STRUCTURAL EQUATION MODELING, *14*(3), 435–463

Detecting Violations of Factorial Invariance Using Data-Based Specification Searches: A Monte Carlo Study

Myeongsun Yoon and Roger E. Millsap
Department of Psychology, Arizona State University

In testing factorial invariance, researchers have often used a reference variable strategy in which the factor loading for a variable (i.e., reference variable) is fixed to 1 for identification. This commonly used method can be misleading if the chosen reference variable is actually a noninvariant item. This simulation study suggests an alternative method for testing factorial invariance and evaluates the performance of the method in specification searches based on the modification index. The results of the study showed that the proposed specification searches performed well when the number of noninvariant variables was relatively small and this performance improved as sample size increased and the size of group differences increased. When the number of noninvariant variables was relatively large, however, the method rarely succeeded in detecting the noninvariant items in the specification searches. Implications of the findings are discussed along with the limitations of the study.

Measurement invariance, also known as measurement equivalence (Cheung & Rensvold, 1998), has recently gained attention in different disciplines of social science, including cross-cultural research (Cheung & Rensvold, 1998; Little, 1997) and organizational behavior research (Byrne, 1991; Marsh & Roche, 1996). The primary reason for this appeal is that measurement invariance is needed to interpret group differences in observed scores. In other words,

Send correspondence to Myeongsun Yoon, Department of Psychology, Arizona State University, Box 871104, Tempe, AZ 85287-1104. E-mail: myoon@asu.edu

researchers must make sure that the relation between observed variables and underlying constructs is the same across groups, or that measurement invariance holds, before comparing observed scores between groups.

A common approach to examining measurement invariance is to apply confirmatory factor analysis (CFA), where a linear relation between observed scores and underlying construct is assumed. Meade and Lautenschlager (2004) conducted a Monte Carlo study to evaluate CFA tests for testing measurement invariance. They showed that CFA tests of measurement invariance performed well under ideal conditions such as large sample size, a sufficient number of manifest indicators, and moderate or high communalities. In the baseline model of the study, they picked one of the invariant items as a reference variable, as invariant items are known in the simulation study. However, invariant items are not known in real data, thus researchers must rely on theoretical grounds to pick an adequate reference variable.

The purpose of this study is to evaluate a method for examining partial factorial invariance without choosing a variable to serve as a reference. By avoiding this choice, the method avoids the problem of choosing a noninvariant reference variable.

MEASUREMENT INVARIANCE

A formal definition of measurement invariance in terms of probabilities is that invariance holds if and only if

$$P(X|W, G) = P(X|W), \tag{1}$$

where $\boldsymbol{X}$ is a vector of observable variables, $\boldsymbol{W}$ is a vector of latent variables underlying $\boldsymbol{X}$, and G represents an indicator for group membership. As a necessary and sufficient condition of measurement invariance, Equation 1 states that the conditional probability of $\boldsymbol{X}$ given $\boldsymbol{W}$ is independent of G. That is, measurement invariance holds when the relation of observed variables to a set of underlying latent variables is independent of group membership (Mellenburgh, 1989; Meredith & Millsap, 1992; Millsap, 1995). In a single-factor case, for example, to fulfill measurement invariance, persons with the same status on an underlying latent variable should have the same probability of achieving any observed score regardless of group membership. If measurement invariance does not hold, group differences on the observed score might be difficult to interpret because different factor structures might be confounded with group differences on the latent variable in producing group differences on the observed score (Millsap, 1997).

FACTORIAL INVARIANCE

A major methodological approach for testing measurement invariance is CFA, where a researcher tests a theory-based factor model under invariance constraints. The measurement model in CFA specifies a linear relation between p observed variables and m common factors, represented by the following equation:

$$X = \tau + \Lambda\xi + \delta, \tag{2}$$

where X refers to the $p \times 1$ vector of observed scores, τ is the $p \times 1$ vector of measurement intercepts, Λ is the $p \times m$ factor pattern matrix, ξ is the $m \times 1$ vector of underlying factor scores, and δ is the $p \times 1$ vector of unique factor scores. For the multiple group case, the corresponding measurement model is

$$X_g = \tau_g + \Lambda_g\xi_g + \delta_g, \tag{3}$$

where g indicates group membership. Given that factor scores (ξ) and unique factors (δ) within each group are assumed to be uncorrelated (i.e., $COV(\xi_g, \delta_g) = 0$), the covariance structure of X in group g is:

$$\Sigma_g = \Lambda_g\Phi_g\Lambda_g' + \Theta_g, \tag{4}$$

where Θ_g is the typically diagonal matrix of unique variances and Φ_g is the factor covariance matrix in group g. Also, assuming that the unique factors have zero means in Equation 3, the expectation of X in each group is:

$$E(X_g) = \tau_g + \Lambda_g\kappa_g, \tag{5}$$

where κ_g is the factor mean in group g. If measurement invariance holds, it follows that $\Theta_g = \Theta$, $\tau_g = \tau$, and $\Lambda_g = \Lambda$, leading to simplifications in Equations 4 and 5 as

$$\Sigma_g = \Lambda\Phi_g\Lambda' + \Theta, \tag{6}$$

$$E(X_g) = \tau + \Lambda\kappa_g \tag{7}$$

Equation 6 states that the difference in covariance structure of observed scores between groups is due to the difference in covariance structure of latent variables (factors) between groups, as long as measurement invariance holds. Equation 7 states that the systematic differences in group means on X are due to group differences in factor means κ_g. Equations 6 and 7 represent the condition of strict factorial invariance (Meredith, 1993).

Conditions of factorial invariance are considered hierarchically in this order: configural invariance, metric invariance, scalar invariance, and strict invariance

TABLE 1
Hierarchical Model of Factorial Invariance

Model	*Condition*
Configural	If λ_{ij} is fixed to zero for one group, then λ_{ij} should be fixed for other groups. If λ_{ij} is freely estimated for one group, then λij should be freely estimated for other groups (variable $i = 1, 2, \ldots p$, factor $j = 1, 2, \ldots, m$)
Partial metric	$\lambda_{ijg} = \lambda_{ij}$ only for some set of i and j.
Metric (pattern)	$\Lambda_g = \Lambda$
Scalar (strong)	$\Lambda_g = \Lambda$, $\tau_g = \tau$
Strict	$\Lambda_g = \Lambda$, $\tau_g = \tau$, $\Theta_g = \Theta$

Note. Λ_g is the $p \times m$ factor pattern matrix in group g, τ_g is the $p \times 1$ vector of measurement intercepts in group g, and Θ_g is the typically diagonal matrix of unique variances in group g.

(Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). Table 1 presents a summary of levels of factorial invariance. The least restrictive model provides configural invariance where the pattern of zero and nonzero loadings should be same across groups (Steenkamp & Baumgartner, 1998; Thurstone, 1947). In other words, if the same pattern of free and fixed loadings with same number of factors across groups can fit the data well, then the measures have configural invariance across groups. This level of invariance provides a baseline model to pursue higher levels of factorial invariance. The next level is metric invariance (also called pattern invariance), which can be defined by identical factor pattern matrices across groups ($\Lambda_g = \Lambda$), but allows differences in unique variances and intercepts (Horn & McArdle, 1992; Millsap, 1997; Thurstone, 1947). If a measure satisfies pattern invariance, then the invariance of intercepts can be tested as the next level of invariance ($\tau_g = \tau$). This level of invariance is called scalar invariance or strong invariance, which makes group mean comparisons meaningful (Meredith, 1993). The last condition of invariance requires invariance in the unique variances across groups ($\Theta_g = \Theta$), leading to strict factorial invariance. Under strict factorial invariance, systematic group differences in means or covariance matrices are due to group differences in common factor score distributions. Recently, testing for the invariance of intercepts (i.e., scalar invariance) has gained attention as a required step for testing invariance across groups (Little, 1997, 2000; Meredith, 1993). The results presented here focus on metric invariance; scalar invariance is not addressed.

