ABSTRACT

JOHNSON, EMILY. Referent Indicators in Tests of Metric Invariance. (Under the direction of Adam Meade.)

Organizations frequently administer surveys and psychological measures to multiple groups (e.g., cultural and demographic groups). However, before making direct cross-group comparisons, researchers need to ensure that the psychometric properties of these measures do not differ by groups. In order to test this hypothesis of measurement invariance, many researchers employ confirmatory factor analytic tests of measurement invariance. These tests 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. These findings underscore the importance of careful RI selection.

REFERENT INDICATORS IN TESTS OF METRIC INVARIANCE

by
EMILY C JOHNSON

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

PSYCHOLOGY

Raleigh, North Carolina

2006

APPROVED BY:

Mark A. Wil — Bart Craig

Mark A. Wilson S. Bartholomew Craig

Adam W. Meade
Chair of Advisory Committee

BIOGRAPHY

Emily Catherine Johnson was born on July 31, 1982 in Lexington, Kentucky where she lived with her parents, Gregory and Marianne Johnson until she completed high school in the spring of 2000.

Between August of 2000 and May of 2004 Emily attended Tulane University in New Orleans, Louisiana, where she earned her Bachelor of Science degree in Psychology. In the fall of 2004, Emily left New Orleans to study Industrial and Organizational Psychology at North Carolina State University, where she is currently a graduate student.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

1. INTRODUCTION

In the social sciences, a considerable amount of research seeks to make comparisons among groups of people. These groups may be defined by variables such as nationality, culture, gender, or race, or they may consist of the same people at different points in time. Regardless of the substantive research questions, a key assumption is that the observable variables on which these comparisons are based, typically sets of items or scales, are related to the same latent variables across groups and, moreover, related to the latent variables in the same way. When this assumption holds, any observed differences across groups can be interpreted as true differences in the unobservable latent variable(s). To assess the validity of this assumption, tests of measurement invariance (MI) may be employed; more precisely, MI is 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.

Two methodological frameworks, one based on item response theory (IRT) and one based on confirmatory factor analysis (CFA) are available for evaluating measurement invariance. While these methods are similar in purpose, the methodologies have evolved and been treated relatively independently of one another. The current study deals with an issue encountered in confirmatory factor analytic 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, while not problematic in single group applications of CFA, the assumptions inherent in the procedure, when violated, 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, I focus on the role of the referent indicator (RI), the item chosen to provide a metric for the latent variable. The remainder of this section will proceed as follows: First, I will provide a brief overview of the general CFA model. Next I will introduce the procedures in which CFA techniques are used to test for measurement invariance. Then, more specifically, I will discuss issues related to model identification and scaling inherent in these procedures and, in particular, the role of the referent indicator. It is argued that referent indicator selection has serious implication for conclusions drawn from tests of measurement invariance; however, the exact nature of these implications has not yet been explored in the literature. Therefore, the following sections detail the methodology and results of a study designed to investigate these issues.

1.1 The CFA Model

Confirmatory factor analysis is a special case of the broader structural equation modeling (SEM) or, alternatively, covariance structure analysis. CFA requires the researcher to make *a priori* predictions regarding the structure of the model to be tested. The analysis provides a quantification of the degree to which this theoretical model 'fits' the observed data. The general model can be represented as:

$$\mathbf{x} = \Lambda\xi + \delta \,, \qquad \textbf{(1)}$$

where **x**, a (q x 1) vector of observed variables, is a function of ξ, a (s x 1) vector of latent variables, and a (q x 1) vector of item uniqueness terms (δ) which represents the specific and random variances associated with each observed variable. Λ is a (q x s) matrix made up of

λ_{ij}s factor loadings, and represents the relationships between the latent variables and the observed variables; that is, for $x_i = \lambda_{ij}\xi_j + \delta_i$, λ_{ij} refers to the number of units change expected in x_i for every unit change in ξ_j.

Before conducting any CFA analyses, the researcher must specify a model. This model represents the researcher's sense as to how latent unobserved variables are reflected by observed variables, or indicators. Model specification allows the researcher to 'fix' and 'free' model parameters. Fixing a parameter constrains it to equal some specified value while freeing allows a parameter to be freely estimated from input data. The model must include both fixed and freed parameters. To specify a two-factor model in which each factor has three associated indicators and each indicator is associated with only one factor, factor loadings would be fixed and freed so that parameters relating indicators to their predicted latent variables may be freely estimated, while parameters relating indicators with the unintended latent variable are fixed to 0. In matrix form these constraints can be thought of as:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{bmatrix} \quad \textbf{(2)}$$

Additionally, a researcher may choose to impose equality constraints. In the above example, it may be believed that indicator variables x_1, x_2, and x_3 measure ξ_1 equally well (i.e., they have equal factor loadings). Constraining λ_{11} through λ_{31} to be equal will allow the parameters to be estimated from input data but requires estimation to produce the same value for all three λs.

The number of available degrees of freedom (df) for the specified model is calculated by multiplying the number of observed variables (ν) by ν+1 and then dividing the quantity by two.

$$df = \frac{\nu(\nu+1)}{2} \qquad \textbf{(3)}$$

In the previous example, the model contains ν = 6 observed variables, so the potential df = [6(6+1)] / 2 = 21.

Once a model is specified, the CFA program can then be used to estimate model parameters that will reproduce the observed covariance matrix **Σ**. To illustrate the relationship between the parameters estimated in CFA and the observed covariance matrix (**Σ**), Bollen (1989) shows that the implied covariance matrix of observed variables, given the model parameters (θ) can be written as:

$$\sum(\theta) = \Lambda_x \Phi \Lambda'_x + \Theta_\delta \qquad \textbf{(4)}$$

Given a one-factor, three-indicator model,

$$\begin{gathered} x_1 = \lambda_{11}\xi_1 + \delta_1 \\ x_2 = \lambda_{21}\xi_1 + \delta_2 \\ x_3 = \lambda_{31}\xi_1 + \delta_3 \qquad \textbf{(5)} \\ \mathrm{E}(\delta_i) = 0 \end{gathered}$$

$$\operatorname{cov}(\xi_1, \delta_i) = 0, \quad \text{for } i = 1, 2, 3$$

$$\operatorname{cov}(\delta_1, \delta_j) = 0, \quad \text{for } i \neq j$$

the associated matrices would be:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \qquad \mathbf{\Lambda_x} = \begin{bmatrix} \lambda_{11} \\ \lambda_{21} \\ \lambda_{31} \end{bmatrix}, \quad \boldsymbol{\xi} = [\xi_1], \qquad \mathbf{\Phi} = [\phi_{11}], \qquad \boldsymbol{\delta} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix}, \tag{6}$$

where the covariances among the δ terms is given as:

$$\mathbf{\Theta}_\delta = \begin{bmatrix} \operatorname{var}(\delta_1) & & \\ 0 & \operatorname{var}(\delta_2) & \\ 0 & 0 & \operatorname{var}(\delta_3) \end{bmatrix} \tag{7}$$

When Σ(θ) is calculated using the above matrices, the result is:

$$\Sigma(\theta) = \begin{array}{lll} \lambda_{11}^2\phi_{11} + \operatorname{var}(\delta_1) & & \\ \lambda_{21}\lambda_{11}\phi_{11} & \lambda_{21}^2\phi_{11} + \operatorname{var}(\delta_2) & \\ \lambda_{31}\lambda_{11}\phi_{11} & \lambda_{31}\lambda_{21}\phi_{11} & \lambda_{31}^2\phi_{11} + \operatorname{var}(\delta_3) \end{array} \tag{8}$$

Given that the covariance matrix of observed variables (Σ) can be written as:

$$\Sigma = \begin{bmatrix} \operatorname{var}(x_1) & & \\ \operatorname{cov}(x_2, x_1) & \operatorname{var}(x_2) & \\ \operatorname{cov}(x_3, x_1) & \operatorname{cov}(x_3, x_2) & \operatorname{var}(x_3) \end{bmatrix} \tag{9}$$

it follows that:

$$\operatorname{var}(x_1) = \lambda_{11}^2\phi_{11} + \theta_{11},$$

$$\operatorname{cov}(x_2, x_1) = \lambda_{21}\lambda_{11}\phi_{11}, \tag{10}$$

and so on.

