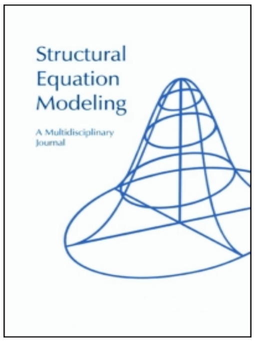


Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: http://www.tandfonline.com/loi/hsem20

Two-Step Approach to Partial Factorial Invariance: Selecting a Reference Variable and Identifying the Source of Noninvariance

Eunju Jung & Myeongsun Yoon

To cite this article: Eunju Jung & Myeongsun Yoon (2017) Two-Step Approach to Partial Factorial Invariance: Selecting a Reference Variable and Identifying the Source of Noninvariance, Structural Equation Modeling: A Multidisciplinary Journal, 24:1, 65-79, DOI: 10.1080/10705511.2016.1251845

To link to this article: http://dx.doi.org/10.1080/10705511.2016.1251845

Published online: 29 Nov 2016.

Structural Equation Modeling: A Multidisciplinary Journal, 24: 65–79, 2017
Copyright © Taylor & Francis Group, LLC
ISSN: 1070-5511 print / 1532-8007 online
DOI: 10.1080/10705511.2016.1251845

Routledge
Taylor & Francis Group

Two-Step Approach to Partial Factorial Invariance: Selecting a Reference Variable and Identifying the Source of Noninvariance

Eunju Jung[1] and Myeongsun Yoon[2]
[1]*Indiana University*
[2]*Texas A & M University*

To date, no effective empirical method has been available to identify a truly invariant reference variable (RV) in testing measurement invariance under a multiple-group confirmatory factor analysis. This study proposes a method that, in selecting an RV, uses the smallest modification index (min-mod). The method's performance is evaluated using 2 models: (a) a full invariance model, and (b) a partial invariance model. Results indicate that for both models the min-mod successfully identifies a truly invariant RV (Study 1). In Study 2, we use the RV found in Study 1 to further evaluate the performance of item-by-item Wald tests at locating a noninvariant variable. The results indicate that Wald tests overall performed better with an RV selected in a partial invariance model than an RV selected in a full invariance model, although in certain conditions their performances were rather similar. Implications and limitations of the study are also discussed.

Keywords: modification index, multiple-group confirmatory factor analysis, partial factorial invariance, reference variable

It is common in educational and psychological research to compare measured constructs across different groups (e.g., different countries, time points, languages, or modes of test delivery; Cheung & Rensvold, 1999; McArdle, 2009; Meredith, 1993; Schmitt & Kuljanin, 2008; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000; Widaman & Reise, 1997). As a necessary condition for making cross-group comparisons of observed scores on measured variables, the establishment of measurement invariance beforehand has been seriously valued and increasingly tested over the last two decades (Putnick & Bornstein, 2016; Schmitt & Kuljanin, 2008; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). *Measurement invariance* is broadly defined as the condition in which measured variables are related to latent constructs in the same way across different groups (McArdle, 2009; Meredith, 1993; Millsap, 2012; Vandenberg & Lance, 2000). The group difference in observed scores can be interpreted as originating from the group difference in latent constructs on established measurement invariance (Meredith, 1993; Millsap, 2012; Millsap & Olivera-Aguilar, 2012; Widaman & Reise, 1997, p. 282). On the other hand, the group difference in observed scores might be compromised by measurement bias unless measurement invariance has been established between the groups (Horn & McArdle, 1992; Meredith, 1993; Millsap, 2012; Millsap & Olivera-Aguilar, 2012). If it is not so established, making cross-group comparisons based on observed scores is discouraged because the measurement at hand would operate differentially across the groups.

Under a structural equation modeling (SEM) framework, a multiple-group confirmatory factor analysis (MCFA) is one of the most popular methods for testing measurement invariance (Cheung & Rensvold, 1999; McArdle, 2009; Meredith, 1993; Schmitt & Kuljanin, 2008; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000; Widaman & Reise, 1997; Yoon & Millsap, 2007). Measurement invariance under a CFA model is called *factorial invariance* (Jung & Yoon, 2016; Yoon & Millsap, 2007). Using an MCFA model, we can flexibly test the equality of factor model parameters (factor loadings,

Correspondence should be addressed to Eunju Jung, Center for Evaluation and Education Policy, Indiana University, Bloomington, IN. E-mail: jungeu@iu.edu

intercepts, and unique variances; Jung & Yoon, 2016; Schmitt & Kuljanin, 2008; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). The observed multiple indicators are linearly related to latent constructs with the factor model parameters mentioned previously. If we have k observed indicators and l latent constructs ($k > l$), an MCFA model can be expressed as follows:

$$X_m = \tau_m + {}_m\xi_m + \delta_m \tag{1}$$

In Equation 1, X represents a vector of $k \times 1$ observed indicator variables; ξ represents a vector of $l \times 1$ latent variables; Λ denotes a $k \times l$ matrix of factor loadings, which are regression weights relating observed variables to latent variables; τ stands for a $k \times 1$ vector of intercepts, which are the observed scores of the point where the latent score is zero; and δ indicates a $k \times 1$ vector of unique factor scores, which are analogous to the errors in a regression model. However, the unique factor score of a variable includes both random error and the unique contribution of the variable. The subscript m denotes an indicator for group membership, which allows factor model parameters (Λ, τ, and δ) to take different values in different groups. Using Equation 1, we can test the invariance of all factor model parameters: factor loadings (Λ_m), intercepts (τ_m), and unique variances (θ_m)—variances of unique factor scores, δ_m. As Widaman and Reise (1997) outlined, factorial invariance tests are typically conducted in a sequential manner via four steps: (a) configural invariance (equal pattern of zero and nonzero factor loadings; Horn & McArdle, 1992), (b) metric invariance (equal factor loadings; Horn & McArdle, 1992), (c) scalar invariance (equal factor loadings and intercepts; Meredith, 1993; Steenkamp & Baumgartner, 1998), and (d) strict invariance (equal factor loadings, intercepts, and unique variances; Widaman & Reise, 1997). When strict invariance is established, any observed group difference in means and variances would be due to group differences in factor means and variances. Yet achieving strict invariance is unnecessary when the major focus is to compare the mean structures across groups (Meredith, 1993; Steenkamp & Baumgartner, 1998; Widaman & Reise, 1997). The following equation expresses the mean structure of Equation 1:

$$E(X_m) = \tau_m + {}_m\kappa_m \tag{2}$$

Here, $E(X_m)$ denotes a mean vector of group m's observed indicator variables, and κ_m denotes a mean vector of group m's latent factor scores. In Equation 2, the mean vector of δ_m is not shown because the mean of δ_m is expected to eventually be zero. Equation 2 implies that cross-group mean comparison is legitimate when scalar invariance ($\Lambda_m = \Lambda_{m'}$ and $\tau_m = \tau_{m'}$) is established (Meredith, 1993; Widaman & Reise, 1997).

In addition to its flexibility in factorial invariance tests, MCFA lets us compare latent means between groups more realistically than traditional, observed score methods (Byrne, Shavelson, & Muthén, 1989; Schmitt & Kuljanin, 2008; Steenkamp & Baumgartner, 1998; Whittaker, 2013). Those methods assume no measurement errors across multiple indicators, whereas MCFA allows modeling measurement errors in the estimation of cross-group latent means (Whittaker, 2013). Moreover, MCFA can further relax the assumption of scalar invariance—a necessary condition for comparing cross-group observed means. In other words, we can compare the latent means of different groups in a partial factorial invariance model in which the invariant constraints of some factor loadings and intercepts are relaxed (Byrne, Shavelson, & Muthén, 1989; Jung & Yoon, 2016; Schmitt & Kuljanin, 2008; Steenkamp & Baumgartner, 1998; Whittaker, 2013).

