ 500). The main reason of the lower perfect recovery rates with larger samples appeared to be the inflated Type I error rates in the PSI conditions. Overall, the forward approach maintained very low model-level Type II error rates while having very high or perfect item-level power except for the conditions with a small sample size (N =100) or with small difference regardless of the significance level. Compared to the PMI conditions, however, the forward approach showed higher power (or lower Type II error rate) in those conditions.

Unexpectedly, the perfect recovery rates in the PSI conditions were relatively low compared to the PMI conditions mainly due to inflated Type I error rates. However, it maintained both model-level and item-level Type I error rates lower or close to Bradley's criteria in the nonuniform difference condition. In terms of confidence levels, higher perfect recovery rates were found with 99% level than with 95% level.

Table 10. Performance of the Forward Approach in PSI Conditions

Condition	N	99% BCBS-CIs					95% BCBS-CIs				
		Perfect recovery	Type I error	Type II error	Power/ # of items	Type I error/ # of items	Perfect recovery	Type I error	Type II error	Power/ # of items	Type I error/ # of items
Small-difference	100	0.201	0.023	0.795	0.603	0.008	0.385	0.099	0.583	0.709	0.030
	250	0.698	0.040	0.280	0.860	0.013	0.766	0.174	0.093	0.954	0.060
	500	0.895	0.094	0.017	0.992	0.031	0.759	0.239	0.002	0.999	0.080
	1000	0.866	0.134	0.000	1.000	0.045	0.641	0.359	0.000	1.000	0.120
Large-difference	100	0.945	0.028	0.031	0.985	0.009	0.902	0.096	0.004	0.998	0.030
	250	0.958	0.042	0.000	1.000	0.014	0.825	0.175	0.000	1.000	0.060
	500	0.907	0.093	0.000	1.000	0.031	0.762	0.238	0.000	1.000	0.080
	1000	0.869	0.131	0.000	1.000	0.044	0.641	0.359	0.000	1.000	0.120
Mixed-size-difference	100	0.544	0.041	0.435	0.782	0.014	0.709	0.120	0.204	0.898	0.040
	250	0.933	0.049	0.021	0.990	0.016	0.817	0.182	0.004	0.998	0.060
	500	0.904	0.096	0.000	1.000	0.032	0.760	0.240	0.000	1.000	0.080
	1000	0.865	0.135	0.000	1.000	0.045	0.641	0.359	0.000	1.000	0.120
Nonuniform-difference	100	0.790	0.002	0.208	0.896	0.001	0.930	0.022	0.053	0.974	0.010
	250	0.992	0.008	0.000	1.000	0.003	0.956	0.044	0.000	1.000	0.010
	500	0.972	0.028	0.000	1.000	0.009	0.905	0.095	0.000	1.000	0.030
	1000	0.950	0.050	0.000	1.000	0.017	0.782	0.218	0.000	1.000	0.070

Note. Perfect recovery rates and item-level power greater than or equal to 0.90 were underlined.

Backward Approach Using the Max-Mod

The performance of the backward approach in which the intercept indicated by the largest modification index (Max-Mod) was sequentially relaxed until no more significant modification index (Mod = 3.841 at $\alpha = .05$ and 6.635 at $\alpha = .01$) was left in the model. Table 11 shows the performance of the backward approach in specifying a PSI model in terms of perfect recovery rates, model-level Type I error rates, model-level Type II error rates, item-level power, and item-level Type I error rates depending on the significance criteria.

The backward approach showed very high perfect recovery rates (greater than .916) using the modification-index value of 6.635 (at $\alpha = .01$) except for two conditions (small-difference PSI condition with N =100; nonuniform-difference PSI condition with N =100). However, it achieved higher perfect recovery rates in such conditions compared to the same PMI conditions. However, similarly in the PMI conditions, the perfect recovery rates (PRs $\leq .835$) decreased much using the lower cut-off value (Mod = 3.841).

As explained before, the model-level nominal Type I error rate of the backward approach is 0.04 at $\alpha = .01$ while it is .20 at .05 in the PSI conditions. If we use Bradley's formula, acceptable model-level Type I error rate is located between .02 and .06 ($\alpha = .01$) or between .10 and .30 ($\alpha = .05$). Across all PSI conditions, model-level Type I error rates were very close to the acceptable level of Bradley's criteria when using Mod = 6.635 ($\alpha = .01$). However, similar to the PMI conditions, Type I error rates

Table 11. Performance of the Backward Approach in PSI Conditions

Condition	N	Mod = 6.635 (α = 0.01)					Mod = 3.841 (α = 0.05)				
		Perfect recovery	Type I	Type II	Power/ # of items	Type I error/ # of items	Perfect recovery	Type I	Type II	Power/ # of items	Type I error/ # of items
Small-difference	100	0.508	0.131	0.448	0.623	0.040	0.634	0.230	0.176	0.803	0.079
	250	0.937	0.054	0.010	0.988	0.016	0.828	0.172	0.000	0.995	0.047
	500	0.935	0.065	0.000	0.998	0.018	0.810	0.190	0.000	0.998	0.052
	1000	0.968	0.032	0.000	1.000	0.008	0.831	0.169	0.000	1.000	0.043
Large-difference	100	0.958	0.042	0.000	0.995	0.014	0.835	0.165	0.000	0.995	0.047
	250	0.960	0.040	0.000	1.000	0.010	0.829	0.171	0.000	1.000	0.044
	500	0.935	0.065	0.000	1.000	0.017	0.810	0.190	0.000	1.000	0.050
	1000	0.968	0.032	0.000	1.000	0.008	0.832	0.168	0.000	1.000	0.043
Mixed-size-difference	100	0.916	0.049	0.035	0.974	0.013	0.825	0.166	0.009	0.989	0.045
	250	0.960	0.040	0.000	1.000	0.010	0.829	0.171	0.000	1.000	0.044
	500	0.935	0.065	0.000	1.000	0.017	0.810	0.190	0.000	1.000	0.050
	1000	0.968	0.032	0.000	1.000	0.008	0.832	0.168	0.000	1.000	0.043
Nonuniform-difference	100	0.883	0.045	0.072	0.955	0.012	0.817	0.167	0.016	0.984	0.045
	250	0.960	0.040	0.000	1.000	0.010	0.829	0.171	0.000	1.000	0.044
	500	0.935	0.065	0.000	1.000	0.017	0.810	0.190	0.000	1.000	0.050
	1000	0.968	0.032	0.000	1.000	0.008	0.831	0.169	0.000	1.000	0.043

Note. Perfect recovery rates and item-level power greater than or equal to 0.90 were underlined.

with Mod = 3.841 (α = .05) were too high while leading to low perfect recovery rates although the item-level Type I error rates appeared not to be very high (below the nominal level). Under both criteria (Mod = 6.635 and 3.841), the backward approach maintained very low Type II error rates except for the small-difference condition with N =100. Compared to the forward approach, it showed lower model-level Type I error rates. The item-level Type I error rates of the backward approach were within Bradley's criteria in most cases except for the small-difference condition with N =100.

