

5-1-2008

Using Exploratory Factor Analysis for Locating Invariant Referents in Factor Invariance Studies

W. Holmes Finch

Ball State University, whfinch@bsu.edu

Brian F. French

Washington State University, frenchb@wsu.edu

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

Recommended Citation

Finch, W. Holmes and French, Brian F. (2008) "Using Exploratory Factor Analysis for Locating Invariant Referents in Factor Invariance Studies," *Journal of Modern Applied Statistical Methods*: Vol. 7 : Iss. 1 , Article 18.
DOI: 10.22237/jmasm/1209615420

Using Exploratory Factor Analysis for Locating Invariant Referents in Factor Invariance Studies

W. Holmes Finch
Ball State University

Brian F. French
Washington State University

Model identification in multi-group confirmatory factor analysis (MCFA) requires an equality constraint of referent variables across groups. Invariance assumption violations make it difficult to locate parameters that actually differ. Suggested procedures for locating invariant referents are cumbersome, complex, and provide imperfect results. Exploratory factor analysis (EFA) may be an alternative because of its ease of use, yet empirical evaluation of its effectiveness is lacking. EFAs accuracy for distinguishing invariant from non-invariant referents was examined.

Key words: Factor analysis, invariance, bias.

Introduction

The use of scores for making decisions about persons, be it for job placement, graduation from high school, acceptance to graduate school, or obtaining a license to operate a motor vehicle, relies on the continued accumulation of empirical and theoretical validity evidence to support such score use (Messick, 1989). One form of empirical validity evidence is measurement invariance or equivalence. An assessment instrument, for example, should have the same psychometric properties across groups to help ensure that measurement of the specified construct is the same across groups. In the absence of such evidence, group comparisons on the ability of interest may be meaningless, as observed differences could be the result of ability differences or measurement differences (i.e., a lack of invariance).

Holmes Finch is an Assistant Professor of research methods and statistics in the Teachers College at Ball State University. E-mail: frenchb@purdue.edu. Brian French is an Associate Professor teaching courses in research methods and statistics in the College of statistics in the College of Education at Washington State University. Email: Email: frenchb@wsu.edu

A common method for assessing factor invariance, a form of measurement invariance (MI), is multi-group confirmatory factor analysis (MCFA). MCFA allows for an *a priori* specified latent structure of an instrument to be assessed for MI across groups or time (e.g., Alwin & Jackson, 1981; Golembiewski, Billingsley, & Yeager, 1976). A powerful feature of MCFA is the ability to compare specific model features (e.g., factor loadings) at the matrix level, as well as individual elements of the matrix under examination.

Invariance testing in MCFA involves comparing increasingly more restricted factor models by sequentially constraining different parameter estimates (e.g., factor loadings, error variances) invariant or equal across groups or time. The presence of MI is determined using differences in the chi-square goodness-of-fit statistics for more and less restrictive models, where a non-significant difference indicates invariance. This procedure has been well documented both in theoretical and applied examples (e.g., Bollen, 1989; Byrne, Shavelson, & Muthén, 1989; Jöreskog & Sörbom, 1996; Maller & French, 2004; Raju, Laffitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993).

There are several procedural aspects of invariance testing that deserve further attention before practitioners and researchers have complete confidence in such results (Little, 2000). One of several unresolved issues in MCFA is the need to constrain a referent

INVARIANT REFERENTS IN FACTOR INVARIANCE STUDIES

indicator equal across groups (Millsap, 2005). Latent factors are constructed on arbitrary coordinate systems making comparison of models across populations difficult because they are not constrained to the same system in relation to the other populations or groups of interest (Wilson, 1981).

The model standardization, or identification, procedure can solve this problem by assigning units of measurement to the latent variables (Jöreskog & Sörbom, 1996), generally by aligning the latent factors to a scale based on the same indicators across groups. To meet identification requirements, per factor, either a factor variance or a factor loading is set to 1.0 across groups. Additional methods have been suggested (e.g., see Drasgow & Kanfer, 1985; Reise, Widaman, & Pugh, 1993), but the factor loading method appears to be used most commonly (Brown, 2006; Vandenberg & Lance, 2000). These procedures require the assumption that the referent variable constrained equal is, in fact, invariant. This assumption cannot be directly tested, however, because only the ratio of individual factor loadings to the referent can be compared across groups (Bielby, 1986; Cheung & Rensvold, 1999; Wilson, 1981). Furthermore, complications arise as different constraint choices may lead to different results in terms of model fit and hypotheses concerning equality of parameters (Millsap, 2001; Steiger, 2002; Wilson).