When a partial factorial invariance (PFI) model is pursued, one should be careful in choosing a reference variable (RV), which has been dominantly used to identify the unknown scale of the latent variable of interest (French & Finch, 2008; Johnson, Meade, & DuVernet, 2009; Whittaker & Kojasteh, 2013; Yoon & Millsap, 2007). In an RV method, the variance–covariance structure of the latent variable is typically identified by fixing the factor loading of a selected RV at one (Whittacker & Kojasteh, 2013; Yoon & Millsap, 2007). Similarly, its mean structure is usually identified by constraining the intercept of a chosen RV to zero (Whittacker & Kojasteh, 2013). As such, the parameters of the chosen RV are assumed to be equal when we are testing factorial invariance under an MCFA model. However, the results of factorial invariance testing might be falsified if the chosen RV is not invariant across groups. Yoon and Millsap (2007) mathematically illustrated how a truly invariant variable can appear to be noninvariant, and vice versa, when one is using a noninvariant variable as an RV. In addition, Johnson et al. (2009) conducted a simulation study in which the role of an RV was investigated for factorial invariance testing. They found that a noninvariant RV affected the results of item-level factorial invariance tests by inflating Type I error rates, whereas scale-level factorial invariance tests were robust to the selection of a noninvariant RV. Nevertheless, the problem of choosing a truly invariant RV has seldom been acknowledged in the applied factorial invariance literature. For example, Johnson et al. (2009) reviewed studies testing partial factorial invariance from 2005 and 2007, and they found that only 4.6% of the factorial invariance studies (N = 153) heeded the problem of noninvariant RVs. Such a low rate is unsurprising in light of the scarcity of effective empirical methods for differentiating between invariant and noninvariant variables. It has been pointed out that no empirical method is currently available to accurately identify an invariant RV (Raykov, Marcoulides, & Li, 2012). Although theories can guide applied researchers in selecting an appropriate RV, such theories are no guarantee either (Millsap & Olivera-Aguilar, 2012).

Methods for Partial Factorial Invariance Without Choosing a Reference Variable

In the factorial invariance literature, there are noteworthy studies that have addressed RV selection problems in testing factorial invariance. First, Rensvold and Cheung (1998) proposed the factor-ratio test for detecting the violation of the factorial invariance without choosing an RV. In the factor-ratio test, every variable serves as an RV while all the remaining variables are being tested. Suppose that there are five variables: X1, X2, X3, X4, and X5. When X1 is selected as an RV, then X2, X3, X4, and X5 are tested for invariance. The major statistical procedure is to compare an unconstrained model (i.e., configural invariance model identified with X1) with a constrained one (i.e., model with one more equality constraint for each of the remaining variables in addition to the configural invariance model) using the difference in the chi-square fit statistics ($\Delta\chi^2$) with 1 *df.* A well-known drawback of the factor-ratio test is that the procedure becomes cumbersome with more indicators because it needs $\{k \times (k - 1)\}/2$ tests for the cases with k indicators (Cheung & Lau, 2011; French & Finch, 2008; Jung & Yoon, 2016; Whittaker & Khojasteh, 2013; Yoon & Millsap, 2007). Detailed procedures on how to conduct the factor-ratio test can be found in several studies (Cheung & Lau, 2011; Cheung & Rensvold, 1999; Rensvold & Cheung, 1998). After a decade from its proposal, French and Finch (2008) evaluated the performance of the factor-ratio test under various simulation conditions. They reported that the factor-ratio test performed adequately, with high power and low false positive rates.

Second, Yoon and Millsap (2007) suggested using the largest modification index sequentially (sequential max-mod) to isolate the source of noninvariance. Under various simulated conditions, they systematically evaluated the performance of the sequential use of the largest modification index. Yoon and Millsap used a fully constrained metric invariance model. The model's variance–covariance structure was identified by fixing the variance of the first group to 1 while freely estimating the variance of the second group with one set of equally constrained factor loadings of the corresponding items between groups. Instead of looking at all retrieved modification indexes, Yoon and Millsap focused on the modification indexes related to the equally constrained factor loadings. They sequentially relaxed the equality constraint of the factor loadings indicated by the largest modification index until there was no modification index greater than the cutoff value (3.84). The results showed that sequential max-mod worked promisingly under ideal conditions (data with low contaminations, large differences in noninvariance variables, and large samples). However, the modification index is known for its inflated false positive rates with large sample sizes and misspecification in the model (Whittaker, 2012; Yoon & Millsap, 2007).

Finally, Raykov, Marcoulides, and Millsap (2013) demonstrated how to use a multiple testing method for factorial invariance. Their method compares two models using the difference in chi-square fit statistics. The baseline model is a fully constrained invariance model in which, for example, every factor-loading pair of like items has an invariant constraint across groups. In the tested model, one pair of factor loadings is relaxed from the fully constrained model. To test all variables in the model, the number of tests needed corresponds to the number of variables. Due to the nature of the multiple testing methods, an inflated false positive is likely to happen. Raykov et al. (2013) adjusted the significance value using the Benjamini–Hochberg multiple testing procedures and demonstrated how to employ that method to test factor loading invariance. However, their study is based on only one simulated data set, and the performance of their proposed method is yet to be evaluated systematically under various data conditions.

Simulation Studies Directly Comparing the Methods for Partial Invariance

Whittaker and Khojasteh (2013) directly compared the methods suggested by Cheung and Rensvold (1999) and Yoon and Millsap (2007) under various partial metric invariance conditions. They varied the study conditions by manipulating the number of indicator variables, sample sizes, magnitude of factor loading differences, frequency of noninvariant indicator variables, and pattern of factor loading differences. The outcome variables of the studies were true positive rates and true negative rates in identifying truly invariant and noninvariant indicators. They noted that each of the methods performed more ideally under certain circumstances and had a different pattern of accuracy. To briefly summarize their findings, the sequential max-mod method had lower false negative rates but higher false positive rates than the factor-ratio test.

Most recently, Jung and Yoon (2016) proposed a forward method using confidence intervals (forward CI method) to address the problems of the other methods (i.e., inflated false positive rate of the sequential use of max-mod and labor-intensive procedure of the factor-ratio test). By the nature of its procedure, it is less susceptible to inflating false positive rates because it compares the model with no invariant constraint with the model with one invariant constraint (Jung & Yoon, 2016). In addition, the multiple model comparisons were simplified using the CIs of newly constructed variables, which are the difference of the tested parameters for invariance between groups. Their study directly compared the performances of the sequential use of max-mod, the factor-ratio test, and the forward CI method. Jung and Yoon added more layers to the previous studies by including partial scalar invariance conditions, by simplifying the procedure of the factor-ratio test using the bias-corrected bootstrapping CIs, and by adding a more conservative cutoff value for the sequential use of max-mod. In addition, they reported model-level false positive and false negative rates, supplying more rigorous standards with which to investigate the relative efficacy of the compared methods. The performance of the three methods was compared under various conditions with different locations of noninvariance (factor loading or scalar), varying degrees and pattern of noninvariance, and sample sizes. The results demonstrated that the forward CI method generally outperformed the sequential max-mod and the factor-ratio test in terms of perfect recovery rates—without both false positives and false negatives. The sequential max-mod performed comparably with the forward method when using conservative criterion (modification index value = 6.645) rather than the typically used criterion (modification index value = 3.841). Using the model-level false negative rates, Jung and Yoon's study was able to catch extremely high false negative rates of the factor-ratio test under the conditions of two noninvariant variables with the same degree of difference in the same direction—rates that previous studies had been unable to detect (French & Finch, 2008; Whittaker & Khojasteh, 2013); these studies had reported only item-level false positive and false negative rates. Although the forward method has shown promise, it requires a prespecified RV, which should be truly invariant across groups. As noted earlier, though, there has been little guidance on how to choose a truly invariant RV empirically.