In sum, the backward approach performed ideally with higher significance value (Mod = 6.635) in the PSI conditions. As in the PMI conditions, the major source low perfect recovery rates was the high model-level Type I error rates when Mod = 3.841 was applied. Except for the conditions with a small sample size, the backward approach did not show much difference in the perfect recovery rates no matter which significance values was used. Generally, the backward approach showed promising performance in specifying the original PSI models when using the significance value at α = .01.

Factor-ratio Test

Table 12 presents the performance of the factor-ratio test in specifying the original partial scalar invariance (PSI) model: the perfect recovery rates, model-level Type I error rates, model-level Type II error rates, item-level power, and item-level Type I error rates using both 99% and 95% BCBS-CIs.

Similar to the PMI condition, the perfect recovery rates of the factor-ratio test were extremely low in small- and large-difference PSI conditions across all sample sizes. In the mixed-size- and nonuniform-difference conditions, it showed improved perfect

recovery rates, particularly, when using 99% BCBS-CIs. Among those conditions, the highest perfect recovery rate (PR = .874) was found in the nonuniform-difference condition with N = 250. Overall, the perfect recovery rates of the factor-ratio test in the PSI conditions appeared to be lower than in the same PMI conditions.

As we observed the inflated Type I error rates of the forward approach in the PSI conditions, the factor-ratio test showed higher model-level and item-level Type I error rates in the PSI conditions than the PMI conditions. Especially, the inflation of the Type I error rates seemed to be related to increase in sample sizes. The patterns of the model-level Type II error rates and item-level power were very similar to those in the PMI conditions. In the small- and large-difference conditions, it had high item-level power while showing extremely poor perfect recovery rates. In the mixed-size- and nonuniform-difference conditions, it showed very high item-level power with larger samples ($\geq$ 250) while demonstrating lower perfect recovery rates which were not that extreme as in the small- and large-difference conditions.

In sum, the factor-ratio test seems not to perform adequately in the PSI conditions. Particularly, its performance was extremely poor when there are two noninvariant intercepts with the same size of difference in the same direction. Although it showed improved performance in the other conditions (i.e., mixed-size- and nonuniform-difference condition), it was not still ideal in the PSI conditions.

Table 12. Performance of the Factor-ratio Test in PSI Conditions

		99% BCBS-CIs					95% BCBS-CIs				
Condition	N	Perfect recovery	Type I error	Type II error	Power/ # of items	Type I error/ # of items	Perfect recovery	Type I error	Type II error	Power/ # of items	Type I error/ # of items
Small-difference	100	0.000	0.065	1.000	0.381	0.013	0.004	0.222	0.996	0.596	0.056
	250	0.004	0.081	0.996	0.794	0.016	0.031	0.275	0.968	0.867	0.069
	500	0.005	0.128	0.994	0.886	0.027	0.029	0.331	0.950	0.894	0.093
	1000	0.008	0.172	0.988	0.890	0.039	0.024	0.444	0.948	0.895	0.127
Large-difference	100	0.008	0.065	0.992	0.874	0.013	0.037	0.221	0.954	0.892	0.056
	250	0.016	0.081	0.983	0.891	0.016	0.043	0.274	0.939	0.896	0.069
	500	0.004	0.126	0.995	0.889	0.027	0.032	0.330	0.951	0.894	0.093
	1000	0.006	0.173	0.990	0.890	0.039	0.032	0.444	0.943	0.895	0.126
Mixed-size-difference	100	0.023	0.065	0.977	0.773	0.013	0.144	0.221	0.837	0.875	0.056
	250	0.422	0.083	0.538	0.939	0.016	0.491	0.274	0.319	0.964	0.069
	500	0.728	0.127	0.165	0.982	0.027	0.631	0.330	0.059	0.993	0.093
	1000	0.825	0.170	0.006	0.999	0.039	0.555	0.444	0.002	1.000	0.127
Nonuniform-difference	100	0.196	0.065	0.804	0.774	0.013	0.466	0.221	0.466	0.908	0.056
	250	0.874	0.081	0.061	0.992	0.016	0.724	0.274	0.006	0.999	0.069
	500	0.873	0.127	0.000	1.000	0.027	0.669	0.331	0.000	1.000	0.093
	1000	0.828	0.172	0.000	1.000	0.039	0.556	0.444	0.000	1.000	0.126

Note. Perfect recovery rates and item-level power greater than or equal to 0.90 were underlined.

Comparisons of the Performances of the Three Approaches: PSI Conditions

Perfect Recovery Rates

We also examined, in a graph of each PSI condition, the perfect recovery rates of the forward approach, backward approach, and factor-ratio test. As in the PMI conditions, we located the perfect recovery rates were displayed under both significance (confidence) levels: $\alpha = .01$ and at $\alpha = .05$ for the backward approach and 99% and 95% BCBS-CIs for the forward approach and factor-ratio test. For convenience, we will use the terms 99% confidence level (or $\alpha = .01$) and 95% confidence level (or $\alpha = .05$) to represent the significant (confidence) level for the three approaches.

The perfect recovery rates of the three approaches in the small-difference PSI condition are demonstrated in Figure 7. In all sample size conditions, the backward approach performed best based on 99% confidence level (or $\alpha = .01$). It maintained very high perfect recovery rates (PRs $\geq .935$) given larger sample sizes ($N \geq 250$). Differently from the PMI conditions, the forward approach presented lower perfect recovery rates even with larger sample ($N \geq 500$) than the backward approach mainly due to the inflated Type I error rates. The factor-ratio test exhibited extremely poor performance across all sample sizes as it did in the small-difference PMI conditions. Under 95% confidence level (or $\alpha = .01$), the backward approach outperformed the forward approach while the factor-ratio almost failed to recover the original PSI model. Overall, among the three approach, the backward approach performed best across all conditions, and it showed ideal perfect recovery rates with a larger sample size ($N \geq 250$) when using the more conservative significance level ($\alpha = .01$).