When the referent parameter is not invariant, estimates of other parameters may be distorted, which can lead to inaccurate conclusions regarding their invariance (Bollen, 1989; Cheung & Rensvold, 1999; Millsap, 2005). A circular situation exists with this assumption 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. Thus, we are back to the original invariant referent assumption. That is, to assess invariance for a given factor loading across groups, for instance, an equality constraint (that is actually true) must already be placed on another factor loading. This circular conundrum is parallel to the ability purification process in detection of differential item functioning (DIF) (e.g., Holland & Thayer, 1988; Lord, 1980), another method commonly

employed to establish MI at the item level. Ability purification in DIF analysis attempts to identify a set of non-DIF items for use as the matching criterion and can lead to more accurate DIF detection (Ackerman, 1992; Clauser et al., 1993). A similar procedure with MCFA would seem appropriate with the expected outcome of more accurate detection of a lack of MI.

A search procedure (i.e., factor-ratio test and the stepwise partitioning procedure) was designed to identify invariant and non-invariant variables (Rensvold & Cheung, 2001). The method uses each variable, in turn, as the referent in a set of models with each other variable constrained to be invariant. The iterative procedure tests all pairs of variables (i.e., $p(p - 1) / 2$ pairs) and becomes quite complex with many indicators, making it not “user-friendly” for practitioners (Vandenberg, 2002). A moderate length instrument (i.e., 30 indicators), for instance, requires 435 individual invariance tests. Furthermore, empirical evaluation of the method demonstrated adequate (e.g., acceptable false and true positives) but far from perfect performance (French & Finch, 2006a).

To overcome these limitations, exploratory factor analysis (EFA) may be a viable alternative for identifying invariant referents, as a purification step prior to a MCFA. That is, if a researcher intends to set one loading invariant across groups, a single EFA could be conducted for each group separately and loading estimates compared to ascertain which loadings appear to be invariant. With an EFA conducted on each group separately, such an analysis may be considered a weak test of factorial invariance (Zumbo, 2003).

EFA is not a formal test of invariance, but instead is a possible method to examine parameter estimates across groups to obtain a sense of the differences in the factor loadings without need of conducting a large number of analyses as is required when using the factor-ratio test. Specifically, pattern coefficients appearing most similar would be eligible for serving as a referent variable in the MCFA. Such use is in accord with suggestions that EFA be used to examine loadings with an “interocular eyeball test” (Vandenberg, 2002) to judge similarity of loadings to identify appropriate

referent variables. However, there does not appear to be empirical evaluation of EFA for locating potentially invariant referents.

The purpose of this study is to assess the utility of EFA in identifying non-invariant or invariant factor loadings between two groups. This procedure would be used prior to the actual MCFA as a “purification” process for identifying a loading that is likely to be invariant for use as the referent parameter. The procedure would simply entail conducting one EFA per group with one reference group and one comparison group. The loadings (i.e., pattern coefficients) from the separate analyses would be compared visually to determine similarity of individual loadings. Loadings that appear markedly different would not be used as a referent, while loadings appearing most similar would be used as the referent. If multiple loadings across groups were equally similar, any of them could serve as the referent.

Methodology

Simulated data were used to control variables that could influence the magnitude of factor loading estimates, with 1000 replications for each combination of conditions described below. Simulations and analyses were completed in SAS, V9.1 (The SAS Institute, 2003).

Number of Factors and Indicators

Data were simulated from both 1- and 2-factor models, with interfactor correlations set at .50 to represent moderately correlated factors. The number of indicators per factor was 6. Data were simulated to reflect simple structure for continuous and normally distributed substest level data.

Sample Size

The necessary sample size to obtain reasonable estimates in factor analysis varies depending on the data conditions. For this reason, three sample size conditions were simulated: 100, 500, and 1000 in order to reflect small, medium and large samples. These values are consistent with other factor analysis simulation studies (e.g., Cheung & Rensvold, 2002; Lubke & Muthén, 2004; Meade & Lautenschlager, 2004), ranging from poor (n=

100) to excellent (n =1000) (Comery & Lee, 1992), and may not be of much concern here as communalities were high (MacCallum, Widaman, Zhang, & Hong, 1999).

