Spring 2007

Methods of Detecting Differential Item Functioning: A Comparison of Confirmatory Factor Analysis Methods

Jenny Chia Yi Kuang
Old Dominion University

METHODS OF DETECTING

DIFFERENTIAL ITEM FUNCTIONING: A COMPARISON OF

CONFIRMATORY FACTOR ANALYSIS METHODS

by

Jenny Chia Yi Kuang
B.A., University of California, Berkeley, May 1997
M.S., Old Dominion University, Norfolk, VA, May 2000

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirement for the Degree of

DOCTOR OF PHILOSOPHY

INDUSTRIAL AND ORGANIZATIONAL PSYCHOLOGY

OLD DOMINION UNIVERSITY
May 2007

Approved by:

Terry L. Dickinson (Director)

Valerian J. Derlega (Member)

James M. Henson (Member)

Robert M. McIntyre (Member)

ABSTRACT

METHODS OF DETECTING DIFFERENTIAL ITEM FUNCTIONING:
A COMPARISON OF CONFIRMATORY FACTOR ANALYSIS METHODS

Jenny Chia Yi Kuang
Old Dominion University, 2007
Director: Dr. Terry L. Dickinson

The present Monte Carlo study compared four confirmatory factor analysis (CFA) methods for detecting differential item functioning (DIF). The four methods were the noniterative and iterative mean and covariance structure analysis (MACS) methods, the modification index (MI) method, and the modification index-divided sample (MI-divided) method. Reference and focal groups responded to 12 items with 3 of the 12 items designed to exhibit DIF. Sample sizes of 250 and 500 were examined. In addition, three types of DIF were examined: DIF on loadings, DIF on thresholds, and DIF on both loadings and thresholds. Results indicated that for sample size 250, all methods had good DIF detection rates for DIF on thresholds and for DIF on loadings and thresholds; all methods were not sensitive to DIF on loadings. For sample size 500, all methods had good DIF detection rates for DIF on thresholds and for DIF on loadings and thresholds. With the greater sample size, the noniterative and iterative MACS methods were more sensitive to DIF on loadings. The MI-divided method improved to a lesser degree, but the MI method did not improve at all. For DIF on thresholds for sample size 500, the noniterative MACS, MI, and MI-divided methods had false positive rates that were greater than expected by chance. Only the iterative MACS method maintained the false positive rates at or below that expected by chance.

This dissertation is dedicated to

Fred, Hui Chin, and Daniel Kuang

ACKNOWLEDGMENTS

There are many people whom I wish to acknowledge for helping me accomplish this goal. I wish to give very special thanks to Dr. Terry L. Dickinson, my advisor, for providing me with invaluable guidance and support. I also wish to thank my dissertation committee members, Dr. Valerian J. Derlega, Dr. James M. Henson, and Dr. Robert M. McIntyre for their assistance and expertise.

Finally, I wish to thank my parents, Fred and Hui Chin Kuang, and my brother, Daniel Kuang, for their unwavering love and support.

TABLE OF CONTENTS

Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Psychological measures are commonly used in applied and research settings to accomplish a variety of purposes. These measures may influence personnel decisions and have consequences for individual careers as well as for organizational performance. The measures may also influence current theories and the development of new theories to guide future research. Using psychological measures for these purposes often requires that the measures are equivalent across groups or subpopulations of individuals (Drasgow & Kanfer, 1985).

Measurement equivalence exists when the relations between observed variables and latent variables (or constructs) are the same across groups (Drasgow & Kanfer, 1985). With measurement equivalence, observed variables from different groups are on the same measurement scale, making the scores of individuals directly comparable. Individuals with the same amount of a construct would be expected to have the same observed score on a measure of that construct regardless of group membership.

Lack of equivalence for a measure is referred to as differential functioning. For items of a measure, lack of equivalence is described as differential item functioning (DIF). An item exhibits DIF when individuals from different groups have the same standing on the construct, but the individuals have different expected observed scores for the item (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Hulin, Drasgow, & Parsons, 1983; Lord, 1980; Raju & Ellis, 2002).

This dissertation adheres to the format of the *Journal of Applied Psychology*.

It is important to note that neither the presence nor absence of DIF implies that groups are equal or unequal in the amount of the construct. Item invariance for an item only stipulates that individuals who are equal on the construct have the same expected observed score on that item, regardless of group membership. It is quite possible that two groups differ significantly in their average score on the construct, and none of the items exhibit DIF. The true difference in construct means is referred to as *impact* (Drasgow & Hulin, 1990). Impact and DIF are different concepts, and neither one necessarily implies the other (Raju & Ellis, 2002).

Employment tests are an example of measures for which it is vital to identify and eliminate DIF items. DIF items may contribute to a mean difference in test scores between groups that is not reflective of a true mean difference in the construct (i.e., impact). This mean difference due to DIF items is merely a lack of measurement equivalence. The DIF items may lead to adverse impact against a non-favored group despite the non-favored group being equally as capable as the favored group (Drasgow & Kanfer, 1985). Such DIF items are a threat to the validity of an employment test and may lead to unfair hiring decisions by an organization. Legal problems may result that could prove costly in time and money for non-favored group members and the organization.

Typically, DIF analyses are done for two groups that differ in demographic characteristics such as gender or race (e.g., Caucasians versus African Americans). The favored group is referred to as the *reference group* (e.g., majority group, males, native speaking), and the non-favored group is referred to as the *focal group* (e.g., minority group, females, non-native speaking).

However, DIF analyses can generalize beyond comparing groups that differ in demographic characteristics (Raju & Ellis, 2002). For instance, in cross-cultural research, such analyses can be used to establish the fidelity of translated measures across cultural groups (Ellis, 1989; Ellis & Kimmel, 1992; Hulin et al., 1983). In an organizational context, DIF analyses may be used to compare the ratings of employees to those of their peers or supervisors (Maurer, Raju, & Collins, 1998).

In summary, the analysis of items to detect DIF is a powerful tool that can be used in applied and theoretical research contexts. Organizations need to remove or revise items that exhibit DIF from their measures. Removal or revision of such items improve an organization's ability to use human resources fairly and effectively and reduces legal problems. Researchers also need to remove or revise items that show DIF to ensure that their measures possess measurement equivalence. The presence of DIF hinders construct comparisons and theory development across groups, cultures, and contexts.

Types of Differential Item Functioning

Mellenbergh (1982) described two types of DIF: uniform DIF and nonuniform DIF. Whether DIF is uniform or nonuniform depends on the type of item parameter that differs across groups. For DIF detection, item difficulty and item discrimination are the two parameters of interest.

Item difficulty is a location parameter that associates the latent variable and the mean item response. The location parameter corresponds to the value on the latent variable scale at which the mean item response is a fixed value. Alternatively, the location parameter corresponds to the mean item response value at which the latent variable is at a fixed value. Depending on the type of item response model, it may be the value on the latent variable scale that is fixed, or it may be the mean item response that is fixed. For Item Response Theory (IRT) models, the mean item response (i.e., the probability of a correct item response) is typically fixed at .50 so that the item difficulty parameter relates to the corresponding value on the latent variable scale. An item difficulty parameter that is higher in value indicates that a higher latent variable level is required if an examinee is to have a .50 probability of endorsing the particular response. For factor analytic item response models, the value on the latent variable scale is typically fixed at 0 so that the item difficulty parameter relates to the corresponding mean item response value. An item difficulty parameter that is higher in value indicates that an examinee with a value of 0 on the latent variable scale would obtain a lower item mean (i.e., probability of correct response) (Chan, 2000; Ferrando, 1996).

Item discrimination involves the extent to which an item is able to distinguish between those individuals who are high on the latent variable from those who are low on the latent variable. An item with a higher discrimination parameter has a narrower range on its latent variable scale within which the item is able to make the distinction between those who are high in the latent variable from those who are low in the latent variable.

Uniform DIF exists when the item difficulty parameter differs across groups, but the item discrimination parameter is the same. Under this circumstance, there is no interaction between ability level and group membership. Nonuniform DIF exists when the item discrimination parameter differs across groups. In the case of nonuniform DIF, there is an interaction between ability level and group membership.