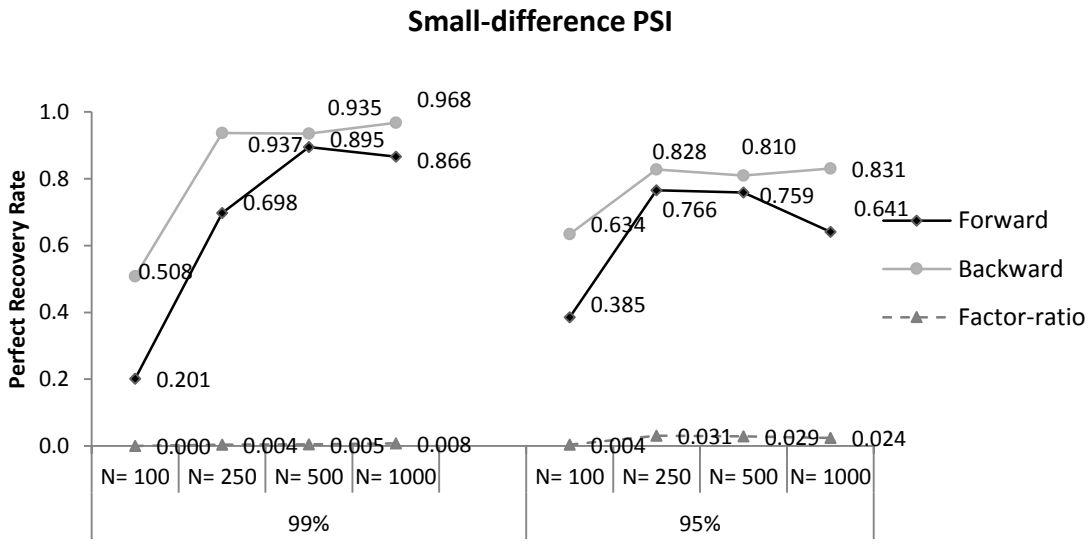

Figure 7. Perfect Recovery Rates in the Small-difference PSI Conditions

As illustrated in Figure 8, the backward approach showed the highest perfect recovery rates regardless of the sample sizes under 99% confidence level (or $\alpha = 0.01$). Although the forward approach could not beat the backward approach, it also presented very high perfect recovery rates with $N \leq 500$. It had the lowest perfect recovery rate with N =1000 because of high Type I error rate. The factor-ratio test demonstrated extremely poor performance across all sample sizes as well. Under 95% confidence level (or $\alpha = 0.01$), the forward approach had the highest perfect recovery rates with N = 100. However, its perfect recovery rates decreased as the sample size grew. The backward approach maintained very similar perfect recovery rates across all sample sizes, however, they are much lower than those under the more conservative significance level ($\alpha = .01$). In sum, the performance of the backward approach was most promising among the three approaches when we use 99% confidence level (or $\alpha = .01$).

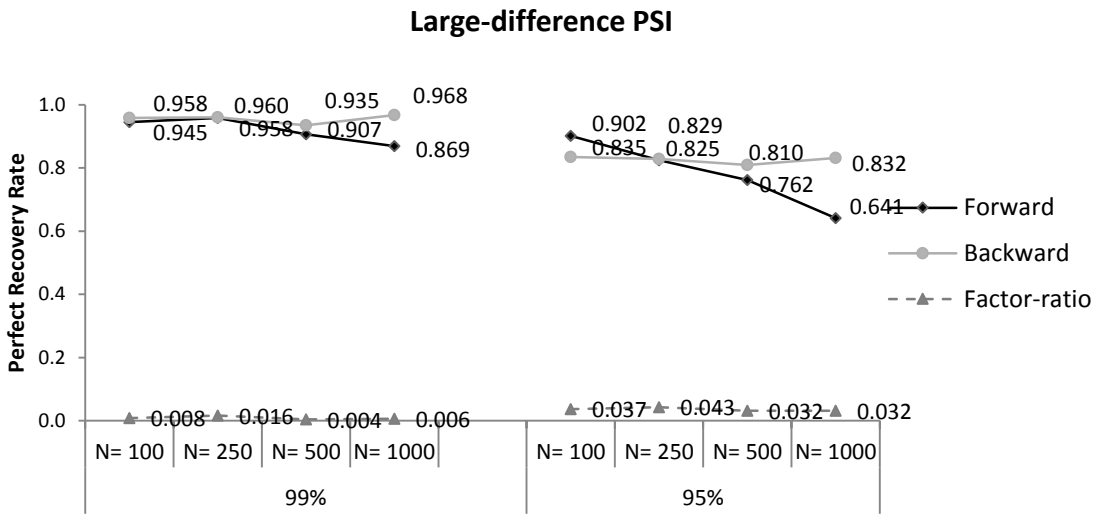

Figure 8. Perfect Recovery Rates in the Large-difference PSI Conditions

Figure 9 displays the perfect recovery rates of the three approaches in the mixed-size-difference condition. As in the small- and large-difference PSI conditions, the backward approach most successfully recovered the original mixed-size-difference PSI model under 99% confidence level (or $\alpha = .01$). Although the forward approach achieved high perfect recovery rates with N =250 and 500, it could not outperform the backward approach. The factor-ratio test showed improved perfect recovery rates compared to the small- and large-difference PSI conditions, but, it still performed worst among the three approaches. Under 95% confidence level (or $\alpha = .05$), the backward approach had the highest perfect recovery rates across all sample sizes, but its performance was not as good as under 99% confidence level (or $\alpha = .01$). The next well-performing approach was the forward approach while the factor-ratio test demonstrated the poorest performance.

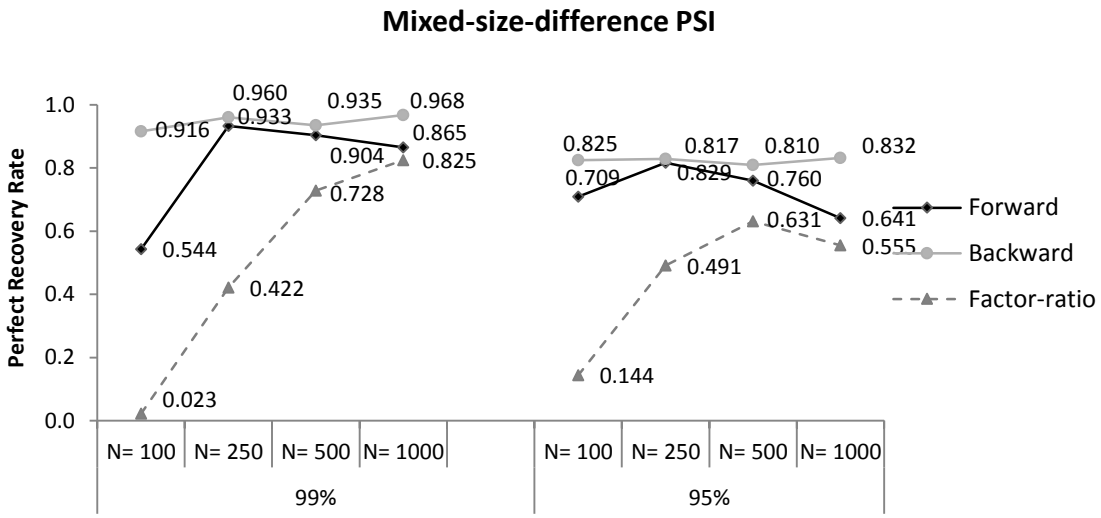

Figure 9. Perfect Recovery Rates in the Mixed-size-difference PSI Conditions

Figure 10 illustrates the perfect recovery rates of the three approaches in specifying the original nonuniform-difference PSI model. Under 99% confidence level (or $\alpha = .01$), the backward approach had the highest perfect recovery rates with N = 100 and N = 1000 while the forward approach presented the highest perfect recovery rates with N =250 and 500. The factor-ratio test showed the poorest performance. When we referred to 95% confidence level (or $\alpha = .05$), the forward approach outperformed the other approaches except for with N =1000. The backward approach had the highest perfect recovery rate with N =1000. The factor-ratio test still presented the lowest perfect recovery rates.