When constraints are such that only one possible set of estimates can be produced, the model is said to be “identified.” For identification to occur, the number of parameters estimated must be equal to or less than the number of available df. When the number of estimated parameters equals the available df, the actual df of the model is equal to zero and model implied covariance matrix will perfectly match the observed covariance matrix (i.e., the model is “just-identified”). For this reason, in order to test model fit, one less parameter must be estimated than available degrees of freedom. Such a model is said to be “over-identified”.

Additionally, model identification requires that some scaling constraint is placed in order to provide a metric for the latent variable. Because it is rarely the case that the latent variable of interest possesses an inherent scale, it is left up to the researcher to assign a scale. This can be done by assigning a value to either the variance of the latent factor or one of the λ_{ij} s. Most commonly, scaling is accomplished by the latter procedure (Bollen, 1989), setting the factor loading of one item to a value of 1.0. When this option is chosen, the effect is such that scores on the latent variable are expressed in the scale of the selected indicator (called a referent indicator, RI); therefore, for each latent variable, factor loadings are interpretable with respect to one another. One advantage of constraining factor variances to a constant is that, in models that contain more than one latent variable, constraining the factor variances to the same constant produces factor loadings that are expressed in the same scale, regardless of latent variable. However, this method assumes that variances really are equal for all latent variables. Differences in the variances of latent variables will be manifest via different factor loadings when standardized factor variances are used to set the metric of the scale. For this reason, this scaling option is generally not recommended in tests of MI (Cheung & Rensvold, 1999). Note that whether model identification is achieved via a referent indicator or standardized factor variances, model fit will not be affected.

Using the specified model and observed covariance matrix as input, maximum likelihood estimation seeks to match the observed data to the target model. The discrepancy between the target and observed matrices can be evaluated statistically to confirm or reject the null hypothesis that the expected model 'fits' the data. The most obvious fit index that can be used is the chi-square (χ^2) statistical significance test,

$$\chi^2 = (N-1)F_{ML} \tag{11}$$

where F_{ML} is the maximum likelihood fit function defined as,

$$F_{ML} = \log|\mathbf{\Sigma}(\theta)| + \mathrm{tr}(\mathbf{S}\mathbf{\Sigma}^{-1}(\theta)) - \log|\mathbf{S}| - (p+q) \tag{12}$$

The fit function represents the degree of difference between the observed covariance matrix and the model-implied covariance matrix (as computed from Formula 4 using parameters estimated to minimize this difference). This statistic, with df equal to the difference between the model's available df and the number of estimated parameters, represents a test of "badness of fit". Thus, a non-significant χ^2 supports the hypothesized model. While the χ^2 statistic has the advantageous property of being a parametric statistic with known distributional properties, the χ^2 statistic is somewhat problematic in that, as the sample size increases, the discrepancy between the implied and observed covariance matrices needed to reject the null hypothesis decreases (Hu & Bentler, 1999). This is evidenced by the incorporation of sample size in the chi-square formula. Because of this shortcoming, many researchers choose to supplement χ^2 with other fit indices, such as the normed fit index (NFI; Bentler & Bonett, 1980), the comparative fit index (CFI; Bentler, 1990), and the root-mean-square error of approximation (RMSEA; Steiger & Lind, 1980). A review of these and other fit indices is beyond the scope of this paper and can be found in Hu and Bentler (1999).

While the above discussion is framed in terms of single group analyses, the CFA procedure can be utilized to evaluate the invariance of model parameters. In such a case, Equation 1 is modified to read,

$$\mathbf{x}^{(g)} = \Lambda^{(g)}\xi^{(g)} + \delta^{(g)} \qquad \textbf{(13)}$$

where the superscript (*g*) denotes that the parameters are group specific. As before **x** represents a (*q* x 1) vector of observed variables, ξ is a (*s* x 1) vector of latent variables, Λ is a (*q* x s) vector of factor loadings relating the *x*s to ξs, and δ is a (*q* x 1) vector of uniqueness terms.

MI is investigated via a series of increasingly restrictive hypotheses regarding the equality of parameter estimates. To test these hypotheses, models are nested such that the more restrictive model is nested in the less restrictive, baseline model. Most commonly, when groups are compared in CFA, fit is evaluated using the difference between the chi-square values of the models being compared, with df equal to the difference in the df of the two models (Bollen, 1989). This test is sometimes called a likelihood ratio test (LRT; Thissen, Steinberg, & Wainer, 1988) as well.

1.2 Tests of Measurement Invariance

The sequential hypothesis tests utilized in multi-group CFA allow MI to be assessed for multiple parts of a given model. In a comprehensive review of MI literature, Vandenberg and Lance (2000) observe various consistencies in the way tests of invariance are conducted. With regard to test sequence, the authors found considerable agreement in the initial steps.

Omnibus test. The typical sequence of invariance hypotheses follows from the Jöreskog (1971) tradition, in which the first step is to conduct an omnibus test of equality of covariance matrices, or a test of the hypothesis,

$$H_0 = \Sigma^{(1)} = \Sigma^{(2)} = \ldots = \Sigma^{(g)} \qquad \textbf{(14)}$$

where superscripts denote group membership. In this test, no model parameters are estimated. Instead, observed covariance matrices for the groups are directly constrained to be equal across groups. If the equality constraints do not result in a significant decrement of fit, the constraints are considered to reflect the true nature of the data. In this test, failure to find a significant difference is considered evidence of MI. In such a case, invariance is considered to be such that the groups can be treated as one (Byrne, Shavelson, & Muthen, 1989) and, therefore, additional MI tests are unnecessary. However, as Cheung and Rensvold point out, this situation is unlikely to occur unless the groups were, in fact, drawn from the same population (1999, p. 6). In the more probable case, the hypothesis of equal covariance structures is rejected, and a series of increasingly restrictive hypotheses are tested to determine the source of invariance. For this reason and others, some researchers choose to omit tests of equal observed covariance matrices.

Configural invariance. In the case where the hypothesis of equal covariance matrices is rejected, the test is followed by a test of factor loading pattern invariance. Invariance at this level is referred to as “configural invariance” (Horn & McArdle, 1992) and represents a necessary condition for further tests of invariance. In this test, no equality constraints are imposed. Rather, a theoretically derived pattern of fixed and freed factor loadings is imposed on the two groups. The least restrictive test of MI, invariance of model form can be thought of the test of the null hypothesis,

$$H_0 = \Lambda_{form}^{(1)} = \Lambda_{form}^{(2)} = ... = \Lambda_{form}^{(g)} \qquad \textbf{(15)}$$

If the model is not rejected, it may be used as a baseline against which more restrictive models are compared in subsequent tests such as metric invariance (Horn & McArdle, 1992), ($\Lambda_x^{(1)} = \Lambda_x^{(2)} = ... = \Lambda_x^{(g)}$), invariant unique variances ($\Theta_\delta^{(1)} = \Theta_\delta^{(2)} = ... = \Theta_\delta^{(g)}$), or invariant factor variances ($\Phi^{(1)} = \Phi^{(2)} = ... = \Phi^{(g)}$). On the other hand, if the configural invariance hypothesis is rejected, other comparisons between groups cannot be interpreted meaningfully.

Metric invariance. When the hypothesis of configural invariance is not rejected, the next test is typically a test of "metric invariance" (Horn & McArdle, 1992) in which the established baseline model is compared to a model in which factor loadings of like-items are constrained to be equal across groups (i.e., $\Lambda^{(1)} = \Lambda^{(2)}$). If the constrained model fits well and does not produce a significant decrement in fit, metric invariance is supported. At this point, the researcher may proceed with comparisons of means or intercepts across groups.