Classical Test Theory and Item Response Theory Methods

Various methods have been developed to detect DIF. The first methods designed to detect DIF were based on classical test theory. These methods include the transformed item difficulty index, adjustments to the transformed item difficulty index, analysis of variance, The Golden Rule procedure, and differences in item point-biserial coefficients (Angoff, 1993; Chan, 2000; Camilli & Shepard, 1994). In their review, Camilli and Shepard (1994) recommended that classical test theory methods should not be used due to their relative ineffectiveness in detecting DIF when compared to item response theory (IRT) methods.

IRT methods for DIF detection are among the most promising, but they have limitations. These methods are most appropriate for longer tests, and they require very large sample sizes. Both requirements can be a challenge to attain in many applied and research situations (Reise, Widaman, & Pugh, 1993).

In the present research, four methods are compared for their capability to detect DIF in a computer simulation. All methods are based on structural equation modeling analysis. In comparison to IRT methods, structural equation modeling methods have the advantage of being more appropriate for shorter tests and smaller sample sizes.

Confirmatory Factor Analysis Methods

Confirmatory factor analysis (CFA) has been noted as a promising method for examining DIF (Camilli & Shepard, 1994). CFA specifies a general model for testing hypotheses regarding relationships between observed and latent variables. Latent variables (i.e., factors, traits, constructs, or abilities) are abstract concepts that are not directly measured. Observed variables are directly measured, and they serve as indicators of the latent variables.

The general CFA model is represented as:

$$X = \Lambda\xi + \delta \tag{1}$$

In this equation, X is a vector of observed variables (e.g., items in a test), Λ is a matrix of structural coefficients (i.e., factor loadings) for the latent variables, ξ is a vector of latent variables, and δ is a vector of measurement errors.

When applying CFA, the relationships among observed and latent variables of the model are evaluated for their fit to item responses. The fitted model is associated with a likelihood ratio statistic that approximately follows the chi-square distribution. This statistic provides an overall indication of model fit to the item responses. CFA also provides modification indexes to indicate how the model might be adjusted to improve its fit to the item responses. Each modification index is a statistic that approximates a chi-square distribution with one degree of freedom (Jöreskog & Sörbom, 1996a). Statistically significant modification indexes suggest parameters to estimate that were formerly fixed. Doing so would improve model fit by a value that is reasonably close to the modification index value.

Four different CFA methods for DIF detection are described next. The first method requires examining chi-square goodness-of-fit indexes and the remaining three methods are based on the use of modification indexes.

Mean and Covariance Structure Analysis Method. The mean and covariance structure analysis (MACS) model proposed by Sörbom (1974) is a method that has been shown to be effective in detecting DIF. Chan (2000) applied the MACS model to detect

DIF between occupational groups and between gender groups on cognitive style scales. Whiteside-Mansell and Corwyn (2003) applied the MACS model to detect DIF between age groups on a self-esteem scale. Everson, Millsap, and Rodriguez (1991) used similar models to address DIF and measurement equivalence. In addition, several simulation studies have been conducted to examine the ability of the MACS model to detect DIF (Meade & Lautenschlager, 2004a, 2004b; Stark, Chernyshenko, & Drasgow, 2005). The studies indicate that the MACS model is effective in detecting DIF for both real and generated data.

Within the MACS model, items are usually measured with a Likert-type rating scale. Such a scale uses a polytomous ordered response format that is considered to approximate responses on a continuous line (Ferrando, 1996). Consider an item response x_{ij} that represents the observed response of individual i to item j where x is a number on a continuous scale. Assume that item responses on the measure are explained by one latent variable ξ (i.e., factor). The MACS model represents the relationship between x and ξ in a linear regression of x on ξ:

$$x_{ij}^{(g)} = \mu_j^{(g)} + \lambda_j^{(g)} \xi_i^{(g)} + e_{ij}^{(g)} \tag{2}$$

In this equation, μ_j is the regression intercept (or the mean response to item j when ξ is 0); λ_j is the regression coefficient (or the factor loading for item j) and is the expected change in the scale response per unit change in ξ; e_{ij} is a stochastic error term; and g refers to group membership.

In the MACS model, the item intercept reflects the item difficulty parameter, and the item factor loading reflects the item discrimination parameter (Ferrando, 1996). Lack of invariance of μ_j and λ_j values across groups implies the existence of DIF. Uniform DIF is reflected by between-group differences in μ_j, and nonuniform DIF is reflected by between-group differences in λ_j (Chan, 2000).

The MACS model may be applied with an iterative or a noniterative procedure. For the iterative procedure, either a constrained-baseline approach or a free-baseline approach may be used. Both approaches require that one item is restricted to be invariant across groups. In a constrained-baseline iterative approach, the initial model restricts all of the remaining (i.e., studied) items' parameters to be invariant across groups. Next, a series of augmented models is formed by freeing parameters across groups for the studied items one at a time. In a free-baseline iterative approach, the initial assessed model restricts only the referent item to be invariant across groups. Then, a series of constrained models is formed by restricting parameters to be invariant across groups for the studied items one at a time. Stark et al. (2005) compared a constrained-baseline MACS approach with a free-baseline MACS approach and found that the free-baseline MACS approach was superior in detecting DIF items regardless of sample size, type, and amount of DIF. The free-baseline approach likely performed better because the baseline model did not contain a large number of DIF items which could negatively impact model fit, and, in turn, adversely affect DIF detection. This study uses a free-baseline iterative MACS approach.

In the free-baseline iterative MACS procedure, a baseline model is estimated first. All items are entered into the model with the referent item's loadings and intercepts set invariant across groups. In the next iteration, one item in addition to the referent item has its loadings and intercepts set invariant across groups. A chi-square difference test is conducted between the chi-square value in the second step and that of the baseline model estimated in the first step. If the chi-square is statistically significant, the item is flagged as exhibiting DIF. In the next iteration, a second item and the referent item have their loadings and intercepts restricted to be invariant across groups. A chi-square difference test is conducted between the chi-square value in the third step and that of the baseline model estimated in the first step. If the chi-square is significant, the second item is flagged as exhibiting DIF. The iterative process is continued until all of the items in a measure have been paired with the referent item, and a chi-square difference test has been performed for each augmented model to investigate the presence of DIF.

A common way to set the metric for latent variables is to select a reference item and set its loading equal to one across groups. In addition, the intercepts for the reference item must be constrained to be equal across groups.

When comparing multiple baseline models, the critical *p*-value may be adjusted using the Bonferroni method to control Type I error. However, Stark et al. (2005) found that with samples of 1,000 or less, Bonferroni corrections are too conservative. The present research does not adjust critical *p*-values.

The MACS model may also be applied with a noniterative procedure. In the noniterative procedure, all items are entered into the model with each item's loading and intercept set (i.e., restricted to be) invariant across groups. Once the model is estimated, modification indexes for the loadings and intercepts are examined to identify items exhibiting DIF. Specifically, items with modifications indexes that are large and statistically significant are flagged as exhibiting DIF, because their intercepts or loadings are not invariant and differ across groups. The nature of the DIF (i.e., uniform or nonuniform) depends on whether intercepts, loadings, or both show DIF.

Modification Index Method. In the modification index (MI) method, two latent variables are specified for the model: ability and group membership. Group membership is specified in the hypothesized model as a latent variable that has no measurement error. All items in the measure are included in the model with all of the items loading on the ability factor. The group membership indicator variable alone loads on the group membership factor. Once the model is estimated, modification indexes on the group membership latent variable are examined to identify items exhibiting DIF. Namely, items with modifications indexes that are large and statistically significant are flagged as exhibiting DIF (Oort, 1992, 1996, 1998).

As mentioned previously, the modification index is approximately a chi-square with one degree of freedom. A modification index reflects the value by which the overall chi-square of a model will decrease if the item parameter is freely estimated. With the MI method, only one model needs to be estimated to acquire the modification index for a suspected item. The modification index for the item parameter on the group membership latent variable provides the value by which the overall chi-square of the model will decrease if the item parameter was freely estimated to load on the group membership latent variable. Thus, the MI method does not require a series of comparisons between a baseline model and augmented models to calculate the difference in their chi-square values. The MI method is very practical and efficient.

Oort (1992, 1996, 1998) proposed the use of an adjusted critical value (AC) for determining the significance of a modification index to adjust for the occurrence of false positives. The adjusted critical value is:

$$AC = \left[\frac{\chi^2}{C + df - 1} \right] C \qquad (3)$$

where C is a critical value (e.g., .01 or .05). A particular item is considered to exhibit DIF if its modification index is greater than the adjusted critical value. The adjusted critical value is especially useful when there are no anchor items, and all items in the hypothesized models are under suspicion for DIF. Because the present research has a small percentage of DIF items, an adjusted critical value is not used.

