Johnson, E. C., & Meade, A. W. (2007, April). *The Role of Referent Indicators in Tests of Measurement Invariance.* Paper presented at the 22nd Annual Meeting of the Society for Industrial and Organizational Psychology, New York.

The Role of Referent Indicators in Tests of Measurement Invariance

Emily C. Johnson and Adam W. Meade
North Carolina State University

Confirmatory factor analytic tests of measurement invariance require a referent indicator (RI) for model identification. This RI is assumed to be perfectly invariant across groups. Using simulated data, results indicate that inappropriate RI selection may be mildly problematic for scale-level invariance tests and highly problematic for item-level tests.

In the social sciences, a considerable amount of research seeks to make comparisons between groups of people. These groups may be defined by nationality, culture, gender, race, or the same people at different points in time. Regardless of the substantive research questions, a key assumption is that the observed variables on which these comparisons are based, typically sets of items or scales, assess the groups in the same way. If items measure the same latent variables and are related to the latent variables in the same way, any observed differences across groups can be interpreted as true differences in the unobservable latent variables. The validity of this assumption is assessed by tests of measurement invariance; more precisely, measurement invariance (MI) can be considered the degree to which measurements conducted under different conditions yield equivalent measures of the same attributes (Horn & McArdle, 1992). If measurement invariance cannot be supported, differences between groups cannot be meaningfully interpreted. It is, therefore, critical that tests of measurement invariance produce valid, unambiguous results.

The current study deals with an issue in confirmatory factor analytic (CFA) tests of measurement invariance, termed the "standardization problem" (Cheung & Rensvold, 1999; Rensvold & Cheung, 1998, 2001). The problem relates to the standardization procedures required with any use of CFA; however, the assumptions inherent in the procedure, when untenable, potentially compromise the validity of conclusions about MI. Thus, the goal of the current study is to determine the conditions under which researchers should be the most wary of tests of MI and those under which conclusions drawn from scale and item-level tests can safely be considered valid. Specifically, we focus on the role of the referent indicator (RI), which is the item chosen to provide a metric for the latent variable.

The CFA Model and Measurement Invariance Tests

The basic CFA model represents a linear relationship between a set of items and one or more latent constructs. Before conducting any CFA analyses, the researcher must ensure the model is identified (see Bollen, 1989 for a review) and that some scaling constraint is present in order to provide a metric for the latent variable. This can be done by assigning a value (typically 1.0) to either the variance of the latent factor or one of the item factor loadings, most commonly by the latter procedure (Bollen, 1989). When this option is chosen, the effect is such that scores on the latent variable are expressed in the scale of the selected item, the referent indicator (RI). Note that whether model identification is achieved via a referent indicator or standardized factor variances, model fit will not be affected.

Typically, tests of MI compare the fit of a number of nested models in a procedure referred to as a likelihood ratio test (LRT). First, data from both groups is analyzed simultaneously, with the pattern of fixed and freed factor loadings held constant. This model serves as a baseline model to which more constrained models are then compared. Next, a test of "metric invariance" (Horn & McArdle, 1992) is conducted in which the established baseline model is compared to a model in which all factor loadings are constrained to be equal across group (i.e., $\Lambda^{(1)} = \Lambda^{(2)}$). If, metric invariance is not supported, the researcher has few choices. One option is that the researcher can stop all additional analyses and declare that the measure is not invariant. In this case, observed and latent scores are not directly comparable and further analysis of the data should not take place. Obviously, discounting all of the efforts of data collection and analysis is not a decision taken lightly in most cases. Thus, most researchers will likely attempt to determine the source of non-invariance. Byrne, Shavelson, and Muthen (1989) and more recently Stark, Chernyshenko, and Drasgow (in press), recommend a

procedure for testing individual items for cross-group equivalence one at a time. If the source of a lack of invariance can be isolated to a small number of items, the researcher can allow those items' factor loadings to differ across groups and continue MI tests, eventually comparing latent mean scores. Note that in the current study, the focus will be on tests of metric invariance at the scale- and item-level, though other tests of other model parameters are certainly possible (see Ployhart & Oswald, 2004; Vandenberg & Lance, 2000 for excellent reviews).