Proposed Method to Identify a Truly Invariant Reference Variable

The value of a modification index indicates the approximate decrement in the chi-square statistics (with 1 *df*) when the indicated constrained parameter is released (Jöreskog & Sörbom, 1993; Muthén & Muthén, 1998–2012). In factorial invariance studies, the modification index has been used to detect noninvariant variables. Based on the prespecified criterion (e.g., 3.84), a variable with a modification index exceeding the criterion was treated as a noninvariant variable, and one with a modification index below the criterion was treated as invariant (Jung & Yoon, 2016; Whittaker & Khojasteh, 2013; Yoon & Millsap, 2007). We hypothesized that the smallest modification index (min-mod) would indicate the smallest difference in the set of constrained parameters for invariance testing, using the idea of the "all others as anchors" (AOAA) method in the item-response theory literature (Meade & Wright, 2012; Woods, 2009). In AOAA, every variable is tested for invariance by comparing two models—one with fully constrained invariance parameters and the other with a single freely estimated parameter and all others fixed as invariant. The model fits of every pair of the fully constrained invariance model and of the less constrained model are compared using likelihood ratio tests; the necessary number of likelihood ratio tests is equal to the number of variables. If the result of a likelihood ratio test is not statistically significant, the tested variable is considered to be invariant. Compared to AOAA, which requires multiple likelihood ratio tests, using min-mod is a much simpler procedure because we need only one data analysis phase for testing a fully constrained invariance model while retrieving all modification indexes of the parameters tested for invariance. However, the value of the modification index in the misspecified model is expected to be less trustworthy than that in the model without misspecification (Whittaker, 2012). Therefore, we chose two baseline models for the purpose of comparison. In one, all the invariance parameters of interest (factor loadings or intercepts) are equally constrained in all groups. Hereafter, this model is referred to as the full invariance model. The other has no significant modification index (3.84) and is produced by sequentially relaxing the invariant constraint with the max-mod from the fully constrained invariance model. Hereafter, this second model is referred to as the partial invariance model. A full invariance model was simpler to use because it required only one data analysis. A partial invariance model was deemed to be less susceptible to the problems arising from misspecification in the model, however.

Study Aim

The primary aim of the study is to propose an empirical method to identify a truly invariant RV using the smallest modification index (min-mod) and to evaluate the accuracy of this method under various partial factorial invariance conditions. It is also important to correctly locate the source of noninvariance using the chosen RV. Thus, we extend the study to evaluate the performance of item-by-item Wald tests using the empirically chosen RV to detect the violation of factorial invariance in terms of perfect recovery rates, power, and false positive rates.

METHOD

We conducted Monte Carlo simulation studies to gauge how accurately the smallest modification index identified a truly invariant RV (Study 1) and to examine the performance of item-by-item Wald tests (Study 2). We generated various partial metric and scalar invariance data conditions using M*plus* 7.0 (Muthen & Muthen, 1998–2012). The detailed generated data conditions are presented next.

Simulated Data Conditions

To balance simplicity with comprehensiveness, some data conditions were fixed and others of more interest were manipulated. We considered only two groups, fixing the number of factors to one and the number of variables to six under all data conditions. In addition, we allowed only two variables to be noninvariant under all data conditions. To obtain stable results, we also set the number of replications to 1,000 under all conditions. For the data generation, we selected three conditions to vary: (a) location of noninvariance, (b) size and pattern of noninvariance, and (c) sample size. Table 1 summarizes the data conditions with respect to the manipulated factors.

Location of noninvariance

An RV should be selected for either factor loadings or intercepts in testing scalar invariance. Although scalar invariance is a necessary condition for testing cross-group latent mean comparisons, studies that investigate methods of testing factorial invariance have more often selected noninvariance in factor loading (French & Finch, 2008; Johnson et al., 2009; Whittaker & Khojasteh, 2013; Yoon & Millsap, 2007). In this study, we imposed noninvariance on both factor loadings and intercepts to evaluate the performance of the min-mod at identifying a truly invariant RV for metric invariance and scalar invariance tests. To simplify the analysis, however, we did not consider conditions under which noninvariance was manipulated for both factor loadings and intercepts.

Size and pattern of noninvariance

To examine the performance of the min-mod comprehensively, we generated four levels of the size and pattern of noninvariance. Depending on the magnitude and direction of the differences, these were (a) small-magnitude, (b) large-magnitude, (c) mixed-magnitude, and (d) mixed-direction conditions. The terms *small magnitude* and *large magnitude* do not imply any absolute magnitude values; we used the terms simply to examine the relative effects of different magnitudes of noninvariance in our study conditions.

Under all partial factor loading invariance conditions, the second and third factor loadings (λ_2 and λ_3) were manipulated to be noninvariant. Also serving as noninvariant parameters, under all partial intercept invariant conditions, were the second and third intercepts (τ_2 and τ_3). Under the small-magnitude condition, the difference between the manipulated parameters in the two groups was set to .2 ($\lambda_{m2} - \lambda_{m'2} = \lambda_{m3} - \lambda_{m'3} = .2$; $\tau_{m2} - \tau_{m'2} = \tau_{m3} - \tau_{m'3} = .2$). Under the large-magnitude condition, the difference between the manipulated parameters was .4 ($\lambda_{m2} - \lambda_{m'2} = \lambda_{m3} - \lambda_{m'3} = .4$; $\tau_{m2} - \tau_{m'2} = \tau_{m3} - \tau_{m'3} = .4$). Under the mixed-magnitude condition, two noninvariant parameters had varying degrees of difference between the groups ($\lambda_{m2} - \lambda_{m'2} = .3$, $\lambda_{m3} - \lambda_{m'3} = .5$; $\tau_{m2} - \tau_{m'2} = .3$, $\tau_{m3} - \tau_{m'3} = .5$). Under the mixed-direction condition, two noninvariant parameters had the same magnitude of difference in opposite directions ($\lambda_{m2} - \lambda_{m'2} = .3$; $\lambda_{m3} - \lambda_{m'3} = -.3$; $\tau_{m2} - \tau_{m'2} = .3$, $\tau_{m3} - \tau_{m'3} = -.3$).

Sample size

Whereas the ratio of sample sizes among the groups was held constant for simplicity, we manipulated the total sample sizes into four levels: $N_{\text{total}} = 200$, 400, 1,000, and 2,000. The resulting sample size per group was $N = 100$, 200, 500, or 1,000. This was done to include a wide range of sample sizes in real research settings.

TABLE 1
Simulation Conditions

		Group 2			
Group 1		Small Magnitude	Large Magnitude	Mixed Magnitude	
Factor loading					
λ_{x1}	0.90	0.90	0.90	0.90	0.90
$\boldsymbol{\lambda_{x2}}$	0.70	0.50	0.30	0.40	0.40
$\boldsymbol{\lambda_{x3}}$	0.60	0.40	0.20	0.10	0.90
λ_{x4}	0.80	0.80	0.80	0.80	0.80
λ_{x5}	0.70	0.70	0.70	0.70	0.70
λ_{x6}	0.60	0.60	0.60	0.60	0.60
Intercept					
τ_{x1}	−0.15	−0.15	−0.15	−0.15	−0.15
$\boldsymbol{\tau_{x2}}$	0.25	0.05	−0.15	−0.05	−0.05
$\boldsymbol{\tau_{x3}}$	0.15	−0.05	−0.25	−0.35	0.45
τ_{x4}	−0.25	−0.25	−0.25	−0.25	−0.25
τ_{x5}	−0.10	−0.10	−0.10	−0.10	−0.10
τ_{x6}	0.10	0.10	0.10	0.10	0.10
Unique variances					
ε_{x1}—ε_{x6}	0.30	0.30			
Factor variance					
φ	1.00	1.30			
Factor mean					
κ	0.00			0.30	

Note. All conditions presented in the table have four levels of sample size ($N = 100$, 200, 500, and 1,000 per group).