On the other hand, if metric invariance is not supported, the researcher is left with few options. One option would be to stop all additional analyses and declare that the measure is not invariant. That is to say observed and latent scores are not directly comparable and, thus, further data analysis should not be conducted. Obviously, discounting all of the efforts of data collection and analysis is not appealing to researchers in most cases. Thus, most will attempt to determine the source of non-invariance. Byrne et al. (1989) and more recently Stark, Chernyshenko, and Drasgow (2006), recommend a procedure for testing individual items for cross-group equivalence one at a time. This process is considered a test of "partial metric invariance." The rationale is that, if the source of the lack of invariance is found to be limited 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. In the current study, the focus will be on tests of metric invariance at the scale- and item-level, though tests of other model parameters are certainly possible (see Ployhart & Oswald, 2004; Vandenberg & Lance, 2000 for excellent reviews).

1.3 Model Identification and Scaling in Multi-Group CFA

Recall from the previous discussion of CFA model estimation that some parameters must be constrained in order to provide a scale for the unobservable latent variables, either by setting the variances of the latent variables equal to a constant (e.g., to 1.0) or, more commonly, by giving them the scale of one of their indicators, by constraining a factor loading to 1.0 (Bollen, 1989). In the latter scenario, the item with a constrained loading is typically referred to as a "referent indicator" (RI) item. In multi-group comparisons, either scaling procedure will implicitly assume some degree of invariance; that is, it is assumed that the entity selected for standardization, either the variance of the latent variable or RI, is invariant across groups (Cheung & Rensvold, 1999). Both standardization procedures have direct consequences for the estimation of factor loadings; therefore, if the standardized entity is not truly equal across groups, factor loadings in different groups will be expressed in different metrics across groups, possibly compromising the validity of conclusions drawn from the test.

To elaborate this point, 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 (or reliability) of the item. The magnitudes of all other relationships between observed variables and the latent variables will be expressed in the scale of the RI. An unpublished manuscript by Hancock, Stapleton, and Arnold-Berkovits provides an illustration. In their example, they 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}. The effect of the scaling constant can be seen upon consideration of the relationship between the observed covariance matrix and the estimated parameters. For $\mathrm{cov}(x_2, x_1) = \lambda_{21}\lambda_{11}\phi_{11}$, the quantity $\lambda_{21}\phi_{11}$ must adjust by 1/*j* to reproduce the observed $\mathrm{cov}(x_2, x_1)$. Likewise, for $\mathrm{cov}(x_3, x_1) = \lambda_{31}\lambda_{11}\phi_{11}$, the quantity $\lambda_{31}\phi_{11}$ must adjust by the same factor, 1/*j*. This leaves the remaining expression, $\mathrm{cov}(x_3, x_2) = \lambda_{31}\lambda_{21}\phi_{11}$, in which λ_{31} and λ_{21} would adjust by *j* and ϕ_{11} by $1/j^2$.

Consider a true model in which a one unit change on the latent variable (ξ_1) is, on average, reflected in a difference of .8 units on the first indicator variable (x_1). In this case, the constraint of 1.0 would be 1/.8, or 1.25 times the actual value of λ_{11}. Thus, in order to reproduce $\mathrm{cov}(x_2, x_1) = \lambda_{21}\lambda_{11}\phi_{11}$ where λ_{11} is adjusted by *j* = 1.25, $\lambda_{21}\phi_{11}$ must in turn adjust by 1/1.25. Similarly, to reproduce $\mathrm{cov}(x_3, x_1) = \lambda_{31}\lambda_{11}\phi_{11}$, the quantity $\lambda_{31}\phi_{11}$ must adjust by 1/1.25. The effect of the adjustments in these two equations, taken together with the final equation, $\mathrm{cov}(x_3, x_2) = \lambda_{31}\lambda_{21}\phi_{11}$, is such that λ_{31} and λ_{21} must adjust by 1.25 and ϕ_{11} by $1/1.25^2$ in order to reproduce the observed covariance matrix.

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, in reality, different across groups, parameter estimations will be adjusted differentially across groups. Several such scenarios are presented in the Hancock et al. paper and can be seen in table form in Table 1. Consider the true situation in which configural invariance is present, but metric invariance is not. In their example, the true population parameters are said to be $\lambda_{11} = \lambda_{21} = \lambda_{31} = .8$ in Group 1 and $\lambda_{11} = \lambda_{21} = \lambda_{31} = .5$ in Group 2. If λ_{11} were selected to equal to 1.0 in both populations, $\lambda_{21}^{(1)}$ and $\lambda_{31}^{(1)}$ would be adjusted by 1/.8 and $\lambda_{21}^{(2)}$ and $\lambda_{31}^{(2)}$ would be adjusted by 1/.5. In this case, λ_{21} and λ_{31} would be estimated to be 1.0 in both populations, creating the appearance of full metric invariance when, in fact, none of the factor loadings were invariant across populations.

In the preceding situation, the selection of any item to serve as a referent indicator would have produced the same result. Alternatively, the authors offer a scenario of partial metric invariance in which x_1 is related to the construct in the same way across populations while x_2 and x_3 are not. In this example, the true parameters are $\lambda_{11} = .8$ for both groups but, while in Group 1 λ_{21} and λ_{31} also equal .8, in Group 2 λ_{21} and λ_{31} equal .6. If λ_{11} was selected as the RI, the constraint would be an appropriate one, and λ_{21} and λ_{31} would be correctly identified as differentially functioning (DF) parameters. Conversely, if the DF λ_{21} parameter was selected instead, estimation would be influenced by a different factor in Group

1 than in Group 2, and λ_{11} would incorrectly be identified as DF. Further, λ_{31} would be identified as invariant, also incorrectly. The adjustments of 1/.8 and 1/.6, or 1.25 and 1.67, yield parameter estimates of $\lambda_{11} = \lambda_{31} = 1.00$ for the first group and $\lambda_{11} = 1.33 \neq \lambda_{31} = 1.00$ for the second. The same result would have been produced by selecting x_3 as the referent indicator. In this example, tests of invariance could indicate that either one or two loadings are invariant, depending on the choice of RI. Clearly, this is problematic at any level of analysis.

As these scenarios illustrate, tests of MI require researchers to assume exactly what it is that they are investigating, namely, that one item is truly invariant across groups. Further, the assumption is not only un-testable but one that, when violated, 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 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 questions about how problematic the practice might be or under which 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 used simulated data to manipulate the magnitude of differential functioning (DF; a lack of invariance) on the RI. Additionally, in some conditions, a manipulation was included in which either two or no non-RI items are specified to function differentially. Of primary interest were the effects of DF of the RI on both scale- and item-level tests of MI. While it is generally understood that when the RI is invariant (i.e., the same across groups), DF of other items will be accurately detected as 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.

The RI determines the scaling of the latent variable, which is then reflected in the estimated factor loadings for all items. As a result, to the extent to which there is DF in the RI, this difference should be reflected in other scale items. Therefore, RI DF may be accurately detected as a lack of invariance at the scale-level but result in inaccurate conclusions at the item-level. Moreover, with larger magnitudes of DF across groups for the RI, the likelihood of accurately detecting DF at the scale-level should increase. In order to investigate the effects of DF of the RIs, conditions were simulated in which the RIs were DF to varying degrees but the other scale items were invariant. In a second set of conditions, both the RI and an additional item on each of two factors were functioned differently across groups. Finally, as Meade and Lautenschlager (2004) have shown that sample size directly affects the power of MI tests, two conditions of sample size were also included. As the use of a RI causes an adjustment in scaling of all item factor loadings, but in somewhat unpredictable ways, I propose:

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

Hypothesis 2: The effects of selecting a DF RI will have a minimal impact on the accuracy of MI conclusions at the scale-level.

Hypothesis 3: The effects of selecting a DF RI will have a large impact at the item-level.