Model Identification and Scaling in Multi-Group CFA

In multi-group comparisons, scaling via RIs is used almost exclusively though both scaling procedures will implicitly assume some degree of invariance. To elaborate, consider the case in which a latent variable is assigned the scale of an RI. By setting the factor loading of the RI to 1.0, the latent variable is standardized to a sample-specific quantity, the difference between the RI's observed and unique variances (Bielby, 1986). Presuming simple structure exists, this quantity will equal the square-root of the communality of the item. The magnitudes of all other relationships between observed variables and the latent variable will be expressed in the scale of the RI. As an illustration, consider a one-factor, three-indicator model. A constraint is placed such that $\lambda_{11} = 1.0$; however, the true situation is one in which 1.0 is *j* times the actual value of λ_{11} (e.g., .7). The effect of the scaling constant can be seen upon consideration of the relationship between the observed covariance matrix and the estimated parameters. For $\text{cov}(x_2, x_1) = \lambda_{21}\lambda_{11}\phi_{11}$, the quantity $\lambda_{21}\phi_{11}$ must adjust by 1/*j* to reproduce the observed $\text{cov}(x_2, x_1)$. Likewise, for $\text{cov}(x_3, x_1) = \lambda_{31}\lambda_{11}\phi_{11}$, the quantity $\lambda_{31}\phi_{11}$ must adjust by the same factor, 1/*j*. This leaves the remaining expression, $\text{cov}(x_3, x_2) = \lambda_{31}\lambda_{21}\phi_{11}$, in which λ_{31} and λ_{21} would adjust by *j* and ϕ_{11} by $1/j^2$.

The standardization procedure, while not problematic in single group contexts, has potential to greatly obscure the true state of invariance in multi-group comparisons. Note that for the comparison to be valid, the selected parameter need not actually have a true value of 1.0 (as is never the case). So long as the parameters are truly invariant, the influence of the RI occurs proportionally in both groups and is not problematic (Bielby, 1986). On the other hand, if the value of the RI is different across groups, parameter estimations will be adjusted differentially across groups.

As a result, tests of MI require researchers to assume exactly what it is that they are investigating - that one item is truly invariant across groups. Further, the assumption is not only un-testable but one that, when not tenable, could greatly obscure the true state of invariance. In order to partially address this issue, Rensvold and Chueng (2001) devised a method to facilitate the choice of an RI. However, their method is somewhat labor intensive and is seldom used in practice. As a result, most researchers have taken to simply acknowledging that the practice of standardization via RIs is problematic, or ignoring it altogether. The acknowledgement speaks little to the questions of how problematic the practice might be and under what circumstances researchers and evaluators of research should be most wary.

This study seeks to explicate the conditions under which researchers can be most confident about the inferences drawn from tests of MI. To examine these issues, the current study uses simulated data to manipulate the magnitude of differential functioning (DF; a lack of invariance) on the RI. Additionally, in some conditions, a manipulation is included in which either two or no non-RI items are specified to function differentially. Of primary interest is the effects of DF of the RI on both scale and item-level tests of MI. While we know that when the RI is invariant (i.e., the same across groups), DF of other items will result in the accurate detection of a lack of MI (Meade & Lautenschalger, 2004), it is less clear how DF of the RI will affect accurate detection of a lack of MI. We propose:

Hypothesis 1: Larger sample sizes will be associated with more frequent detection of a lack of MI.

Hypothesis 2: When the RI is invariant, the LRT will accurately detect a lack of invariance at both the scale and item-levels for those non-RI items that are DF.

The effects of DF of the RI are less intuitive. The choice of RI determines the scaling of the latent variable, which is then reflected in the estimated factor loadings for all items. Thus, to the extent to which there is DF in the RI, this difference should be reflected in other scale items. Moreover, the larger the magnitude of DF across groups for the RI, the more likely it should be to detect invariance at the scale level. In order to investigate the effects of DF of the RI, we simulated some condition in which the RI was DF to varying degrees but the other scale items were invariant. In a second set of conditions, both the RI and two other scale items were functioned

differently across groups. As the use of a RI causes an adjustment in scaling of all item factor loadings, but in somewhat unpredictable ways, we propose:

Hypothesis 3: The effects of selecting a DF RI will have a minimal impact on the accuracy of MI conclusions at the scale-level but a large impact at the item-level.