PARTIAL METRIC INVARIANCE

Among the different levels of factorial invariance, metric invariance has been most discussed in the literature (Vandenberg & Lance, 2000). However, *full*

metric invariance (i.e., invariance of all factor loadings) is untenable in some cases. Instead of full metric invariance, *partial metric invariance* (i.e., invariance of some of factor loadings) has been discussed for more than a decade (Byrne, Shavelson, & Muthén, 1989; Cheung & Rensvold, 1998; Reise, Widaman, & Pugh, 1993; Steenkamp & Baumgartner, 1998).

Under partial metric invariance, noninvariant items can be retained with varied loadings across groups (Cheung & Rensvold, 1998). Byrne et al. (1989) argued that full metric invariance is not a prerequisite step for testing more restrictive models such as scalar invariance and strict invariance, or for comparing factor means across groups. However, the effect of the proportion (or number) of invariant items is not clear. Reise et al. (1993) proposed that a majority of items on a given latent variable should have invariant loadings across groups to ensure the nonarbitrariness of the group comparisons. Cheung and Rensvold (1998) stated that noninvariant items usually constitute only a small portion of the model and thus have little effect on group comparisons. On the other hand, Steenkamp and Baumgartner (1998) pointed out that partial invariance is acceptable if the compared groups have invariant loadings across groups for at least one item other than the reference variable for which loadings are fixed to one for all groups.

In addition to the proportion of invariant items across groups, there is another important issue to consider before comparing the means of a particular measure with partial metric invariance across groups: the choice of constraints for model identification across groups. Two commonly used methods for identifying the factor model in single group studies are to either fix the variance of each factor to one, or fix a loading for each factor to one. Additional constraints are required, which typically include setting subsets of loadings to zero. When conducting multiple group comparisons, however, fixing the variance of a factor is not recommended. If the factor variance is fixed to one in each group, the resulting implicit standardization would alter the loadings in each group, and would do so in ways that differ across groups if the factor variances differ. As a result, factor pattern matrices that are truly invariant would not appear to be so because of these standardizations. Thus, fixing one of the loadings is preferred for model identification, with the same loadings being fixed to one in each group.

In practice, full metric invariance may be rejected, leading to questions about which factor loadings differ across groups. To answer such questions, models representing different configurations of invariant and noninvariant loadings (i.e., partial metric invariance) can be tested in hopes of finding which loadings differ across groups. Here the use of a fixed loading for identification purposes can interfere with the process of finding the correct model of partial invariance (Cheung & Rensvold, 1999). If the chosen reference variable is actually one with varying loadings across groups, the forced fixed loading can distort the

pattern of invariant and noninvariant loadings and can lead to choice of the wrong model.

Mathematical Illustration of the Reference Variable Problem

To illustrate the problem of choosing which loading to fix, suppose that we have a single-factor case with two groups. Let Λ_1 and Λ_2 denote factor loading vectors in the first group and the second group, respectively. Suppose that the factor loading vector in each group can be written

$$\Lambda_1 = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \cdot \\ \lambda_m \\ \lambda_{m+1,1} \\ \cdot \\ \lambda_{p,1} \end{bmatrix} \text{ and } \Lambda_2 = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \cdot \\ \lambda_m \\ \lambda_{m+1,2} \\ \cdot \\ \lambda_{p,2} \end{bmatrix}, \tag{8}$$

In Equation 8, the first m variables are invariant and the last $(p - m)$ variables are noninvariant.

Let Σ_1 and Σ_2 denote covariance matrices in the first group and the second group, respectively. Then, the covariance matrix for each group can be specified as

$$\Sigma_1 = \Lambda_1 \phi_1 \Lambda_1' + \Theta_1 \text{ and } \Sigma_2 = \Lambda_2 \phi_2 \Lambda_2' + \Theta_2 \tag{9}$$

Without any constraints on factor structure, there is an indeterminacy among the factor loadings, factor variances, and unique variances. For example, the covariance matrix for each group can be rewritten as

$$\begin{aligned} \Sigma_1 &= \Lambda_1^* \phi_1^* \Lambda_1^{*'} + \Theta_1, \\ \Sigma_2 &= \Lambda_2^* \phi_2^* \Lambda_2^{*'} + \Theta_2 \end{aligned} \tag{10}$$

where $\Lambda_1^* = \alpha_1 \Lambda_1$, $\phi_1^* = \alpha_1^{-2} \phi_1$, $\Lambda_2^* = \alpha_2 \Lambda_2$, and $\phi_2^* = \alpha_2^{-2} \phi_2$ when α_1 and α_2 are scalars. Given that α_1 and α_2 are arbitrary numbers, an indeterminacy exists.

We need to pick one of the variables for the reference variable to identify factor structure. If one of the invariant variables is chosen for a reference variable, the estimated factor loadings will have same invariant and noninvariant

loading pattern across groups. For example, if the first variable is picked as a reference variable, then $\alpha_1 = \alpha_2 = \lambda_1^{-1}$, which leads to

$$\Lambda_1^* = \begin{bmatrix} 1 \\ \lambda_2\lambda_1^{-1} \\ \cdot \\ \lambda_m\lambda_1^{-1} \\ \lambda_{m+1,1}\lambda_1^{-1} \\ \cdot \\ \lambda_{p,1}\lambda_1^{-1} \end{bmatrix} \text{ and } \Lambda_2^* = \begin{bmatrix} 1 \\ \lambda_2\lambda_1^{-1} \\ \cdot \\ \lambda_m\lambda_1^{-1} \\ \lambda_{m+2,1}\lambda_1^{-1} \\ \cdot \\ \lambda_{p,2}\lambda_1^{-1} \end{bmatrix}$$

The first m variables are still invariant and the last $(p - m)$ variables are still noninvariant.

Different results are obtained if we choose one of the noninvariant variables as a reference variable. Suppose that the last variable is the chosen reference variable. Then, we will have $\alpha_1 = \lambda_{p,1}^{-1}\alpha_2 = \lambda_{p,2}^{-1}$ in Equation 10 and the resulting factor loading vector for each group is

$$\Lambda_1^* = \begin{bmatrix} \lambda_1\lambda_{p,1}^{-1} \\ \lambda_2\lambda_{p,1}^{-1} \\ \cdot \\ \lambda_m\lambda_{p,1}^{-1} \\ \lambda_{m+1,1}\lambda_{p,1}^{-1} \\ \cdot \\ 1 \end{bmatrix} \text{ and } \Lambda_2^* = \begin{bmatrix} \lambda_1\lambda_{p,2}^{-1} \\ \lambda_2\lambda_{p,2}^{-1} \\ \cdot \\ \lambda_m\lambda_{p,2}^{-1} \\ \lambda_{m+2,1}\lambda_{p,2}^{-1} \\ \cdot \\ 1 \end{bmatrix}$$

It is shown that the first m variables are noninvariant in the new loading vector by picking a noninvariant variable as the reference variable. This illustrates the problem of the reference variable strategy in the single-factor case when researchers are not confident about which variables are truly invariant.

PREVIOUS RESEARCH ON IDENTIFICATION PROBLEMS IN PARTIAL INVARIANCE

To date, there have been very few studies emphasizing the identification problem. Reise et al. (1993) illustrated three possible baseline models with the same degrees of freedom for a two-group comparison: (a) by constraining one of the loadings to 1 for both groups and freeing the other loadings and factor variance for both groups; (b) fixing factor variance to 1 for both groups and freeing all loadings in each group; or (3) fixing the factor variance to 1 in the first group, constraining one loading to invariance across groups, and estimating factor variance in the second group and the other loadings in each group separately. They

argued that the resulting estimates would be in an easily interpreted metric in the third model. We adopt a similar strategy for models that stipulate metric invariance: Fix the factor variance to 1 in one group, constrain all factor loadings to invariance, and fix no parameters in the second group. The advantage of this approach is that one does not have to decide a priori which item to use as a reference variable for identification.