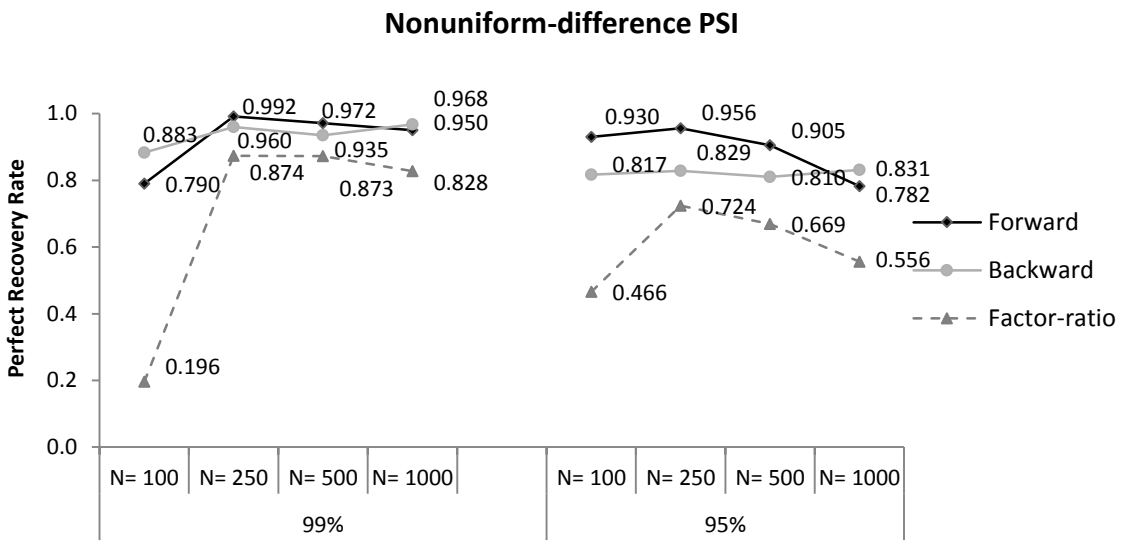

Figure 10. Perfect Recovery Rates in the Nonuniform-difference PSI Conditions

Model-level Type I Error Rate

As defined before, model-level Type I error rate indicates any occurrence of falsely specifying invariant intercept as noninvariant in the final PSI model. Interestingly, we observed inflated Type I error rates for both the forward approach and factor-ratio test with larger samples (see Figure 11). The Type I error inflation got severe under 95% confidence level (or $\alpha = .05$). Different from our expectation, model-level Type I error rates of the backward approach did not seem to be associated with larger sample sizes. Instead, it presented the highest model-level Type I error rates in the condition with small-difference combined the small sample size (N =100). In sum, the backward approach performed best in terms of model-level Type I error rate when using the more conservative significance level ($\alpha = .01$).

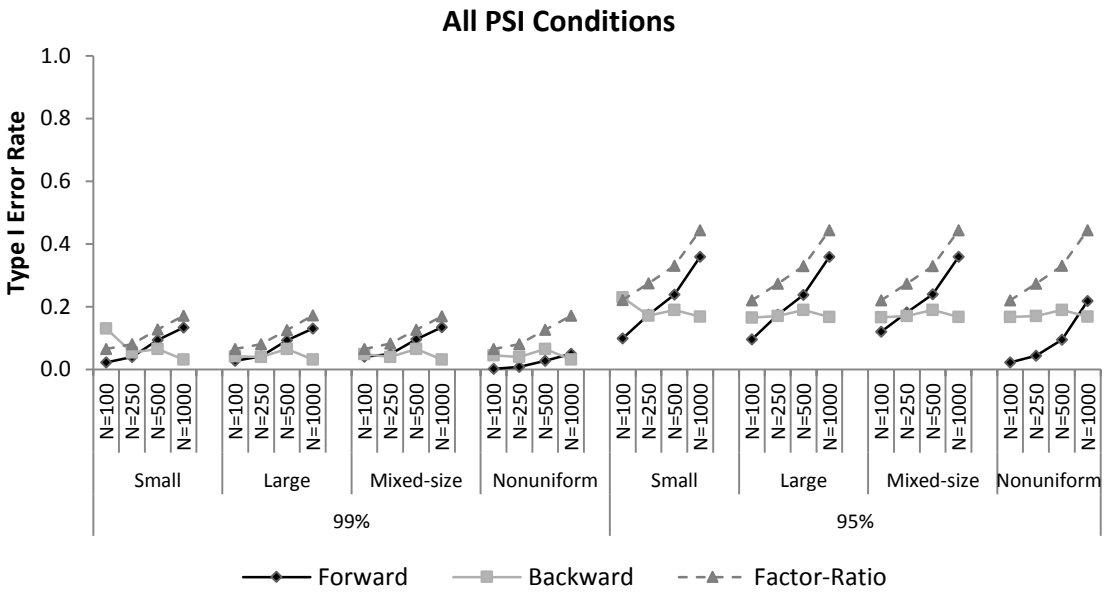

Figure 11. Model-level Type I Error Rates across PSI Conditions

Model-Level Type II Error Rate

Model-level Type II error is defined as the failure of detecting any truly noninvariant intercept in the finally specified PSI model. Hence, the error rate was calculated by dividing any occurrence of Type II error in the final PSI model by the number of replications (= 1000). Regardless of the chosen confidence (significance) levels, the backward approach maintained very low model-level Type II error rates except for some conditions in which either the same size is small or the size of noninvariance was small (see Figure 12). The forward approach showed the similar pattern with the backward approach, but it had higher Type II error rates in the conditions with small sample sizes and/or small difference. The factor-ratio test showed extremely high Type II error rates in both small- and large-difference PSI conditions while its Type II error rates were lower in the remaining PSI conditions with larger samples. All three approaches presented lower error rates under 95% confidence level (or $\alpha = .05$) than under 99% confidence level (or $\alpha = .01$) in the conditions exhibiting high model-level Type II error rates. In sum, the backward approach was least prone to model-level Type II error no matter which significance level was given.