Hypothesis 4: The effects of selecting a DF RI on the accuracy of MI conclusions should be more severe when the magnitude of this DF is large.

Method

In this study, data with known properties was simulated to represent various conditions of non-invariance. Data properties were simulated to represent "Group 1" data, and then some of these properties were changed in order to create several different conditions of Group 2 data.

Sample Size

Data were simulated to represent sample sizes of 150, 250, and 350. Given the nature of the simulated data in this study and the recommendations of previous studies (MacCallum, Widaman, Zhang, & Hong, 1999; Meade & Lautenschlager, 2004) these sample sizes were selected to represent a condition of minimally adequate power and conditions of larger sample, as might be expected in practice.

Nature of the Model

For all conditions, the model simulated (Figure 1) was one with two latent variables specified to correlate at .3 (cf. Meade & Kroustalis, 2006). For each latent variable, four indicator variables were simulated. While in practice the number of indicators varies considerably across studies, a review of published studies using CFA to test MI on non-simulated data yielded a median and mode of four indicators per latent variable.[1] The population factor variances for both factors were set to 1.0 (cf. Meade & Lautenschalger, 2004). Factor loading values used for Group 1 data (see Table 2) were determined based on the estimated loadings from a large sample (N=686) of undergraduate respondents on two scales of the Occupational Personality Questionnaire (OPQ-32; SHL, 2000). One hundred sample replications were simulated for each study condition.

Factor Loading Differences

Five conditions were simulated to represent varying degrees of DF for the RI between Group 1 and Group 2: a control condition of true RI invariance (no differences in RI values beyond that of sampling error) and differences of 05, .1, .2, and .4. For each condition of RI DF, two conditions of non-RI DF were specified: one in which the non-RI items were truly invariant and one in which two non-RI items (Items 3 and 7) were specified to have a difference of .25 in factor loadings. Population factor loadings simulated in Group 2 for the various conditions are presented in Table 3.

Model Parameter Simulation

Initial structural models were simulated for the various conditions outlined in Table 1 using the PRELIS program which accompanies the LISREL 8.51 software package (Jöreskog & Sörbom, 1996). Group 1 data was simulated to represent the 8-indicator, 2-factor model (Table 2) and was analyzed in all conditions while Group 2 data was modified to simulate conditions of a lack of invariance by subtracting the specified amount of DF (see Table 3).

Data Analysis

A model of equivalent factor patterns served as a baseline model to which the subsequent tests of metric invariance were compared. In this model, the correct pattern of factor loadings was specified and model parameters were freely estimated in each group. Nested model chi-square difference tests (i.e., LRTs) were used to evaluate the decrement in fit resulting from imposing factor loading equality constraints. Item-level tests of factor loading invariance were also conducted. In these analyses, the fit of the baseline model was compared to a model in which the factor loading of a single item was constrained to be equal for the two groups for each non-RI item (see Stark et al., in press, for the merits of this type of item-level test).

Outcome Measures

The outcome of interest in this study was the performance of both scale- and item-level tests of invariance. For each condition, the results of tests of scale-level metric invariance are reported as the percentage of the 100 data replications in each condition that indicate a statistically significant lack of invariance. Because the results of the metric invariance tests are expressed as a dichotomous, significant/non-significant dependent variable,

[1] A literature search was conducted to identify studies which used the CFA framework to test real data for measurement invariance. Search terms were the following: factorial invariance, measurement invariance, measurement equivalence, and alpha, beta, and gamma change. Journals included in the search were: Journal of Applied Psychology, Psychological Methods, and Educational and Psychological Measurement.

logistic regression was used to evaluate the effects of the study variables.

At the item level, results were reported in two metrics. First, true positive (TP; number of truly DF items detected as DF by the item-level analyses) and false positive (FP; the number of truly invariant items falsely detected as DF by the item-level analyses) values were computed. TP and FP rates were computed for each of the 100 replications, and then averaged across these replications for each condition. Second, we also report the percentage of the one-hundred replications that were significant for each item.