Magnitude of Difference with the Non-Invariant Indicators

Six levels of factor loading values for the non-invariant indicator were simulated. A baseline condition was established where no differences in loadings were present, so that the first indicator had a loading value of 0.75, as did the other variables. The remaining 5 conditions were characterized by declines in the target loading from 0.10 to 0.50 in increments of 0.10 (i.e., 0.65, 0.55, 0.45, 0.35, and 0.25). These levels were selected as there is no effect size, at least to the knowledge of the authors, for what represents a meaningful factor loading difference (Millsap, 2005) and the range covers values used in previous MCFA simulation work (e.g., French & Finch, 2006b; Meade & Lautenschlager, 2004).

Contamination

The location of invariant parameters may be influenced by the number of indicators that lack invariance (Millsap, 2005). Thus, the presence of a factor loading exhibiting a difference from 0.75 other than that for the target indicator was varied as either present or absent. In other words, for half of the simulated conditions only the target indicator loading was contaminated, while for the other half of the simulations a second target indicator loading also was contaminated at the same difference as the target indicator. This allowed assessment of the influence of additional contaminated variables.

Analysis

All analyses were conducted by group using principal axis factoring with PROMAX rotation in the 2-factor condition. These settings follow recommendations for using EFA for a referent indicator search and are more consistent with educational and psychological data (e.g., presence of measurement error, correlated factors; Vandenberg, 2002).

INVARIANT REFERENTS IN FACTOR INVARIANCE STUDIES

Evaluation Criteria

Factor loadings (i.e., pattern coefficients) obtained from the EFA for the target variable were compared with 0.75, which was the population value for the reference group. The assumption of this study was that if a researcher were to use EFA to identify invariant indicators, the observed loadings would be compared between the two groups, as described above. Therefore, performance could be judged by how well EFA would estimate factor loading values for the second group both when they differ in the population from that of the reference group, and when they do not. Three statistics across replications were used to operationalize this overall outcome: (a) the mean loading for the target variable (loading bias), (b) the standard deviation of the target loading, and (c) the percent of replications for which the observed loading was within 10% of the baseline loading of 0.75; i.e. between 0.675 and 0.825. This latter criterion was selected because of suggestions that bias values less than 10-15% may not be considered serious in many latent variable modeling situations (Muthén, Kaplan, & Hollis, 1987).

Results

Factor loading bias

Based on the Analysis of variance ANOVA ($\alpha = 0.05$) used to identify the manipulated variables and their interactions that were associated with factor loading bias, the 3-way interaction of magnitude of difference by number of factors by contamination was the highest order significant term. Other 2-way interactions involving combinations of these three variables also were statistically significant, as were the main effects of number of factors and the magnitude of the difference. The 3-way interaction had an η^2 value of only 0.02, while the magnitude of loading difference had an η^2 of 0.94. Thus, while the interaction should not be ignored, it is clear that the most important factor in determining the mean loading is the magnitude of the difference from the baseline of 0.75. For this reason, both terms are discussed below.

The means of factor loading estimates across the magnitude of difference, number of

factors and level of contamination appear in Table 1.

These values demonstrate that EFA, using principal axis factoring and PROMAX rotation, accurately estimates the population factor loading of 0.75 for both the 1- and 2-factor conditions when all other loadings also are 0.75. Furthermore, the estimates also were very close to the population value of 0.75 when a loading other than that for the target variable was set at 0.65 (i.e., contaminated condition). When the target loading was different from 0.75 in the population, the sample estimate was generally very close to the actual population value in the 1-factor case, regardless of whether other factor loadings were contaminated. This result was mirrored in the 2-factor case with no contamination. However, when non-target loadings were contaminated, the means of the target loadings reflect overestimation except when the target was 0.65. As expected due to high communalities, sample size was not significantly related to the mean value of the estimated factor loadings.

Standard Deviation

The ANOVA identified the interaction of sample size by number of factors by magnitude of the difference as statistically significant for the standard deviation of loading estimates. In addition, the main effects of magnitude of difference, sample size, number of factors, and contamination were also statistically significant. It should be noted that the sample size accounted for 75.5% of the variation in the standard deviation, while none of the other terms in the model accounted for more than 4%.