Research Question 1: To what extent does the number of DF items affect the accuracy of MI conclusions?

2. Method

In this study, data with known properties were 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. One hundred sample replications were simulated for each of the study conditions described below. A summary of the conditions is presented in Table 2.

2.1 Sample Size

Data will be simulated to represent sample sizes of 150 and 500. 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 a condition of a larger sample, as might be expected in practice.

2.2 Nature of the Model

Because orthogonal factors are unlikely to be encountered in practice, for all conditions, the model simulated was one with two latent variables specified to correlate at .3 (cf. Meade & Kroustalis, 2006). For each latent variable, four indicator variables were simulated, as represented in Figure 1. 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 3) 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).

2.3 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 factor loading differences of .25. Population factor loadings simulated in Group 2 for the different conditions are presented in Table 4.

2.4 Model Parameter Simulation

Initial structural models were simulated for the various conditions outlined in Table 2 using the PRELIS program which accompanies the LISREL 8.51 software package (Jöreskog & Sörbom, 1996). Group 1 data were simulated to represent the 8-indicator, 2-factor model (Table 3) and were analyzed in all conditions while Group 2 data were modified to simulate conditions of a lack of invariance by subtracting the specified amount of DF (see Table 4).

2.5 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 (in addition to the RI) was constrained to be equal; this was repeated until each of the non-RI items had been constrained (see Stark et al., 2006, for the merits of this type of item-level test).

2.6 Outcome Measures

The outcomes of interest in this study were 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, logistic regression was used to determine the probability of significant versus non-significant lack of invariance based on the level of the various study variables (*N,* RI DF, and non-RI DF). The Cox and Snell Index and the Nagelkerke Index were used to assess the acceptability of the proposed model and Wald statistics were used to evaluate the significance of the βs relating each of the predictor variables to the dichotomous significant/non-significant dependent variable. A significant Wald statistic signals the statistical significance of the associated predictor.

At the item level, results were reported in two metrics. First, true positive (TP; the 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, also reported for each condition are the percentages of the one-hundred replications that were significant for each item, and whether significance represented TP or FP depending on the true invariance status of the particular item.

3. Results

Across the study conditions, for tests of scale-level metric invariance, the percentage of 100 data replications indicating a statistically significant lack of invariance ranged from 5 to 100. As seen in Table 5, conditions of larger sample sizes, greater RI DF, and greater non-RI DF exhibited larger percentages of significant replications. In order to provide a more thorough investigation of these results, logistic regression analyses were conducted with significant versus non-significant lack of scale-level MI as the dependent variable and magnitude of RI DF (0, .05, .1, or .2) and DF on two non-RI items (0 or .25) as the independent variables, in order to determine whether a statistically significant relationship existed between the MI results and the study variables. Though originally hypothesized to be an additional predictor variable in this relationship, it was not possible to include sample size in the analysis, as discussed in more detail below.

3.1 Scale-level Tests of Measurement Invariance

The overall logistic regression model for the scale-level analyses was significant (Wald = 237.71, $p < .0001$). To assess the fit of the model, I looked at two *"pseudo-R^2"* statistics (Cohen, Cohen, West & Aiken, 2003), the Cox and Snell Index and the Nagelkerke Index. The Cox and Snell index is intended to be a logistic analogue to the multiple correlation in OLS regression. However, the Cox and Snell index is problematic in that it has a maximum value of .75 and thus, I also present the Nagelkerke Index, which is adjusted such that a maximum value of 1.00 can be attained. The Cox and Snell R^2 value for the model was .37 and the Nagelkerke R^2 statistic was .50, indicating that 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 6. Below, I discuss how the logistic regression results relate to the study hypotheses.

Hypothesis 1. The first research hypothesis predicted that larger sample sizes would be associated with a more frequent detection of a lack of invariance. In order to investigate this hypothesis, the sample size (N=150 vs. N=500) was included in the logistic regression analyses as a predictor variable, such that a significant main effect for sample size would indicate that a lack of MI is more likely to be detected in samples of 500 than in samples of 150. However, the detection of lack of invariance in the N=500 sample size was so frequent that it was not possible to include these data in the logistical regression analyses. That is, when the N=500 data was included in the analyses, the combination of the sample size variable and the DF=.25 level of the non-RI DF variable almost always resulted in significance (see Table 5), thus resulting in a separation of data points. In logistic regression, when a separation of data points such as this occurs, maximum likelihood estimation cannot converge on a solution. Therefore, data simulated under the N=500 condition was dropped from the regression analyses. Notably, though, with the exception of conditions in which no DF was present in the model, the percentage of replications in which significant DF was detected was always greater when N=500 than when N=150. Taking the DF conditions together, the mean percentage of replications indicating significant lack of metric invariance across study conditions was significantly larger for the N=500 conditions than the mean for the N=150 conditions, (m_{150}=52, m_{500}=81.13; t(16), =1.95, p<.05).

Hypothesis 2. Significant Wald statistics for all levels of the RI DF variable (Table 6) indicate that the simulated RI DF is accurately detected at the scale level. As the severity of RI DF increases, the likelihood of detecting a lack of invariance also increases. The top half of Table 5, which presents the percentages of replications found significant across the various study conditions, provides further evidence of this trend, with the percentage of significant replications increasing with increasing RI DF in both sample size conditions. Pictorially, this monotonic increase is illustrated by the lighter line (nonRI DF=0) in Figures 2 and 3; detection of a lack of MI was more likely when the magnitude of RI DF was large.

Research Question 1. The final research objective for the scale level analyses was to evaluate the extent to which the number of DF items affects the accuracy of invariance tests (Research Question 1) for both scale- and item-level MI tests. In logistic regression terms, this is a question of whether or not the presence or absence of non-RI DF predicts the likelihood of significant invariance test results, as indicated by a significant Wald statistic. Results of these analyses demonstrated a significant relationship, such that when two non-RI items were specified to function differentially, the likelihood of rejecting the hypothesis of invariance was significantly greater than when all non-RI items were truly invariant (*Wald*=55.70, *p*<.01). Furthermore, RI DF interacted significantly with the non-RI DF at all levels of RI DF except for that where RI DF=.1 (Table 6). Figure 2 provides a visual representation of this interaction for the *N*=150 condition. For conditions of no DF on non-referent indicator variables, the likelihood of rejecting the metric invariance hypothesis increases as RI DF increases. By inappropriately constraining the DF RI to be invariant across groups, the true differences are forced onto the other scale items. When these transferred differences are large, the lack of invariance is more frequently detected at the scale level. However, when DF of .25 is simulated on two additional items, the probability of detecting DF decreases as RI DF decreases to a point, and then begins to increase. In other words, while the line relating magnitude of RI DF to the probability of significant invariance test results increases monotonically when no other DF is present in the model, when two items non-RI items are DF, the relationship is displayed as a u-shaped line, where the probability of detecting a significant lack of invariance decreases for small magnitudes of RI DF and then increases for larger magnitudes. In this case, when the transferred differences are small, they actually minimize the effects of the differences introduced by specifying non-RI DF; in terms of the logistic regression results, the amount of RI DF does not significantly predict measurement invariance test results when RI DF magnitude is .1 and non-RI DF is present. When the transferred differences are larger (i.e. RI DF = .2 or .4), however, they interact with non-RI DF, augmenting the difference between groups, evidenced by the greater number of tests indicating a lack of invariance. When *N*=500 (Figure 3), the line relating magnitude of RI DF to the probability of significant invariance test results, like in the *N*=150 condition, increases monotonically when no DF is present on non-RI items. However, because of the increased power to detect a lack of invariance associated with the larger sample size, when DF is present on other items in the model, the line steadily indicates 100 percent detection, demonstrating that, across levels of RI DF, nearly all tests of MI indicated a significant lack of invariance.

3.2 Item-Level Tests of Invariance

In order to assess the accuracy of item-level MI tests, true positive (TP; the 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 for each study condition. TP and FP rates were computed for each of the 100 replications, and then averaged across these replications for each condition.