Cheung and Rensvold (1999) approached the reference variable problem in measurement invariance studies by using the factor-ratio test. The factor-ratio test is a systematic examination of all possible combinations of referents and arguments across groups, where referent is the variable selected as a reference, and argument is the variable being tested for invariance. Thus, there are a total of $p*(p-1)/2$ factor ratio tests required assuming a single-factor model, where p is the number of variables to be tested. With three variables, for example, the three combinations of tests (i.e., $3*(3-1)/2 = 3$) are conducted as shown here:

1. When choosing the first variable as the reference, perform a chi-square difference test between the invariance constraint for the second variable and no constraint for the second variable.
2. When choosing the first variable as the reference, perform a chi-square difference test between the invariance constraint for the third variable and no constraint for the third variable.
3. When choosing the second variable as the reference, perform a chi-square difference test between the invariance constraint for the third variable and no constraint for the third variable.

As the number of variables increases, the Cheung and Rensvold (1999) method becomes less attractive. For instance, 45 tests are required if 10 variables are studied and the chi-square difference tests must be hand-calculated because no current software does this automatically. Also, to date, there has been no large simulation study to show that this approach works accurately.

PURPOSE OF THIS STUDY

Consistent with Reise et al. (1993) and Cheung and Rensvold (1999), this study examines an approach to avoiding or minimizing the problem of choosing a reference variable in multiple-group invariance studies. To narrow the scope of this study, we focus on the two-group condition. The results are expected to generalize to conditions containing three groups or more. To identify the model under metric invariance for a two-group case, the following method is suggested: (a) the variance of the factor is fixed to 1 in one group but not in the other group, and (b) all loadings are constrained to invariance across groups. The

purpose of these procedures is to examine whether full metric invariance holds across groups when configural invariance is assumed. If the model of metric invariance cannot be accepted based on fit indexes, invariance constraints can be relaxed based on the modification index (MI) values sequentially until the model reaches adequate fit. In doing so, we only allow loading constraints to be freed; off-diagonal elements of the unique factor covariance matrix are always set to zero.

Other researchers have used MI values as a basis for relaxing the invariance constraints of appropriate loadings (Byrne et al., 1989; Oort, 1998; Reise et al., 1993; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). The MI value computed in LISREL gives the expected drop in the likelihood ratio chi-square statistic when a constrained parameter is freed, and so it is said to be a measure of how poorly a particular parameter constraint is chosen (Jöreskog & Sörbom, 1996a; Muthén & Muthén, 2001). In general, this type of post-hoc model fitting strategy (i.e., based on MI) has been discouraged because it is data driven and therefore can give misleading results (MacCallum, 1986; MacCallum, Roznowski, & Necowitz, 1992). Clearly, any model modification strategy should be validated on new data. On the other hand, a sequential MI strategy showed no problems in some cases (MacCallum et al., 1992; Sörbom, 1989). In addition, the purpose of this study is not to explore all possible alternative models based on the MI, but only a well-defined subset of models (i.e., those with partial invariance in loadings). Therefore, it is worthwhile to see how the MI strategy works for this proposed identification method. The purpose of this study is to examine whether one could use MIs to accurately discern the partial invariance. To our knowledge, there have been no published simulation studies that examine the accuracy of this method. We examine the proposed method within a single-factor model with covariance structure for two populations. Neither multiple underlying factors nor mean structures are considered here. However, the results from this study are expected to extend to those cases.

METHODS

Simulation Conditions

Four major variables were manipulated for this study: (a) the number of measured variables, (b) sample size, (c) the number of loadings that truly differ between two groups, and (d) the extent of the group difference in each loading.

Number of measured variables. We considered two conditions for the number of measured variables. In one condition, 6 variables were simulated and 12 variables were simulated in the other condition. These conditions have

been used previously in a simulation study of factorial invariance (Meade & Lautenschlager, 2004). In both conditions, one common factor was assumed to underlie these observed variables.

Sample size. Sample sizes were varied with $N = 200$ and $N = 500$ per group. MacCallum, Widaman, Zhang, and Hong (1999) did a simulation study to examine the impact of varying sample sizes in factor analysis. They showed that the required sample size decreased as communalities increased and the ratio of variables to factors increased. According to MacCallum et al., $N = 200$ is needed to adequately recover a population model in factor analysis when communalities are low and ratio of variables to factors is 20:3, which is close to our six-variable condition. We included a larger sample size ($N = 500$) than the minimum to see the effect of sample size, if any.

Frequency of noninvariant items. Either one-third or two-thirds of the variables are noninvariant in loadings across groups. In the *low frequency* condition, one-third of variables are noninvariant, whereas two-thirds of variables are noninvariant in the *high frequency* condition. The detection of the noninvariant items in the high frequency condition is expected to be more difficult due to the limited basis for comparisons across groups under these conditions.

Size of difference between groups. Group differences in loadings were created by keeping the same set of loadings in the first group under all conditions, while varying the loadings in the second group. Although it would have been possible to create the second group loadings by simply subtracting a constant amount from selected first group loadings, this approach was not adopted because a fixed loading difference might have different meanings depending on loading size. For example, the difference between loadings of .9 and .8 might not have the same importance or practical significance as the difference between loadings of .5 and .4. For this reason, loading differences were instead created via proportional rescalings in a two-step process. In Step 1, the first Group 2 loading was created by subtracting a fixed constant from .7, which was the first loading in Group 1. Here a difference of .1 was considered to be a small group difference, a difference of .2 was considered a medium difference, and a difference of .3 was considered large. In Step 2, the remaining noninvariant loadings in Group 2 were created as being proportional to their counterparts in Group 1, using a constant of proportionality that reflected the proportional drop in Step 1. For example, if the drop in Step 1 went to .6 from .7, the proportional drop was (.7 − .6)/.7 or 1/7. Hence all other noninvariant loadings in Group 2 were set to $b - b/7$ or $b(6/7)$, where b is the counterpart loading in Group 1. If the drop in Step 1 went to .5 from .7, the proportional drop was (.7 − .5)/.7 or 2/7. The other noninvariant loadings in Group 2 were set to $b - b(2/7)$ or $b(5/7)$.

Other conditions. Under the preceding conditions, the loading differences across groups for noninvariant variables were proportional. As shown in Appendix A, this proportionality can lead to difficulties in locating which items are noninvariant. To address this problem, conditions in which the loadings for the noninvariant variables differed across groups but were not proportionally different were added. Also, the case where noninvariant loadings have a mixed pattern in which some loadings were higher in one group and some loadings were lower in that group was added. Based on the previous study (Meade & Lautenschlager, 2004) in which testing factorial invariance had higher power in the mixed condition than the uniformly lower (or higher) condition, we expect that our specification searches perform better in the mixed pattern of loading condition. To shorten the procedure, we tested these two conditions in the case where two thirds of variables are noninvariant (high frequency), sample size is large, the loading difference is large, and the number of variables is six. In all other conditions, noninvariant loadings are uniformly and proportionally lower in the second group.

In sum, the main portion of this study has four manipulated variables with $2 * 2 * 2 * 3 = 24$ different conditions. We investigated whether the proposed approach performed differently in the mixed pattern of loadings, and whether the result will be changed if loadings in noninvariant items are not proportional to each other. We evaluated these two conditions for only the $N = 500$ and six-variable case, making the total number of simulation conditions 26.