Oort (1996, 1998) used MI methods to detect DIF in a simulation study and concluded that they are promising methods. MI methods were found to be comparable to IRT in detecting DIF for dichotomous items. In addition, MI methods were found to be better than IRT in detecting DIF for polytomous responses. MI methods were superior to IRT in conditions involving seven-point items, large sample sizes, small differences in ability between groups, equal group sizes, and large amounts of DIF.

Modification Index-Divided Sample Method. The modification index divided sample (MI-divided) method (Wanichtanom, Dickinson, & Coates, 2003) applies the logic of a Mantel-Haenszel method for DIF detection (Mazor, Clauser, & Hambleton, 1994).

The Mantel-Haenszel method is not based on a latent variable model; it only involves observed variables. The method is very popular because it is appropriate for small sample sizes and is easily implemented in statistical program packages. One disadvantage of the method, however, is its lack of sensitivity to nonuniform DIF (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan &

Rogers, 1990). To address this weakness, Mazor et al. (1994) proposed a variation of the Mantel-Haenszel method that is effective in detecting nonuniform DIF.

Mazor et al.'s (1994) approach involves three analyses of items. First, the total sample is analyzed for the presence of DIF. The total sample is then divided into high-scoring and low-scoring subgroups. Each subgroup is analyzed separately for the presence of DIF. If one of the three analyses indicates DIF for an item, then that item is flagged as exhibiting DIF. Mazor et al. (1994) demonstrated that this divided sample method can increase detection of DIF from 68% to 82% without increasing the Type I error rate.

In the MI-divided method, Wanichtanom et al. (2003) applied Mazor et al.'s (1994) approach to CFA for DIF detection for dichotomous items. The MI method for CFA described earlier is first applied to the total sample. Then, the MI method is applied to the high-scoring group and the low-scoring group. If an item exhibits DIF in any of the three analyses, the item is flagged. The method successfully detected both uniform and nonuniform DIF in dichotomous items (Wanichtanom et al., 2003). In particular, the MI-divided method had DIF detection rates that were better than the MI method and were comparable to IRT analysis. The present research will also examine the MI-divided method.

Variables Considered

Level of Differential Item Functioning. A common approach to DIF detection research is the use of Monte Carlo simulation that allows a researcher to create data with "known" item parameters. Item parameters for a reference group may be defined to represent a range of item difficulty and item discrimination values. These item parameter values would also be used for the focal group, but a subset of the items would be specified to contain DIF for the focal group. The parameters for this subset of items would be modified by adding or subtracting values (e.g., .5) to create DIF. Item responses would be generated from the parameters for each group. The responses for the two groups would then be analyzed for the influence of studied variables (e.g., sample size) that are hypothesized to affect the detection of the items that were designed to exhibit DIF.

There are two ways by which DIF can be designed for item parameter values. Most existing research uses a procedure wherein constant values are added or subtracted to item parameters to define an amount of DIF. This approach works well for studies investigating uniform DIF (i.e., only the difficulty parameters are different between groups). This approach does not work well, however, when a study is investigating nonuniform DIF because adding and subtracting constant values to item discrimination parameters is not linearly related to the amount of DIF.

A second procedure is based on the IRT literature and involves the use of item characteristic curves (ICC). IRT has two basic postulates (Suen, 1990). First, the performance of a respondent on a measure can be predicted by a set of latent variables called traits, constructs, or abilities. Second, the relationship between a respondent's performance on an item and the set of latent variables underlying item performance can be expressed as a monotonically increasing function; this function is the ICC.

Plots of ICCs are useful in DIF detection. Each plot typically contains the ICCs for two groups—a reference group and a focal group. The area between the two ICCs reflects the type and amount of DIF for an item. ICCs that overlap completely such that there is no area between the ICCs reflect an item with no DIF. If the ICCs do not overlap, however, the item exhibits DIF, and the greater the area between the curves, the greater the amount of DIF. Raju (1988) provided an exact formula for calculating the area between ICCs:

$$Area = (1-c)\left|\frac{2(a_2-a_1)}{Da_1a_2}\right|\ln\left[1+e^{Da_1a_2(b_2-b_1)/(a_2-a_1)}\right]-(b_2-b_1) \quad (4)$$

In the equation, b is the difficulty parameter, a is the discrimination parameter, c is the pseudo-chance level parameter (or guessing parameter), and D is a scaling constant that is usually set equal to 1.7 or 1.702 (Hulin et al., 1983). Designing DIF using ICCs involves using the area formula to find parameter values that will hold the area between the ICCs constant (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993). This research will use the area between ICCs as a means to set the type and amount of DIF to be detected.

Previous research shows that DIF can generally be detected when the area between ICCs is at least .5. For example, in the IRT literature, one study that set the levels of DIF at .5 and at 1.0 had success in detecting most of the items with DIF (Cohen, Kim, and Baker, 1993). In the CFA literature, one study that set the level of DIF at .5 had success in detecting uniform DIF across several methods (e.g., MI, MI-divided), but the MI-divided method was the only CFA method that performed well in detecting nonuniform DIF (Wanichtanom et al., 2003). This study will fix the DIF level at .50 by using Raju's (1988) formula to calculate the area between ICCs. The method section describes the use of the area formula.

Type of Differential Item Functioning. Previous research has shown that both uniform and nonuniform DIF exist in real data. For example, a study that examined a translation of NEO-PI items found that a large proportion (40%) of the items functioned differently across American and Philippine samples (Huang, Church, & Katigbak, 1997). Both difficulty and discrimination parameters varied between the two groups.

Previous research has also shown that uniform and nonuniform DIF can successfully be detected. For example, Wanichtanom et al. (2003) successfully detected uniform DIF with MI and MI-Divided methods. However, only the MI-Divided method was successful in detecting nonuniform DIF. Stark et al. (2005) successfully detected both uniform and nonuniform DIF using MACS methods.

This study examined uniform and nonuniform DIF. Three types of DIF were examined: DIF on the difficulty parameter only (uniform DIF), DIF on the discrimination parameter only (nonuniform DIF), and DIF on both the difficulty and discrimination parameters (nonuniform DIF).

Sample Size. Previous studies have examined the effect of sample size on DIF detection using CFA methods. Meade and Lautenschlager (2004a) simulated 150, 500, and 1,000 respondents, and sample sizes of 500 and 1,000 were particularly successful for detecting DIF. Stark et al. (2005) successfully detected DIF using the MACS method with sample sizes of 500 and 1,000, and Wanichtanom et al. (2003) successfully detected DIF using the MI-divided method with a sample of size of 1,000.

One of the proposed advantages of CFA methods is the ability to detect DIF using smaller sample sizes. The present study uses sample sizes of 250 and 500. These sizes were selected to be small enough to assess the ability of the CFA methods to detect DIF in smaller sample sizes and also to be large enough to provide sufficient power for adequate detection.

Test Length. Previous research on DIF detection using CFA methods examined various test lengths. Number of items varied from 6 to 50 items (Meade & Lautenschlager, 2004a, 2004b; Oort, 1998; Stark et al., 2005; Wanichtanom et al., 2003). In Meade and Lautenschlager's (2004a) study which examined test lengths of 6 and 12 items, there was greater success in detecting DIF for the longer test length of 12 items. Stark et al. (2005) had success in detecting DIF when using a test length of 15 items.

The present study uses a test length of 12 items that is representative of many unidimensional attitudinal and personality measures that are used in practice and that has been shown to be adequate for detecting DIF. The data represent a single scale measuring a single unidimensional construct (Meade & Lautenschlager, 2005). In practice, if a multidimensional measure contains multiple scales measuring different constructs, each scale can be examined in turn for DIF using the CFA methods examined in the present research.

Proportion of Items with Differential Item Functioning. One disadvantage of internal methods of DIF detection is that the test score itself serves as the criterion. DIF items may not be detected when a large proportion of items in the test exhibits DIF because the test score itself is largely determined by these DIF items. Previous research indicates that as the percentage of DIF items increases on a test, detection rates decrease and false negatives increase (Kim and Cohen, 1992; Oshima and Miller, 1992). In addition, Cohen and Kim (1993) found that false negative rates tend to increase when the proportion of items with DIF is large. In the present research, the proportion of items with DIF is fixed at 25% with 3 of the 12 items designed to exhibit DIF. This percentage is judged to be appropriately low based on previous research.