Hypothesis 1. To determine the extent to which Hypothesis 1, that larger sample sizes were associated with more frequent detection of a true lack of MI, was supported, TP rates were compared across the two sample size conditions. The results indicate that, as expected, the *N*=500 sample size condition was characterized by a higher rate of TPs than the *N*=150 condition (See Table 7) as shown in Figures 4 and 6.

Hypothesis 3. Hypothesis 3 predicted that the effects of selecting a DF RI on detection of true MI should be severe for item-level tests of MI. Support for this hypothesis would be indicated a high rate of FP detection of DF items. As displayed in Figures 5 and 7, the percentage of replications which resulted in FPs increased with increasing magnitudes of RI DF in both sample size conditions, supporting the hypothesis that RI DF leads to inaccurate conclusions at the item-level. Specific results of item-level tests giving rise to FP as well as TP rates can be seen in Table 7. It is evident that, when no DF is present on the non-RI items, the DF introduced by the RI is transferred to other non-DF items, as demonstrated by their erroneous detection as DF items (FPs) a sizable percentage of the time.

Research Question 1. To evaluate Research Question 1 regarding the effect of DF of non-RI items on the accuracy of MI conclusions at the item level, TPs and FPs for the conditions in which DF was specified on additional items were compared to those in which no additional DF was specified (Table 7). Interestingly, when DF is present on non-RI items, DF on the RI results in the significant (and erroneous) detection of *non-DF* items (FPs), while truly DF items (Items 3 and 7) are erroneously *not* detected as DF (False Negatives). This trend is apparent in both sample size conditions.

4. Discussion

The results of this study illustrate the effects of referent indicator (RI) selection on the validity of conclusions drawn from measurement invariance tests conducted at the scale- and item-level. In this study, various conditions of sample size, RI DF, and DF of non-RI items were simulated. The results generally supported the research hypotheses, namely, that larger sample size would result in more frequent detection of a true lack of invariance and that the effects of selecting a DF RI have a minimal impact on the accuracy of MI conclusions at the scale-level but a large impact at the item-level. However, the findings with regard to the effects of non-RI DF complicate the interpretation somewhat. DF on non-RI items, in conjunction with RI DF, was found to affect the conclusions drawn from scale-level tests of measurement invariance while at the item-level, RI DF seems to be the primary determinate in the accuracy of MI conclusions. These findings and the implications for researchers conducting tests of measurement invariance on data with unknown properties are discussed in the following sections.

4.1 Sample Size

Sample size has been shown to directly affect the power of tests of invariance (Meade & Lautenschlager, 2004); specifically, when sample size is large, smaller deviations from the baseline model are required to reject the null hypothesis that the two models are not different than when sample size is small. This is a result of the hypothesis testing procedure's reliance on the chi-square statistic, which is calculated by multiplying *N* by the degree of difference between the observed covariance matrix and the model-implied covariance matrix (the maximum likelihood fit function). Thus, Hypothesis 1 predicted that larger sample size would result in a more frequent detection of DF. The results indicated that this was the case and, in fact, when other DF items were included in the model, significant DF was detected between 99% and 100% of replications, across magnitudes of RI DF (see Figure 3). Thus, because of the separation of data points, data from any condition in which *N*=500 was not included in the regression analyses.

4.2 Scale Level Impact of RI DF

This study provides an illustration of the comparative effects of selecting truly invariant and DF items to serve as the RI for tests of measurement invariance. Further, of particular interest is the extent to which the magnitude of RI DF affects the results of measurement invariance tests. The results of this study suggest that, when a truly invariant RI is selected, valid conclusions can be drawn regarding the invariance status of other items in the model and the validity of these conclusions is not affected by DF in other items. Regardless of sample size, when no items functioned differentially, selecting a truly invariant RI resulted in the erroneous detection of a lack of invariance in only 5% of the replications (i.e. the α level of the LRT). When DF was present on other items, selecting a truly invariant RI resulted in the accurate detection of a lack of invariance in 70% of replications of the *N*=150 condition and 100% of the *N*=500 condition. In other words, the results demonstrate that when an appropriate RI is selected, that is one that does not function differentially across groups, tests will accurately reflect the true state of invariance.

When the selected RI did function differentially, and all non-RI items were truly invariant, the likelihood of rejecting the scale-level hypothesis of metric invariance increased as the magnitude of RI DF increased. Put differently, differences between groups on the RI were transferred to other scale items via the constraints on the RI to be equal to 1.0 in both groups. The larger these transferred differences were (i.e. greater magnitude RI DF), the greater the likelihood that these differences will be accurately detected at the scale level. The results indicate that the scale-level test accurately reflects the lack of invariance (while not providing specific information about the source of this lack of invariance). Thus, when metric invariance is rejected, the source of DF could be any of the items in the model, including the RI. When researchers find support for metric invariance at the scale level, these results suggest that they can be somewhat confident that RI DF was not present. However, the effect of selecting a DF RI when additional items also function differentially seems to be somewhat less straightforward.

When small magnitudes of DF were present on the RIs, the frequency with which a lack of invariance was detected was less than when only non-RI items functioned differentially (no RI DF). On the other hand, when larger magnitudes of RI DF were present, the frequency of detection increased to greater than that found in the truly invariant RI condition. Thus, it seems that the adjustments introduced by relatively low amounts of RI DF serve to minimize differences across the two groups' factor loadings. The adjustments introduced by larger magnitudes of RI DF, on the other hand, appear to compound the effects of non-RI DF, producing more readily detectable scale-level differences between groups. In practice, the true amount of RI DF will not be known and, therefore, it is impossible to discern whether finding support for metric invariance should be interpreted as true invariance or a situation in which small amounts of RI DF have minimized the effects non-RI DF. On the other hand, the findings indicate that large amounts of RI DF will be detected at the scale level, whether or not additional items function differentially.

4.3 Item Level Impact of RI DF

At the item level, the results of the current study suggest that the use of inappropriate RIs has serious ramifications for the researcher seeking to pinpoint differences in factor loadings when the hypothesis of metric invariance is not supported. Tests of "partial measurement invariance," (Byrne et al., 1989) are believed to provide a means for identifying the source of DF by independently testing the invariance of each factor loading. However, as in tests of full metric invariance, this method implicitly assumes RI invariance. The results of the current study suggest that the tenability of this assumption will have consequences for the performance of this method and the conclusions drawn from it. Specifically, as the magnitude of RI DF increases, the likelihood of non-DF items being erroneously detected as DF increases while the likelihood of truly DF items being accurately detected decreases. These findings suggest that researchers seeking to make use of partial invariance tests should proceed with caution. The procedure has obvious appeal in that, rather than abandoning further analyses when a lack of metric invariance is detected, specific items can be identified as DF and allowed to vary across groups, enabling the comparison of latent means. However, the procedure is only useful to the extent that identification of DF and non-DF items is possible and, given the results of the current study, may only be appropriate when RI DF is minimal.

4.4 The Case of DF on Non-RI Items

Scale level impact. As discussed above, in conditions where DF was simulated for two of the non-RI items, the likelihood of detecting DF at the scale-level actually decreased with increasing magnitude of RI DF when the magnitude of simulated RI DF was small (.05 and .1). While Hypothesis 2 (the effect of RI DF will be minimal at the scale level) was supported when all non-RI items were invariant, when non-RI DF was present, small amounts of RI DF obscured the true amount of scale-level DF, leading to fewer accurate conclusions than when RI DF was large. Thus, the results do support the notion that RI DF is less problematic at the scale-level than it is at the item-level, in the case where non-RI items function differentially, RI DF may still obscure the true state of invariance.