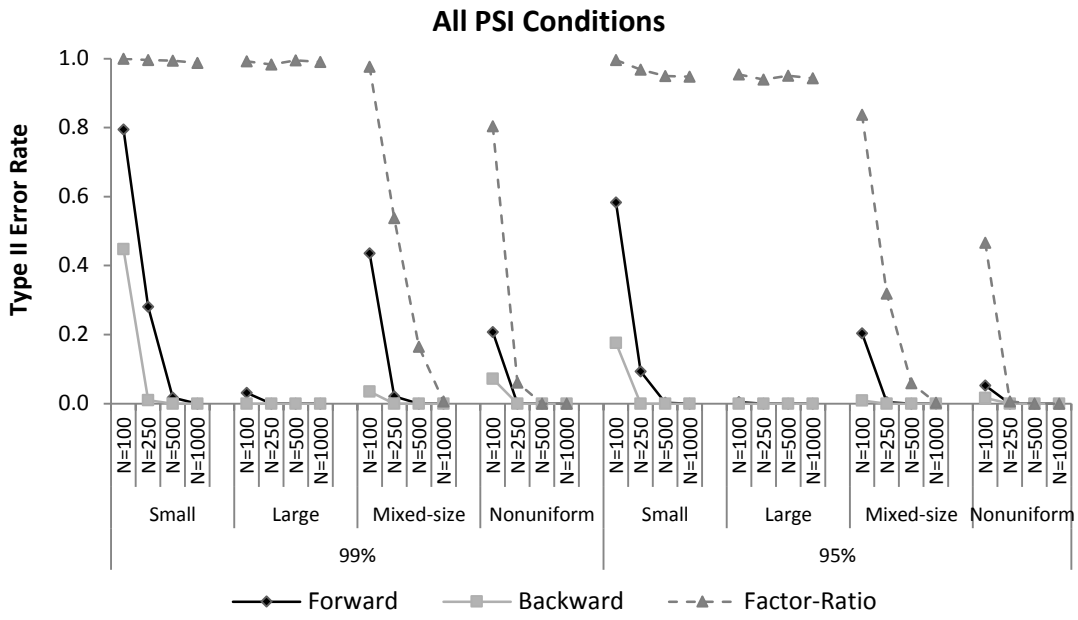

Figure 12. Model-level Type II Error Rates across PSI Conditions

Design Effects on the Performance of the Three Approaches

The effects of the methods and simulation conditions on the perfect recovery rates, model-level Type I error rates, and model-level Type II error rates were examined through the analysis of variance (ANOVA) with respect to the location of nonivaraince. Because the sample size was very large for each analysis (N = 32000), all p-values yielded to a significant value. Hence, only the effect size (η^2) were reported to see the contributing variance of each factor in the total variance of the outcome variables (i.e., perfect recovery rates, Type I error rates, and Type II error rates). The design factors were the specification methods (e.g., forward approach, backward approach, and factor-ratio test), the size/patterns of noninvariance (e.g., small-, large-, mixed-size-, and nonuniform-difference), and the sample size per group (e.g., 100, 250, 500, and 1000).

With respect to the location of noninvariance (factor loadings or intercepts) we conducted ANOVA for each outcome only under 99% confidence level (or $\alpha = .01$) since the results of all three approaches yielded , generally, better performance under the confidence (or significance) level.

Partial Metric Invariance Conditions

Table 13 shows the proportion of the variance explained by each method and design factor and the two-way and three-way interactions among them in the perfect recovery rates, model-level Type I error rates, and model-level Type II error rates separately in specifying a true partial metric invariance (PMI) model in various PMI conditions.

First, the perfect recovery rates of the PMI conditions across the three methods were mainly accounted for by the method, the sample size, the size/pattern of noninvariance, and the two-way interaction between the method and the size/pattern of noninvariance. The method explained 37.4% of the variance in the perfect recovery rates. The next influencing factor was the sample size which explained 18.9% of variance in the perfect recovery rates. Then, the size/pattern and the two-way interaction between the method and the size/pattern accounted for 15.6% and 14.1% variance in the perfect recovery rates. Second, the model-level Type I error rates of the simulated PMI conditions across the three methods were mainly explained by the method ($\eta^2 = .743$) among the three factors and their interactions. Although the remaining variance in the model-level Type I error explained by the remaining factors and all types of interactions, their effects were not very large ($\eta^2 \leq .061$). Third, the variance in the model-level Type

II error was explained by the method (η^2 = .364), the sample size (η^2 = .187), the size/pattern of noninvariance (η^2 = .158), and the two-way interaction between the method and the size/pattern of noninvariance (η^2 = .147). The pattern of the effect on the model-level Type II error rate was very similar with that on the perfect recovery rates.

Table 13. Effect Size (η^2) of Each Method and Design Factor in PMI conditions

Design Factor	Perfect recovery	Type I error	Type II error
Method	0.374	0.743	0.364
Size/Pattern	0.156	0.013	0.158
Sample size	0.189	0.059	0.187
Method* Size/Pattern	0.141	0.054	0.147
Method*Sample size	0.002	0.061	0.003
Size/Pattern*Sample size	0.050	0.016	0.049
Method*Size/Pattern*Sample size	0.088	0.054	0.092

Partial Scalar Invariance Conditions

Table 14 presents the effects of the design factors on the performance of the three methods in specifying the true partial scalar invariance (PSI) models. In the table, the proportion of the variance in the perfect recovery rates, model-level Type I error rates, and model-level Type II error rates are shown by the method, the size/pattern of noninvariance, the sample size, and all two- and three-way interactions among them.

First, the perfect recovery rate was mainly explained by the method (η^2 = .544), the size/pattern of noninvariance (η^2 = .126), and their two-way interaction (η^2 = .130). Second, the majority of the variance in the model-level Type I error rates was explained by the method (η^2 = .398), the sample size (η^2 = .338), and their interaction (η^2 = .158).

Third, a total of 86.9% of the variance in the Type II error rates was explained by the method (η^2 = .487), the size/pattern of noninvariance (η^2 = .125), the sample size (η^2 = .104), and the two-way interaction between the method and the size/pattern of noninvariance (η^2 = .153).

Table 14. Effect Size (η^2) of Each Method and Design Factor in PSI Conditions

Design Factor	Perfect recovery	Type I error	Type II error
Method	0.544	0.398	0.487
Size/Pattern	0.126	0.030	0.125
Sample size	0.085	0.338	0.104
Method* Size/Pattern	0.130	0.056	0.153
Method*Sample size	0.011	0.158	0.017
Size/Pattern*Sample size	0.035	0.007	0.035
Method*Size/Pattern*Sample size	0.068	0.013	0.078

CHAPTER V

SUMMARY AND CONCLUSION

Summary

The motivation for the two studies in this dissertation was to address unanswered problems in measurement invariance literature, particularly, related to partial measurement invariance. Once full measurement invariance is rejected, investigating the source of noninvariance might be the most commonly chosen next step no matter how the information is used later. To identify the source of noninvariance accurately, according to Johnson et al. (2009), identification of a multi-group confirmatory factor analysis model should be achieved through a truly invariant reference variable (RV). However, it has been recognized that a well-performing methodological approach has yet to accurately identify a truly invariant reference variable (Raykov & Marcoulides, 2012). At best, the most common recommendation is turning to theories. To avoid erroneously selecting a noninvariant RV, two approaches have been suggested: the factor-ratio test (Cheung & Rensvold, 1998) and backward approach using the largest modification index (Yoon & Millsap, 2007). Those approaches do not require a specific reference variable. However, they have not yet been directly compared. Given the limitations of each method, it was also imperative to come up with a new method that could address the problems of both methods. Thus, two simulation studies were conducted to address the unresolved problems in the partial measurement invariance literature. Study I investigated the accuracy of the smallest modification index (Min-

Mod) in identifying a truly invariant RV within the equally constrained parameter sets under various partial factorial invariance (PFI) conditions. Study II compared the three approaches in specifying a true PFI models. The summaries of findings and discussions for each study follow.