Purpose of the Research

The present research compares four different CFA methods for detecting DIF: the iterative MACS method, the noniterative MACS method, the MI method, and the MI-divided method.

Previous research has shown the MI-divided method to be successful in detecting DIF, outperforming the MI method in the detection of nonuniform DIF. For the MACS method, previous research has shown the iterative MACS method to be very successful in detecting DIF. This study introduces the noniterative MACS method and examines its efficacy in detecting DIF.

The present research hypothesizes that the iterative MACS method and the MI-divided method will perform comparably well in detecting DIF, as both have performed well when compared to item response theory methods (Stark et al., 2005; Wanichtanom et al., 2003). If the noniterative DIF detection approaches are found to be efficacious in detecting DIF, this would support the use of desirable alternatives to the more laborious iterative MACS approach.

CHAPTER II

METHOD

Design

A simulation was conducted to compare the efficacy of four CFA methods of DIF detection for uniform and nonuniform DIF using a unidimensional 5-point Likert-type scale: iterative and noniterative MACS methods, the modification index (MI) method, and the modification index-divided sample (MI-divided) method. The three types of DIF are defined by (1) group differences on difficulty parameters only (uniform DIF), (2) group differences on the discrimination parameter only (nonuniform DIF), and (3) group differences on both the difficulty and discrimination parameters (nonuniform DIF). Two sample sizes of 250 and 500 simulated respondents were examined. Test length and number of DIF items were fixed at 12 and 3, respectively.

Data were simulated for 50 replications to compare reference and focal groups. Thus, the total number of data sets was 400: 50 data sets for each of the two sample sizes for each of the three types of DIF (or focal groups), and 50 data sets for each of the 2 sample sizes for the reference group (see Table 1).

The iterative and noniterative MACS methods, the MI method, and the MI-divided method each require 12, 1, 1, and 3 analytic programs, respectively. One of the 3 analytic programs for the MI-divided method is identical to the analytic program for the MI method, so for the present research, the MI-divided method requires 2 analytic programs. With 2 sample sizes, 3 types of DIF, and 50 replications per condition, the total number of data analytic programs was 4,800. See Table 2 for the number of unique analytic programs required per replication.

Table 1
Number of Data Sets Generated

Sample Size	Reference Group: No DIF	Focal Group: λ DIF	Focal Group: γ DIF	Focal Group: λ and γ DIF	Total
250	50	50	50	50	200
500	50	50	50	50	200
Total	100	100	100	100	400

Note. λ=loadings; γ=thresholds.

Table 2
Data Analyses: Number of Programs per Replication

Sample Size	Type of DIF	Method of Detection				Total
		MACS NI	MACS I	MI	MI-Div	
250	λ	1	12	1	2	16
	γ	1	12	1	2	16
	λ and γ	1	12	1	2	16
500	λ	1	12	1	2	16
	γ	1	12	1	2	16
	λ and γ	1	12	1	2	16
	Total	6	72	6	12	96

Note. λ=loadings; γ=thresholds. With 50 replications for each condition, the number of analytic programs for data analyses totaled 4,800. The MI-divided method required 2 rather than 3 analytic programs for the present research because 1 of the 3 sets of results required for analyses was supplied by the MI method.

Data Generation

Polytomous data for the simulation conditions described in Table 1 were generated with the use of the linear common factor model (Stark et al., 2005). The single common factor model can be expressed as:

$$x_{ij} = \lambda_i \xi_j + \beta_i \varepsilon_{ij} \quad (5)$$

In the equation, the subscript i reflects the item, j reflects the simulated respondent (simulee), ξ_j is the factor score for respondent j, λ_i is the loading of item i on the common factor ξ, ε_{ij} is the unique factor score for respondent j on item i, β_i is the loading of item i on the unique factor (given by $\sqrt{1-\lambda_i^2}$), and x_{ij} is a respondent's item propensity (i.e., tendency to endorse a particular response option).

Analysis of response data collected by Benn and Dickinson (2004) using the NEO-PI (Costa & McCrae, 1992) were used to set threshold values. Responses from 234 respondents on the Neuroticism subscale of the NEO-PI provided realistic values for the common factor loadings (λ_i), unique factor loadings (β_i), and item threshold values (γ_i). The Neuroticism subscale is a 12-item measure with 5 response options per item. The λs were obtained by principal axis factoring and the βs were calculated as $\sqrt{1-\lambda_i^2}$.

To generate polytomous data, four item thresholds were needed as boundaries for the five response options. To obtain the four item thresholds for each item in the measure, the cumulative proportion of endorsement of each item response category was computed and transformed to a standard normal metric. To compute 50 item propensity values for each simulee, common factor loadings and unique factor loadings were used in conjunction with common factor scores and unique factor scores that were sampled from normal distributions. For each simulee, each item propensity value was compared to the corresponding four item thresholds. If the propensity value exceeded the highest response category's threshold, the response was scored as a 5. If the propensity value was between the next highest response category's threshold and the highest response category's threshold, the response was scored as a 4. A similar pattern was followed to assign scores of 2 and 3 to propensity values. Propensity values that fell below the lowest response category's threshold were scored as a 1. Loadings (λs) and item thresholds (γs) for data generation can be found in Tables 3, 4, 5 and 6. Reference group parameters are presented in Table 3, and focal group parameters are presented in Tables 4, 5, and 6.

Table 3
Parameters for Reference Group Data Generation

		Reference Group: No DIF			
Item	λ	γ_1	γ_2	γ_3	γ_4
1	0.465	-1.681	-0.665	-0.454	0.705
2	0.555	-0.746	0.514	0.976	1.955
3	0.609	-1.563	-0.625	-0.197	0.789
4	0.740	-1.495	-0.143	0.122	1.433
5	0.660	-1.046	0.306	0.587	1.775
6	0.776	-0.538	0.419	0.705	1.775
7	0.600	-1.639	-0.208	0.143	1.463
8	0.548	-1.299	-0.016	0.419	1.274
9	0.570	-1.123	0.090	0.408	1.376
10	0.628	-1.495	-0.133	0.273	1.463
11	0.492	-0.454	0.665	1.274	1.888
12	0.608	-0.732	0.154	0.454	1.563

Note. λ=loading values; γ=threshold values. Item 6 was used as the referent item.

Table 4
Parameters for Focal Group Data Generation for DIF on Loadings

		Focal Group: λ DIF			
Item	λ	γ_1	γ_2	γ_3	γ_4
1	0.465	-1.681	-0.665	-0.454	0.705
2	0.555	-0.746	0.514	0.976	1.955
3	0.609	-1.563	-0.625	-0.197	0.789
4	**0.548**	-1.495	-0.143	0.122	1.433
5	0.660	-1.046	0.306	0.587	1.775
6	0.776	-0.538	0.419	0.705	1.775
7	**0.456**	-1.639	-0.208	0.143	1.463
8	0.548	-1.299	-0.016	0.419	1.274
9	0.570	-1.123	0.090	0.408	1.376
10	**0.474**	-1.495	-0.133	0.273	1.463
11	0.492	-0.454	0.665	1.274	1.888
12	0.608	-0.732	0.154	0.454	1.563

Note. λ=loading values; γ=threshold values. Item 6 was used as the referent item. Items with simulated DIF are in bold.

Table 5
Parameters for Focal Group Data Generation for DIF on Thresholds

		Focal Group: γ DIF			
Item	λ	γ_1	γ_2	γ_3	γ_4
1	0.465	-1.681	-0.665	-0.454	0.705
2	0.555	-0.746	0.514	0.976	1.955
3	0.609	-1.563	-0.625	-0.197	0.789
4	0.740	**-0.995**	**0.357**	**0.622**	**1.933**
5	0.660	-1.046	0.306	0.587	1.775
6	0.776	-0.538	0.419	0.705	1.775
7	0.600	**-1.139**	**0.292**	**0.643**	**1.963**
8	0.548	-1.299	-0.016	0.419	1.274
9	0.570	-1.123	0.090	0.408	1.376
10	0.628	**-0.995**	**0.367**	**0.773**	**1.963**
11	0.492	-0.454	0.665	1.274	1.888
12	0.608	-0.732	0.154	0.454	1.563

Note. λ=loading values; γ=threshold values. Item 6 was used as the referent item. Items with simulated DIF are in bold.