Analytic Procedure

Study 1: Selecting a reference variable using the smallest modification index

In Study 1, we evaluated the accuracy of min-mod in identifying a truly invariant RV. As references to the min-mod, we selected the following two models: (a) a full invariance model, and (b) a partial invariance model. An RV for factor loadings was selected only in the conditions that imposed noninvariance on factor loadings. To select an RV for intercepts, we used the data conditions under which only intercepts were manipulated for noninvariance.

As an outcome, we examined the rate of accuracy in choosing a truly invariant variable as an RV. When the chosen factor loading or intercept belonged to the truly invariant parameter group, it was coded as accurate. When it did not, it was coded as inaccurate. The mean accuracy was calculated across 1,000 replications for each condition, by sample size and with respect to the location of noninvariance, as follows:

$$\text{Mean accuracy} = \frac{\text{Accurate cases}}{\text{Total cases } (= 1000)}$$

We also examined the congruency level of the chosen RV using the min-mod under each model. We calculated the level by dividing the number of cases in which the chosen RVs were the same by the number of replications (N = 1,000), as follows:

$$\text{Congruency level} = \frac{\begin{array}{c}\text{Cases in which the same variable}\\ \text{was chosen as an RV}\end{array}}{\text{Total cases } (= 1000)}$$

Study 2: Locating noninvariant variables

The fundamental reason for choosing an invariant RV is to correctly differentiate noninvariant variables from invariant variables when full factorial invariance is rejected. As a subsequent procedure, to locate the source of noninvariance, we conducted item-by-item Wald tests. The baseline model for the Wald test was a configural invariance model that had one equality constraint for identification across groups using the chosen RV. The compared model had one more equality constraint across groups for testing invariance. Because we conducted multiple Wald tests, to test every variable except the chosen RV, we used the critical value adjusted by the Bonferroni correction. The resulting critical value was α = 0.01 because we divided α = 0.05 by five, which is the number of variables we tested. We simplified the process of carrying out multiple Wald tests by using the “Model Test” command (see Appendix A).

For outcome variables, we calculated perfect recovery rates, power, and false positive rates in the final specified model. Perfect recovery is a case in which truly invariant variables are not detected as noninvariant and truly noninvariant variables are detected as such. In other words, a perfectly recovered model includes neither false positives nor false negatives. We calculated the perfect recovery rate by dividing the number of perfectly recovered cases by 1,000 (the number of replications). To isolate the source of imperfect recovery, we also calculated power and false positive rates through 1,000 replications under each condition.

RESULTS

Study 1: Selecting an Invariant Reference Variable Using the Min-Mod

To identify an invariant RV, we used the min-mod in two models: (a) a full invariance model, in which every invariance parameter was equally constrained in all groups, and (b) a partial invariance model without any significant modification index (> 3.84). We refer to the selected RV in a full invariance model as an RVF and the selected RV in a partial invariance model as an RVP. Table 2 shows the accuracy of the min-mod according to the location of noninvariance, size and pattern of noninvariance, and sample size. Overall, its accuracy was nearly perfect regardless of the model except for some conditions with N = 100. The average accuracy of RVF in identifying a truly invariant factor loading was 0.998 (*SD* = 0.009) and that of RVP was 0.994 (*SD* = 0.016). Except for some conditions with small samples combined with small magnitude of differences, both methods perfectly or almost perfectly identified a truly invariant factor loading as an RV. Both methods had the lowest accuracy in the small-magnitude condition with N = 100. However, the error rate was still very low for both methods (RVF = 3.5%, RVP = 6.1%). In identifying a truly invariant intercept as an RV, the average accuracy of RVF was 0.995 (*SD* = 0.014) and that of RVP was 0.994 (*SD* = 0.016). In comparison to the baseline accuracy rate of guessing (0.667, given two noninvariant variables among the six variables), the accuracies of the RVF and RVP were very high. We could observe a similar pattern of accuracy in identifying an RV for factor loading. Both methods showed the lowest accuracy in the small-magnitude condition with N = 100 while maintaining perfect or almost perfect accuracy in most conditions. The maximum error rates were 5.5% and 7.1% for RVF and RVP.

We were also interested in whether the selected RV in either model was congruent. Table 2 shows the congruency levels of the RVF and RVP. The average congruency in selecting an RV for factor loadings was 0.192 (*SD* = 0.115), and for intercepts 0.251 (*SD* = 0.187). Interestingly, the congruency was higher under the mixed-direction condition than under the others. In the mixed-direction condition, the average congruency for factor loading was 0.346 (*SD* = 0.104), and for intercepts 0.547 (*SD* = 0.042).

TABLE 2
Accuracy of Selecting a Truly Invariant Reference Variable

		Accuracy		
	Sample Size	RVF	RVP	Congruency
Factor loading				
Small magnitude	100	0.965	0.939	0.336
	200	0.999	0.992	0.123
	500	1.000	1.000	0.069
	1,000	1.000	1.000	0.117
Large magnitude	100	1.000	1.000	0.125
	200	1.000	1.000	0.094
	500	1.000	1.000	0.134
	1,000	1.000	1.000	0.171
Mixed magnitude	100	0.999	0.999	0.132
	200	1.000	1.000	0.082
	500	1.000	1.000	0.137
	1,000	1.000	1.000	0.176
Mixed direction	100	1.000	0.980	0.441
	200	1.000	1.000	0.443
	500	1.000	1.000	0.328
	1,000	1.000	1.000	0.208
Intercept				
Small magnitude	100	0.945	0.929	0.300
	200	0.996	0.991	0.097
	500	1.000	0.998	0.092
	1,000	1.000	1.000	0.147
Large magnitude	100	0.999	0.996	0.059
	200	1.000	0.999	0.119
	500	1.000	1.000	0.167
	1,000	1.000	1.000	0.232
Mixed magnitude	100	0.988	0.999	0.083
	200	0.999	1.000	0.140
	500	1.000	1.000	0.175
	1,000	1.000	1.000	0.220
Mixed direction	100	1.000	0.992	0.483
	200	1.000	1.000	0.567
	500	1.000	1.000	0.564
	1,000	1.000	1.000	0.572

Note. N = sample size per group; RVF = reference variable selected under the full invariance model; RVP = reference variable selected under the partial invariance model. The baseline accuracy rate by guessing is 0.667 given four invariant factor loadings among the six variables.

Figure 1 shows how each model selected an RV. The min-mod using a full invariance model tended to select more variables with lower factor loadings for both factor loading and intercept conditions than the variables with higher factor loadings, except under the mixed-direction conditions. This tendency was stronger with larger samples. However, the min-mod using a partial invariance model tended to select every invariant variable as an RV proportionately.

Study 2: Locating Noninvariant Variables Using Wald Test

Although the min-mod performed adequately in selecting a truly invariant RV, our ultimate objective in testing factorial invariance was to locate the source of noninvariance. In Study 2, we tested every variable for invariance through a Wald test with one of the selected RVs and the min-mod found on the different models. We examined the perfect recovery rate of each approach to see how each method perfectly recovered the original data conditions. We also examined the power and false positive rates of each approach. Figure 2 also demonstrates perfect recovery rates, power, and false positive rates Wald test with an RVF (Wald-RVF) and with an RVP (Wald-RVP).

Perfect recovery rates

Table 3 presents the perfect recovery rates of Wald-RVF and Wald-RVP, by the location of noninvariance, size and pattern of noninvariance, and sample size. When the noninvariance variables were located in factor loadings, the average perfect recovery rates of Wald-RVF and Wald-RVP were 0.779 (SD = 0.295) and 0.841 (SD = 0.266). When the noninvariant variables existed in intercepts, the mean perfect recovery rates of Wald-RVF and Wald-RVP were 0.719 (SD = 0.325) and 0.817 (SD = 0.273), respectively. Wald-RVP had a higher perfect recovery rate than Wald-RVF in most conditions. Both methods generally showed the lowest perfect recovery rates across all conditions with N = 100.