Table 2 presents population loading parameters, unique variances, and factor variances for each condition. For example, to construct the population covariance matrix for low frequency with six variables, the specified structure of the first group is

$$\Lambda_1 = \begin{bmatrix} .7 \\ .9 \\ .5 \\ .6 \\ .8 \\ .3 \end{bmatrix}, \quad \phi_1 = 1.0, \quad \Theta_1 = \begin{bmatrix} .8 \\ 1.3 \\ .4 \\ .5 \\ .9 \\ .2 \end{bmatrix},$$

and the specified structure of the second group is

$$\Lambda_2 = \begin{bmatrix} .6 \\ .77 \\ .5 \\ .6 \\ .8 \\ .3 \end{bmatrix}, \quad \phi_2 = 1.3, \quad \Theta_2 = \begin{bmatrix} .8 \\ 1.3 \\ .4 \\ .5 \\ .9 \\ .2 \end{bmatrix}.$$

TABLE 2
Simulation Condition for Six-Variable Cases

Condition	*Loading Difference*	*Population Parameter Values*
Low frequency	Small	Λ_2 = [.6, .77, .5, .6, .8, .3]
		$\Theta_{1=2}$ = [.8, 1.3, .4, .5, .9, .2]
	Medium	Λ_2 = [.5, .64, .5, .6, .8, .3]
		$\Theta_{1=2}$ = [.7, 1.3, .4, .5, .9, .2]
	Large	Λ_2 = [.4, .51, .5, .6, .8, .3]
		$\Theta_{1=2}$ = [.7, 1.2, .4, .5, .9, .2]
High frequency	Small	Λ_2 = [.6, .77, .5, .51, .69, .3]
		$\Theta_{1=2}$ = [.8, 1.3, .4, .6, .9, .2]
	Medium	Λ_2 = [.5, .64, .5, .43, .57, .3]
		$\Theta_{1=2}$ = [.7, 1.3, .4, .5, .9, .2]
	Large	Λ_2 = [.4, .51, .5, .34, .46, .3]
		$\Theta_{1=2}$ = [.7, 1.2, .4, .5, .9, .2]
High frequency/mixed	Large	Λ_2 = [.4, .51, .5, .86, 1.14, .3]
		$\Theta_{1=2}$ = [.7, 1.2, .4, 1.1, 1.8, .2]
High frequency/nonproportional	Large	Λ_2 = [.25, .65, .5, .45, .45, .3]
		$\Theta_{1=2}$ = [.5, 1.2, .4, .4, .9, .2]

Note. The factor variances are set to 1 and 1.3 in the first and second group, respectively, for all conditions. Λ_2 is the vector of population factor loadings in the second group and $\Theta_{1=2}$ is the vector of population-unique variances that applied to both groups. The population factor loadings for the first group are same across all conditions, Λ_1 = [.7, .9, .5, .6, .8, .3].

Factor loading values are arbitrarily chosen in the first group and factor loadings for noninvariant variables in the second group were proportionally lower than the first group. In the small loading difference case, for example, the loading of the second item in the second group is $.9(6/7) = .77$ and the loading in the second group for the medium difference case is $.9(5/7) = .64$. In the mixed design, the loadings of the first and second items were lower and the loadings of the fourth and fifth items were higher in the second group than the first group. For the calculation of loadings higher than the first group, the value of loading of the fourth variable is $.6 \times (10/7) = .86$ and the loading of the fifth variable is $.8 \times (10/7) = 1.14$. In the nonproportional loading pattern condition, four noninvariant loadings in the second group lowered by .15, .25, .35, and .45 with a mean of .3 difference. In this case, the variables with noninvariant loadings were randomly selected.

Across all conditions, factor variances are set to 1 in the first group and 1.3 in the second group. In this study, we generated samples with low communalities between .1 and .5. The unique variances are assigned to be the same for two groups and to have adequate communality (h_j^2) for each measured variable,

based on following equation (i.e., the communality in variable j):

$$h_j^2 = \frac{\lambda_j^2 \phi}{\lambda_j^2 \phi + \theta_j} \tag{11}$$

The communalities used for the six-variable case are shown in Table 3. The parameter values and communalities are chosen in the same way for the 12-variable case. For example, population parameter values for low frequency in the 12-variable case are the following:

$$\Lambda_1 = \begin{bmatrix} .7 \\ .9 \\ .5 \\ .6 \\ .8 \\ .3 \\ .7 \\ .9 \\ .5 \\ .6 \\ .8 \\ .3 \end{bmatrix}, \quad \Lambda_2 = \begin{bmatrix} .6 \\ .77 \\ .5 \\ .6 \\ .8 \\ .3 \\ .6 \\ .77 \\ .5 \\ .6 \\ .8 \\ .3 \end{bmatrix}, \quad \Theta_{1=2} = \begin{bmatrix} .8 \\ 1.3 \\ .4 \\ .5 \\ .9 \\ .2 \\ .8 \\ 1.3 \\ .4 \\ .5 \\ .9 \\ .2 \end{bmatrix}, \quad H_1 = \begin{bmatrix} .38 \\ .38 \\ .38 \\ .42 \\ .42 \\ .31 \\ .38 \\ .38 \\ .38 \\ .42 \\ .42 \\ .31 \end{bmatrix}, \quad H_2 = \begin{bmatrix} .37 \\ .37 \\ .45 \\ .48 \\ .48 \\ .37 \\ .37 \\ .37 \\ .45 \\ .48 \\ .48 \\ .37 \end{bmatrix},$$

$$\phi_1 = 1.0, \quad \phi_2 = 1.3,$$

where H_1 is communalities in the first group and H_2 = communalities in the second group.

Simulation Procedure

Data were simulated through two main procedures. First, the population covariance matrix was calculated based on Equation 4 for each condition. In Equation 4, the factor loadings and the factor variances were fixed at the values of interest and unique variances were chosen to yield communalities between .1 and .5 as described earlier. Then, the LISREL program was used to calculate the population covariance matrix of each group for each condition. The fitted covariance matrix for each group in the LISREL output is the population covariance matrix for each group and each condition.

Second, following the steps in Jöreskogand Sörbom (1996b, pp. 189–194), 100 samples of size N from the appropriate multivariate normal distribution were simulated for each condition and each group. Two steps were taken here.

TABLE 3
Communalities in the Case of Six Variables

Condition	*Loading Difference*	*Communalities*
Low frequency	Small	H_1 = [.38, .38, .38, .42, .42, .31]
		H_2 = [.37, .37, .45, .48, .48, .37]
	Medium	H_1 = [.41, .38, .38, .42, .42, .31]
		H_2 = [.32, .29, .45, .48, .48, .37]
	Large	H_1 = [.41, .40, .38, .42, .42, .31]
		H_2 = [.23, .22, .45, .48, .48, .37]
High frequency	Small	H_1 = [.38, .38, .38, .38, .42, .31]
		H_2 = [.37, .37, .45, .36, .41, .37]
	Medium	H_1 = [.41, .38, .38, .42, .42, .31]
		H_2 = [.32, .29, .45, .32, .32, .37]
	Large	H_1 = [.41, .40, .38, .42, .42, .31]
		H_2 = [.23, .22, .45, .23, .23, .37]
High frequency/mixed	Large	H_1 = [.41, .40, .38, .25, .26, .31]
		H_2 = [.23, .22, .45, .47, .48, .37]
High frequency/nonproportional	Large	H_1 = [.49, .40, .38, .47, .42, .31]
		H_2 = [.14, .31, .45, .40, .23, .37]

Note. H_1 = vector of communalities in the first group; H_2 = vector of communalities in the second group.