Table 6
Parameters for Focal Group Data Generation for DIF on Loadings and Thresholds

	Focal Group: λ and γ DIF				
Item	λ	γ_1	γ_2	γ_3	γ_4
1	0.465	-1.681	-0.665	-0.454	0.705
2	0.555	-0.746	0.514	0.976	1.955
3	0.609	-1.563	-0.625	-0.197	0.789
4	**0.563**	**-1.245**	**0.107**	**0.372**	**1.683**
5	0.660	-1.046	0.306	0.587	1.775
6	0.776	-0.538	0.419	0.705	1.775
7	**0.467**	**-1.389**	**0.042**	**0.393**	**1.713**
8	0.548	-1.299	-0.016	0.419	1.274
9	0.570	-1.123	0.090	0.408	1.376
10	**0.486**	**-1.245**	**0.117**	**0.523**	**1.713**
11	0.492	-0.454	0.665	1.274	1.888
12	0.608	-0.732	0.154	0.454	1.563

Note. λ=loading values; γ=threshold values. Item 6 was used as the referent item. Items with simulated DIF are in bold.

Using the method described above, 400 data sets were generated with PRELIS 2 (Jöreskog & Sörbom, 1996b) in total for the reference and focal groups (See Table 1). New values for the common factor scores and for the unique factor scores were obtained using a different seed for each replication. In addition, DIF was simulated for the focal group data on Items 4, 7, and 10 by shifting loadings, thresholds, or both loadings and thresholds by specific amounts. DIF item loadings and thresholds are indicated in bold in Tables 4, 5, and 6. There is a direct correspondence between factor analytic loadings (λ)

and thresholds (γ) and the IRT discrimination (a) and difficulty (b) parameters (Lord & Novick, 1968). Specifically:

$$a = \frac{\lambda}{\sqrt{1-\lambda^2}} \tag{6}$$

and

$$b = \frac{\gamma}{\lambda} \tag{7}$$

These formulas were used in conjunction with Raju's (1988) area formula to determine the specific values by which loadings and thresholds for DIF items were shifted to simulate a fixed DIF level of .50. Specifically, to create DIF due to difference in loadings (λs), focal group loadings were decreased by .192 for Item 4, by .144 for Item 7, and by .154 for Item 10. To create DIF due to difference in thresholds (γs), focal group thresholds were increased by .500 for all three items. To create DIF due to difference in both loadings and thresholds simultaneously, focal group loadings were decreased by .177 for Item 4, by .133 for Item 7, and by .142 for Item 10, and focal group thresholds were increased by .250 for all three items.

Population data (i.e., N=100,000) were simulated for reference and focal groups using the above parameters to describe the nature of observed scores. Table 7 provides the differences in population means and standard deviations between the reference and focal groups across the 3 different types of DIF. DIF on thresholds yielded the greatest difference in means and standard deviations followed by DIF on both thresholds and loadings. DIF on loadings alone yielded little difference in either means or standard deviations. These values illustrate that observed score differences in means and standard deviations are not particularly informative of DIF.

Table 7
The Effect of .05 Area Value of DIF on Population Means and Standard Deviations

		Means			Standard Deviations		
Type of DIF	Item	Ref. Mean	Focal Mean	Mean Diff.	Ref. Std. Dev.	Foc. Std. Dev.	Std. Dev. Diff.
	4	3.013	3.013	0	1.151	1.152	-0.001
λ	7	3.046	3.045	0.001	1.107	1.106	0.001
	10	2.952	2.952	0	1.119	1.120	-0.001
	4	3.013	2.494	0.519	1.151	1.098	0.053
γ	7	3.046	2.542	0.504	1.107	1.058	0.049
	10	2.952	2.445	0.507	1.119	1.053	0.066
	4	3.013	2.749	0.264	1.151	1.136	0.015
λ and γ	7	3.046	2.790	0.256	1.107	1.095	0.012
	10	2.952	2.693	0.259	1.119	1.097	0.022

Note. λ=loading values; γ=threshold values. Data is based on simulation of population data (i.e., N=100,000).

DIF Detection

Several analytic steps were required by each of the four CFA methods. For all methods, the LISREL 8.7 computer program (Jöreskog & Sörbom, 1996a) was used to perform analyses.

The steps for the iterative MACS method are as follows:

1. Estimate a "compact," baseline model. All items (Items 1-12) are entered into the model with the referent item's (Item 6) loadings and intercepts set invariant across groups. The CFA provides C(1), a chi-square goodness-of-fit value.

2. Estimate an "augmented" model. One item in addition to the referent item has its loadings and intercepts set invariant across groups (e.g., Item 1). The CFA provides C(2), a chi-square goodness-of-fit value.
3. Compute the difference C(2)-C(1). This difference approximately follows a chi-square distribution with 2 degrees of freedom.
4. If the difference exceeds the critical value of a chi-square distribution with 2 degrees of freedom at $p<.05$ (i.e., 5.99), the item is detected as showing DIF.
5. Repeat steps 1 to 4 for the remaining 10 items.

The steps for the noniterative MACS method are as follows:

1. All items (Items 1-12) are entered into the model and each item's loading and intercept are set to be invariant across groups. The CFA provides a modification index for the loading and intercept for each group (reference and focal) for each item resulting in 4 modification indexes of interest for each item. Each modification index is a statistic that approximates a chi-square distribution with one degree of freedom.
2. If one of the four modification indexes of interest for an item exceeds the critical value of a chi-square distribution with 1 degree of freedom (i.e., 3.84), the item is detected as showing DIF. Because Item 6 served as a referent item in the iterative MACS method analyses, it is included in the model but is not examined for DIF in the noniterative MACS method analyses. This maintains an equal number of items that are examined for DIF between methods.

The steps for the MI method are as follows:

1. All items (Items 1-12) are included in the model. All of the items load on the ability factor. Group membership is specified in the hypothesized model as a latent variable that has no measurement error, and only the group membership indicator variable loads on the group membership factor. The CFA provides a modification index for each item on the group membership variable. Each modification index is a statistic that approximates a chi-square distribution with 1 degree of freedom.
2. If an item has a modification index that exceeds the critical value of a chi-square distribution with 1 degree of freedom (i.e., 3.84), the item is detected as showing DIF. Because Item 6 served as a referent item in the iterative MACS method analyses, it is included in the model but is not examined for DIF in the MI method analyses. This maintains an equal number of items that are examined for DIF between methods.

The steps for the MI-Divided method are as follows:

1. The steps of the MI method are repeated for three groups of respondents: (1) total group, (2) high scoring group (i.e., total score was equal to or greater than the median), and (3) low scoring group (i.e., total score was less than the median).
2. If an item has a modification index that exceeds the critical value of a chi-square distribution with 1 degree of freedom (i.e., 3.84) for one or more of the three groups, the item is detected as showing DIF. Because Item 6 served as a referent item in the previous analyses, it is included in the model but is not

examined for DIF in the MI-divided method analyses. This maintains an equal number of items that are examined for DIF between methods.

See Table 8 for a summary of the procedures for identifying items exhibiting DIF. See Table 9 for a summary of the simulation design and data analyses.

Table 8
Identifying Items Exhibiting Differential Item Functioning

Method of DIF Detection	Parameters	Examine	Item Exhibits DIF if:
MACS NI	1 MI-λ 1 MI-τ 2 Groups: Foc. & Ref.	4 MIs for each item: 1 MI for loading and 1 MI for intercept for each group (reference and focal)	1 or more of the 4 MIs exceeds the critical value
MACS I	1 χ^2 baseline model 1 χ^2 augmented model	Chi-square difference value between augmented model and baseline model	Chi-square difference exceeds the critical value
MI	1 MI-λ	MI for each item on the group membership latent variable	MI exceeds the critical value
MI-divided	1 MI-λ Total 1 MI-λ High Score 1 MI-λ Low Score	MI for each item on the group membership latent variable for: (1) total group, (2) high scoring group, and (3) low scoring group	1 or more of the 3 MIs exceeds the critical value

Table 9
Simulation Design and Data Analyses

	Simulation Design			Data Analyses			
	N	Type of DIF Simulated	Simulated Datasets per DIF Condition	Method of DIF Detection	Analytic Programs per Replication	Total Analytic Programs	
Count	2	3	50	4	1-12		
Desc.	250	λ		MACS NI	1	2x3x50x1	=300
	500	γ		MACS I	12	2x3x50x12	=3600
		λ and γ		MI	1	2x3x50x1	=300
				MI-Div	2	2x3x50x2	=600
Total						300+3600+300+600=4800	

Note. λ=loading; γ=threshold. The MI-divided method required 2 rather than 3 analytic programs for the present research because 1 of the 3 sets of results required for analyses was supplied by the MI method.