Table 2 contains the standard deviations of the factor loading estimates by the number of factors, sample size, and magnitude of difference between the target loading and 0.75. An examination of these results suggests that in general, larger sample sizes were associated with lower variation in the estimates. In addition, as the magnitude of the difference increased, the standard deviation did as well. This effect was slightly more pronounced for smaller samples. Finally, the difference in standard deviations by sample size was slightly greater in the 2-factor case. Again, it is important to note that while this interaction was found to be statistically

FINCH & FRENCH

Table 1. Mean of Factor Loadings across Replications by Number of Factors, Sample Size and Population Target Loading Value

<u>Loading for Group 2</u>	<u>Contaminated</u>	
<u>1 Factor</u>	<u>No</u>	<u>Yes</u>
0.75	0.742	0.741
0.65	0.647	0.645
0.55	0.551	0.550
0.45	0.453	0.454
0.35	0.353	0.356
0.25	0.255	0.271
	<u>2 Factor</u>	
0.75	0.735	0.731
0.65	0.641	0.580
0.55	0.548	0.578
0.45	0.451	0.490
0.35	0.353	0.390
0.25	0.255	0.339

INVARIANT REFERENTS IN FACTOR INVARIANCE STUDIES

Table 2. Standard deviation of Factor Loadings across Replications by Sample Size and Population Target Loading Value

<u>Loading for Group 2</u>	<u>Sample size</u>		
<u>Factor 1</u>	100	500	1000
0.75	0.049	0.022	0.016
0.65	0.065	0.029	0.021
0.55	0.076	0.034	0.024
0.45	0.089	0.038	0.028
0.35	0.097	0.043	0.029
0.25	0.105	0.047	0.033
<u>Factor 2</u>			
0.75	0.062	0.028	0.026
0.65	0.084	0.037	0.023
0.55	0.092	0.041	0.029
0.45	0.103	0.046	0.033
0.35	0.114	0.052	0.036
0.25	0.147	0.052	0.037

significant, it accounted for less than 5% of the variance, whereas sample size accounted for 75% of the variance in the standard deviation values.

Logistic regression was used to identify significant main effects and interactions that were associated with loadings being within a 10% range of 0.75. In this case, the outcome for each replication was coded as either 1 (loading was within 10% of 0.75) or 0 (loading was not within this range). It is important to keep in mind that when the target loading for the second group was simulated to be 0.75, the proportion

of cases not within this 10% range could be considered a false positive (incorrect identification of difference when no difference existed). On the other hand, when the target loading was simulated to be some other value (e.g., 0.25), the proportion of cases outside of the 10% range represent a true positive (i.e., correct identification of differences between the groups' loadings). Two separate logistic regression models were used: 1) Examining only those cases where the target loading was set at 0.75 (Model 1) and 2) Examining all other target loading conditions (Model 2).

Table 3. Proportion of Factor Loadings within 10% of 0.75 (0.675, 0.825) when loading was 0.75 across Replications by Number of Factors and Sample Size

<u>Sample size</u>	<u>Factors</u>	
	<u>1</u>	<u>2</u>
100	0.875	0.771
500	0.998	0.979
1000	1.000	0.996

In the case of Model 1, the only statistically significant effects were the interaction of the number of factors by sample size and the main effect of sample size. Table 3 contains the proportion of cases within 10% of 0.75 when the target loading was in fact 0.75 in the population, by sample size and the number of factors. The results show that for the 1-factor case, over 87% of the factor loading estimates were within the correct range, regardless of sample size. In contrast, for the 2-factor case, the smallest sample size was associated with a somewhat lower proportion of cases within the 10% range of the 0.75 value compared to the 1-factor case, otherwise the results across factor models were nearly identical.

In the case of Model 2, the logistic regression analysis identified the 3-way interactions of number of factors by contamination by magnitude of difference and number of factors by sample size by contamination as significantly associated with the proportion of cases within 10% of the 0.75 loading value. In this context being outside of this range would be correct, given that the population values for the simulated loadings were less than 0.75.

Table 4 contains the proportion of replications within 10% of 0.75 by the magnitude of the difference, the number of factors and contamination. For the 1-factor case, regardless of contamination, the larger the target loading was in the population (i.e., less of a

difference), the greater the proportion of replications for which the estimated value was within 10% of 0.75. The largest proportion of values within this range occurred for the population loading of 0.65 across the number of factors and level of contamination.