First, the LISREL program was used to find T in the triangular factoring $\Sigma = TT'$, where Σ represents the population covariance matrix. Then, the PRELIS program was used to generate multivariate normal variables using T in the LISREL output of the previous step.

Analysis

Simulated data were analyzed using LISREL (Version 8.53). We fit the full metric invariance as an initial model, and this model was adjusted sequentially, freeing the loading that had the largest MI until the largest MI was no longer significant at the .05 significance level. These sequential adjustments were done automatically using the AM option in LISREL, which runs a sequence of models by freeing one constraint at a time based on the largest MI (Jöreskog & Sörbom, 1996a). All loadings were constrained to be invariant between groups in the initial model but unique variances were freely estimated without equality constraints at any time in these analyses. Covariances among unique factor scores (off-diagonals of Θ) were always fixed to zero.

The results will be summarized in terms of the number of true and false detections. In the search of noninvariant variables based on the MI, *true detection*

is defined as detection of a noninvariant item and *false detection* is a detection of an invariant item as noninvariant. Three primary dependent measures were evaluated to check the performance of suggested procedure.

Perfect recovery rate. It is important to know what percentage of samples recovered the true model through the model modification sequence. The term *perfect recovery rate* is used for this percentage of interest. To put it another way, perfect recovery happened when all noninvariant variables were detected in the final estimated model without falsely detecting any invariant variable as noninvariant along the way. Logistic regression was used to decide which of the manipulated variables had a statistically significant effect on the perfect recovery rate.

True and false detection. It is also important to know the number of true and false detections along with the perfect recovery rate. We looked at the number of true detections to see how well the proposed procedure performed and we also checked the number of false detections to evaluate how badly it performed.

RMSEA. We checked the root mean square error of approximation (RMSEA) for the initial model where all factor loadings were constrained to be invariant. The RMSEA value at each subsequent step was also recorded. We expected the RMSEA to be higher across study conditions as the loading difference increased and frequency of noninvariant variables increased.

RESULTS

Perfect Recovery Rate

Perfect recovery rate in each condition is shown in Table 4.

Low frequency. In the low frequency condition, specification searches based on the MI were more successful as the size of loading difference increased, the sample size increased, and the number of variables decreased. The best performance occurred in the condition of the large loading difference with $N = 500$ and six items in which 83% of samples led to perfect recovery where all noninvariant items were detected in the specification searches without detecting any invariant item as noninvariant.

TABLE 4
Perfect Recovery Rate

Condition	*Loading Difference*	*Sample Size* 200	500
Low frequency/6 variables	Small	.00	.10
	Medium	.12	.55
	Large	.65	.83
Low frequency/12 variables	Small	.00	.00
	Medium	.08	.56
	Large	.41	.61
High frequency/6 variables	Small	.00	.00
	Medium	.00	.00
	Large	.02	.01
High frequency/6 variables/mixed	Large		.26
High frequency/6 variables/nonproportional	Large		.16
High frequency/12 variables	Small	.00	.00
	Medium	.00	.00
	Large	.00	.00

Note. In mixed condition, half of noninvariant variables were higher and the other half were lower in the first group than the second group. Nonproportional represents the condition where noninvariant loadings are not proportional to each other.

High frequency. In the high frequency condition, however, the same pattern was not observed. The specification searches were very bad in those cases. In the six-variable case with large loading difference, only 2 samples out of 100 led to perfect recovery in the $N = 200$ condition and 1 sample in the $N = 500$ condition. In the 6-variable case with small and medium loading difference and in the 12-variable case, perfect recovery was not produced even once out of 100 samples. Both the mixed design and nonproportionality increased the perfect recovery rate in the high frequency condition, although it was still far from being perfect. In the mixed design with six variables, 26% of samples had the perfect recovery. When noninvariant loadings are not proportional, 16% of samples reached perfect recovery.

Logistic regression results confirmed significant effects of the size of loading differences ($p < .001$), the frequency ($p < .001$), the number of measured variables ($p < .001$), and sample size ($p < .001$), on the perfect recovery rate. We calculated the approximate pseudo r^2 statistic for each of the effects. The Nagelkerke r^2 statistic in each logistic regression with single factor was .173, .261, .009, and .039 for the effects of size of loading differences, the frequency, the number of measured variables, and the sample size, respectively. Not surprisingly, the frequency of noninvariant items had the highest Nagelkerke r^2.

True Detection: Detecting Noninvariant Variables as Noninvariant

Both possible and maximum numbers of true detections were different across conditions. In the six-variable condition, the possible number of true detections are 0, 1, and 2 (maximum) for the low frequency condition and 0, 1, 2, 3, and 4 (maximum) for the high frequency condition. In the 12-variable condition, the maximum number of true detections is 4 for the low frequency condition and 8 for the high frequency condition. To simplify Table 5, we present only the cases where zero or all of the noninvariant variables were detected, as those cases are of the most interest (see Appendix B for more detailed tables). Table 5 gives the proportions of samples with zero (on the left) and maximum (on the right) true detections.

Low frequency. In the low frequency condition, the pattern of true detections looked different across the conditions of loading differences and sample size. For both the 6- and 12-variable conditions, the modification search

TABLE 5
The Proportion of Zero and Maximum True Detections

		Zero Detection		*Maximum Detection*	
		Sample Size		*Sample Size*	
Condition	*Loading Difference*	*200*	*500*	*200*	*500*
Low frequency/6 variables	Small	.80	.61	.00	.11
	Medium	.40	.14	.15	.66
	Large	.10	.01	.70	.98
Low frequency/12 variables	Small	.61	.45	.00	.00
	Medium	.11	.00	.09	.80
	Large	.01	.00	.62	1.00
High frequency/6 variables	Small	.80	.77	.00	.00
	Medium	.70	.70	.00	.00
	Large	.63	.67	.02	.01
High frequency/6 variables/mixed	Large		.00		.26
High frequency/6 variables/nonproportional	Large		.00		.16
High frequency/12 variables	Small	.50	.44	.00	.00
	Medium	.41	.52	.00	.00
	Large	.49	.48	.00	.00

Note. In the mixed condition, half of noninvariant variables were higher and the other half were lower in the first group than the second group. Nonproportional represents the condition where noninvariant loadings are not proportional to each other.

procedure performed better with the larger sample size and larger loading difference. More samples had maximum true detections and fewer samples had zero true detections as sample size increased and as the loading difference increased. In the large loading difference with $N = 500$, 98 samples had all 2 noninvariant variables detected in the 6-variable condition and all 100 samples had all 4 noninvariant variables detected in the 12-variable condition, whereas in the small loading difference with $N = 200$, 80 and 61 samples had zero true detections for the 6- and 12-variable conditions, respectively.

High frequency. When two-thirds of variables were noninvariant (high frequency), maximum true detections were rarely observed, especially when the noninvariant variables had uniformly and proportionally lower or higher loadings in one group, as only 3 samples out of 1,200 had maximum true detections and most of these had no detection of a noninvariant variable. Mixed (and proportional) pattern of loadings and nonproportional (and nonmixed) pattern of loadings conditions increased the proportion of maximum true detections up to .26 and .16, respectively. In both conditions, at least one true detection was found.

False Detection: Detecting Invariant Variables as Noninvariant

As in the true detection case, the possible and maximum numbers of false detections were different across conditions. In the six-variable condition, the possible numbers of false detections were 0, 1, 2, 3, and 4 (maximum) for the low frequency condition and 0, 1, and 2 (maximum) for the high frequency condition. In the 12-variable condition, the maximum number of false detections is 8 for low frequency and 4 for high frequency. As in the true detection case, we present only the cases where none or all of the invariant variables were detected as noninvariant (see Appendix B for more detailed tables). Table 6 gives the proportions of samples with zero (on the left) and maximum (on the right) false detections.