CHAPTER III

RESULTS

Overview

Tables 10 through 12 summarize the results for the simulation study. True positive and false positive rates for the detection of DIF by the methods are provided. True positive rates represent the proportion of DIF items that were correctly identified as exhibiting DIF across the 50 replications in each condition. False positive rates represent the proportion of times an item designed not to have DIF was incorrectly identified as exhibiting DIF.

Please note that in examining DIF on loadings (λs), 1 of the 50 analytic programs did not converge to yield results for the MI-divided method applied to the low-scoring group for sample size 250. Therefore, true positive and true negative rates for the MI-divided method's detection of DIF on loadings were based on a total of 49 rather than 50 replications. In addition, in examining DIF on loadings and thresholds (λs and γs), 3 of the 50 analytic programs did not converge to yield results for the MI-divided method applied to the low-scoring group for sample size 250. Therefore, true positive and true negative rates for the MI-divided method's detection of DIF on loadings and thresholds were based on a total of 47 rather than 50 replications.

The detection rates of each method for each type of DIF are now described.

Detection Rates for DIF on Loadings

The three items that show discrimination DIF for the focal group have loading (λ) values that are a different magnitude than those of the reference group. The difficulty or threshold (γ) values for the focal group, however, are identical to those of the reference group (see Table 4). The loading and threshold values for the three items achieve the same amount of DIF with regard to the area (i.e., .5) between the ICCs of the reference and focal groups. These three items exhibit DIF that is nonuniform in nature.

Table 10 displays the detection rates for DIF on loadings by the four methods across the two sample sizes. For a sample size of 250, detection of DIF was poor. The noniterative MACS method fared the best with a true positive rate of .16. The iterative MACS and MI-divided methods had true positive rates of .05 and .01, respectively. The MI method did not detect any items with DIF on loadings. All methods did not detect any false positives.

Table 10
True Positive and False Positive Rates of Detection of DIF on Loadings (λs)

		Type of DIF: λ	
N	Method	True Positive	False Positive
250	MACS NI	0.16	0.00
	MACS I	0.05	0.00
	MI	0.00	0.00
	MI-Div	0.01	0.00
500	MACS NI	0.60	0.00
	MACS I	0.55	0.00
	MI	0.00	0.00
	MI-Div	0.13	0.00

Note. True positive and false positive rates for the MI-divided method for sample size 250 was based on a total of 49 rather than 50 replications due to lack of convergence on one of the analytic programs.

For a sample size of 500, detection of DIF improved for all methods except for the MI method. With the increased sample size, the noniterative MACS and iterative MACS methods' true positive rates increased considerably with true positive rates of .60 and .55, respectively. The MI-divided true positive rate increased but was still poor (.13). However, the MI method still did not detect any items with DIF on loadings with the larger sample size. All methods did not detect any false positives.

It is evident that increasing the sample size from 250 to 500 improved the accuracy of DIF detection for the MACS methods while still holding false positives to a minimum. The MI-divided method showed some improvement in DIF detection, but the MI method did not improve to any degree.

Detection Rates for DIF on Thresholds

The three items that show difficulty DIF for the focal group have threshold (γ) values that are a different magnitude than those of the reference group. The discrimination or loading (λ) values for the focal group, however, are identical to those of the reference group (see Table 5). The loading and threshold values for the three items achieve the same amount of DIF with regard to the area (i.e., .5) between the ICCs of the reference and focal groups. These three items exhibit DIF that is uniform in nature.

Table 11 displays the detection rates for DIF on thresholds by the four methods across the two sample sizes. For a sample size of 250, all four methods exhibited sensitivity to DIF on thresholds with a true positive rate of 1.00. In addition, all four methods held false positive rates to an acceptable amount (i.e., below .05). The noniterative MACS, MI, and MI-divided methods had false positive rates of .02, .02, and .03, respectively. The iterative MACS method detected no false positives.

Table 11
True Positive and False Positive Rates of Detection of DIF on Thresholds (γs)

N	Method	Type of DIF: γ True Positive	False Positive
250	MACS NI	1.00	0.02
	MACS I	1.00	0.00
	MI	1.00	0.02
	MI-Div	1.00	0.03
500	MACS NI	1.00	0.44
	MACS I	1.00	0.00
	MI	1.00	0.45
	MI-Div	1.00	0.46

For a sample size of 500, all four methods again exhibited true positive rates of 1.00. However, the false positive rates for three of the methods increased considerably with the greater sample size. The noniterative MACS, the MI, and the MI-divided methods had false positive rates of .44, .45, and.46, respectively. The iterative MACS method again detected no false positives.

All four methods, regardless of sample size, performed well in detecting DIF on thresholds. With an increase in sample size from 250 to 500, however, all methods with the exception of the iterative MACS method had false positive rates that were well above what is expected by chance.

Detection Rates for DIF on Loadings and Thresholds

The three items that show discrimination and difficulty DIF for the focal group have both loading (λ) and threshold (γ) values that are different in magnitude than those of the reference group (see Table 6). The loading and threshold values for the three items achieve the same amount of DIF with regard to the area (i.e., .5) between the ICCs of the reference and focal groups. These three items exhibit DIF that is nonuniform in nature.

Table 12 displays the detection rates for DIF on both loadings and thresholds by the four methods across the two sample sizes. For a sample size of 250, all four methods were successful in detecting DIF on both loadings and thresholds. The noniterative MACS, the iterative MACS, the MI, and the MI-divided methods had true positive rates of .93, .99, .95, and .97, respectively. All four methods did not detect any false positives.

Table 12
True Positive and False Positive Rates of Detection of DIF on Loadings (λs) and Thresholds (γs)

		Type of DIF: λ and γ	
N	Method	True Positive	False Positive
250	MACS NI	0.93	0.00
	MACS I	0.99	0.00
	MI	0.95	0.00
	MI-Div	0.97	0.00
500	MACS NI	1.00	0.00
	MACS I	1.00	0.00
	MI	1.00	0.00
	MI-Div	1.00	0.00

Note. True positive and false positive rates for the MI-divided method for sample size 250 was based on a total of 47 rather than 50 replications due to lack of convergence on three of the analytic programs.

For a sample size of 500, all four methods were again successful in detecting DIF on both loadings and thresholds with all four methods having a true positive rate of 1.00. All four methods also did not detect any false positives.

All four methods, regardless of sample size, performed well in detecting DIF on both loadings and thresholds. With an increase in sample size from 250 to 500, all four methods held false positive rates to a minimum.

CHAPTER IV

CONCLUSIONS

General Findings

The present research compared four different CFA methods for detecting DIF: the iterative MACS method, the noniterative MACS method, the MI method, and the MI-divided method.

Based on previous research, it was expected that the MI-divided method would perform better than the MI method in detecting nonuniform DIF. The results of the study indicate that at a sample size of 500, the MI-divided method was, indeed, more sensitive than the MI method to DIF on loadings with true positive rates of .13 and 0, respectively. However, both methods did not perform well in detecting DIF on loadings. At a sample size of 250, detection rates for both methods were also poor (.01 and 0, respectively).

For DIF on both loadings and thresholds, however, the MI method had very similar results to the MI-divided method for a sample size of 250 with true positive rates of .95 and .97, respectively. For a sample size of 500, both methods had a detection rate of 1.00. Neither method detected false positives for either sample size. In contrast to the present study, Wanichtanom et al. (2003) found that the MI-divided method was more sensitive to detecting DIF on both discrimination (loadings) and difficulty (thresholds) than the MI method. Although both the present study and Wanichtanom et al.'s (2003) study simulated the same amount of DIF (.5 area between ICCs), Wanichtanom et al.'s (2003) study examined detection of DIF in dichotomous data. The present study examined detection of DIF in polytomous data. The results indicate that the MI and MI-divided methods may both perform well in detecting DIF on both loadings and thresholds in polytomous data, whereas the MI-divided method performs better than the MI method at detecting DIF in dichotomous data.