Indeed, the results for the 1-factor cases (both contaminated and not) and the 2-factor uncontaminated case were all very comparable. However, in the contaminated 2-factor condition with a population loading of 0.65, the proportion of replications within 10% of 0.75 (i.e., 0.055) was much lower than in the other 3 conditions ($M = 0.192$). For the other loading values, the results for the contaminated 2-factor case were just slightly higher than for the others simulated.

Table 5 displays the proportion of replications within 10% of 0.75 by the number of factors, sample size and contamination condition. Overall, the proportions decline in conjunction with increasing sample sizes. For 1-factor these proportions were very comparable regardless of whether another loading was different from 0.75 (contaminated condition). While the pattern of changes in the proportion declined with increasing sample sizes in the 2-factor case, there was a slightly greater difference in the proportions between the contaminated and uncontaminated conditions, leading to the significant interaction described above.

INVARIANT REFERENTS IN FACTOR INVARIANCE STUDIES

Table 4. Proportions of Factor Loadings within 10% of 0.75 (0.675, 0.825) across Replications by Number of Factors, Sample Size and Population Target Loading Value

<u>Loading for Group 2</u>	<u>Contaminated</u>	
<u>1 Factor</u>	<u>No</u>	<u>Yes</u>
0.65	0.194	0.185
0.55	0.013	0.015
0.45	0	0.001
0.35	0	0
0.25	0	0
	<u>2 Factor</u>	
0.65	0.197	0.055
0.55	0.031	0.047
0.45	0.002	0.011
0.35	0	0.001
0.25	0	0.001

Discussion

The results reported in this study suggest that in many instances EFA may be a useful tool for identifying potential indicator variables with invariant loadings across groups for use in a subsequent MCFA. Across most of the conditions simulated here, the factor loading estimates provided by principal axis EFA with PROMAX rotation were very close to the population values. Indeed, the only instances where simulated values were not approximated occurred with 2 factors in conjunction with the contamination of one other factor loading. These generally positive results would seem to suggest that practitioners using EFA can be confident that the sample estimate of loadings are unbiased

in conditions such as those simulated here.

The amount of variation in sample estimates was largely a function of sample size. While loading estimates had greater variability across replications for smaller loading values in the population, there were more marked differences in variation across the three sample size conditions. In addition, this difference in variability was largely mitigated by sample size, so that for 100 participants the standard deviation increased by as much as 0.8 (2 factors) as the population loading value declined, while for 500 or 1000 participants, this increase was always less than 0.03.

In short, with sufficient sample size, a researcher using EFA to identify invariant factor loadings can be almost as confident in their result whether the loading is at or near 0.75 or

closer to 0.25. Note that a certain sample size (e.g., $N = 500$) requirement is not being recommended, but rather that the sample size be sufficient given the data conditions (e.g., number of variables, communalities), as sample sizes requirements for accurate estimate can depend on data conditions (e.g., MacCallum et al., 1999).

In terms of the identification of false positives (obtaining a sample estimate that was more than 10% different than 0.75 when it was not in the population), EFA appears to have performed better for larger sample sizes, particularly in the 2-factor case. Indeed, with a sample size of 500 or greater with the population loading set at 0.75, the likelihood of making a false positive was essentially 0.02 or less. That is, the sample estimate was within the expected range 98 % of the time or higher. In contrast, the rate of sample estimates being within 10% of 0.75 when they should not have been (i.e. the population loading was not 0.75) declined as (a) the value of the population loading declined increasingly from 0.75, and (b) as sample size increased. Given that the results have shown generally little or no bias in loading estimates, this outcome is not a surprise. Indeed, if the target loading was 0.55 or lower in the population, the sample estimates were within 10% of 0.75 in fewer than 5% of cases, regardless of contamination condition. Thus, supporting that EFA could quite accurately detect a non-invariant loading.

The identification of an invariant referent loading is a crucial step in MCFA. As described above, a failure to accurately select an invariant parameter value in the model identification step could lead to severely biased parameter estimates (e.g., factor loadings) which in turn could compromise other analyses, such as the comparison of latent means. The primary method suggested in the literature for identifying invariant indicators, or sets of indicators, is the factor-ratio test and SP procedure (Rensvold & Cheung, 2001), which involves a fairly complex and time consuming multi-step analysis. While this approach appears to work reasonably well for fairly limited models it can become intractably time consuming with increasing model complexity (French & Finch, 2006a).