Results

Metric Invariance

The overall logistic regression model for the scale-level analyses was significant (Wald = 835.15, $p < .0001$). The Cox and Snell R^2 value for the model was .48 and the Nagelkerke R^2 statistic was .65. Thus, a moderately large proportion of variance in the metric invariance test results was accounted for by the study conditions (the variance not accounted for is due to sampling error). Wald significance statistics, standardized parameter estimates, odds-ratios, and their associated confidence intervals for individual study variables can be found in Table 4

Significant main effects were found for RI DF, non-RI DF, and sample size (see Table 4). In addition, the three-way interaction of study variables as well as all two-way interactions, with the exception of the RI DF*non-RI DF interaction, were also statistically significant. These results are illustrated pictorially in Figure 2. It is clear that, for all DF conditions, larger sample sizes resulted in a more frequent detection of DF. In addition, for conditions of no DF on non-referent indicator variables, the likelihood of rejecting the metric invariance hypothesis increased as RI DF increased. However, when DF of .25 wais simulated on two additional items, the probability of detecting DF remained more or less constant across levels of RI-DF, displaying a somewhat u-shaped function in the smaller sample size conditions. Taken together, these results provide support for the first two study hypotheses.

Item-Level Tests of Invariance

At the item level, results indicate that as the amount of DF simulated on the RI increases, the likelihood of detecting TPs decreases while the likelihood of FP increases. This trend is presented pictorially in Figures 3 and 4. Specific results of item-level tests giving rise to these TP and FP rates for the N=250 condition can be seen in Table 5. In the top half of Table 5, it is evident that the DF due to the RI is transferred to other non-DIF items resulting in their erroneous detection as DF items a sizable percentage of the time. Results in the bottom half of Table 5 are extremely interesting in that they illustrate how DF on the RI results in the significant (and erroneous) detection of *non-DF* items, while truly DF items (Items 3 and 7) are erroneously *not* detected as DF. These results, taken in tandem with those of the scale-level analyses, support Hypothesis 3. As expected, as the magnitude of DF of the RI increases, the effects of this DF are more apparent (i.e., Hypothesis 4 is supported).

Discussion

The results of this study illustrate the effects of referent indicator (RI) selection on the validity of conclusions drawn from measurement invariance tests for a number of conditions. First, as predicted (Hypothesis 1), larger sample sizes will be associated with more frequent detection of a lack of MI (see Figure 2). Moreover, when a truly invariant RI is selected, valid conclusions can be drawn with the presence of DF in other items (Hypothesis 2). Importantly however, when a DF RI is selected, our results suggest that the invariance status of the other items in the model matters a great deal. That is, when all non-RI items are truly invariant, the likelihood of rejecting the hypothesis of metric invariance increases as the magnitude of RI DF increases. Put differently, differences between groups on the RI are transferred to other scale items via the constraints on the RI to be equal to 1.0 in both groups. These transferred differences are then accurately detected at the scale level. However, item-level tests for these data illustrate how misleading a poor choice of RI can be. Specifically non-DF items are erroneously detected as DF while DF items are erroneously *not* detected. These findings are extremely troubling for cases in which the researcher attempts to isolate a lack of invariance found at the scale level to individual items so that partial invariance can be allowed, as is often the case.

We also found that in conditions with simulated DF on two non-RI items, the likelihood of detecting DF at the scale-level actually decreased slightly with increasing magnitude of RI DF in the conditions where simulated RI DF was small (.05 and .1). The reason for this effect is still a point for investigation; it may be that the effect represents some sort of 'masking' or is the result of unreliability introduced into the model through the lowering of factor loadings to simulate DF. While more work needs to be done, these results seem to suggest a situation somewhat more complicated than that

predicted by Hypothesis 4 (the effect of RI DF will be greatest when RI DF is largest). While Hypothesis 4 was supported at the item-level when all non-RI items were invariant, when non-RI DF was present, small amounts of RI DF obscured the true scale-level DF, leading to fewer accurate conclusions than when RI DF was large.

At the item-level, the results of this study suggest that the accuracy of MI conclusions decrease with increasing RI DF whether or not non-RI items exhibit DF. That is, item-level tests become less effective with larger magnitudes of RI DF, whether or not DF is present on other, non-RI items. Our results also indicate that the rate of detection item level DF, both accurate and inaccurate, increases with increasing sample size.

