Structural Equation Modeling, 15:96–113, 2008
Copyright © Taylor & Francis Group, LLC
ISSN: 1070-5511 print/1532-8007 online
DOI: 10.1080/10705510701758349

Psychology Press
Taylor & Francis Group

Multigroup Confirmatory Factor Analysis: Locating the Invariant Referent Sets

Brian F. French
Department of Educational Studies
Purdue University

W. Holmes Finch
Department of Educational Psychology
Ball State University

Multigroup confirmatory factor analysis (MCFA) is a popular method for the examination of measurement invariance and specifically, factor invariance. Recent research has begun to focus on using MCFA to detect invariance for test items. MCFA requires certain parameters (e.g., factor loadings) to be constrained for model identification, which are assumed to be invariant across groups, and act as referent variables. When this invariance assumption is violated, location of the parameters that actually differ across groups becomes difficult. The factor ratio test and the stepwise partitioning procedure in combination have been suggested as methods to locate invariant referents, and appear to perform favorably with real data examples. However, the procedures have not been evaluated through simulations where the extent and magnitude of a lack of invariance is known. This simulation study examines these methods in terms of accuracy (i.e., true positive and false positive rates) of identifying invariant referent variables.

In the analysis of structural equation models, the importance of measurement invariance (MI), particularly factor invariance, cannot be overstated. For instance, when questions concerning latent mean differences across groups are of central

Correspondence should be addressed to Brian F. French, Purdue University, Department of Educational Studies, 100 N. University Street, West Lafayette, IN 47907. E-mail: frenchb@purdue.edu

interest, evidence of factor invariance must be provided before differences can be accurately examined and interpreted (Bollen, 1989). In fact, a hierarchy of MI (Little, 1997) requires the psychometric properties of an instrument be equivalent (i.e., configural, metric, measurement error, and scalar invariance; see Bollen, 1989; Horn & McArdle, 1992; Jöreskog, 1971; Meredith, 1993; Thurstone, 1947) before group differences in latent means are evaluated (e.g., Bollen, 1989; Sörbom, 1974). Examination of methods to establish MI is important as interest in latent means analysis continues to grow given that observed-means analyses (a) are inconsistent with examining group differences when latent factor mean differences are of interest (Hancock, 2004) and (b) can lead to inaccurate results (Cole, Maxwell, Avery, & Salas, 1993). Furthermore, the focus on the establishment and importance of MI in research examining differences between groups has been insufficient (Rensvold & Cheung, 2001).

A common method for examining factor invariance is multigroup confirmatory factor analysis (MCFA). MCFA allows for testing an a priori latent structure theory across groups (Alwin & Jackson, 1981) or time (Golembiewski, Billingsley, & Yeager, 1976), which yields comparisons of specific factor model features (e.g., factor loadings). The CFA model, following Jöreskog and Sörbom (1996), can be written as:

$$x = \Lambda_x \xi + \delta, \tag{1}$$

where x is a vector of observed variables (e.g., subtest scores), Λ is a matrix of factor loadings that relate the factor to the observed variables, ξ is a vector of underlying factors, and δ is a vector of measurement errors. Equation 1 does imply:

$$\Sigma = \Lambda \Phi \Lambda' + \Psi \tag{2}$$

where Σ is the covariance matrix of the observed variables, Λ is defined as in Equation 1, Φ is a covariance matrix of the underlying factors, and Ψ is a covariance matrix of measurement errors. In actual practice, these population values are estimated (e.g., $\hat{\Sigma}$) using sample data. For MCFA and factor invariance testing, the values in Equations 1 and 2 would be estimated for each group and then compared to determine whether they are invariant between groups. If latent mean differences, and not just factor structure invariance, are of interest, scalar values (e.g., intercepts) must be added to the model and tested for invariance before estimating latent means. In the presence of noninvariant intercepts, expected observed score differences might reflect group differences in the underlying trait (e.g., math ability), or systematic measurement bias (Cole et al., 1993; Hancock, 1997). However, our focus in this study was on factor structure invariance not including intercept invariance.

Invariance testing involves comparing increasingly more restricted factor models by sequentially constraining different parameter estimates (e.g., factor loadings, error variances) invariant across groups. See Maller and French (2004) for an applied example. The presence (or absence) of MI is determined by the examination of differences in the chi-square goodness-of-fit statistics for more and less restrictive models. A nonsignificant difference in these chi-square values indicates invariance. If a significant decline in fit occurs, then each component of a matrix (e.g., a factor loading) is constrained to be equal between groups in an attempt to locate the source of noninvariance (i.e., a specification search; Millsap, 2005). Recommendations have been made for researchers to continue to evaluate the accuracy of these methods (e.g., Millsap, 2005) as the assumption that these procedures work well under a wide variety of conditions might not be tenable. Researchers have addressed this issue to some degree (e.g., French & Finch, 2006; Meade & Lautenschlager, 2004) and identified conditions in which these methods appear most effective.