Study I

In Study I, we examined the performance of the smallest modification index (Min-Mod) to identify a truly invariant set of factor loadings or intercepts. In the four partial factorial invariance scenarios, a fully constrained metric or scalar invariance model served as a baseline model in identifying a truly invariant factor loading while a fully constrained intercept or scalar invariance model was used for selecting an invariant intercept. The chosen identification method was the variation of reference variable identification method (VRV-IM). Within the tested set of equally constrained parameters (factor loadings or intercepts), the smallest modification index (Min-Mod) was hypothesized to indicate the smallest difference in the parameters between the two groups.

In Study I, we could observe only little variation in the accuracy of the Min-Mod to detect a truly invariant RV with very high accuracy levels. The results indicated that the accuracy of the Min-Mod was almost perfect across all PFI scenarios except for some conditions. For example, all conditions with N = 100 presented slightly lower accuracy rates compared to the conditions with ≥ 250 in all PFI scenarios. In addition, we observed the lowest accuracy rates in the mixed-size-difference condition of the partial metric and scalar invariance of the same variable (PMSI-S). In such condition, if

we focus on larger sample sizes (≥ 250), the Min-Mod showed slightly higher accuracy rates in the model in which only the targeted parameters were constrained to be invariant (i.e., metric invariance model for selecting a factor loading; intercept invariance model for selecting an intercept).

In this study, we focused on identifying a set of parameters which was supposed to have the least difference across two groups rather than searching for all possible invariant factor loadings or intercepts. The results of the study supported the idea that the Min-Mod was able to identify, almost perfectly, a truly invariant RV. The findings of this study are very promising to guide researchers who lack a theoretical guideline in selecting an appropriate RV in testing measuring invariance under a multi-group confirmatory factor analysis model. Even for those who already have a theoretical guideline to select an RV, they can also, through this empirical guideline (Min-Mod), be provided evidence of the adequacy of the chosen RV. As commonly pointed out, this Monte Carlo study simulated only limited partial metric or partial scalar invariance conditions. Therefore, the results can be generalized to only similar situations investigated in this study.

Study II

Study II aimed to evaluate the performances of the forward approach using BCBS-CIs, the backward approach using the largest modification index (Max-Mod), and the factor-ratio test (FR-T) in correctly specifying various partial metric or scalar invariance models. As the most interesting outcome, we examined the perfect recovery

rates of each method. We also looked at the model-level Type I error, model-level Type II error, item-level power, and item-level Type I error.

Forward Approach

The forward approach was newly proposed in this dissertation. Actually, this approach had been thought to be unrealistic since no empirical method was available to search for a truly invariant reference variable (RV). Although the issue of selecting a truly invariant RV (Raykov & Marcoulides, 2012) is known to be an unresolved one, the results of Study I indicated that the smallest modification index (Min-Mod) within the constrained parameter set could accurately identify a truly invariant pair of parameters. Hence, the chosen variable from Study I served as an RV in Study II.

The results indicate that the forward approach using the BCBS-CIs performed well with very high perfect recovery rates in most partial metric invariance conditions. If the size of noninvariance and the sample size were small, however, it was not able to correctly detect existing noninvariance. For those conditions, the main source of errors was Type II errors, which means failing to detect truly noninvariant parameters as so they were. In previous studies related to the power to detect noninvariance in full measurement invariance levels, noninvariance could not be detected successfully with small difference and small sample size. Therefore, it might not constitute such a large problem having low perfect recovery rates of the forward approach in the small sample size and small difference condition might not be a big problem. To put it into another way, it is very unlikely to reject full measurement invariance given small difference and small sample sizes, and thus, it is unlikely to require the further step to investigating

partial measurement invariance. More importantly, the performance of the forward approach was almost perfect in most conditions with substantially larger sample sizes with larger differences. In those conditions, it had negligible Type I and Type II errors.

However, the forward approached showed inflated Type I error rates as the sample size grew in specifying a true partial scalar invariance model. As a result, the perfect recovery rates got to be lower than those in the PMI conditions while it was outperformed by the backward approach. Future study may be necessary to investigate why such inflated Type I error rates happened.

Overall, the newly proposed forward approach looked very promising if our targeted parameters are factor loadings using 99% confidence intervals. The researchers can accurately determine the source of noninvariance when metric invariance is rejected using the forward approach. If the samples sizes are not too large (≤ 500), the forward approach also can serve as the method to identify noninvariant intercepts. In addition to its high perfect recovery rates, the forward approach can be conducted very simply. When a full factorial invariance model is rejected, we need only two data analysis procedures. One is to select an RV using the Min-Mod as in Study I. The other is to specify a partial factorial invariance model with the chosen RV using the forward approach which needs only one data analysis phase.

Backward Approach

For the backward approach, the baseline model was either a fully constrained metric or scalar invariance model with the variation of the reference variable identification method (VRV-IM). From the baseline model, a set of equally constrained

parameters (factor loadings or intercepts) with the largest modification index (Max-Mod) greater than 3.84 (a significant modification index for one parameter given α = .05) was relaxed once at a time until no significant modification index left in the model. In addition, we also tried the model modification using the significant value of the modification index at α = .01. Although it showed high Type I error rates consistent with the previous study (Yoon & Millsap) when using the modification-index value at α = .05 (Mod = 3.841), the Type I error rates substantially decreased with the larger value (Mod = 6.635 at α = .01). Different from our expectation, the backward approach was not prone to inflated Type I error rates when we adjusted the significance value, particularly, when there is no misspecification left in the model (see Appendix E and Appendix F). The finding from Study II indicates that we can confidently use the backward approach (iterative process) in identifying the source of noninvariance for both factor loadings and intercepts, but, the significance value should be Mod = 6.635 rather than Mod = 3.841.

Factor-ratio Test

The factor-ratio test was to test to test every pair of a reference variable (RV) and the other variable, and thus, we need $\frac{p\,(P-1)}{2}$ (p = number of parameters to be tested for invariance) tests. In Study II, the procedure was simplified using the BCBS-CIs as in the forward approach. Since we simulated only one factor model with six variables, the number of retrieved BCBS-CIs was 15. The results of Study II indicate the factor-ratio test performed extremely poorly when the size of noninvariance was the same across groups, as in the small and large difference PMI or PSI conditions. Regardless of the sample sizes, it could detect almost none of the noninvariant parameters, and the high

Type II error rates directly affected the low perfect recovery rates. However, its perfect recovery rates were even higher than the backward approach in most mixed-difference and nonuniform-difference conditions with N =500 and 1000, but it could not outperform the forward approach. In terms of Type I error rates, it performed ideally in all PMI conditions. However, it had higher Type I error rates in the PSI conditions. The factor-ratio can perform well only in the conditions with the different degree of noninvariance. In addition, the factor-ratio test requires subsequent procedures to discriminate noninvariant sets from invariant sets. Yet in some cases the subsequent procedures produced more than one (Cheung & Rensvold, 1999; Rensvold & Cheung, 2001; French & Finch, 2008; Cheung & Lau, 2011). The inconsistent performances and ambiguity of the factor-ratio test discourage its possibility as a specification search approach for partial factorial invariance.