Implications & Recommendations

When conducting measurement invariance tests using CFA techniques, care must be taken to choose RIs that are truly invariant. While we found the DF of the RI can have some effect on power at the scale level, we found that the invariance of the RI is imperative for item-level tests. As stated previously, when scale-level MI tests indicate DF of factor loadings, researchers may either stop all further analyses or attempt to identify the source of the invariance. Given the extensive effort required in collecting and analyzing data, cessation of data analysis is undesirable. Thus, it is important to be able to accurately detect the particular item(s) responsible for a lack of scale-level invariance (Byrne et al., 1989). Our results indicate that when RI items display DF, the detection of true source of DF will be unlikely.

We recommend that researchers carefully choose potential RIs. Ideally, theory will be a guide in RI selection. In practice, however, it seems somewhat unlikely that researchers could anticipate which items may or may not be invariant across groups. For this reason, we believe that procedures such as that illustrated by Rensvold and Cheung (2001) should be given much further consideration. While we did not evaluate the efficacy of their method in this study, any procedure that would increase the likelihood of identifying the true source of invariance may prove promising for dealing with the RI issue.

Limitations and Future Directions

As in any study, there are a number of limitations. Due to the very nature of simulation studies, the generalizability of our findings is restricted. While we did investigate the effects of including other DF items in our model, we did not vary the magnitude of this DF. We simulated DF by subtracting the specified amount from the Group 1 population loadings; however, it is possible that the lower reliability (see Fornell & Larcker, 1981) introduced by this practice could have had an effect on the results. While we attempted to bolster the external validity of the study by using factor loadings obtained from real data, these numbers represent only one of an infinite number of possible models. It is important to note that our goal was not to determine the precise power of the LRT under different conditions but rather to illustrate the potential effects of DF for the RI. We do not expect that the general pattern of results that we have found would differ in other studies, but certainly the percentages of samples in which DF was detected would vary for different model conditions.

It is our hope that this study will serve to draw increased attention to the largely ignored 'standardization problem'. Despite its limited scope, the study suggests that the tenability of the assumptions inherent in standardization procedures does affect the validity of conclusions drawn from measurement invariance tests.

References

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

Bollen, K. A. (1989). *Structural equations with latent variables*. Oxford, England: John Wiley and Sons.

Byrne, B. M., Shavelson, R. J., & Muthen, 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.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*, 39-50.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3-4), 117-144.

Jöreskog, K., & Sörbom, D. (1996). *LISREL 8: Users reference guide.* Chicago: Scientific Software International.

Meade, A. W., & Kroustalis, C. M. (2006) Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods, 9*, 369-403.

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

Ployhart, R. E., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods, 7*, 27-65.

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 (Vol. 1): Equivalence in measurement* (pp. 21-50). Greenwich, CT: Information Age.

SHL (2000). Notes accompanying the OPQ 32 survey. Boulder, Colorado.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (in press). Detecting differential item functioning with CFA and IRT: Toward a unified strategy. *Journal of Applied Psychology*.

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.

Author Contact Info:

Emily C. Johnson
Department of Psychology
North Carolina State University
Campus Box 7650
Raleigh, NC 27695-7650
Phone: 919-515-2251
Fax: 919-515-1716
E-mail: ecjohnso@ncsu.edu

Adam W. Meade
Department of Psychology
North Carolina State University
Campus Box 7650
Raleigh, NC 27695-7650
Phone: 919-513-4857
Fax: 919-515-1716
E-mail: awmeade@ncsu.edu

Table 1

Summary of Manipulated Conditions

Condition	Manipulation
Sample size	150, 250, 350
Magnitude of RI difference	0, 0.05, 0.1, 0.2, 0.4
Magnitude of non-RI difference	0, .25

Note. Non-RI refers to Items 3 and 7 population factor loadings. For all non-RI items, population factor loadings are equal across groups in all conditions.