Power

Table 4 shows the power of each method to detect noninvariant variables according to the location of noninvariance, size and pattern of noninvariance, and sample size. For factor-loading noninvariance, the mean powers of Wald-RVF and Wald-RVP were 0.805 (SD = 0.308) and 0.852 (SD = 0.269), respectively. For intercept noninvariance, the mean powers of Wald-RVF and Wald-RVP were 0.736 (SD = 0.342) and 0.828 (SD = 0.277). Overall, the Wald-RVF had the lower average power. This was due mainly to lower powers in the conditions with small magnitudes of difference and smaller sample sizes than the Wald-RVP.

False positives

Table 5 gives the false positive rates of Wald-RVF and Wald-RVP. When noninvariance was manipulated for factor loadings, the average false positive rates of Wald-RVF and Wald-RVP were 0.030 (SD = 0.022) and 0.015 (SD = 0.006), respectively. When noninvariance was imposed in intercepts, the mean false positive rates of Wald-RVP were 0.034 (SD = 0.023) and 0.019 (SD = 0.011). Wald-RVP had lower false positive rates than Wald-RVF under most conditions other than the mixed-direction one. Although Wald-RVP had false positive rates below the nominal level ($\alpha = 0.05$) under all conditions, Wald-RVF also showed promising false positive rates. Although its false positive rates were greater than the nominal levels in some conditions, they were very close to those levels.

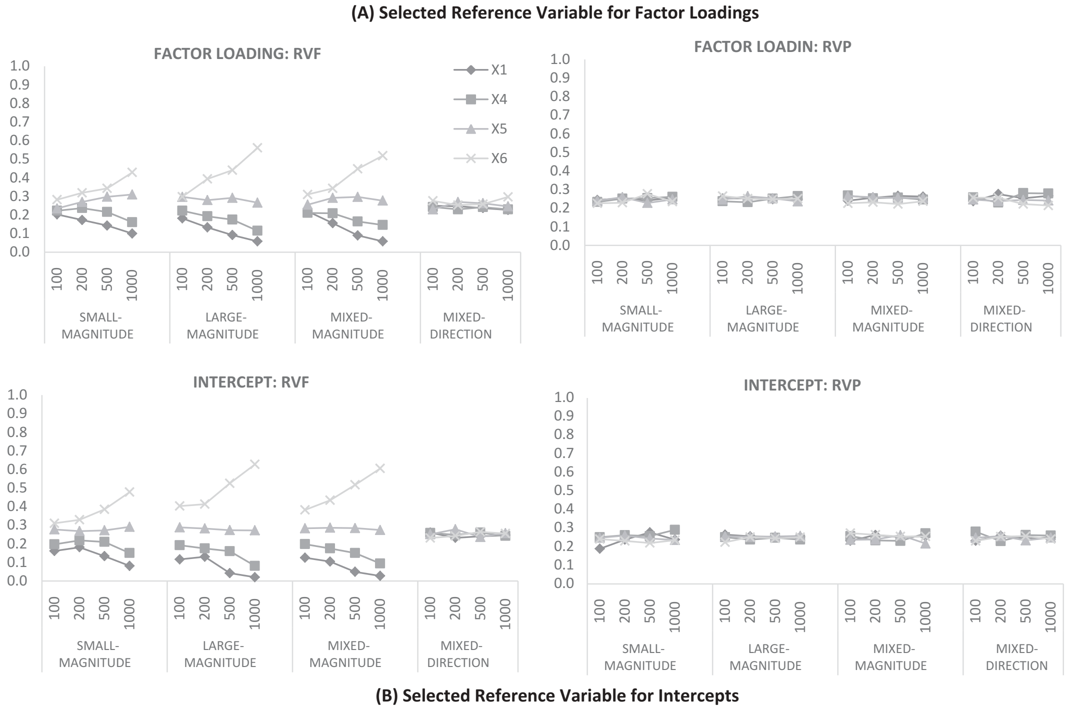


FIGURE 1 Selected reference variable. *Note*. RVF = the selected reference variable under the fully constrained invariance model; RVP = the selected reference variable under the partially constrained invariance mode.

DISCUSSION

The establishment of measurement invariance has been widely practiced among educational and psychological researchers (Sass, 2011). In particular, full scalar invariance has been generally accepted as a prerequisite for cross-group mean comparisons (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). When the measurement at hand does not hold scalar invariance for all items, it might be a more feasible decision for applied researchers to use the imperfect measurement than to throw away the whole test. There are two options: (a) comparing latent means made by only the subset of the items that were found to be invariant, and (b) comparing latent means made by all items in an MCFA model in which the invariant constraint of the identified noninvariant items is relaxed. Whichever option is chosen, accurate detection of invariant and noninvariant items is of substantial importance. As Johnson et al. (2009) demonstrated, a truly invariant RV plays a crucial role in doing this. The majority of applied researchers, however, have conducted factorial invariance with little caution about how to choose a truly invariant RV (Johnson et al., 2009). In addition, no empirical method is yet available for accurately selecting a truly invariant RV (Raykov et al., 2012). The results suggest that this study adequately fills the gap in the factorial invariance literature.

In this study, we proposed a two-step approach for testing factorial invariance. In Study 1, we examined the performance of the smallest modification index (min-mod) in identifying a truly invariant RV. The min-mod was examined using two models: a full invariance model and a partial invariance model. The results indicated that the min-mod shows promise for selecting a truly invariant RV. Generally, it successfully selected a truly invariant RV using both models. Only very low error rates were found under some conditions with small sample sizes (N = 100 and 200) in combination with small differences. One interesting finding was that the congruency level of selected RVs from the two models was quite low, except under the mixed-direction conditions. That is, the min-mod tended to select a different RV depending on the degree of misspecification in a given model. As shown in Figure 1, the min-mod using a full invariance model was more likely to choose the variable with the smallest factor loading (X6) than the variable with the largest factor loading (X1), whereas the min-mod using a partial invariance model tended to choose every invariant variable more evenly than its counterpart. One possible explanation for this is that the size of the factor loading is more likely to contribute to the size of the corresponding modification index when a certain degree of misspecification exists (as in a full invariance model) than when misspecification is rare (as in a partial invariance model).

In Study 2, we explored the performance of item-level Wald tests in detecting noninvariant variables using the RV selected in the first step. As shown in Figure 1, the Wald tests using a selected RV in a partial invariance model