One potentially problematic issue in MCFA involves the need to constrain a referent indicator to be equal across groups (Millsap, 2005). This standardization or model identification procedure serves to assign units of measurement to the latent variables (Jöreskog & Sörbom, 1996) by either (a) setting the factor variances to 1.0 (e.g., $\Phi_{11} = \Phi_{11}$ across groups) or (b) setting a factor loading not being tested for invariance to 1.0 (e.g., $\lambda_{11} = \lambda_{11}$ across groups). Cheung and Rensvold (1999) provided analytical examples of each standardization procedure. The latter method of standardization, which is more commonly used (Vandenberg & Lance, 2000) and popular in the applied settings, as pointed out by an anonymous reviewer, assumes the referent factor loading is invariant. This assumption is not directly testable because only the ratio of factor loadings can be tested across groups (see Cheung & Rensvold, 1999, for a complete description). These equality constraints are generally thought of as being for model identification purposes only (Steiger, 2002) and selection of loadings to be held invariant is typically somewhat arbitrary. It should be noted, however, that despite this lack of a systematic selection procedure, different constraint choices can lead to quite different model fit results (Millsap, 2001; Steiger, 2002). It would appear, therefore, that the selection of which loading to constrain might be more important than would appear at first blush.

If the referent factor loading invariance assumption is violated, parameter estimates can be distorted, which could lead to inaccurate conclusions regarding invariance for the other loadings being tested (Bollen, 1989; Cheung & Rensvold, 1999; Millsap, 2005). Other variants of constraining loadings across groups to achieve standardization have been suggested (Drasgow & Kanfer, 1985; Reise, Widaman, & Pugh, 1993) but are not the focus here. Given the necessity of this assumption to conduct invariance testing, a circular situation exists where (a) the referent variable must be invariant, (b) invariance cannot be established without estimating a model, and (c) model estimation requires an invariant referent, which brings the process back to the original invariant referent assumption.

This circular conundrum is parallel to that found with the purification process recommended with differential item functioning (DIF) analysis, another method used to establish item-level MI. Purification in DIF analysis makes an effort to identify a set of non-DIF items for use as the matching criterion in DIF detection. Matching on a criterion or referent comprised of DIF items can lead to inaccurate DIF identification (Clauser, Mazor, & Hambleton, 1993). Ability purification has been recommended (e.g., Holland & Thayer, 1988; Lord, 1980; Marco, 1977) for such situations and can lead to more accurate DIF detection (Ackerman, 1992; Clauser et al., 1993). Thus, a similar procedure with MCFA would seem appropriate.

As a way of dealing with this circular problem in MI testing, the factor ratio test and the stepwise partitioning (SP) procedure in combination have been suggested to test for referent invariance in MCFA (Cheung & Rensvold, 1999; Rensvold & Cheung, 1998, 2001). This combination approach is an extension of earlier suggestions for dealing with this problem (Byrne, Shavelson, & Muthén, 1989). However, the proposed method uses each variable, in turn, as the referent in a set of models with each other variable constrained to be invariant. The iterative search procedure tests all such pairs of variables (i.e., $p(p-1)/2$ pairs) in the attempt to identify invariant variable sets. For instance, consider a case with six observed variables. Using this method for identifying sets of invariant variables would require conducting 15 separate tests to examine all of the possible variable pairs. This procedure would become quite labor intensive as the number of variables increases. An invariant set is a group of variables "defined by the property that every item passes the test of invariance (as an argument) when every other member of the set is used as a referent" (Rensvold & Cheung, 1998, p. 1023). Clearly, when the invariant variable sets are found, the noninvariant sets are also identified. In this case, the noninvariant sets are defined as those where one of the items does not pass the test of invariance when another item serves as the referent. To control Type I error with such a procedure, adjustment of the alpha level by dividing the chosen alpha by the number of sequential tests is suggested (Bollen, 1989; Rensvold & Cheung, 1998).