Table 2

Population Factor Loadings for Simulated Group 1 Data

Item	***Factor 1***	***Factor 2***
1	**.82**	**-**
2	**.78**	**-**
3	**.76**	**-**
4	**.63**	**-**
5	**-**	**.81**
6	**-**	**.77**
7	**-**	**.73**
8	**-**	**.70**

Table 3

Population Factor Loadings for Simulated Group 2 Data

	Condition 0a		**Condition 1a**		**Condition 2a**		**Condition 3a**		**Condition 4a**	
Item	**Factor 1**	**Factor 2**	**Factor 1**	**Factor 2**	**Factor 1**	**Factor 2**	**Factor 1**	**Factor 2**	**Factor 1**	**Factor 2**
1	**.82**	**-**	**.77**	**-**	**.72**	**-**	**.62**	**-**	**.42**	**-**
2	**.78**	**-**	**.78**	**-**	**.78**	**-**	**.78**	**-**	**.78**	**-**
3	**.76**	**-**	**.76**	**-**	**.76**	**-**	**.76**	**-**	**.76**	**-**
4	**.63**	**-**	**.63**	**-**	**.63**	**-**	**.63**	**-**	**.63**	**-**
5	**-**	**.81**	**-**	**76**	**-**	**.71**	**-**	**.61**	**-**	**.41**
6	**-**	**.77**	**-**	**.77**	**-**	**.77**	**-**	**.77**	**-**	**.77**
7	**-**	**.73**	**-**	**.73**	**-**	**.73**	**-**	**.73**	**-**	**.73**
8	**-**	**.70**	**-**	**.70**	**-**	**.70**	**-**	**.70**	**-**	**.70**
	Condition 0b		**Condition 1b**		**Condition 2b**		**Condition 3b**		**Condition 4b**	
Item	**Factor 1**	**Factor 2**	**Factor 1**	**Factor 2**	**Factor 1**	**Factor 2**	**Factor 1**	**Factor 2**	**Factor 1**	**Factor 2**
1	**.82**	**-**	**.77**	**-**	**.72**	**-**	**.62**	**-**	**.42**	**-**
2	**.78**	**-**	**.78**	**-**	**.78**	**-**	**.78**	**-**	**.78**	**-**
3	**.51**	**-**	**.51**	**-**	**.51**	**-**	**.51**	**-**	**.51**	**-**
4	**.63**	**-**	**.63**	**-**	**.63**	**-**	**.63**	**-**	**.63**	**-**
5	**-**	**.81**	**-**	**.76**	**-**	**.71**	**-**	**.61**	**-**	**.41**
6	**-**	**.77**	**-**	**.77**	**-**	**.77**	**-**	**.77**	**-**	**.77**
7	**-**	**.48**	**-**	**.48**	**-**	**.48**	**-**	**.48**	**-**	**.48**
8	**-**	**.70**	**-**	**.70**	**-**	**.70**	**-**	**.70**	**-**	**.70**

Note. Underlined values indicate DF item.

Table 4

Results of Logistic Regression of Metric Invariance Test on Study Variables

Parameter	β	SE	Wald Chi Square	Wald 95% Confidence Interval	
Intercept	-2.42	0.30	64.34**	-3.017	-1.832
RIdf	5.40	1.81	8.93**	1.859	8.946
N	0.01	0.00	20.81**	0.003	0.008
nonRIdf	-0.81	0.30	7.24**	-1.406	-0.221
RIdf*N	0.03	0.01	12.03**	0.012	0.045
RIdf*nonRIdf	2.12	1.81	1.38	-1.419	5.668
N*nonRIdf	-0.01	0.00	29.03**	-0.009	-0.004
RIdf*N*nonRIdf	0.03	0.01	8.91**	0.009	0.041

Note. * $p<.05$, ** $p<.01$. $df = 1$ for all analyses. For nonRIdf, df=.25 was the reference category.

Table 5

Results of Item-Level Tests of Metric Invariance for N=250

	RIdif	i2	i3	i4	i6	i7	i8
nonRI DF = 0							
	0	6	7	6	7	6	8
	.05	9	9	6	8	3	9
	.10	17	16	10	13	9	13
	.20	49	53	38	47	44	45
	.40	97	95	90	93	94	93
nonRI DF = .25							
	0	8	<u>79</u>	6	7	<u>77</u>	9
	.05	7	<u>60</u>	4	7	<u>61</u>	8
	.10	14	<u>46</u>	10	14	<u>42</u>	13
	.20	46	<u>15</u>	35	36	<u>8</u>	42
	.40	94	<u>20</u>	87	93	<u>14</u>	93

Note. Results are expressed as percentage of 100 replications found significant. Underlined values are true positives, all others are false positives.