Item level impact. At the item level, the results of this study suggest that the accuracy of MI conclusions decreases with increasing RI DF whether or not non-RI items exhibit DF. To answer the research question posed in this study, "to what extent does the number of DF items affect the accuracy of MI conclusions," the rates of FPs across levels of RI DF when non-RI DF was present were compared to those for conditions where non-RI DF was absent. As depicted in Figures 5 and 7, when the magnitude of DF on the RI is small, the tendency of the test to identify truly invariant items as DF is comparable to that when RI DF is absent, regardless of the invariance status of other items. As the magnitude of RI DF increases, so does the rate of FPs; however, differences between conditions of non-RI DF remain minimal. Moreover, this pattern of results is apparent regardless of sample size, though the tendency of the test to produce FPs was severe when the sample size was large.

4.5 Implications & Recommendations

When conducting measurement invariance tests using CFA techniques, care must be taken to choose RIs that are truly invariant. While this study found the DF of the RI can have some effect on power at the scale level, the invariance of the RI was imperative for item-level tests. Averaging across items, the percentage of tests conducted in the *N*=150 condition detecting truly invariant items as DF when all non-RI items were truly invariant ranged from 5.83% when the RI was likewise invariant to 76.67% in the largest magnitude RI DF condition. Moreover, the average percentage of accurate detections of non-RI DF decreased from 54% when a truly invariant RI was selected to 14% in the largest magnitude RI DF condition. When a larger sample size was simulated, these results were even more pronounced, with the rate of false positives reaching 100% when RI DF was large. 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.

I 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, I believe that procedures such as those illustrated by Rensvold and Cheung (2001) should be given much further consideration.

Rensvold and Cheung's (2001) factor-ratio test procedure has the benefit of utilizing all, versus one, item as referent indicators. Using an iterative procedure, each indicator is specified as the RI and one additional indicator is constrained, for all possible pairings. By comparing each of these models with the unconstrained model, non-invariant indicator pairs are identified. Once all non-invariant pairs have been identified, steps can then be taken to identify invariant subsets of indicators. While I 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.

4.6 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 these findings is restricted. While I did investigate the effects of including other DF items in our model, I did not vary the magnitude of this DF. I simulated DF by subtracting .25 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. That is, the practice of subtracting from Group 1 factor loadings to form those of Group 2 results in lower reliability for the group model; greater RI DF would have resulted in lower Group 2 reliability. Because lower factor reliabilities can produce decrements the power of CFA tests, the results of the current study are likely somewhat less pronounced than those that would be produced had reliability been held constant across conditions.

While I 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 my 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. I do not expect that the general pattern of results that I 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 my 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.

5. References

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-600.

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.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (Eds.). (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Hancock, G. R., Stapleton, L. M., & Arnold-Berkovits, I. The tenousness of invariance tests within multiple sample covariance and mean structure models.

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.

Hu, L. T., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409-426.

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

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84-99.

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. (2006). Detecting differential item functioning with CFA and IRT: Toward a unified strategy. *Journal of Applied Psychology, 91,* 1292-1306.

Steiger, J. H., & Lind, J. C. (1980). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Thissen, D., Steinberg, L., & Wainer, H (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity*, (pp. 147-

169). Hillsdale, NJ: Erlbaum.

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.

6. Footnotes

[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.

Table 1

Effects of RI Selection for Three-Indicator CFA Examples

	$\lambda_{11}^{(1)}$	$\lambda_{21}^{(1)}$	$\lambda_{31}^{(1)}$	$\lambda_{11}^{(2)}$	$\lambda_{21}^{(2)}$	$\lambda_{31}^{(2)}$
True value	$\underline{.8}$	.8	.8	$\underline{.5}$	.5	.5
Adjustment	$.8^{-1}$	$.8^{-1}$	$.8^{-1}$	$.5^{-1}$	$.5^{-1}$	$.5^{-1}$
Estimated value	1.0	1.0	1.0	1.0	1.0	1.0
True value	$\underline{.8}$	.8	.8	$\underline{.8}$	.6	.6
Adjustment	$.8^{-1}$	$.8^{-1}$	$.8^{-1}$	$.8^{-1}$	$.8^{-1}$	$.8^{-1}$
Estimated value	1.0	1.0	1.0	1.0	.75	.75
True value	.8	$\underline{.8}$	.8	.8	$\underline{.6}$	.6
Adjustment	$.8^{-1}$	$.8^{-1}$	$.8^{-1}$	$.6^{-1}$	$.6^{-1}$	$.6^{-1}$
Estimated value	1.0	1.0	1.0	1.33	1.0	1.0

Note. Superscripts indicate group. Underlined values represent RIs.

Table 2

Summary of Manipulated Conditions

Condition	Manipulation
Sample size	150, 500
Magnitude of RI difference	0, 0.05, 0.1, 0.2, 0.4
Magnitude of non-RI difference	0, .25

Note. Non-RI difference refers to a difference of .25 for Items 3 and 7 population factor loadings. For all others, population factor loadings are equal across groups.

Table 3

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 4

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 5

Results of Scale Level Tests of Metric Invariance

	RI DF	*N=150*	*N=500*
nonRI DF = 0			
	0	5	5
	.05	7	13
	.10	14	40
	.20	43	98
	.40	95	100
nonRI DF = .25			
	0	70	100
	.05	57	99
	.10	42	99
	.20	65	100
	.40	93	100

Note. Results are expressed as percentage of 100 replications found significant.

Table 6

Results of Logistic Regression of Metric Invariance Test on Study Variables

			Wald	Odds	Odds Ratio 95%	
Parameter	*β*	SE	Statistic	Ratio	Confidence Interval	
Intercept	-0.96	0.26	13.42**			
RIdif .05	-1.65	0.35	22.67**	1.43	0.44	4.67
RIdif .1	-0.88	0.28	9.92**	3.09	1.07	8.95
RIdif .2	0.66	0.23	8.18**	14.33	5.37	38.39
RIdif .4	3.88	0.39	97.65**	361.00	101.20	>999.99
nonRIdif	1.90	0.25	55.70**	44.33	16.38	120.01
RIdif .05 * nonRIdif	0.99	0.40	6.24*	0.40	0.11	1.49
RIdif .1 * nonRIdif	0.30	0.34	0.79	0.20	0.06	0.67
RIdif .2 * nonRIdif	-0.98	0.30	10.36**	0.06	0.02	0.18
RIdif .4 * nonRIdif	-2.23	0.51	19.24**	0.02	0.00	0.07

Note. * *p*<.05, ***p*<.01. *df* = 1 for all analyses. For all study variables, no DF was the reference category.

Table 7

Percentage of Replications Yielding Significant Results for Item-Level Tests of Metric Invariance

N=150

RIdif	i2	i3	i4	i6	i7	i8	FP Rate	TP Rate
nonRI DF 0								
0	4	10	1	7	5	8	5.83	
.05	4	9	4	6	7	8	6.33	
.10	14	14	4	10	10	12	10.67	
.20	40	34	26	26	32	27	30.83	
.40	78	80	68	83	79	72	76.67	
nonRI DF= .25								
0	4	53	1	8	55	6	4.75	54.00
.05	5	40	4	8	38	9	6.50	39.00
.10	11	33	5	10	26	13	9.75	29.50
.20	31	16	23	24	11	25	25.75	13.50
.40	72	17	63	79	11	70	71.00	14.00

N=500

RIdif	i2	i3	i4	i6	i7	i8	FP Rate	TP Rate
nonRI DF=0								
0	7	4	7	8	1	5	5.33	
.05	11	14	10	10	7	9	10.17	
.10	31	39	29	37	21	23	30.00	
.20	86	85	67	80	79	72	78.17	
.40	100	100	100	100	100	100	100.00	
nonRI DF= .25								
0	10	98	6	7	98	6	7.25	98.00
.05	11	84	9	9	90	7	9.00	87.00
.10	26	72	25	29	69	23	25.75	70.50
.20	79	12	64	76	17	70	72.25	14.50
.40	100	39	100	99	30	100	99.75	34.50

Note. Underlined values are detection of true DF, all others are invariant items detected as DF.