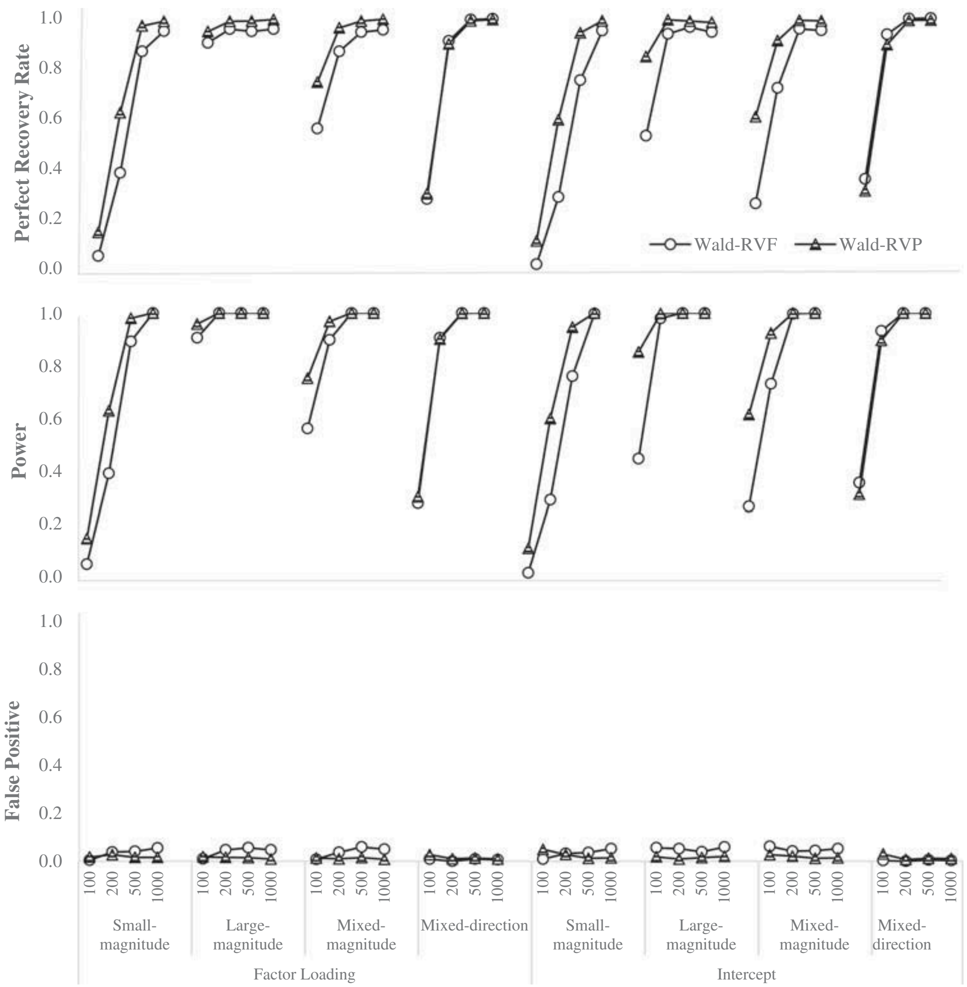


FIGURE 2 Perfect recovery rates, power, and false positive rates. *Note*. Wald-RVF = Wald test using a reference variable selected in a fully constrained invariance model; Wald-RVP = Wald test using a reference variable selected in a partially constrained invariance mode.

(Wald-RVP) generally showed higher perfect recovery rates than the Wald tests using an RV selected in a full invariance model (Wald-RVF). Both methods seem to work inadequately for most of the conditions with $N = 100$. Although the differences in the perfect recovery rates between Wald-RVF and Wald-RVP appeared not to be very prominent, the two methods yielded quite different perfect recovery rates under some conditions, with small samples combined with small differences. In the partial metric invariance conditions, the largest differences were found in the small-magnitude condition with $N = 200$ where Wald-RVP had 24% higher perfect recovery rates than Wald-RVF. In the partial scalar invariance conditions, the greatest difference was found in the mixed-magnitude condition with $N = 100$ in which the perfect recovery rate of Wald-RVP was 35% higher than that of Wald-RVP. Under those conditions, the major sources of differences in the perfect recovery rates were false negative rates. That is, the Wald-RVF had higher false negative rates than did the Wald-RVP, a fact that is associated with the size of the factor loading of the selected RV. As noted earlier, the chosen RV was more likely to be the variable with the smallest factor loading when a full invariance model rather than a partial invariance model was used. As far as the association between the selected RV and the false negative rates is concerned, we found that the false negative rates were higher in general when the chosen RV had a smaller factor loading.

To examine the source of differences in the false negative rates between the Wald-RVF and the Wald-RVP, we conducted a post hoc analysis in which we closely analyzed the pattern of false negative rates under only some conditions with substantial differences in false negative rates (e.g., small-magnitude factor loading conditions with $N = 200$ and $N = 500$, and small-magnitude intercept conditions with $N = 200$ and $N = 500$). However, we excluded the conditions of $N = 100$ because it is believed that $N = 200$ is the minimum sample size necessary to achieve adequate power (as discussed in previous studies; e.g., Kim &

TABLE 3
Perfect Recovery Rates

	Sample Size	Wald-RVF	Wald-RVP
Factor loading			
Small magnitude	100	0.116	0.237
	200	0.379	0.619
	500	0.864	0.966
	1,000	0.945	0.984
Large magnitude	100	0.894	0.919
	200	0.953	0.984
	500	0.944	0.985
	1,000	0.953	0.991
Mixed magnitude	100	0.676	0.792
	200	0.864	0.958
	500	0.941	0.985
	1,000	0.95	0.992
Mixed direction	100	0.192	0.238
	200	0.906	0.894
	500	0.991	0.986
	1,000	0.994	0.991
Intercept			
Small magnitude	100	0.014	0.107
	200	0.282	0.591
	500	0.749	0.938
	1,000	0.947	0.986
Large magnitude	100	0.527	0.843
	200	0.934	0.990
	500	0.961	0.985
	1,000	0.941	0.98
Mixed magnitude	100	0.257	0.603
	200	0.718	0.907
	500	0.954	0.988
	1,000	0.948	0.986
Mixed direction	100	0.355	0.307
	200	0.932	0.891
	500	0.995	0.988
	1,000	0.997	0.989

Note. Wald-RVF = Wald test using a reference variable selected in a fully constrained invariance model; Wald-RVP = Wald test using a reference variable selected in a partially constrained invariance model.

TABLE 4
Power to Detect Noninvariant Variables

	Sample Size	Wald-RVF	Wald-RVP
Factor loading			
Small magnitude	100	0.047	0.144
	200	0.390	0.628
	500	0.892	0.980
	1,000	1.000	1.000
Large magnitude	100	0.906	0.958
	200	1.000	1.000
	500	1.000	1.000
	1,000	1.000	1.000
Mixed magnitude	100	0.561	0.752
	200	0.898	0.968
	500	1.000	1.000
	1,000	1.000	1.000
Mixed direction	100	0.277	0.301
	200	0.906	0.901
	500	1.000	0.999
	1,000	1.000	1.000
Intercept			
Small magnitude	100	0.014	0.107
	200	0.290	0.599
	500	0.760	0.946
	1,000	0.998	1.000
Large magnitude	100	0.446	0.852
	200	0.980	0.998
	500	1.000	1.000
	1,000	1.000	1.000
Mixed magnitude	100	0.265	0.614
	200	0.731	0.924
	500	0.997	1.000
	1,000	1.000	1.000
Mixed direction	100	0.355	0.311
	200	0.932	0.896
	500	1.000	1.000
	1,000	1.000	1.000

Note. Wald-RVF = Wald test using a reference variable selected in a fully constrained invariance model; Wald-RVP = Wald test using a reference variable selected in a partially constrained invariance model.

Yoon, 2011; MacCallum, Widaman, Zhang, & Hong, 1999; Meade & Bauer, 2007; Meade, Johnson, & Braddy, 2008). We also focused on the conditions under which the chosen RVs were not congruent and the results of the Wald-RVF and Wald-RVP differed in false negative rates. For example, each method chose different variables as RVs in 87.7% of the small-magnitude factor loading conditions with $N = 200$ and the small-magnitude intercept conditions with $N = 200$. Under those conditions, the Wald tests yielded different results in 254 and 315 cases, respectively. Specifically, the Wald-RVF could not detect noninvariant variables, but the Wald-RVP detected them in the majority of cases (246 of 254 cases in the small-magnitude factor loading condition with $N = 200$; 312 of 315 cases in the small-magnitude intercept conditions with $N = 200$). In sum, the Wald-RVF was prone to having an RV with a smaller factor loading, which in turn might lead to higher false negative rates than the Wald-RVP.

As another post hoc investigation, we explored the effect of the number of RVs for the condition with the smallest perfect recovery rates (mainly due to the low power). We conducted Wald tests using two RVs (those with the smallest and second-smallest measurement invariances), which were selected in a partial invariance model. In the small-magnitude partial metric invariance condition with $N = 200$, the perfect recovery rate was 0.665 (only a 4.6% increase over the Wald test with one RV). Although the false negative rates decreased approximately 20.5%, the false positive rates increased 17.6%. In the small-magnitude partial scalar invariance condition with $N = 200$, the perfect recovery rate was 0.591 (only a 0.2% increase over the Wald test with one RV). Although the false negative rates decreased approximately 2.2%, the false positive rates increased 2.1%. Therefore, it is not worth using more than one RV to test partial factorial invariance. Our findings indicate that, for higher power, the magnitude of the factor loading of the selected RV is more important than the number of RVs.