Figure 1. Measurement Model for All Study Conditions

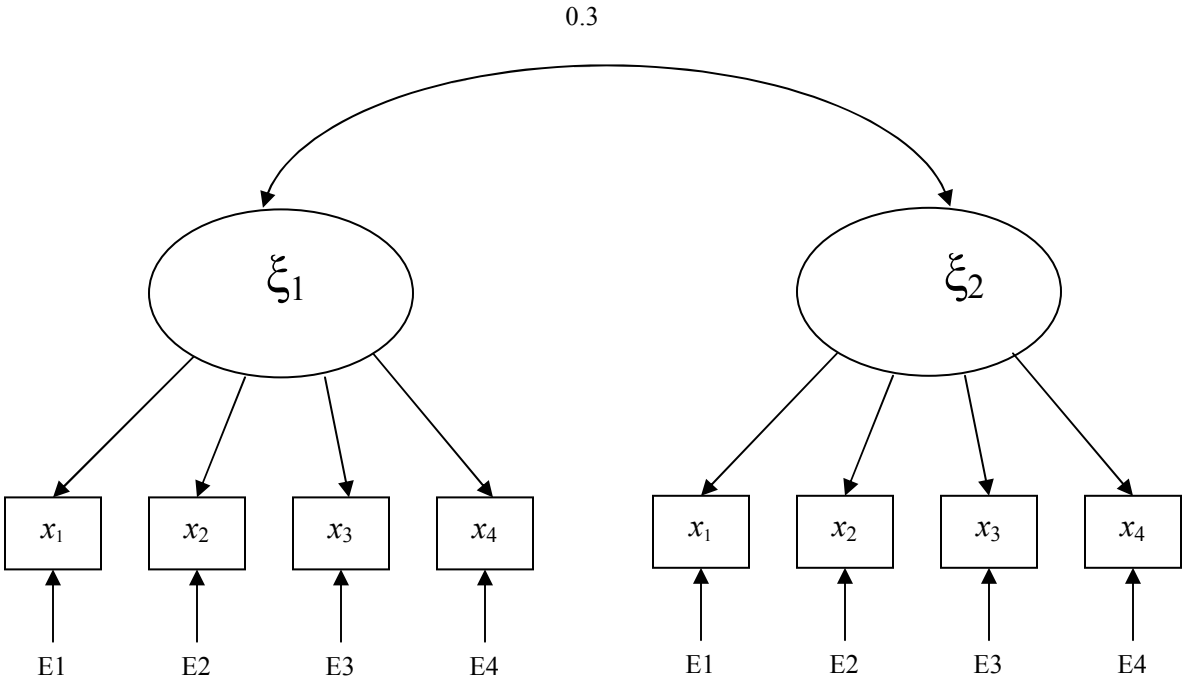

Figure 2. Effect of Study Variables on Scale-Level Tests of Metric Invariance

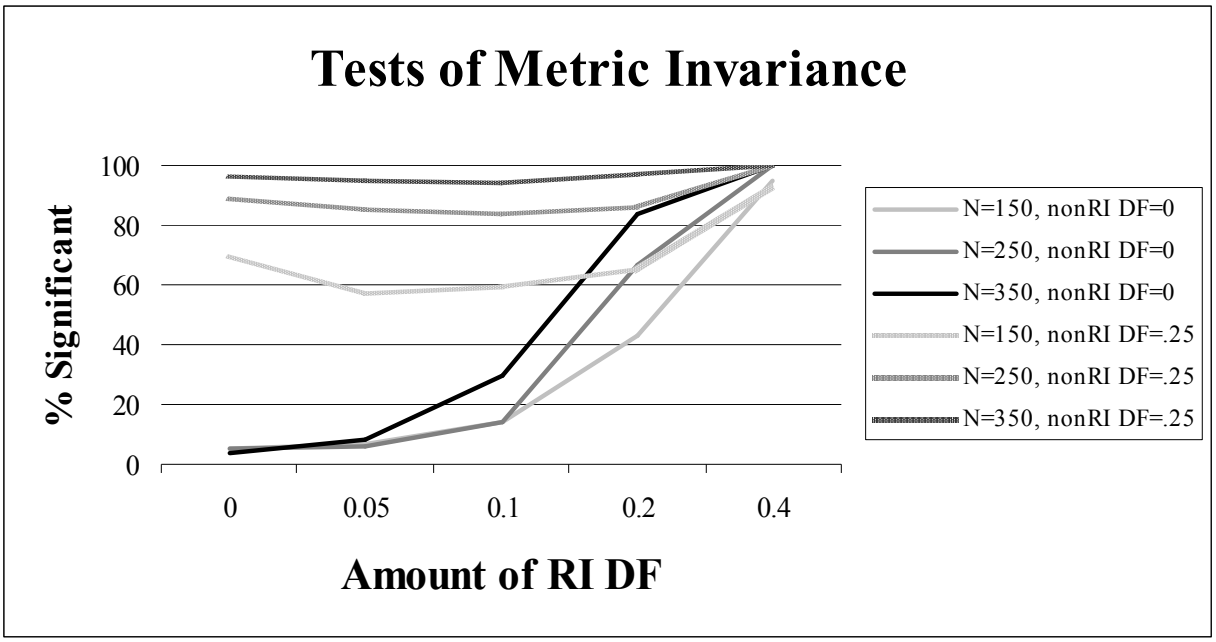