Low frequency. When one-third of the variables are noninvariant (low frequency), more than half had no false detections across the different conditions of loading difference and sample size. This proportion seemed higher with 6 variables than in the 12-variable condition. The proportion of samples with no false detections ranged from .69 to .84 in the 6-variable condition and .51 to .67 in the 12-variable condition. There were none that had maximum false detections.

TABLE 6
The Proportion of Zero and Maximum False Detections

Condition	Loading Difference	Zero Detection		Maximum Detection	
		Sample Size		Sample Size	
		200	500	200	500
Low frequency/6 variables	Small	.84	.75	.00	.00
	Medium	.69	.76	.00	.00
	Large	.84	.84	.00	.00
Low frequency/12 variables	Small	.51	.58	.00	.00
	Medium	.60	.67	.00	.00
	Large	.62	.61	.00	.00
High frequency/6 variables	Small	.82	.71	.00	.06
	Medium	.39	.12	.30	.71
	Large	.18	.03	.68	.96
High frequency/6 variables/mixed	Large		.56		.21
High frequency/6 variables/nonproportional	Large		.33		.45
High frequency/12 variables	Small	.62	.47	.00	.05
	Medium	.19	.00	.16	.77
	Large	.02	.01	.71	.97

Note. In the mixed condition, half of noninvariant variables were higher and the other half were lower in the first group than the second group. Nonproportional represents the condition where noninvariant loadings are not proportional to each other.

High frequency. When two-thirds of the variables were noninvariant (high frequency), there was more variability in the proportion of samples with no false detection. This proportion decreased as the size of loading difference increased and the sample size increased. When the loading difference is large and the sample size is 500, only four samples had no false detections involved (three for the 6-variable condition and one for the 12-variable condition). The mixed (and proportional) pattern of loadings and the nonproportional (and nonmixed) pattern of loadings increased the proportion of no false detections to .56 and .33, respectively. There was also variability in the proportion of samples with maximum false detections. The proportion of maximum false detections increased as the size of the loading difference increased and the sample size increased. This proportion went up to .96 in the 6-variable condition and .97 in the 12-variable condition when the loading difference is large and the sample size is 500. The mixed (and proportional) pattern of loadings and the nonproportional (and nonmixed) pattern of loadings decreased the proportion of maximum false detections to .21 and .45, respectively.

TABLE 7
Mean and Standard Deviation of RMSEA in the Initial Model

		Sample Size			
		200		*500*	
Condition	*Loading Difference*	*M*	*SD*	*M*	*SD*
Low frequency/6 variables	Small	.018	.020	.019	.014
	Medium	.033	.023	.034	.015
	Large	.057	.021	.058	.011
Low frequency/12 variables	Small	.015	.014	.013	.008
	Medium	.021	.014	.025	.008
	Large	.039	.012	.040	.006
High frequency/6 variables	Small	.019	.020	.015	.014
	Medium	.033	.024	.032	.013
	Large	.052	.022	.057	.011
High frequency/6 variables/mixed	Large			.078	.010
High frequency/6 variables/nonproportional	Large			.073	.011
High frequency/12 variables	Small	.015	.014	.013	.010
	Medium	.025	.012	.025	.009
	Large	.039	.011	.040	.006

RMSEA

Table 7 gives the means and standard deviations of RMSEA in the initial model where all loadings were constrained to be invariant. Surprisingly, the means of RMSEA were not very high even when two thirds of the variables were noninvariant (high frequency condition) with large loading differences. The largest RMSEA averages for the nonmixed and proportional loading conditions were .058. In the mixed pattern of loadings and the nonproportional pattern of loadings, the RMSEA averages were .078 and .073, respectively. As expected, higher means of RMSEA were obtained as the loading difference increased. However, the frequency of noninvariant variables did not greatly increase the magnitude of the RMSEA in the studied conditions.

DISCUSSION

The most commonly used identification method when testing factorial invariance has several serious weaknesses when researchers are not sure which items are invariant or noninvariant. An alternative identification method is to form the baseline model by constraining all loadings to be invariant, fixing the factor variance of one group to unity. A potential problem with this alternative method

is that we must rely on the MIs or other fit information to find noninvariant variables, given the invariance of any variables not known by a theory. The purpose of this study was to evaluate specification searches based on the MI for accuracy in finding noninvariant variables in the study of partial invariance.

The most interesting feature of our results is the striking difference in the performance of specification searches between the low and high frequency conditions. In low frequency conditions, specification searches were quite successful in detecting noninvariant variables. This performance got better as the loading difference increased and sample size increased. In the high frequency condition, however, the specification searches consistently failed to detect noninvariant items, even though mixed pattern or nonproportionality of loadings improved the performance to some degree.

The results of this study suggest that the proposed method can be accurate if the set of studied variables includes more invariant variables than noninvariant variables, or in the low frequency condition. The performance of the method improved as sample size increased and as the loading difference between groups increased. On the other hand, the proposed method would not work well when more noninvariant variables are present than invariant variables, or the high frequency condition. In this high frequency condition, the specification searches based on the MI had difficulty in detecting noninvariant items and had specified invariant items as noninvariant in most cases.

One possible explanation for the strikingly different results between the low and high frequency conditions is that there might be a readjustment of scale in the high frequency condition, as described in Appendix A. In the high frequency condition, noninvariant items are the majority and invariant items are the minority, so the scale might be readjusted toward the majority of noninvariant items. The invariant items then become noninvariant based on the new scale if the loading differences are proportionally equal in a condition. Under this hypothesis, the specification searches successfully detected the newly defined "noninvariant" variables based on the adjusted scale. In this study, differences of loadings were designed to be proportional to adjust for the impact of the absolute difference as it varied across various loading sizes. An unfortunate consequence of this proportionality, however, is that when the noninvariant loadings are in the majority, a reversal of invariance status is possible, as illustrated in Appendix A. The increased performance of the searches in mixed design (but with still proportional loadings) and nonproportionality of loading (but still in a nonmixed pattern) conditions partially supports this reasoning.

The results for the RMSEA in the initial model reveal that small RMSEA values are still consistent with misspecified levels of invariance. In most studied conditions, the average estimates of RMSEA were small enough that many investigators would not introduce further modifications. This was even true in the large loading difference condition, where the largest RMSEA averages were .058. In

practice, many researchers would consider retaining models with RMSEA values below .05, possibly leading to failures to detect group differences in loadings. It seems clear that if specification searches for violations of metric invariance are to be pursued, sole reliance on indexes such as the RMSEA can result in Type II errors under conditions resembling those simulated here.

Limitations and Future Research

There were several limitations to this study that can be addressed in future research. First, this study was limited to the single-factor case with two groups. Although studies of factorial invariance often involve more than one factor and more than two groups, this study aimed to show how this method would work with the simplest case possible. If the method does not perform well under simple conditions, it is unlikely to perform well in more complex cases. We would expect that extending the method to more than two groups would add to the number of possible steps needed to explore partial invariance, but would not change the fundamental results found in the two-group case. The extension to multiple factors is less clear, as the results would probably depend on the structure of the additional factors. For example, models in which individual measures are determined by multiple factors might raise difficulties not found when all variables are determined by a single factor.

Second, this study covered the continuous variable case, but did not address the case where observed variables are ordinal and highly discrete. Previous research has shown that the detection of biased items is less successful with dichotomous items than continuous items (Oort, 1998). Appropriate factor models for discrete ordinal items differ from those used in the continuous case (Millsap & Tein, 2004). These models include new parameters (e.g., threshold parameters) and often require large samples for effective use, depending on the choice of estimation procedure. Separate studies of detection accuracy using these models are required before we can have confidence in generalizing the current results to the discrete ordinal case.