Once noninvariant pairs are identified with the factor ratio test, the SP procedure identifies invariant subsets in the following four steps. Table 1 illustrates three examples of the procedure, assuming a single underlying factor. Step 1 involves simply listing all variables. In Step 2 the variables are sorted into noninvariant pairs, with one indicator serving as the referent and the other as the tested indicator (i.e., argument), based on the factor ratio test results (i.e., those for which the change in chi-squares of more and less constrained models was significant). If a noninvariant pair is not contained in the subset, the subset passes to the next stage without change. Otherwise, new subsets are created where the first subset contains all variables except the first variable in the noninvariant pair

TABLE 1
Examples of Stepwise Partitioning Procedure

Example 1							
1 2 3 4	(1,4)[a]	2 3 4	(3, 4)[a]	2 4	2 4**		
				2 3	1 2 3**		
		1 2 3	(3,4)[a]	1 2 3			
Example 2							
1 2 3 4	(1,4)[a]	2 3 4	(2,3)[a]	3 4[b]			
				2 4[b]			
		1 2 3	(2,3)[a]	1 3[b]			
				1 2[b]			
Example 3							
1 2 3 4	(1,2)[a]	2 3 4	(1,3)[a]	2 3 4	(1,4)a	2 3 4	2 3 4[b]
		1 3 4	(1,3)[a]	3 4	(1,4)[a]	3 4	
				1 4	(1,4)[a]	4	
						1	

Note. In Example 3, if pair 1, 4 was not identified as a noninvariant pair, the pair would be considered invariant. This would be an incorrect conclusion based on the simulated results in this study.

[a]Noninvariant pair. [b]Final invariant sets.

and the second subset contains all variables except the second variable in the noninvariant pair. At Step 3, the new subsets are used to apply Step 2 again using the next noninvariant pair, continuing until no more noninvariant pairs remain. In Step 4, all subsets of larger combinations are eliminated and only invariant sets remain.

To provide greater insights into the workings of the method studied here, consider Example 1 in Table 1. The pair of 1 and 4 were found to be noninvariant by the factor ratio test, so that the original set of four indicators is divided into two subsets (2, 3, 4 and 1, 2, 3) so that these variables are kept apart. The other pair found to be noninvariant was 3, 4, leading to the rearrangement of the indicators into three subsets containing only invariant pairs (2, 4; 2, 3; and 1, 2, 3). Finally, these subsets can be rearranged again because the 2, 3 pair is also contained in the 1, 2, 3 set, leading to a final solution with two invariant sets (2, 4 and 1, 2, 3).

Ideally, the resulting invariant sets would leave noninvariant variables clearly easy to identify. For instance, in Table 1, Example 3, variable 1 would appear to be the culprit related to the underlying factor. However, Examples 1 and 2 are possible where more than one invariant set is the result and they might not be exclusive. Thus, the single underlying factor is now "split" into sets that are supposedly measured the same across groups. Interpretation of the results, as in most invariance studies, relies on theory and content to determine why a difference would occur between groups, potentially revealing interesting group differences for subsets of the observed variable loadings. One assumption could be made that variables appearing in multiple sets (the nonexclusive case) might be the only ones that are really invariant. Of course, noninvariant variables or sets of variables can be simply deleted from the scale or one could continue working with the scale under the assumption of partial MI (Byrne et al., 1989; Rensvold & Cheung, 1998; Riese et al., 1993).

Although this procedure has been applied to real data as an example (Rensvold & Cheung, 2001), simulation research where the magnitude and amount of noninvariance is known has not been used to evaluate its accuracy (Millsap, 2005; Vandenberg & Lance, 2000). Thus, this Monte Carlo study applied the factor ratio test to simulated data under known conditions to evaluate false positive and true positive rates of identifying invariant referents. In this context, a false positive would be the identification of a variable pair as noninvariant when in fact it was invariant, and true positive refers to the correct identification of a variable pair as noninvariant. Note that only the factor ratio test was conducted as its results directly influence the SP procedure (see Table 1 for examples). That is, any inaccuracies in the factor ratio test would result in incorrect results in the SP procedure.

METHOD

Real and simulated data were used to establish a baseline and control factors that can influence the procedures, respectively. Replications ($N = 1{,}000$) for the baseline data (Set 1) and each combination of conditions with the simulated data (Set 2) ensured stable results. As described later, the baseline condition refers to simulated data conditions based on Rensvold and Cheung's (2001) applied example. The second simulation involved data that were simulated with known model differences across two groups under varying conditions of sample size, number of factors, number of indicators per factor, and percent of noninvariant indicators. Simulations were completed in PRELIS (Jöreskog & Sörbom, 2002) and LISREL (Jöreskog & Sörbom, 2005). We next describe the conditions used in the simulation of Set 2.