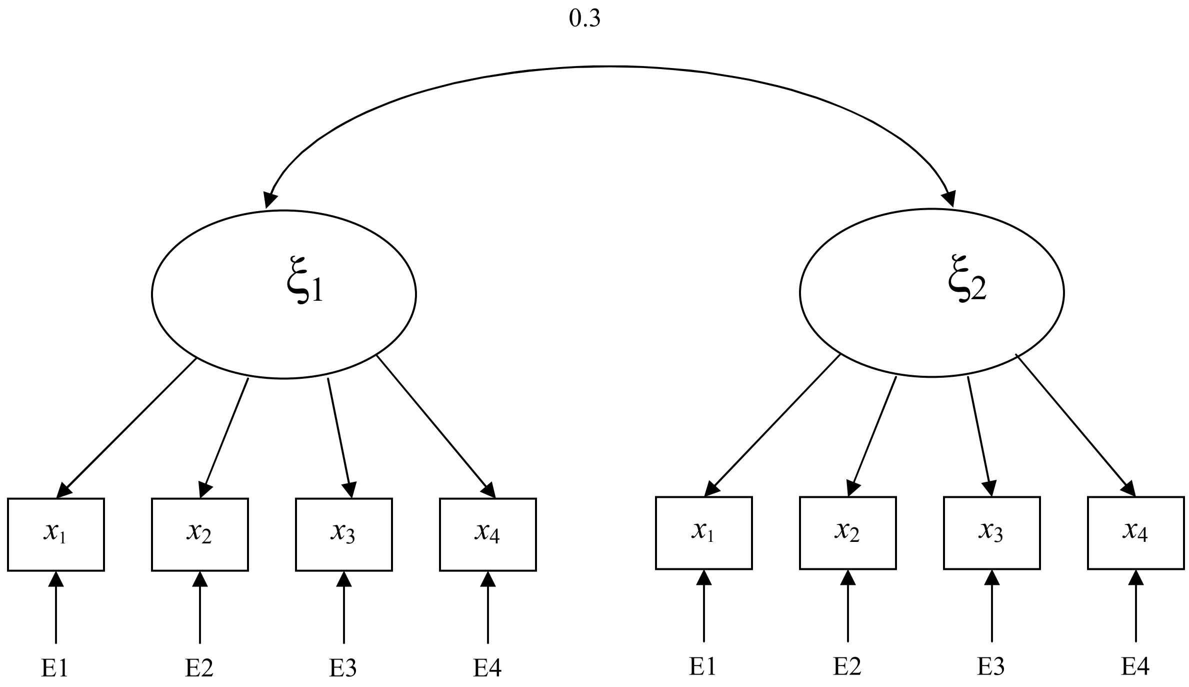

Figure 1. Measurement Model for All Study Conditions

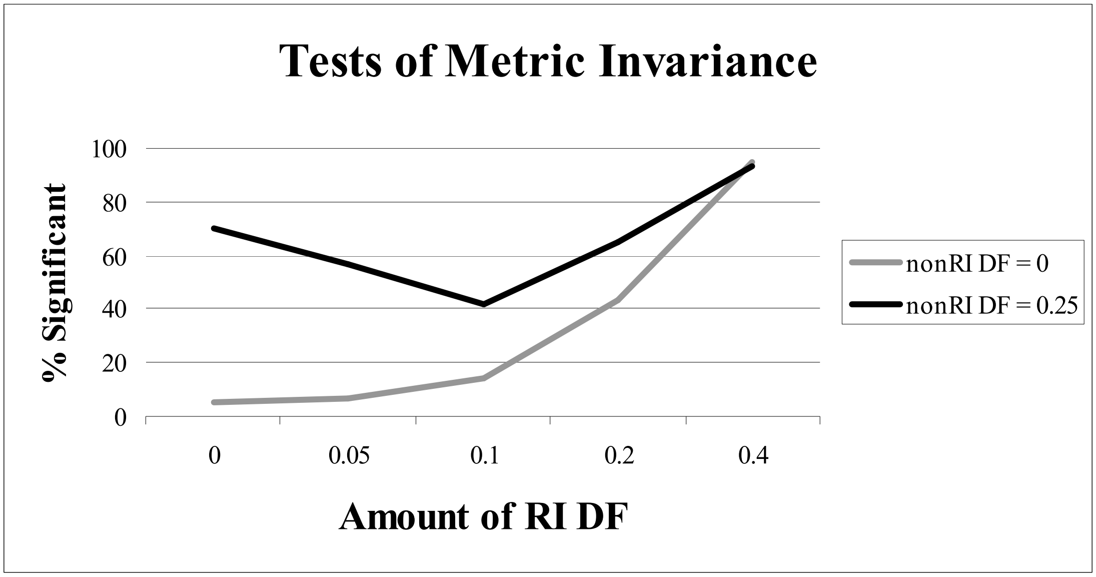

Figure 2. Effect of DF on Scale-Level Tests of Metric Invariance, *N*=150

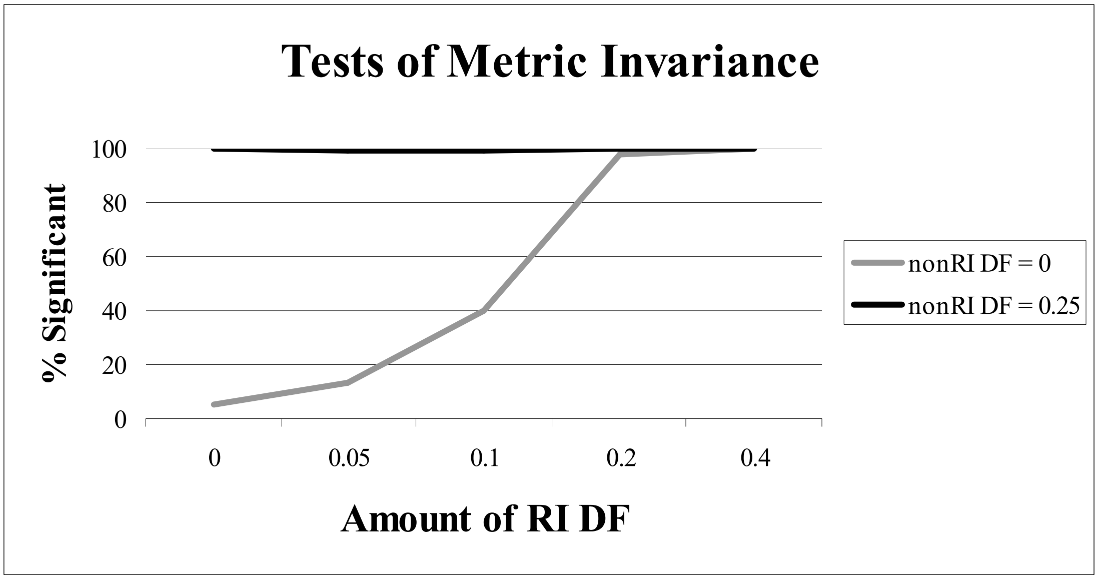

Figure 3. Effect of DF on Scale-Level Tests of Metric Invariance, *N*=500

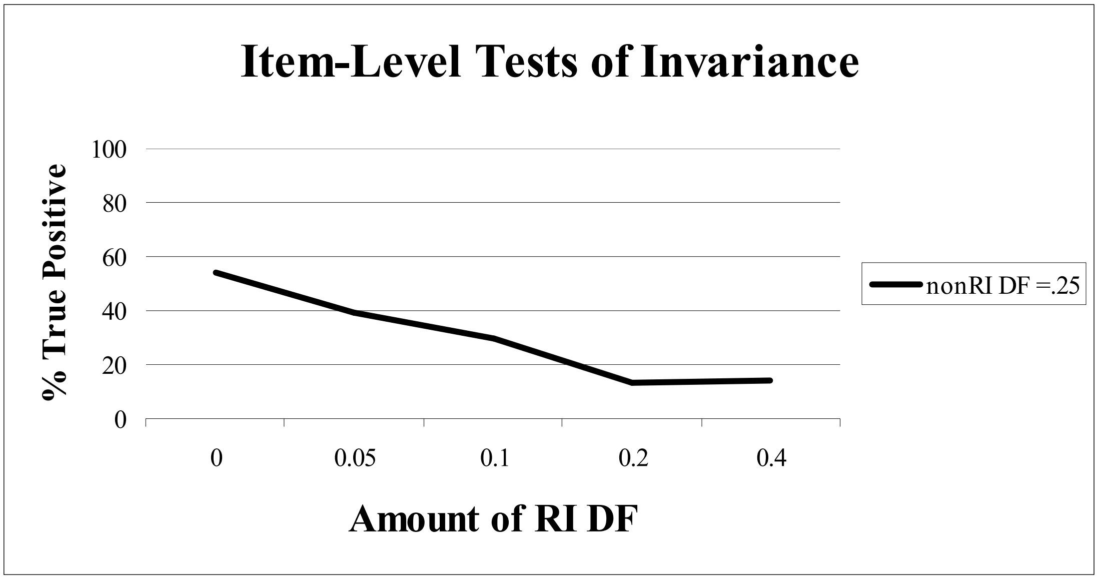

Figure 4. Effects of DF on True Positive Rates of Item-Level Invariance Tests, *N*=150