Figure 3. Effects of DF on True Positive Rates of Item-Level Invariance Tests

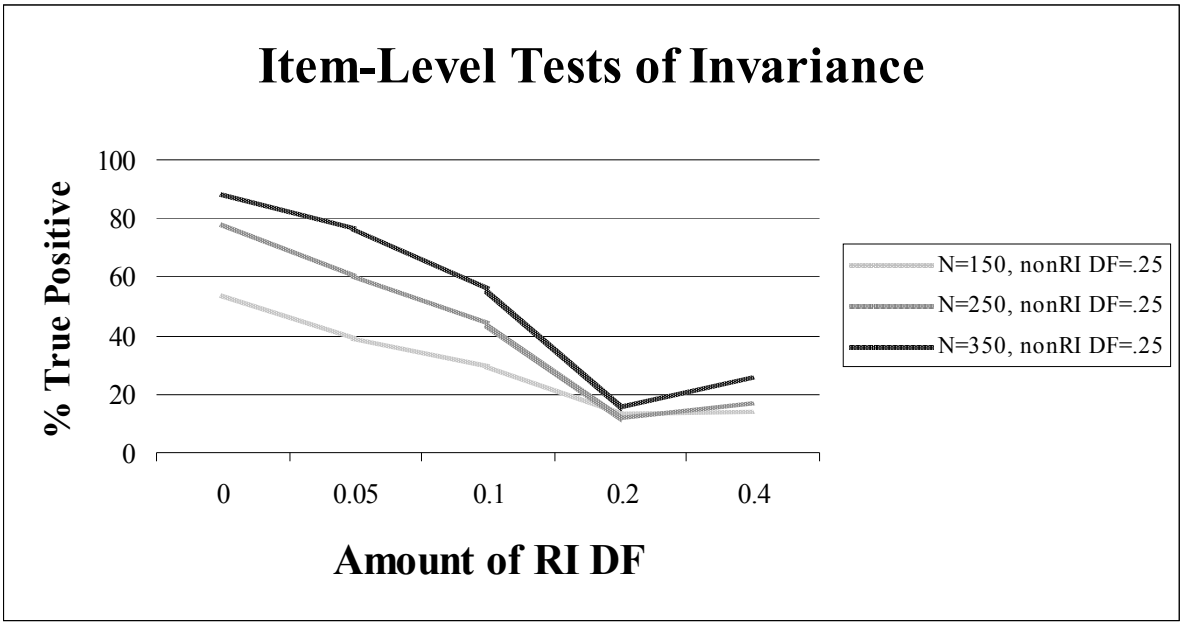

Figure 4. Effects of DF on False Positive Rates of Item-Level Invariance Tests