Number of Factors and Indicators

Data for Set 2 were simulated from both one- and two-factor models, with interfactor correlations set at .50 to represent moderately correlated factors. Correlations were not varied to avoid confounds. Three combinations of number of factors and number of indicators per factor were simulated: one factor, 6 indicators; two factors, 6 indicators (three per factor); and two factors, 12 indicators (six per factor).

Sample Size

The necessary sample size to obtain adequate power in factor analysis varies depending on the data conditions. Therefore, two group sample sizes were employed, 250 and 500, resulting in two sample size combinations: 250/250, 500/500, excluding the baseline condition with real data (i.e., $n = 755$, $n = 678$). These sample sizes are consistent with previous MI simulation research (e.g., Lubke & Muthén, 2004; Meade & Lautenschlager, 2004) and reflect real data conditions.

Percent of Noninvariant Indicators

Three levels of factor loading differences across groups were simulated. To assess false positives (i.e., false identification of a lack of invariance), the case of complete invariance (i.e., no loading differences across groups) was simulated. Additionally, to assess true positive rates (i.e., correct identification of noninvariance), 17% (i.e., low contamination condition) and 34% (i.e., high contamination condition) of the factor loadings differed across groups. Percentages were chosen (a) for practical reasons (i.e., resulted in a whole number of differing loadings) and (b) to reflect what might be found in actual data.

Data Generation and Analysis

Baseline simulated data: Set 1. The variance–covariance matrices from Rensvold and Cheung's (2001) example were used for initial examination of the procedure across replications for (a) baseline conditions of real data and (b) a check of the simulated data procedures employed in this study. Each matrix contained four variables representing a one-factor model measuring organizational commitment with samples from the United States and Japan. See Rensvold and Cheung (2001) for data details. These data allowed evaluation of how well the factor ratio test recovered the original findings (i.e., identification of two noninvariant pairs) from Rensvold and Cheung. Identification of any other sets would give a different result with the SP procedure. To evaluate the false positive rate, one group's data were used to create two groups with identical matrices. Thus, any identified noninvariant pair would represent a false positive. True positive rates were evaluated using the same data structure employed for the false positive condition except that one variable was simulated to be noninvariant with a .37 difference in the factor loadings between groups (the difference in the variables' factor loading in the original data). True positive rates for the initial replication of the real data were defined as the proportion of noninvariant variables identified (i.e., 1, 2; 1, 3; 1, 4 as indicated by Rensvold & Cheung, 2001).

Simulated data: Set 2. In addition to the simulations based on the observed data structure presented by Rensvold and Cheung (2001), a second set of simulations based on data with known covariance matrices, factor loadings, and interfactor correlations was also generated, where Group 1 data represented the initial model and Group 2 differed on specified factor loadings. Individual factor indicators were simulated to have normal distributions to reflect data at a subtest level. Factor loadings (i.e., lambdas) were set at 0.60, variances (i.e., Phi) were set to 1.0, and error terms (i.e., theta-deltas) for the observed variables were defined as 1.0 minus the square of the factor loadings. These values are consistent with previous MCFA simulation work (e.g., French & Finch, 2006; Meade & Lautenschlager, 2004). The error value definition assumes no specific variance (i.e., all error variance) is present, was tenable to avoid potential confounding factors with the results, and was not central to the questions in this study. A difference of .25 in the factor loadings for the noninvariant variables was employed (i.e., loadings for one group were set to .85, whereas those of the other remained .60) and is consistent with previous simulation work with MCFA invariance testing (French & Finch, 2006; Meade & Lautenschlager, 2004). All differences in indicator loadings favored the same group.

Invariance Testing

Two primary models were tested for each simulation: Model 1 was the baseline with all parameters allowed to vary across groups, and resulted in the first chi-square value for comparison with Model 2, which imposed the equality of factor loadings constraint across groups. The difference in the Model 2 and Model 1 chi-squares was used to evaluate overall invariance. Finally, single-indicator invariance detection occurred where each variable was tested for invariance using each of the other indicators, in turn, as invariant referents. To ensure that statistically significant results were not due to model misfit, a variety of fit indexes were examined, including the chi-square goodness-of-fit test, comparative fit index (CFI), Goodness-of-Fit Index (GFI), and standardized root mean squared residual (SRMR). Data were modeled assuming appropriate specification, so that these statistics should indicate adequate model fit. As described earlier, the false positive rate for Set 2 was defined as the proportion of incorrect identifications of invariant variables as noninvariant, whereas true positives for Set 2 were defined as the proportion of noninvariant pairs that were correctly identified. To determine which of the manipulated factors were significantly related to false positive and true positive rates, variance components analysis was used.