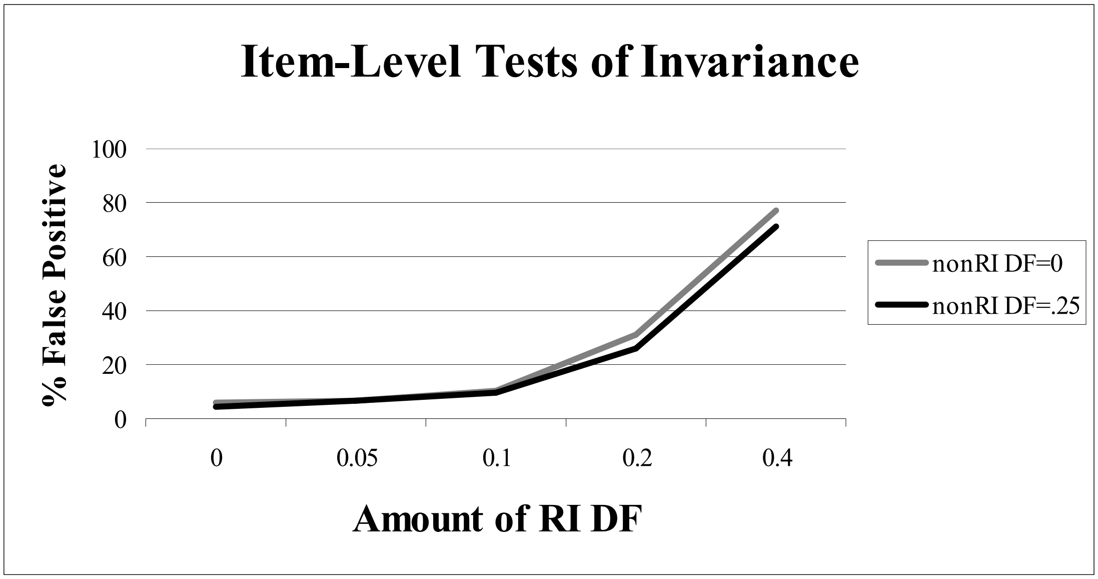

Figure 5. Effects of DF on False Positive Rates of Item-Level Invariance Tests, *N*=150

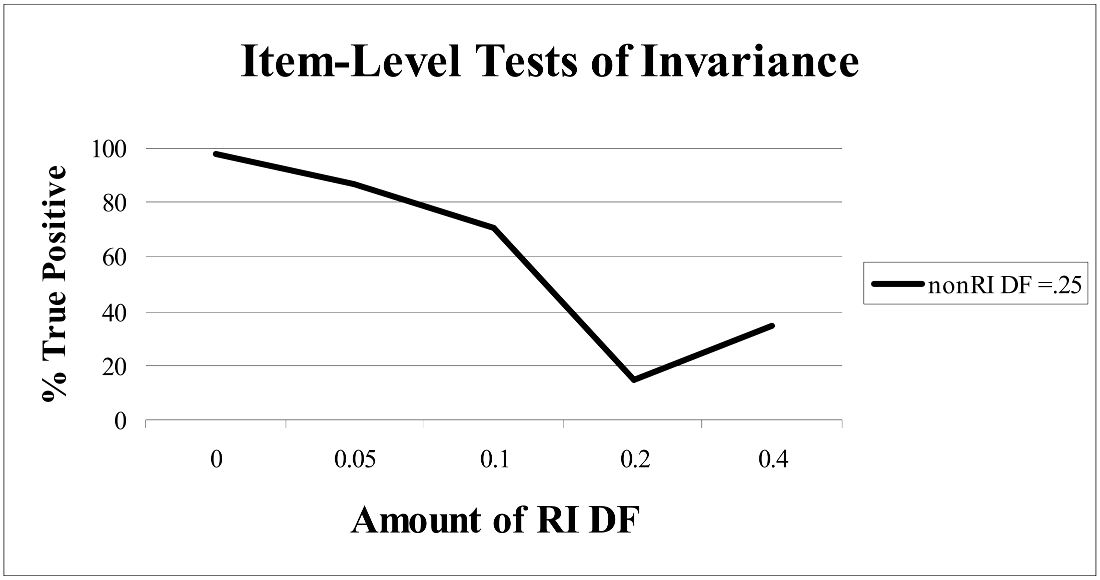

Figure 6. Effects of DF on True Positive Rates of Item-Level Invariance Tests, *N*=500

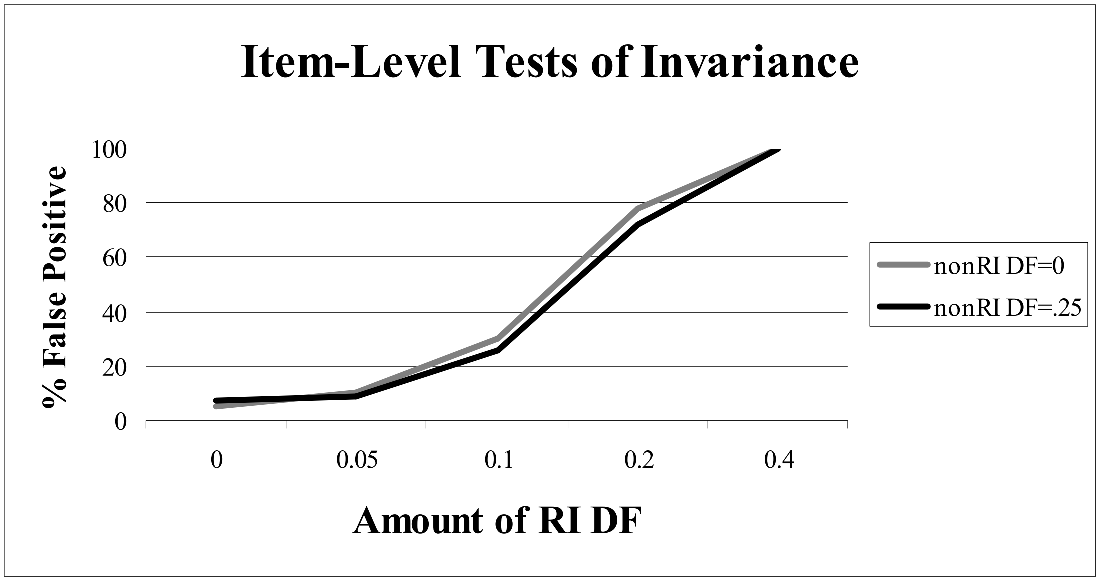

Figure 7. Effects of DF on False Positive Rates of Item-Level Invariance Tests, *N*=500