Limitations and Future Directions

Similar to any simulation studies, we only examined limited conditions. Therefore, the results can be generalized to only the data conditions similar to those in this dissertation study. For example, we examined only partial factorial invariance of one factor model with six indicators. Although we expect the results found in these studies to be generalized to simpler or more complex models, it is hard to say they will before testing with those conditions. Next, we simulated only balanced sample size conditions, and therefore, we cannot be sure that the results can be applied to the cases with substantially imbalanced sample sizes across groups. In addition, we examined only the conditions with continuous indicators, and it is unclear whether the results can be

generalized to cases with categorical indicators (e.g., dichotomous or polytomous). To address the limitations mentioned above, future studies are necessary with more various factor models, imbalanced sample sizes, and categorical variables. Finally, we simulated noninvariance in the models without any type of misspecifications. However, in reality, the confirmatory factor analysis model might have a certain degree of misfit from the beginning. Thus, another possible future study will be investigating the performance of the three approaches in the models with misspecification in other parts.

Conclusion

In this dissertation, we explored two unanswered problems in partial factorial invariance literature. The first study examined the accuracy of the smallest modification index (Min-Mod) to identify a truly invariant reference variable (RV) in a fully constrained factorial invariance model. The Min-Mod almost perfectly selected a truly invariant set of factor loadings or intercepts which can serve as an RV for either metric invariance or scalar invariance test. If the data condition is similar to one of the conditions simulated in the first study, a researcher can confidently choose an invariant RV using the Min-Mod and expect the error rates to be very low. The second study indicated that the forward approach using the 99% BCBS-CIs could specify a true partial metric invariance (PMI) with very high perfect recovery rates. However, inflated Type I error rates might be concern in specifying a partial scalar invariance (PSI) model. Overall, the backward approach performed very adequately in both PMI and PSI conditions when we used the significance values at $\alpha = .01$.

REFERENCES

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

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456-466.

Cheung, G. W., & Lau, R. S. (2011). A direct comparison approach for testing measurement invariance. *Organizational Research Methods*, *15*(2), 167-198.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*, 1-27.

Chou, C., & Bentler, P. M. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, Lagrange multiplier, and Wald tests. *Multivariate Behavioral Research, 25*(1), 115-136.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling*, *13*(3), 378-402.

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

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling, 16*, 642-657.

Jung, E. & Yoon, M**.** (2012, April). Problems of standardization identification method in testing measurement invariance through a multigroup confirmatory factor analysis. Paper presented at the annual meeting of *the American Educational Research Association*, Vancouver, British Columbia, Canada.

Khalid, M. N. (2011). A Comparison of top-down and bottom-up approaches in the identification of differential item functioning using confirmatory factor analysis. *The International Journal of Educational and Psychological Assessment, 7*(2), 1-18.

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

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance Using MIMIC: likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, *72*(3), 469-492.

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

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*(4), 361-388.

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

Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. Hoyle, D. Kaplan, G. A. Marcoulides, & S. West (Eds.), *Handbook of Structural Equation Modeling*. New York, NY: Guilford Press.

Muthén, B. O., & Muthén, L. K. (1998-2012). *Mplus User's Guide (7th ed.)*. Los Angeles, CA: Muthen & Muthen.

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

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552-566.

Rensvold, R. B., & Cheung, G. W. (1998). Testing measurement models for factorial invariance: A systematic approach. *Educational and Psychological Measurement*, *58*(6), 1017-1034.

Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim & L. L. Neider (Eds.), *Research in management*: Vol. 1. Equivalence in measurement (pp. 21-50). Greenwich, CT: Information Age.

Rock, D. A., Charles, E. W., & Ronald, L. F. (1978). The use of analysis of covariances structure for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research, 13*, 403-418.

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

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.

Steenkamp, J. B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78-90.

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), 4-70.

Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *The Journal of Experimental Education, 80* (1), 26-44.

Whittaker, T. A. (2013). The impact of noninvariant intercepts in latent means models. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(1), 108-130.

Woods, C. M. (2009a). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement, 33*, 42–57.

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

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

Zieky, M. J., (2013). Fairness review in assessment. In K. F. Geisinger (Ed). *APA Handbook of Testing and Assessment in Psychology* (Vol. 1. P. 293-302). The American Psychological Association.

APPENDIX A

MPlus Syntax of the FR-T Using the BCBS-CIs: Metric Invariance

```
TITLE: Factor-Ratio Tests for Factor Loadings using BCBS-CIs
DATA: File is ABC.dat;
VARIABLE:
Names are x1-x6 group;
Usev =x1-x6;
Grouping = group (1=g1 2=g2);

MODEL:
f1 by x1@1; f1 by x2 (LA2);
f1 by x3 (LA3); f1 by x4 (LA4);
f1 by x5 (LA5); f1 by x6 (LA6);
f1*;

MODEL g2:
f1 by x1@1; f1 by x2 (LB2);
f1 by x3 (LB3); f1 by x4 (LB4);
f1 by x5 (LB5); f1 by x6 (LB6);;
f1*;

MODEL CONSTRAINT:
NEW (FL12 FL13 FL14 FL15 FL16 FL23 FL24 FL25 FL26 FL34 FL35 FL36 FL45 FL46 FL56);

FL12=LA2 - LB2;
FL13=LA3 - LB3;
FL14=LA4 - LB4;
FL15=LA5 - LB5;
FL16=LA6 - LB6;
FL23=LA3/LA2 - LB3/LB2;
FL24=LA4/LA2 - LB4/LB2;
FL25=LA5/LA2 - LB5/LB2;
FL26=LA6/LA2 - LB6/LB2;
FL34=LA4/LA3 - LB4/LB3;
FL35=LA5/LA3 - LB5/LB3;
FL36=LA6/LA3 - LB6/LB3;
FL45=LA5/LA4 - LB5/LB4;
FL46=LA6/LA4 - LB6/LB4;
FL56=LA6/LA5 - LB6/LB5;

ANALYSIS: Bootstrap =1000;

OUTPUT: Cinterval (BCBootstrap);
```

"MODEL CONSTRAINT' command allows us to create new parameters for testing every pair of an RV and argument for factor loadings.

The parameter "FL12" is defined as the difference of the second factor loadings across groups when the first factor loading serves as a RV.

Entering "Bootstrap =1000" results in 1000 bootstrapping samples.

Bias-corrected confidence intervals are retrieved into the output file by putting "Cinterval (BCBootstrap)" subcommand.