RESULTS

Prior to examining the false positive and true positive rates, it was necessary to establish that the structural models fit the data properly. Acceptable model fit is important because if the latent structure is misspecified, resulting false positive and true positive rates might reflect model misfit as well as or instead of identification of noninvariant referent variables. Such a result would make it difficult, if not impossible, to determine whether a significant result for an invariance test truly reflects group differences on a proposed referent variable or simple model misfit. Given that these simulations were conducted with properly specified models, the results presented here cannot be generalized to situations in which the model has been misspecified. To ensure that the models were correctly specified, several fit indexes were calculated for the real data and for each replication under each simulated condition, and then averaged to provide the data for Tables 2 and 3 for the baseline (Set 1) and simulated (Set 2) data, respectively. All values suggest good fit for the models, based on guidelines for interpretation provided by Kline (2005) and Mueller (1996). Therefore, it is appropriate to interpret the false positive and true positive rates, which are described later for the Set 1 and Set 2 data, respectively.

TABLE 2
Average Model Fit Indexes Across 1,000 Replications for the General Form for the Baseline Data

	χ^2	*df*	*RMSEA*	*CFI*	*SRMR*
United States[a]	9.49	2	0.091	0.99	0.028
Japanese[b]	6.42	2	0.052	0.99	0.020
Combined	15.91*	4	0.076	0.99	0.020
Japanese[c]	6.27	2	0.045	1.00	0.018
Combined	12.69	4	0.049	1.00	0.020
Japanese[d]	6.33	2	0.013	0.97	0.042
Combined	12.75	4	0.097	0.98	0.042

Note. RMSEA = root mean squared error of approximation; CFI = comparative fit index; SRMR = standardized root mean squared residual.

[a] $N = 755$. [b] $N = 678$. [c] Japanese data simulated to be equal to the original Japanese data. [d] Japanese data simulated with one variable noninvariant.

$^*p < .05$.

TABLE 3
Average Fit Indexes Across 1,000 Replications for the General Form of the Models Across Conditions

		χ^2	*df*	*p Value*	*SRMR*	*GFI*	*CFI*
One factor, 6 indicators							
0%[a]	250[b]	18.31	18	.49	0.003	0.99	0.99
	500[c]	27.27	18	.51	0.002	0.99	0.99
17%	250	18.37	18	.48	0.003	0.99	0.99
	500	18.07	18	.49	0.002	0.00	0.99
34%	250	18.37	18	.48	0.002	0.99	0.99
	500	17.92	18	.50	0.002	0.99	0.99
Two factors, 6 indicators							
0%	250	15.95	16	.50	0.011	0.99	0.99
	500	16.14	16	.49	0.007	0.99	0.99
17%	250	16.28	16	.49	0.011	0.99	0.99
	500	16.21	16	.49	0.008	0.99	0.99
34%	250	15.79	16	.51	0.009	0.99	0.99
	500	16.09	16	.50	0.007	0.99	0.99
Two factors, 12 indicators							
0%	250	108.52	106	.45	0.013	0.96	0.99
	500	107.08	106	.48	0.009	0.98	0.99
17%	250	108.92	106	.44	0.013	0.96	0.99
	500	106.92	106	.48	0.009	0.98	0.99
34%	250	108.87	106	.45	0.011	0.97	0.99
	500	107.42	106	.47	0.008	0.98	0.99

Note. SRMR = standardized root mean squared residual; CFI = comparative fit index; GFI = Goodness-of-Fit Index.

[a]Percentage of noninvariant indicators. [b,c]Sample size per group.

Baseline Data for Initial Evaluation: Set 1

To ensure that the simulations based on the real data are appropriate, we first replicated the work of Rensvold and Cheung (2001) with the U.S. and Japanese data. Given the model fit results averaged across replications reported in Table 2, it appears that the simulation of the model mirrors results reported by Rensvold and Cheung. The equality constraint of the factor loadings resulted in a significant decline in fit—for example, the mean $\chi^2_{difference}$ (3, N = 1,433) = 28.15, $p < .001$—from the unconstrained model, indicating noninvariance between the groups, and was comparable to their results, $\chi^2_{difference}$ (3, N = 1,433) = 24.23, $p < .001$. The true positive rate was .807 for identifying the same pairs as Rensvold and Cheung (2001) across replications. These results were expected as they replicate the outcome in the original analysis. Note this difference between the Japanese and U.S. data was not simulated and provided a confirmation check on following their procedures.