Third, the simulations reported here employed communality values that were relatively low in all conditions. Low communalities are realistic when the measured variables are test or questionnaire items, for example. Higher communalities might be expected when the measured variables are subtests or whole tests. In unreported preliminary simulations with large communalities, we found that the modification index strategy presented here achieved a much higher level of accuracy than is found in the low communality conditions reported here. Future studies might examine the high communality case more thoroughly.

Finally, this study focused on the detection of violations of invariance for factor loadings only. Although violations of metric invariance are nearly always of interest, it would be useful to explore the same set of questions in relation

to tests of intercept invariance (i.e., strong factorial invariance). This type of investigation could be done under conditions of metric invariance or conditions of partial metric invariance in which some loadings vary across groups. It would also be useful to examine tests of partial invariance for unique variances, although no confounds with constraints needed for identification will be present for the unique variances.

REFERENCES

Byrne, B. M. (1991). The Maslach Burnout Inventory: Validating factorial structure and invariance across intermediate, secondary, and university educators. *Multivariate Behavioral Research, 23,* 361–375.

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.

Cheung, G. W., & Rensvold, R. B. (1998). Cross-cultural comparisons using non-invariant measurement items. *Applied Behavioral Science Review, 6,* 93–110.

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

Horn, J. L., & McArdle, J. J. (1992). A practical guide to measurement invariance in research on aging. *Experimental Aging Research, 18,* 117–144.

Jöreskog, K. G., & Sörbom, D. (1996a). *LISREL 8: User's reference guide.* Chicago: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1996b). *PRELIS 2: User's reference guide.* Chicago: Scientific Software.

Little, T. D. (1997). Mean and covariance structures analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32,* 53–76.

Little, T. D. (2000). On the comparability of constructs in cross-cultural research: A critique of Cheung and Rensvold. *Journal of Cross-Cultural Psychology, 31,* 213–219.

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

MacCallum, R. C., Roznowski, M., & Necowitz, L. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111,* 490–504.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84–99.

Marsh, H. W., & Roche, L. A. (1996). Structure of artistic self-concepts for performing arts and non-performing arts students in a performing arts high school: "Setting the stage" with multiple group confirmatory factor analysis. *Journal of Educational Psychology, 88,* 461–477.

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling, 11,* 60–72.

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

Meredith, W. (1993). Measurement invariance, factor analysis, and factor invariance. *Psychometrika, 58,* 525–543.

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

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research, 30,* 577–605.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2,* 248–260.

Millsap, R. E., & Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39,* 479–516.

Muthén, B., & Muthén, L. (2001). *M*plus *user's guide.* Los Angeles: Statmodel.

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

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,* 552–566.

Sörbom, D. (1989). Model modification. *Psychometrika, 54,* 371–384.

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

Thurstone, L. L. (1947). *Multiple-factor analysis.* Chicago: University of Chicago Press.

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,* 4–70.

APPENDIX A

This appendix illustrates an indeterminacy in tests of pattern invariance in the single-factor case when certain proportionality relations hold among the loadings across groups.

In the single-factor case, let Λ_1 and Λ_2 denote factor loading vectors in the first group and the second group, respectively. Suppose that the factor loading vector in each group can be written

$$\Lambda_1 = \begin{bmatrix} \Lambda \\ \Lambda_{11} \end{bmatrix} \text{ and } \Lambda_2 = \begin{bmatrix} \Lambda \\ D_2\Lambda_{11} \end{bmatrix}, \tag{A1}$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \cdot \\ \lambda_i \end{bmatrix}, \Lambda_{11} = \begin{bmatrix} \lambda_{i+1} \\ \lambda_{i+2} \\ \cdot \\ \lambda_p \end{bmatrix}, \text{ and } D_2 = \begin{bmatrix} \alpha_2 & 0 & \cdot & 0 \\ 0 & \alpha_2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \alpha_2 \end{bmatrix}$$

In Equation A1, the first i variables are invariant and the last $(p - i)$ variables are noninvariant. The loadings of noninvariant items in the second group are proportional with α_2 to those in the second group. Here α_2 is a scalar constant.

The factor loading vector for the second group in Equation A1 can be rewritten by

$$\Lambda_2 = D\Lambda_1, \tag{A2}$$

where

$$D = \begin{bmatrix} I_1 & Z' \\ Z & D_2 \end{bmatrix} = \begin{bmatrix} I_1 & Z' \\ Z & \alpha_2 I_2 \end{bmatrix} = \alpha_2 \begin{bmatrix} \alpha_2^{-1} I_1 & Z' \\ Z & I_2 \end{bmatrix}, \tag{A3}$$

and I_1 is an $i \times i$ identity matrix, Z is an $i \times (p - I)$ zero matrix, Z' is the transpose of Z, and I_2 is a $(p - I) \times (p - I)$ identity matrix.

Let Σ_1 and Σ_2 denote covariance matrices in the first group and the second group, respectively. Then, the covariance matrix for the first group can be specified as

$$\Sigma_1 = \Lambda_1 \phi_1 \Lambda_1' + \Theta_1 \tag{A4}$$

and the covariance matrix for the second group is

$$\Sigma_2 = D\Lambda_1 \phi_2 \Lambda_1' D + \Theta_2 \tag{A5}$$

Equation A5 can be rewritten

$$\Sigma_2 = \alpha_2 D^* \Lambda_1 \phi_2 \Lambda_1' D^* \alpha_2 + \Theta_2, \tag{A6}$$

where

$$D^* = \begin{bmatrix} \alpha_2^{-1} I_1 & Z' \\ Z & I_2 \end{bmatrix}$$

And Equation A6 can be rewritten again,

$$\Sigma_2 = D^* \Lambda_1 \Theta_2^* \Lambda_1' D^* + \Phi_2, \tag{A7}$$

where

$$\phi_2^* = \alpha_2 \phi_2 \alpha_2 = \alpha_2^2 \phi_2$$

Finally, it is shown that the new loading vector in the second group,

$$D^* \Lambda_1 = \begin{bmatrix} \alpha_2^{-1} \Lambda \\ \Lambda_{11} \end{bmatrix},$$

which indicates that the first i variables are noninvariant and the other $(p - i)$ variables are invariant. This is exactly reversed from starting point and illustrates indeterminacy of detecting invariant and noninvariant variables in the single factor case when the loadings of noninvariant items in one group are proportional to those in the other group. Under these conditions, investigators could easily be misled regarding which subset of variables has invariant loadings.

This result might explain a puzzling finding reported by Meade and Lautenschlager (2004). In this finding, within a partial invariance condition similar

to our $p = 6$, high frequency condition, data were simulated to have invariant factor variances across groups. The results showed poor detection of the factor loading differences, but spurious detection of factor variance differences across groups. The likely explanation for this result lies in the indeterminacy described earlier. The noninvariant loadings were probably nearly proportional across groups, permitting the type of reversal described earlier, along with a rescaling of the factor variance in one group. This rescaling resulted in a group difference in factor variance.

APPENDIX B

Appendix B contains the more detailed results for the proportion of samples having different numbers of true (NT) and false detections (NF).