TABLE 5
False Positive Rates

	Sample Size	Wald-RVF	Wald-RVP
Factor loading			
Small magnitude	100	0.004	0.017
	200	0.038	0.027
	500	0.041	0.016
	1,000	0.055	0.016
Large magnitude	100	0.010	0.018
	200	0.047	0.016
	500	0.056	0.015
	1,000	0.047	0.009
Mixed magnitude	100	0.008	0.013
	200	0.037	0.010
	500	0.059	0.015
	1,000	0.050	0.008
Mixed direction	100	0.009	0.027
	200	0.000	0.010
	500	0.009	0.013
	1,000	0.006	0.009
Intercept			
Small magnitude	100	0.009	0.049
	200	0.032	0.028
	500	0.036	0.012
	1,000	0.052	0.014
Large magnitude	100	0.056	0.018
	200	0.052	0.009
	500	0.039	0.015
	1,000	0.059	0.020
Mixed magnitude	100	0.062	0.027
	200	0.042	0.021
	500	0.044	0.012
	1,000	0.052	0.014
Mixed direction	100	0.003	0.030
	200	0.000	0.007
	500	0.005	0.012
	1,000	0.003	0.011

Note. Wald-RVF = Wald test using a reference variable selected in a fully constrained invariance model; Wald-RVP = Wald test using a reference variable selected in a partially constrained invariance model.

Additionally, we examined the commonly reported alternative fit indexes (AFIs): comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). Appendix C presents the average CFIs, RMSEAs, and SRMRs of the configural invariance model and the models with one set of equally constrained factor loadings for the tested variable with respect to the partial metric invariance conditions. Appendix D provides the average CFIs, RMSEAs, and SRMRs of the metric invariance model and the model with one pair of an equally constrained intercept for the tested variable by the partial scalar invariance condition. As expected, there were obvious drops in the CFIs and increments in both RMSEAs and SRMRs when either factor loadings or intercepts of the noninvariant variables (X2 and X3) were equally constrained. Although several studies (Chen, 2007; Cheung & Rensvold, 2002; Rutkowski & Svetina, 2014) suggested guidelines for using the changes in AFIs to determine the level of full factorial invariance, none of them is directly applicable to testing partial factorial invariance. Hence, we cannot evaluate the performance of the two-step method in relation to the changes in the AFIs because that is beyond the scope of this study. Instead, we expect future studies to evaluate the performances of the AFIs for partial factorial invariance and to suggest appropriate guidelines for using the AFIs under various partial invariance scenarios.

Limitations

As with any simulation study, we examined only limited conditions, and our results can be generalized only to data conditions similar to those in our study. For example, we examined only partial factorial invariance of a single factor model with six indicators. Although we expect the results found in this study to be generalizable to simpler and more complex models, it is hard to say they will be before we have tested those conditions. Second, we simulated only balanced sample-size conditions, so we cannot be sure that the results apply to cases with substantially imbalanced sample sizes between groups. Another limitation is related to the number of groups. Because we simulated only two group conditions, the variables indicated by the smallest modification index were the same between groups. However, we do not suppose that the choice of RV using the smallest modification index with more groups is as simple as this study suggests. Thus, we would like to confine the generalization of these results to cases with only two groups. We also examined only 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 all these limitations, future studies are necessary with more varied factor models, imbalanced sample sizes, larger numbers of groups, and categorical variables. In addition, we simulated noninvariance in models without any type of misspecification. In reality, however, data are likely to have a certain degree of misfit irrelevant to noninvariance. Another possible future study would thus investigate the performance of the minmod under the models with misspecification irrelevant to the source of noninvariance. Finally, we want to add one more caveat related to sampling errors. Our study results are based on the simulated data with known conditions, and possible sampling errors were also modeled using 1,000 replications in our simulation study. Although the results indicate that the error rates in selecting a truly invariant RV were very low due to sampling errors, we cannot guarantee that similar results would be yielded from real data with various sources of sampling errors. To put it another way, the actual impact of the sampling errors in real data cannot be fully addressed in this study.

Conclusion and Recommendation

The findings of this study are very promising for researchers who lack a theoretical guideline in selecting an appropriate RV to test measurement invariance under a MCFA model. Even those who already have a theoretical guideline for selecting an RV can use this empirical guideline (min-mod) to provide evidence of the adequacy of the chosen RV. Although the min-mod using both partial and full invariance models resulted in fairly high accuracy in identifying a truly invariant RV, the selected RVs differed between the two models. The impact of the difference was prominent in the detection of noninvariant variables using the Wald tests. Generally, the Wald-RVP had higher perfect recovery rates, lower false negative rates, and lower false positive rates than the Wald-RVF. Across all the simulated data conditions, the highest gap was found between the two methods when the magnitude of noninvariance was small and the sample size was small. In particular, the lower perfect recovery rates of the Wald-RVF were due mainly to the higher false negative rates that were associated with the factor-loading size of the chosen RV. Therefore, we recommend using the Wald-RVP rather than the Wald-RVF, because it generally produced better results in detecting noninvariant variables, even though more procedures are required to conduct it. In our post hoc analysis, the number of RVs did not increase the power to detect noninvariant variables under conditions in which the small magnitude of noninvariance was combined with the smallest sample size. Thus, it is unnecessary to choose more than one RV to increase the power to detect noninvariant variables. When either full metric or scalar invariance is rejected, the recommended practice for applied researchers is as follows:

1. Choose the variable having the smallest modification index as an RV using a partial invariance model after sequentially relaxing invariant constraints until there remains no more significant modification index (3.84).
2. Conduct item-by-item Wald tests under a model identified by the selected RV in Step 1. Refer to Appendix A to test factor loadings and Appendix B to test intercepts. Use a Bonferroni-adjusted *p* value for the Wald tests. If the *p* value is below the adjusted criterion, the variables are considered noninvariant and vice versa. After conducting multiple Wald tests for all variables (except for the chosen RV), researchers can determine which variables are noninvariant and which are invariant.

As a final remark, based on the results of the study, the two-step approach proposed here is not recommended for research data with small samples, such as $N = 100$.

REFERENCES

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*, 456–466. doi:10.1037/0033-2909.105.3.456

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464–504. doi:10.1080/10705510701301834

Cheung, G. W., & Lau, R. S. (2011). A direct comparison method for testing measurement invariance. *Organizational Research Methods*, *15*, 167–198. doi:10.1177/1094428111421987

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*, 1–27. doi:10.1177/014920639902500101

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

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling*, *15*, 96–113. doi:10.1080/10705510701758349

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117–144. doi:10.1080/03610739208253916

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. doi:10.1080/10705510903206014

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago, IL: Scientific Software International.

Jung, E., & Yoon, M. (2016). Comparisons of three empirical methods for partial factorial invariance: Forward, backward, and factor-ratio tests. *Structural Equation Modeling*, *23*, 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. doi:10.1080/10705511.2011.557337

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. doi:10.1037/1082-989X.4.1.84

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*, 577–605. doi:10.1146/annurev.psych.60.110707.163612

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling*, *14*, 611–635. doi:10.1080/10705510701575461

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*, 568–592. doi:10.1037/0021-9010.93.3.568

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, *97*, 1016–1031. doi:10.1037/a0027934

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543. doi:10.1007/BF02294825

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 380–392). New York, NY: Guilford.