The false positive rate was evaluated by simulating identical matrices from the Japanese data across replications, as there would be no noninvariant variables. The factor loading equality constraint did not result in a significant decline in fit averaged across replications, $\chi^2_{difference}$ (3, N = 1,356) = 2.97, $p > .05$. A false positive was the identification of a variable as noninvariant. The error rate was .041. The true positive rate was evaluated where data from the false positive condition were used except Variable 1 was simulated to be noninvariant between the two simulated groups. The equality constraint of the factor loadings resulted in a significant decline in fit averaged across replications, $\chi^2_{difference}$ (3, N = 1,356) = 13.72, $p = .003$. The false positive rate was .017 for those indicators whose loadings were invariant, and the true positive rate was .79 for one noninvariant pair identified.

Simulated Data: Set 2

False positives. The results of the variance components analysis indicated that only the number of indicators was significantly related to the false positive rate ($p = .042$) and accounted for 46.6% of the variance in this outcome. Table 4 includes the false positive rates by all combinations of conditions.

Although it does appear that the false positive rates are somewhat lower for 12 indicators, it also is clear that in no case is the rate very elevated above the Bonferroni corrected alpha values of .002 (12 indicators) or .003 (6 indicators). Indeed, the difference in Type I rates by number of indicators could have been a result of the aforementioned Bonferroni corrected alpha values for the two conditions. In other words, significant differences between the observed false positive rates of the two conditions (6 and 12 indicators) might simply reflect the systematic difference between the nominal false positive rates that is due to the Bonferroni correction. As is evident in Table 4, the presence of noninvariant indicators (contamination conditions of 17% and 34%) did not appear to influence the false positive rate for the invariant indicators, nor did the number of participants or factors. In summary, it appears that the false positive rate of the procedure is consistently accurate across virtually all conditions included in this study.

True positives. The variance components analysis for the true positive rate of the factor ratio test in detecting noninvariant indicators identified the number of factors ($p = .048$) and level of contamination ($p < .001$) as statistically significant. The number of noninvariant indicators and the number of factors accounted for 67.6% and 14% of the variance in true positives, respectively. The true positive results by the manipulated variables appear in Table 5.

Across all other conditions, true positives were lower when there were a greater number of noninvariant indicators. In addition, there were higher true

TABLE 4
False Positive Rates Across Conditions
for 1,000 Replications

Type I		
One factor, 6 indicators		
0%[a]	250[b]	0.002
	500[c]	0.004
17%	250	0.004
	500	0.003
34%	250	0.003
	500	0.002
Two factors, 6 indicators		
0%	250	0.002
	500	0.002
17%	250	0.004
	500	0.004
34%	250	0.004
	500	0.003
Two factors, 12 indicators		
0%	250	0.002
	500	0.002
17%	250	0.002
	500	0.002
34%	250	0.002
	500	0.002

[a] Percentage of noninvariant indicators.
[b,c] Sample size per group.

positive rates in the one-factor case than for either of the two factor conditions when 34% of the indicators were not invariant, though true positives were 1.0 for both six-indicator conditions with 17% contamination. This result follows suggestions that locating noninvariance might be easier when only a few variables lack invariance (Millsap, 2005). In general, the true positive rates of the procedure were relatively high except with two factors and high contamination, in which case it was below .7. As was the case for false positives, the sample size did not appear to have an influence on true positives.

The SP procedure was not carried out in this study because, as seen in the examples in Table 1, the SP results rely on the ability of the factor ratio test to correctly identify the noninvariant variables. As the results presented here illustrate, the SP procedure would work well in the conditions with less

TABLE 5
True Positive Rates Across Conditions for 1,000 Replications

		True Positive
	One factor, 6 indicators	
17%[a]	250[b]	1.0
	500[c]	1.0
34%	250	0.89
	500	0.89
	Two factors, 6 indicators	
17%	250	1.0
	500	1.0
34%	250	0.66
	500	0.66
	Two Factors, 12 indicators	
17%	250	0.89
	500	0.89
34%	250	0.57
	500	0.57

[a] Percentage of noninvariant indicators.
[b,c] Sample size per group.

contamination and simpler models (e.g., one factor, or less indicators), as the true positive rate was 1.0. The SP procedure would have diminished accuracy in the other conditions where the true positive rate of the factor ratio test was much lower. In relation to false positives, it appears that levels were low enough that using the procedure where invariance holds would result in relatively few noninvariant sets being identified, an encouraging finding given the amount of noninvariance is not known a priori.