TABLE B1
Six Variables With Low Frequency

	N = 200						*N = 500*					
	NF						*NF*					
NT	*0*	*1*	*2*	*3*	*4*	*Total*	*0*	*1*	*2*	*3*	*4*	*Total*
						Small Loading Difference						
0	.67	.12	.01	.00	.00	.80	.39	.19	.03	.00	.00	.61
1	.17	.03	.00	.00	.00	.20	.26	.02	.00	.00	.00	.28
2	.00	.00	.00	.00	.00	.00	.10	.01	.00	.00	.00	.11
Total	.84	.15	.01	.00	.00	1.00	.75	.22	.03	.00	.00	1.00
						Medium Loading Difference						
0	.23	.15	.02	.00	.00	.40	.07	.04	.02	.01	.00	.14
1	.34	.11	.00	.00	.00	.45	.14	.05	.01	.00	.00	.20
2	.12	.03	.00	.00	.00	.15	.55	.11	.00	.00	.00	.66
Total	.69	.29	.02	.00	.00	1.00	.76	.20	.03	.01	.00	1.00
						Large Loading Difference						
0	.03	.01	.04	.02	.00	.10	.00	.00	.00	.01	.00	.01
1	.16	.04	.00	.00	.00	.20	.01	.00	.00	.00	.00	.01
2	.65	.05	.00	.00	.00	.70	.83	.14	.01	.00	.00	.98
Total	.84	.10	.04	.02	.00	1.00	.84	.14	.01	.01	.00	1.00

Note. The possible maximum number of true detections was 2 and the possible maximum number of false detections was 4 in this case.

TABLE B2
Twelve Variables With Low Frequency

	N = 200							*N* = 500				
	NF							*NF*				
NT	*0*	*1*	*2*	*3*	*4*	*5*	*Total*	*0*	*1*	*2*	*3*	*Total*
						Small Loading Difference						
0	.23	.27	.09	.02	.00	.00	.61	.19	.11	.09	.06	.45
1	.19	.07	.02	.00	.00	.00	.28	.20	.07	.05	.01	.33
2	.07	.02	.00	.00	.00	.00	.09	.10	.02	.00	.00	.12
3	.02	.00	.00	.00	.00	.00	.02	.09	.01	.00	.00	.10
4	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Total	.51	.36	.11	.02	.00	.00	1.00	.58	.21	.14	.07	1.00
						Medium Loading Difference						
0	.03	.03	.02	.01	.01	.01	.11	.00	.00	.00	.00	.00
1	.14	.09	.02	.01	.01	.00	.27	.00	.02	.01	.01	.04
2	.19	.06	.03	.01	.00	.00	.29	.01	.00	.00	.00	.01
3	.16	.07	.01	.00	.00	.00	.24	.10	.04	.01	.00	.15
4	.08	.00	.01	.00	.00	.00	.09	.56	.22	.02	.00	.80
Total	.60	.25	.09	.03	.02	.01	1.00	.67	.28	.04	.01	1.00
						Large Loading Difference						
0	.00	.01	.00	.00	.00	.00	.01	.00	.00	.00	.00	.00
1	.00	.00	.01	.01	.00	.00	.02	.00	.00	.00	.00	.00
2	.03	.02	.03	.02	.00	.00	.10	.00	.00	.00	.00	.00
3	.18	.05	.02	.00	.00	.00	.25	.00	.00	.00	.00	.00
4	.41	.19	.02	.00	.00	.00	.62	.61	.33	.03	.03	1.00
Total	.62	.27	.08	.03	.00	.00	1.00	.61	.33	.03	.03	1.00

Note. The possible maximum number of true detections was 4 and the possible maximum number of false detections was 8 in this case. NF greater than 5 did not have any case in $N = 200$ and NF greater than 3 did not have any case in $N = 500$, so those conditions are not shown in the table to save space.

TABLE B3
Six Variables With High Frequency

	N = 200				*N = 500*			
	NF				*NF*			
NT	*0*	*1*	*2*	*Total*	*0*	*1*	*2*	*Total*
	Small Loading Difference							
0	.63	.17	.00	.80	.49	.22	.06	.77
1	.15	.01	.00	.16	.19	.01	.00	.20
2	.04	.00	.00	.04	.03	.00	.00	.03
3	.00	.00	.00	.00	.00	.00	.00	.00
4	.00	.00	.00	.00	.00	.00	.00	.00
Total	.82	.18	.00	1.00	.71	.23	.06	1.00
	Medium Loading Difference							
0	.19	.24	.27	.70	.03	.09	.58	.70
1	.15	.06	.03	.24	.04	.08	.13	.25
2	.05	.01	.00	.06	.04	.00	.00	.04
3	.00	.00	.00	.00	.01	.00	.00	.01
4	.00	.00	.00	.00	.00	.00	.00	.00
Total	.39	.31	.30	1.00	.12	.17	.71	1.00
	Large Loading Difference							
0	.02	.07	.54	.63	.00	.00	.67	.67
1	.03	.04	.13	.20	.00	.01	.28	.29
2	.08	.03	.01	.12	.00	.00	.01	.01
3	.03	.00	.00	.03	.02	.00	.00	.02
4	.02	.00	.00	.02	.01	.00	.00	.01
Total	.18	.14	.68	1.00	.03	.01	.96	1.00
	Mixed Pattern of Loadings (N = 500)				*Nonproportional Loadings (N = 500)*			
0	.00	.00	.00	.00	.00	.00	.00	.00
1	.00	.00	.00	.00	.01	.02	.25	.28
2	.14	.22	.21	.57	.08	.17	.20	.45
3	.16	.01	.00	.17	.08	.03	.00	.11
4	.26	.00	.00	.26	.16	.00	.00	.16
Total	.56	.23	.21	1.00	.33	.22	.45	1.00

Note. The possible maximum number of true detections was 4 and the possible maximum number of false detections was 2 in this case.

TABLE B4
Twelve Variables With High Frequency

	N = 200						*N = 500*					
	NF						*NF*					
NT	*0*	*1*	*2*	*3*	*4*	*Total*	*0*	*1*	*2*	*3*	*4*	*Total*
						Small Loading Difference						
0	.28	.16	.04	.02	.00	.50	.14	.13	.08	.07	.02	.44
1	.23	.10	.05	.00	.00	.38	.23	.09	.08	.01	.02	.43
2	.09	.01	.00	.00	.00	.10	.08	.01	.00	.00	.01	.10
3	.01	.00	.00	.00	.00	.01	.02	.00	.00	.00	.00	.02
4	.01	.00	.00	.00	.00	.01	.00	.01	.00	.00	.00	.01
5	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
6	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
7	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
8	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Total	.62	.27	.09	.02	.00	1.00	.47	.24	.16	.08	.05	1.00
						Medium Loading Difference						
0	.01	.05	.16	.08	.11	.41	.00	.00	.00	.07	.45	.52
1	.06	.08	.08	.08	.03	.33	.00	.00	.04	.05	.26	.35
2	.05	.09	.02	.00	.02	.18	.00	.01	.00	.02	.06	.09
3	.04	.01	.00	.00	.00	.05	.00	.00	.01	.01	.00	.02
4	.02	.00	.00	.00	.00	.02	.00	.01	.00	.00	.00	.01
5	.00	.00	.00	.00	.00	.00	.00	.01	.00	.00	.00	.01
6	.01	.00	.00	.00	.00	.01	.00	.00	.00	.00	.00	.00
7	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
8	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Total	.19	.23	.26	.16	.16	1.00	.00	.03	.05	.15	.77	1.00
						Large Loading Difference						
0	.00	.00	.00	.03	.46	.49	.00	.00	.00	.00	.48	.48
1	.00	.00	.00	.03	.29	.32	.00	.00	.00	.00	.34	.34
2	.00	.01	.03	.03	.06	.13	.00	.00	.00	.00	.15	.15
3	.00	.01	.00	.03	.00	.04	.00	.00	.00	.02	.00	.02
4	.01	.00	.00	.00	.00	.01	.00	.00	.00	.00	.00	.00
5	.01	.00	.00	.00	.00	.01	.00	.00	.00	.00	.00	.00
6	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
7	.00	.00	.00	.00	.00	.00	.01	.00	.00	.00	.00	.01
8	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Total	.02	.02	.03	.12	.71	1.00	.01	.00	.00	.02	.97	1.00

Note. The possible maximum number of true detections was 8 and the possible maximum number of false detections was 4 in this case.