APPENDIX B

MPlus Syntax of the FR-T Using BCBS-CIs: Scalar Invariance

```
TITLE: Factor-Ratio Tests for Intercepts using BCBS-CIs
DATA: File is ABC.dat;
VARIABLE:
Names are x1-x6 group;
Usev =x1-x6;
Grouping = group (1=g1 2=g2);

MODEL:
f1 by x1@1; f1 by x2 (LA2);
f1 by x3 (LA3); f1 by x4 (LA4);
f1 by x5 (LA5); f1 by x6 (LA6);
f1*; [f1*];
[x1@0]; [x2] (IA2);
[x3] (IA3); [x4] (IA4);
[x5] (IA5); [x6] (IA6);
MODEL g2:
f1 by x1@1;f1 by x2 (LB2);
f1 by x3 (LB3); f1 by x4 (LB4);
f1 by x5 (LB5); f1 by x6 (LB6);
f1*; [f1*];
[x1@0]; [x2] (IB2);
[x3] (IB3); [x4] (IB4);
[x5] (IB5); [x6] (IB6);

MODEL CONSTRAINT:
NEW (IT12 IT13 IT14 IT15 IT16 IT23 IT24 IT25 IT26 IT34 IT35 IT36 IT45 IT46 IT56);

IT12 = IA2-IB2;
IT13 = IA3-IB3;
IT14 = IA4-IB4;
IT15 = IA5-IB5;
IT16 = IA6-IB6;
IT23= IT13-LA3/LA2*IT12;
IT24= IT14-LA4/LA2*IT12;
IT25= IT15-LA5/LA2*IT12;
IT26= IT16-LA6/LA2*IT12;
IT34= IT14-LA4/LA3*IT13;
IT35= IT15-LA5/LA3*IT13;
IT36= IT16-LA6/LA3*IT13;
IT45= IT15-LA5/LA4*IT14;
IT46= IT16-LA6/LA4*IT14;
IT56= IT16-LA6/LA5*IT15;

ANALYSIS: Bootstrap =1000;
OUTPUT: Cinterval (BCBootstrap);
```

"MODEL CONSTRAINT' command allows us to create new parameters for testing every pair an RV and argument for intercepts.

The parameter "IT12" is defined as the difference of the second intercepts between groups when the first intercept serves as a RV.

APPENDIX C

MPlus Syntax of the Forward Approach: Metric Invariance

```
TITLE: Forward Approach using BCBS-CIs (Factor loadings)
DATA: File is ABC.dat;
VARIABLE:
Names are x1-x6 group;
Usev =x1-x6;
Grouping = group (1=g1 2=g2);

MODEL:
f1 by x1@1; ! Reference Variable
f1 by x2 (LA2);
f1 by x3 (LA3);
f1 by x4 (LA4);
f1 by x5 (LA5);
f1 by x6 (LA6);
f1*;

MODEL g2:
f1 by x1@1; ! Reference Variable
f1 by x2 (LB2);
f1 by x3 (LB3);
f1 by x4 (LB4);
f1 by x5 (LB5);
f1 by x6 (LB6);
f1*;

MODEL CONSTRAINT:
NEW (FL2 FL3 FL4 FL5 FL6);

FL2=LA2 - LB2;
FL3=LA3 - LB3;
FL4=LA4 - LB4;
FL5=LA5 - LB5;
FL6=LA6 - LB6;

ANALYSIS: Bootstrap =1000;

OUTPUT: Cinterval (BCBootstrap);
```

"MODEL CONSTRAINT' command allows us to create new parameters for testing every pair of corresponding factor loadings.

The parameter "FL2" is defined as the difference of the second factor loadings across groups.

Bias-corrected confidence intervals are retrieved into the output file by putting "Cinterval (BCBootstrap)" subcommand.

APPENDIX D

MPlus Syntax of the Forward Approach: Scalar Invariance

```
TITLE: Forward approach using BCBS-CIs (Intercepts)

DATA: File is ABC.dat;

VARIABLE:
Names are x1-x6 group;
Usev =x1-x6;
Grouping = group (1=g1 2=g2);

MODEL:
f1 by  x1-x6* (L1-L6);
f1*; [f1*];
[x1@0]; ! Reference Variable
[x2] (IA2);
[x3] (IA3);
[x4] (IA4);
[x5] (IA5);
[x6] (IA6);
MODEL g2:
f1 by  x1-x6* (L1-L6);
f1*; [f1*];
[x1@0]; ! Reference Variable
[x2] (IB2);
[x3] (IB3);
[x4] (IB4);
[x5] (IB5);
[x6] (IB6);

MODEL CONSTRAINT:
NEW (IT2 IT3 IT4 IT5 IT6);

IT12 = IA2-IB2;
IT13 = IA3-IB3;
IT14 = IA4-IB4;
IT15 = IA5-IB5;
IT16 = IA6-IB6;

ANALYSIS: Bootstrap =1000;

OUTPUT: Cinterval (BCBootstrap);
```

"MODEL CONSTRAINT' command allows us to create new parameters for testing every set of corresponding intercepts.

The parameter "IT2" is defined as the difference of the intercepts for X2 between groups.

APPENDIX E

Decrement of the Size of Modification Index: Partial Metric Invariance Conditions

Condition	N	Max1	Max2	Max3	Max4	Max5
Small-difference	100	7.63	4.20	**1.61**	0.48	0.04
	250	15.14	10.83	**2.23**	0.70	0.19
	500	26.96	23.23	**2.18**	0.64	0.05
	1000	48.94	47.59	**2.35**	0.71	0.05
Large-difference	100	21.85	16.85	**2.31**	0.66	0.05
	250	51.25	45.96	**2.31**	0.70	0.06
	500	98.45	95.80	**2.19**	0.64	0.05
	1000	188.61	194.84	**2.35**	0.72	0.05
Mixed-size-difference	100	28.11	12.58	**2.23**	0.64	0.05
	250	71.17	31.05	**2.31**	0.71	0.06
	500	143.25	61.34	**2.19**	0.63	0.05
	1000	282.84	121.07	**2.35**	0.72	0.05
Nonuniform-difference	100	19.64	8.50	**2.05**	0.62	0.11
	250	44.60	22.33	**2.30**	0.71	0.06
	500	84.99	46.88	**2.18**	0.65	0.32
	1000	165.67	98.80	**2.33**	0.70	0.05

Note. The bold values indicates the largest modification index values when there is no more noninvariant factor loading left in the model.

APPENDIX F

Decrement of the Size of Modification Index: Partial Scalar Invariance Conditions

Condition	N	Max1	Max2	**Max3**	Max4	Max5
Small-difference	100	11.50	8.43	**2.06**	0.61	0.04
	250	24.00	22.81	**2.43**	0.72	0.06
	500	42.70	47.50	**2.61**	0.75	0.06
	1000	81.11	99.55	**2.24**	0.66	0.05
Large-difference	100	30.75	33.43	**2.43**	0.69	0.06
	250	71.81	88.08	**2.30**	0.71	0.06
	500	135.93	179.54	**2.49**	0.72	0.06
	1000	265.38	367.34	**2.24**	0.66	0.05
Mixed-size-difference	100	29.96	19.37	**2.26**	0.65	0.06
	250	74.79	47.45	**2.30**	0.71	0.06
	500	147.45	93.07	**2.49**	0.72	0.06
	1000	295.44	186.91	**2.24**	0.66	0.05
Nonuniform-difference	100	30.70	14.56	**2.21**	0.65	0.06
	250	70.70	39.39	**2.30**	0.71	0.06
	500	136.72	81.98	**2.49**	0.72	0.06
	1000	267.65	169.29	**2.24**	0.66	0.05

Note. The bold values indicates the largest modification index values when there is no more noninvariant intercept left in the model.