Muthén, B. O., & Muthén, L. K. (1998–2012). *M*plus *user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. doi:10.1016/j.dr.2016.06.004

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*, 954–974. doi:10.1177/0013164412441607

Raykov, T., Marcoulides, G. A., & Millsap, R. E. (2013). Factorial invariance in multiple populations: A multiple testing procedure. *Educational and Psychological Measurement*, *73*, 713–727. doi:10.1177/0013164412451978

Rensvold, R. B., & Cheung, G. W. (1998). Testing measurement models for factorial invariance: A systematic approach. *Educational and Psychological Measurement*, *58*, 1017–1034. doi:10.1177/0013164498058006010

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*, 31–57. doi:10.1177/0013164413498257

Sass, D. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, *29*, 347–363. doi:10.1177/0734282911406661

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*, 210–222. doi:10.1016/j.hrmr.2008.03.003

Steenkamp, J. B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90. doi:10.1086/209528

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. doi:10.1177/109442810031002

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. doi:10.1080/00220973.2010.531299

Whittaker, T. A. (2013). The impact of noninvariant intercepts in latent means models. *Structural Equation Modeling*, *20*, 108–130. doi:10.1080/10705511.2013.742397

Whittaker, T. A., & Khojasteh, J. (2013). A comparison of methods to detect invariant reference indicators in structural equation modelling. *International Journal of Quantitative Research in Education*, *1*, 426–443. doi:10.1504/IJQRE.2013.058310

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

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42–57. doi:10.1177/0146621607314044

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling*, *14*, 435–463. doi:10.1080/10705510701301677

APPENDIX A

M*PLUS* SYNTAX FOR ITEM-BY-ITEM WALD TESTS: FACTOR LOADING

```
Tile: Wald test using "Model Test" command
Data: File is example.dat;
Variable:
Names are x1-x6 g;
Usev = x1-x6;
Grouping = g (1 = g1 2 = g2);
Model:
F1 by x1-x6;
Model g1:
F1 by x1@1: ! If the selected reference variable is x1.
F1 by x2-x6 (A1-A5);
F1*;
Model g2:
F1 by x1@1;
F1 by x2-x6 (B1-B5);
F1*;
Model Test:
A1 = B1; ! If you want to test invariance of the factor loading of x2.
```

Note. You can test each of the five factor loadings (factor loading of x2-x6) by only replacing the highlighted part (assigned names of tested parameters) under the "Model Test" Command. For example, you can test the invariance of the factor loading of x3 by replacing "A1 = B1" with "A2 = B2".

APPENDIX B

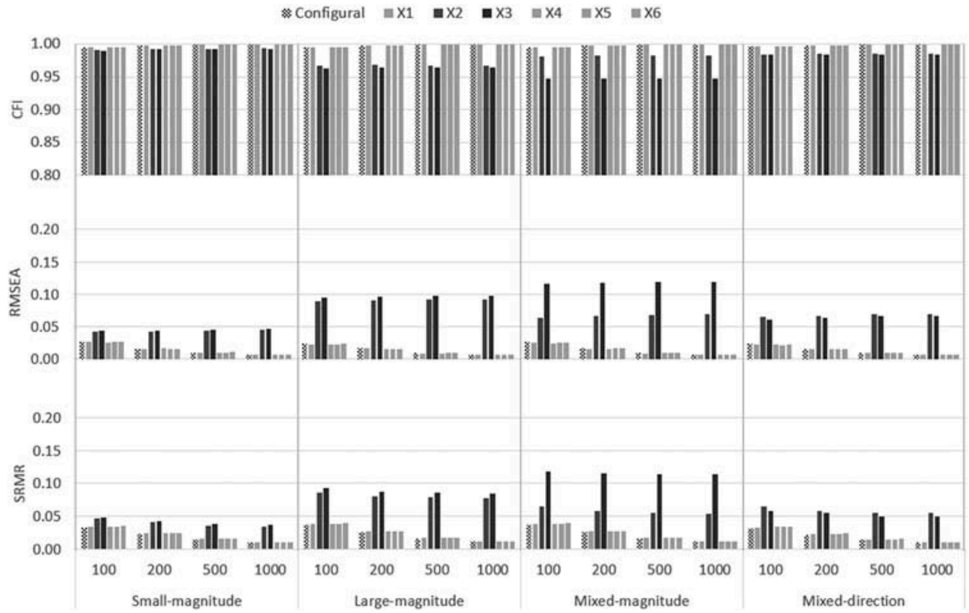
M*PLUS* SYNTAX FOR ITEM-BY-ITEM WALD TESTS: INTERCEPT

```
Tile: Wald test using "Model Test" command
Data: File is example.dat;
Variable:
Names are x1-x6 g;
Usev = x1-x6;
Grouping = g (1 = g1 2 = g2);
Model:
F1 by x1-x6;
Model g1:
F1 by x1-x6 (L1-L6);
F1@1;
[x1@0]; ! Selected reference variable
[x2-x6] (A1-A5);
[F1@0];
Model g2:
F1 by x1-x6 (L1-L6);
F1*;
[x1@0]; ! Selected reference variable
[x2-x6] (B1-B5);
[F1*];
Model Test:
A1 = B1; ! If you want to test invariance of the intercept of x2.
```

Note. You can test each of the five intercepts (intercepts of x2–x6) by only replacing the highlighted part (assigned names of tested parameters) under the "Model Test" Command. For example, you can test the invariance of the intercept of x3 by replacing "A1 = B1" with "A2 = B2".

APPENDIX C

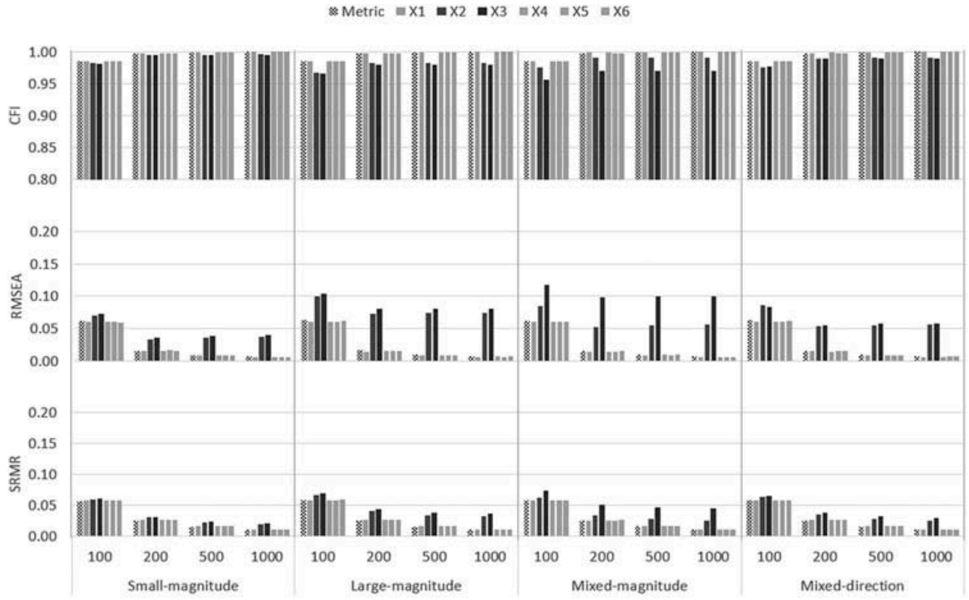
AVERAGE MODEL FIT INDEXES OF PARTIAL METRIC INVARIANCE CONDITIONS



Note. Configural = Model fit indexes of the configural invariance model with a reference variable for factor loadings selected under partially constrained metric invariance model (RVP); X*i* = model fit indees of the model having an additional set of equally constrained factor loadings of X*i*; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

APPENDIX D

AVERAGE MODEL FIT INDEXES OF PARTIAL SCALAR INVARIANCE CONDITIONS



Note. Metric = model fit indexes of the metric invariance model with a reference variable for intercepts selected under partially constrained scalar invariance model (RVP); X*i* = model fit indexes of the model having an additional set of equally constrained intercepts of X*i*; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.