CONCLUSIONS

The results presented in this study support the notion that the factor ratio test and the stepwise portioning procedure described by Rensvold and Cheung (2001) for identifying invariant sets of variables to be used in a more complete model invariance analysis maintains the nominal false positive rate across a variety of conditions. Note that these results only apply to the situation when factor loadings are used as the referents. Indeed, the only factor appearing to significantly influence the false positive rate was the number of indicators, which might have been due more to the fact that the nominal alpha for the 12 indicators setting is lower (.002) than for 6 indicators (.003), and therefore the associated observed false positive rates also differed systematically. It is particularly interesting that the false positive rate for invariant indicators was not influenced by the presence of other, noninvariant variables in the data. This result suggests that if a mixture of invariant and noninvariant indicators is present in the data, the factor ratio test and the SP procedure will correctly identify the invariant variables in conditions similar to those studied here.

The true positive rate of the procedure to correctly identify noninvariant indicators was above .80 except for two-factor models with 34% noninvariant indicators. The fact that it was more difficult to identify noninvariant indicators in higher contamination conditions (i.e., more noninvariant indicators) might not be particularly surprising. The estimate of the latent trait being measured could be contaminated by the presence of other noninvariant variables, leading to problems in estimation of the factors. It appears, therefore, that there is a higher probability of selecting a noninvariant variable as the referent within the higher contamination conditions. This can lead to partial metric invariance models that fit poorly to the data due to the use of a noninvariant referent variable. In turn, findings of a lack of metric invariance might not reflect actual differences between groups, as the latent variable was standardized to different metrics based on the referent variable having a different relation to the latent variable across groups. This is in accord with Millsap's (2005) suggestion that it might be easier to locate a lack of invariance when only a few variables lack invariance. Of particular concern are cases where the model is relatively more complex (e.g., two factors) and a third of the indicators are contaminated. In these cases, less than 70% of the noninvariant indicators were correctly identified, and with 12 indicators the true positive rate is below .6.

In addition to the proportion of noninvariant indicators, the number of factors also appeared to significantly influence true positives. Specifically, the presence of more factors was associated with lower true positives, and in this study the two-factor, 12-indicator condition had somewhat lower true positives than the two-factor, 6-indicator condition. This result suggested that the factor ratio test might be somewhat sensitive to model complexity; that is, for more complicated models it is increasingly difficult to detect noninvariant variables. As with the false positive rate, sample size did not have any discernible influence on the ability of the procedure to identify the noninvariant indicators.

The results for the true positives and false positives presented here generally paint a positive picture of the factor ratio test and SP procedure's utility for identifying noninvariant indicators. It is clear that a researcher using this method will be very unlikely to identify an invariant indicator as differing between groups. In addition, when the proportion of indicators that exhibit noninvariance is relatively low, the procedure effectively identifies noninvariant indicators in the vast majority of cases. On the other hand, it does appear that when the level of contamination among the indicators is higher (34% in this study), the procedure does have diminished true positive rates for identifying noninvariant variables, with the rate being .57 in the worst case.

Although the factor ratio test (and by association the SP procedure) generally exhibits positive performance, it should be noted that the method is nontrivial to carry out. As the number of indicators increases, the number of tests to conduct with the factor ratio test also increases. For instance, a relatively short instrument (i.e., six indicators) would require 15 individual invariance tests and a moderate-length instrument (i.e., 25 indicators) would require 300 individual invariance tests. In cases such as the latter, carrying out the SP procedure becomes quite complicated and rather time intensive unless one were to automate the process, which might be too time intensive for many practitioners.

Limitations and Directions for Future Research

A few limitations should be kept in mind when interpreting the results of this study. First, the scope of the models examined was somewhat constrained, with the most complicated design including two factors and 12 indicators. It should be noted, however, that the method outlined by Rensvold and Cheung (2001) requires a great many analyses, and that more complicated models require exponentially more time and resources than those examined here. Second, only properly specified baseline models were simulated. Thus, results might not generalize to situations when baseline models are improperly specified. Third, the sample sizes examined were selected to represent those often seen in educational and psychological assessments that are not associated with national or statewide testing programs. It is recognized, however, that in some cases smaller samples might be encountered, especially with low-incidence populations. Fourth, data were drawn from a normally distributed population and with use of maximum likelihood estimation. In some contexts, invariance studies are used with questionnaires or cognitive instruments in which the indicators are item responses with a restricted range of responses (e.g., ordinal or dichotomous). However, given the results were encouraging for the procedure, future simulation work might begin to focus on more complex situations (e.g., model misfit and complexity, data structures, estimation procedures, etc.), with an eye toward the development of an automated mechanism for the practitioner to use in conducting these invariant search analyses.