With regard to uniform DIF (DIF on γ values), both the MI and MI-Divided methods performed well with true positive rates of 1.00 for both sample sizes. The false positive rates were low for a sample size of 250 (.02 and .03, respectively), but false positive rates for a sample size of 500 were far above expected levels at .45 and .46 respectively. In contrast, Wanichtanom et al. (2003) found false positive rates for the MI and MI-divided methods that were smaller than those of the present study (.11 and .18, respectively). The results indicate that the MI and MI-divided methods may both have greater false positive rates in detecting DIF on thresholds for polytomous data than for dichotomous data.

Based on previous research, it was also expected that the iterative MACS method and the MI-divided method would perform comparably well as they have both performed well when compared to IRT methods in previous research (Stark et al, 2005; Wanichtanom et al., 2003). For DIF on loadings, this was not the case. For a sample size of 500, the iterative MACS method had a true positive rate of .55, whereas the MI-divided method had a true positive rate of .13. For DIF on thresholds, both methods detected DIF equally well with true positive rates of 1.00 for both sample sizes of 250 and 500. However, for a sample size of 500, the iterative MACS method detected no false positives, whereas the MI-divided method had a false positive rate of .46. For a sample size of 250, false positive rates were held at acceptable levels by both methods (0 and .03). With regard to DIF on both loadings and thresholds, both methods performed well for both sample sizes in detecting DIF and held false positive rates to a minimum.

Overall, the results indicate that for a lower sample size of 250, the iterative MACS method and the MI-divided method produce comparable results in DIF detection. However, at a higher sample size of 500, the iterative MACS method is more sensitive to DIF on loadings and produces fewer false positives for DIF on thresholds than the MI-divided method. Both methods perform well for DIF on both loadings and thresholds.

The positive results for the iterative MACS method replicates previous research by Stark et al. (2005) for DIF detection for polytomous data. Stark et al. (2005) had added or subtracted constant values to item parameters to define the different types of DIF, and, as a result, there is a confounding of type of DIF with the magnitude of DIF in that research. In contrast, in the present study, Raju's (1988) area formula was used to maintain a constant amount of DIF of .5. The present research provides support for the efficacy of the iterative MACS method with results that do not confound type of DIF with magnitude of DIF.

The present study introduced the noniterative MACS method as a possible alternative to the more laborious iterative MACS method. For a smaller sample size of 250, both methods had similar results. For DIF on loadings, the noniterative MACS method was more sensitive to DIF than the iterative MACS method with true positive rates of .16 and .05, respectively. Both methods detected DIF on thresholds equally well with true positive rates of 1.00. Both methods were also sensitive to DIF on both loadings and thresholds with a true positive rate of .93 for the noniterative MACS method and .99 for the iterative MACS method. False positive rates were held at expected levels for both methods across all types of DIF. Although, both methods were not sensitive to DIF on loadings, if circumstances require use of DIF detection for a smaller sample size such as 250, the noniterative MACS method would be the more desirable alternative as it provides similar or better results than the more laborious iterative MACS method.

For a sample size of 500, the noniterative and iterative MACS methods had similar results. For DIF on loadings, the noniterative and iterative MACS methods had true positive rates of .60 and .55, respectively. False positive rates for DIF on loadings were held at expected levels for both methods. For DIF on both loadings and thresholds, both methods had a true positive rate of 1.00. False positive rates for DIF on both loadings and thresholds were held at expected levels for both methods. For detection of DIF on thresholds, both methods had true positive rates of 1.00. However, the noniterative MACS method had a false positive rate much greater than expected (.44), whereas the iterative MACS method detected no false positives. With similar DIF detection rates, the noniterative MACS method is a very promising alternative to the more laborious iterative MACS method. However, its high false positive rate for detecting DIF on thresholds for larger sample sizes is a weakness. Large false positive rates are undesirable, as valuable items would be discarded through misidentification.

Future Research

The present research did not examine a number of variables that occur in real situations. In the present study, several variables were fixed that have an impact on detection rates. Variables such as proportion of DIF items and level of DIF were set at specific levels that previous research indicated would be optimal for DIF detection rather than to reflect what may occur in real situations. For example, the level of DIF was fixed at .5, a moderate amount of DIF deemed reasonable for comparison of the DIF detection methods. An operational measure may contain test items with varying amounts of DIF with some items exhibiting weaker DIF (e.g., .2) and some items exhibiting stronger DIF (e.g., .8). Future simulation research could examine tests containing items with varying amounts of DIF.

Another limitation of the study is the use of a referent item in the iterative MACS method that is known not to be a DIF item. Future research could examine through a simulation study the effect on DIF detection when using a DIF item as a referent item.

The present study did not find the MI-divided method to be as effective as expected based on previous research (Wanichtanom et al., 2003). As mentioned previously, this could be due to the present study examining polytomous data rather than dichotomous data. In addition, Wanichtanom et al.'s study used a different data simulation method and examined a different sample size than the present research. Data were simulated using the 2 parameter logistic model, and a sample size of 1,000 was examined. Future research can use a simulation study to directly compare the efficacy of the MI-divided method in detecting DIF for polytomous data versus dichotomous data. In addition, research should compare results from the two simulation methods for the same sample size to sort out potential confounding of method and sample size.

Finally, the noniterative MACS method appears to be a promising alternative to the more laborious iterative MACS method, but it has the disadvantage of higher false positive rates for detecting DIF on thresholds with increased sample size. Future research could identify strategies to reduce the false positive rate for increasing sample sizes. Stark et al. (2005) found that using a Bonferroni corrected critical value for samples of

1,000 or less is too conservative for the iterative MACS method, but perhaps using an adjusted critical value for detecting DIF on thresholds could prove beneficial for controlling false positive rates for the noniterative MACS method. Given the results of this research, however, the iterative MACS method would be a good option for detecting DIF in applied situations if an adequate sample size is available.

REFERENCES

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Benn, N. A., & Dickinson, T.L. (2004). *The influence of item wording on questionnaire scale scores.* Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, April 2-4, Chicago.

Camilli, B., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Chan, D. (2000). Detection of differential item functioning on the Kirton Adaption-Innovation Inventory using multiple-group mean and covariance structure analyses. *Multivariate Behavioral Research, 35*, 169-199.

Cohen, A. S., & Kim, S. H. (1993). A comparison of Lord's χ^2 and Raju's area measures in detection of DIF. *Applied Psychological Measurement, 17,* 39-52.

Cohen, A. S., Kim, S. H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17,* 335-350.

Costa, P. T. Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory(NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional manual.* Odessa, FL: Psychological Assessment Resources.

Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology*. Palo Alto, CA: Consulting Psychologists Press.

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

Ellis, B. B. (1989). Differential item functioning: Implications for test translation. *Journal of Applied Psychology, 74,* 912-921.

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology, 77,* 177-184.

Everson, H. T., Millsap, R. E. & Rodriguez, C. M (1991). Isolating gender differences in test anxiety: A confirmatory factor analysis of the Test Anxiety Inventory. *Educational and Psychological Measurement, 51,* 243-251.

Ferrando, P. J. (1996). Calibration of invariant item parameters in a continuous item response model using the extended LISREL measurement submodel. *Multivariate Behavioral Research, 31,* 419-439.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Norwell, MA: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Thousand Oaks, CA: Sage.

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: Differential item functioning in the NEO personality inventory. *Journal of Cross-Cultural Psychology, 28,* 192-218.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications of psychological measurement.* Homewood, IL: Dow Jones-Irwin.

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

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

Kim, S., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29,* 51-66.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Maurer, T. J., Raju, N. S., & Collins, W. C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology, 83,* 693-702.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54,* 284-291.

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

Meade, A. W., & Laugtenschlager, G. J. (April, 2004b). *Same question, different answers: CFA and two IRT approaches to measurement invariance*. Symposium presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago.

Meade, A. W., & Lautenschlager, G. J. (2005). *Sensitivity of DFIT Tests of Measurement Invariance for Likert data*. Paper presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20,* 257-274.

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika, 6,* 150-166.