EFA is one approach that has been advocated for use in practice and involves comparison of factor loading estimates between two groups (Vandenberg, 2001; Zumbo, 2003). While this method does not have the advantage of significance testing that is offered by the factor-ratio test, it is much simpler to conduct. The results of this study seem to indicate that in conditions such as those simulated here, EFA generally provides unbiased estimates of factor loadings, which can in turn be compared to a target value (such as those of another group in the MCFA context).

Therefore, practitioners interested in identifying loadings that are invariant across groups may find that this simple approach works quite well in conditions similar to those simulated here. It does seem that greater confidence can be placed in EFA factor loading estimates that are based on larger sample sizes, particularly with respect to false negative outcomes when the population loadings for the groups differ by 0.10 or more. Under such conditions, the EFA approach appears to have low false negative rates (below 0.05). In addition, the lack of bias and the lower standard deviations at sample sizes of 500 or more appear to contribute to the ability of EFA to accurately estimate loadings within 10% of the target value. Study limitations and directions for future research

As with many simulation studies, the generalizability of the results is limited due to the conditions under study, which should be remembered when interpreting these results. First, the factor models simulated were not as complex as seen in some invariance studies. While the EFA worked well for these somewhat simpler models, it will be necessary to assess its performance with more complex problems (e.g., greater number of factors, different variables, various levels of communalities). Second, a related area that deserves attention is the combination of loadings for the observed variables. In this study, all of the loadings were set at 0.75 (unless contaminated). Given that this is one of the first (if not the first) Monte Carlo investigations to examine the use of EFA to accurately identify invariant referent variables, clarity of result interpretation was considered

INVARIANT REFERENTS IN FACTOR INVARIANCE STUDIES

Table 5. Proportion of Factor Loadings within 10% of 0.75 (0.675, 0.825) across Replications by Number of Factors, Sample Size, and Contamination

<u>Sample size</u>	<u>Contaminated</u>	
<u>1 Factor</u>	<u>No</u>	<u>Yes</u>
100	0.079	0.080
500	0.030	0.027
1000	0.015	0.013
	<u>2 Factor</u>	
100	0.090	0.066
500	0.033	0.002
1000	0.014	0.000

paramount, and thus non-target loadings were not varied. However, it is unclear whether the results obtained here would hold for a more complex combination of loading values and factor models, as well as data conditions (e.g., ordinal variables). Thus, although EFA appears to be a promising screening or purification tool prior to MCFA analysis, future research should extend the current work by investigating a broader combination of conditions before the tool is applied unequivocally.

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: Sage.

Bielby, W. T. (1986). Arbitrary metrics in multiple-indicator models of latent variables. *Sociological Methods & Research, 15*, 3-23.

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

Brown, T.A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: The Guilford Press.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures. *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.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Hillsdale, N.J.: Erlbaum.

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. (2006a, June). *Locating the Invariant Referent in Multi-Group Confirmatory Factor Analysis*. Paper presented at the International Psychometric Society meeting in Montreal, Canada.

French, B. F., & Finch, W. (2006b). 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.

- 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: Erlbaum.
- Jöreskog, K.G. & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago: ScientificSoftware.
- 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.
- 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.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- MacCallum, R.C., Widaman, K.F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*, 84-99.
- Maller, S. J., & French, B. F. (2004). Factor invariance of the UNIT across deaf and standardization samples. *Educational and Psychological Measurement, 64*, 647-660.
- 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.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*, 5-11.
- 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.
- 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.
- Muthén, B., Kaplan, D. & Hollis, M (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*, 431-462.
- Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory, *Journal of Applied Psychology, 87*, 517-529.
- 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 Publishing.
- SAS Institute (2004) *SAS version 9.1.3*. Cary, NC: SAS Institute inc.
- 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.
- 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.
- Wilson, K. L. (1981). On population comparisons using factor indexes or latent variables. *Social Science Research, 10*, 301-313.
- Zumbo, B. D. (2003). Does Item-Level DIF Manifest Itself in Scale-Level Analyses?: Implications for Translating Language Tests. *Language Testing, 20*, 136-147.