The continued development and evaluation of procedures to detect invariance are warranted (Millsap, 2005; Vandenberg & Lance, 2000) as MI is a prerequisite for many analyses (e.g., between-group or time comparisons, cross-cultural research, etc.). For instance, evidence of MI must be provided before group comparisons are conducted on scores produced by instruments. That is, for comparisons to be meaningful, evidence that the trait being measured has similar meaning across groups or time must be presented. It is the view of some that MI has not received enough attention from those examining group differences (Rensvold & Cheung, 2001). If the methods to detect MI are not accurate, then subsequent analyses and decisions based on these analyses are meaningless and, more important, potentially harmful for the individuals affected by such decisions. For this reason, there must be approaches available that can successfully identify noninvariant indicators to avoid the use of these as referents in an MI study.

Rensvold and Cheung's (2001) proposed method appeared to function well under many of the studied conditions. At the same time, using this approach is nontrivial, requiring multiple analyses of the data, and can be very labor intensive when there are a number of factors and indicators to be tested. Rensvold and Cheung (1998) indicated that a program for assisting with such analyses has been developed. Such a program would make the procedure less labor intensive. Additionally, alternative methods, such as the use of exploratory factor analysis (see Millsap, 2001; Vandenberg, 2002) should be investigated and compared to the factor ratio test and SP procedure to find the most effective method that would also be relatively easy to use for the practitioner who selects the factor loading constraint as the method to assist with model identification. Nonetheless, the results of this study do suggest that if practitioners are willing to invest the time and effort to conduct such analyses, they should be rewarded by a method that controls the false positive rate and provides reasonable true positives under many conditions.

REFERENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67–91.

Alwin, D. F., & Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. In D. Jackson & E. Borgatta (Eds.), *Factor analysis and measurement in sociological research: A multi-dimensional perspective* (pp. 249–279). Beverly Hills, CA: Sage.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

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. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25,* 1–27.

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel–Haenszel procedure. *Applied Measurement in Education, 6,* 269–279.

Cole, D. A., Maxwell, S. E., Avery, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin, 114,* 174–184.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70,* 662–680.

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

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science, 12,* 133–157.

Hancock, G. R. (1997). Structural equation modeling methods of hypothesis testing of latent variable means. *Measurement and Evaluation in Counseling and Development, 30,* 91–105.

Hancock, G. R. (2004). Experimental, quasi-experimental and nonexperimental design and analysis with latent variables. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 317–334). Thousands Oaks, CA: Sage.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Holland & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 57,* 409–426.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL8: User's reference guide.* Chicago: Scientific Software International.

Jöreskog, K., & Sörbom, D. (2002). PRELIS (Version 2) [Computer software]. Chicago: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (2005). LISREL (Version 8.72) [Computer software]. Chicago: Scientific Software International.

Kline, R. B. (2005). *Principles and practice of structural equation modeling*. New York: Guilford.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 31,* 53–76.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11,* 514–534.

Maller, S. J., & French, B. F. (2004). Factor invariance of the UNIT across deaf and standardization samples. *Educational and Psychological Measurement, 64,* 647–660.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139–160.

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.

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

Millsap, R. E. (2005). Four unresolved problems in studies of factorial invariance. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 153–172). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Millsap, R. E. (2001). When trivial constraints are not trivial: The choice of uniqueness constraints in confirmatory factor analysis. *Structural Equation Modeling, 8,* 1–17.

Mueller, R. O. (1996). *Basic principles of structural equation modeling*. New York: Springer.

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

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

Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim & L. L. Neider (Eds.), *Research in management: Equivalence in measurement* (pp. 25–50). Greenwich, CT: Information Age.

Sörbom, D. (1974). A general model for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology, 27,* 229–239.

Steiger, J. H. (2002). When constraints interact: A caution about reference variables, identification constraints, and scale dependencies in structural equation modeling. *Psychological Methods, 7,* 210–227.

Thurstone, L. L (1947). *Multiple factor analysis; A development and expansion of the vectors of the mind.* Chicago: University of Chicago Press.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5,* 139–158.

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,* 4–69.