Oort, F. J. (1996). *Using restricted factor analysis in test construction.* Unpublished doctoral dissertation, University of Amsterdam, Amsterdam.

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

Oshima, T. C., & Miller, D. M. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement, 16,* 237-248.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495-502.

Raju, N. S., & Ellis, B. B. (2002). Differential item and test functioning. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations.* San Franciso: Jossey-Bass.

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

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting DIF. *Applied Psychological Measurement, 17,* 105-116.

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

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). *Detecting differential item functioning with CFA and IRT: Toward a unified strategy.* Symposium presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles.

Suen, H. K. (1990). *Principles of test theories.* Hillsdale: Lawrence Erlbaum.

Swaminathan, H., & Rogers, H. J. (1990). Detecing DIF using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Wanichtanom, R., Dickinson, T. L., & Coates, G. D. (2003). *Differential item functioning: A comparison of item response theory and confirmatory factor analysis.* Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando.

Whiteside-Mansell, L., & Corwyn, R. F. (2003). Mean and covariance structures analysis: An examination of the Rosenberg Self-Esteem Scale among adolescents and adults. *Educational and Psychological Measurement, 63,* 163-173.

APPENDIX A

SAMPLE PROGRAMS FOR DATA GENERATION

```
TITLE Generation Program: Reference Group for Sample Size of 500
DA NO=500
CO ALL
NE V1=NRAND
NE V2=NRAND
NE V3=NRAND
NE V4=NRAND
NE V5=NRAND
NE V6=NRAND
NE V7=NRAND
NE V8=NRAND
NE V9=NRAND
NE V10=NRAND
NE V11=NRAND
NE V12=NRAND
NE V13=NRAND
NE ABILITY = V1
NE I1 = .4650*ABILITY + .8853*V2
NE I2 = .5550*ABILITY + .8319*V3
NE I3 = .6090*ABILITY + .7932*V4
NE I4 = .7400*ABILITY + .6726*V5
NE I5 = .6600*ABILITY + .7513*V6
NE I6 = .7760*ABILITY + .6307*V7
NE I7 = .6000*ABILITY + .8000*V8
NE I8 = .5480*ABILITY + .8365*V9
NE I9 = .5700*ABILITY + .8216*V10
NE I10 = .6280*ABILITY + .7782*V11
NE I11 = .4920*ABILITY + .8706*V12
NE I12 = .6080*ABILITY + .7939*V13
RE I1 OLD = -9.00--1.681,-1.681--0.665,-0.665--0.454,-0.454- 0.705, 0.705- 9.00 NEW
= 1,2,3,4,5
RE I2 OLD = -9.00--0.746,-0.746- 0.514, 0.514- 0.976, 0.976- 1.955, 1.955- 9.00 NEW
= 1,2,3,4,5
RE I3 OLD = -9.00--1.563,-1.563--0.625,-0.625--0.197,-0.197- 0.789, 0.789- 9.00 NEW
= 1,2,3,4,5
RE I4 OLD = -9.00--1.495,-1.495--0.143,-0.143- 0.122, 0.122- 1.433, 1.433- 9.00 NEW
= 1,2,3,4,5
RE I5 OLD = -9.00--1.046,-1.046- 0.306, 0.306- 0.587, 0.587- 1.775, 1.775- 9.00 NEW
= 1,2,3,4,5
RE I6 OLD = -9.00--0.538,-0.538- 0.419, 0.419- 0.705, 0.705- 1.775, 1.775- 9.00 NEW
= 1,2,3,4,5
RE I7 OLD = -9.00--1.639,-1.639--0.208,-0.208- 0.143, 0.143- 1.463, 1.463- 9.00 NEW
= 1,2,3,4,5
RE I8 OLD = -9.00--1.299,-1.299--0.016,-0.016- 0.419, 0.419- 1.274, 1.274- 9.00 NEW
= 1,2,3,4,5
RE I9 OLD = -9.00--1.123,-1.123- 0.090, 0.090- 0.408, 0.408- 1.376, 1.376- 9.00 NEW
```

= 1,2,3,4,5
RE I10 OLD = -9.00--1.495,-1.495--0.133,-0.133- 0.273, 0.273- 1.463, 1.463- 9.00 NEW = 1,2,3,4,5
RE I11 OLD = -9.00--0.454,-0.454- 0.665, 0.665- 1.274, 1.274- 1.888, 1.888- 9.00 NEW = 1,2,3,4,5
RE I12 OLD = -9.00--0.732,-0.732- 0.154, 0.154- 0.454, 0.454- 1.563, 1.563- 9.00 NEW = 1,2,3,4,5
NE GP = 1
SD V1 - V13 ABILITY
OU RA=Ref500.DAT IX=123456

```
TITLE Generation Program: a DIF for Focal Group for Sample Size of 500
DA NO=500
CO ALL
NE V1=NRAND
NE V2=NRAND
NE V3=NRAND
NE V4=NRAND
NE V5=NRAND
NE V6=NRAND
NE V7=NRAND
NE V8=NRAND
NE V9=NRAND
NE V10=NRAND
NE V11=NRAND
NE V12=NRAND
NE V13=NRAND
NE ABILITY = V1
NE I1 = .4650*ABILITY + .8853*V2
NE I2 = .5550*ABILITY + .8319*V3
NE I3 = .6090*ABILITY + .7932*V4
NE I4 = .5480*ABILITY + .8365*V5
NE I5 = .6600*ABILITY + .7513*V6
NE I6 = .7760*ABILITY + .6307*V7
NE I7 = .4560*ABILITY + .8900*V8
NE I8 = .5480*ABILITY + .8365*V9
NE I9 = .5700*ABILITY + .8216*V10
NE I10 = .4740*ABILITY + .8805*V11
NE I11 = .4920*ABILITY + .8706*V12
NE I12 = .6080*ABILITY + .7939*V13
RE I1 OLD = -9.00--1.681,-1.681--0.665,-0.665--0.454,-0.454- 0.705, 0.705- 9.00 NEW
= 1,2,3,4,5
RE I2 OLD = -9.00--0.746,-0.746- 0.514, 0.514- 0.976, 0.976- 1.955, 1.955- 9.00 NEW
= 1,2,3,4,5
RE I3 OLD = -9.00--1.563,-1.563--0.625,-0.625--0.197,-0.197- 0.789, 0.789- 9.00 NEW
= 1,2,3,4,5
RE I4 OLD = -9.00--1.495,-1.495--0.143,-0.143- 0.122, 0.122- 1.433, 1.433- 9.00 NEW
= 1,2,3,4,5
RE I5 OLD = -9.00--1.046,-1.046- 0.306, 0.306- 0.587, 0.587- 1.775, 1.775- 9.00 NEW
= 1,2,3,4,5
RE I6 OLD = -9.00--0.538,-0.538- 0.419, 0.419- 0.705, 0.705- 1.775, 1.775- 9.00 NEW
= 1,2,3,4,5
RE I7 OLD = -9.00--1.639,-1.639--0.208,-0.208- 0.143, 0.143- 1.463, 1.463- 9.00 NEW
= 1,2,3,4,5
RE I8 OLD = -9.00--1.299,-1.299--0.016,-0.016- 0.419, 0.419- 1.274, 1.274- 9.00 NEW
= 1,2,3,4,5
RE I9 OLD = -9.00--1.123,-1.123- 0.090, 0.090- 0.408, 0.408- 1.376, 1.376- 9.00 NEW
```

```
= 1,2,3,4,5
RE I10 OLD = -9.00--1.495,-1.495--0.133,-0.133- 0.273, 0.273- 1.463, 1.463- 9.00 NEW
= 1,2,3,4,5
RE I11 OLD = -9.00--0.454,-0.454- 0.665, 0.665- 1.274, 1.274- 1.888, 1.888- 9.00 NEW
= 1,2,3,4,5
RE I12 OLD = -9.00--0.732,-0.732- 0.154, 0.154- 0.454, 0.454- 1.563, 1.563- 9.00 NEW
= 1,2,3,4,5
NE GP = 0
SD V1 - V13 ABILITY
OU RA=aDIF500.DAT IX=123456
```

```
TITLE Generation Program: b DIF for Focal Group for Sample Size of 500
DA NO=500
CO ALL
NE V1=NRAND
NE V2=NRAND
NE V3=NRAND
NE V4=NRAND
NE V5=NRAND
NE V6=NRAND
NE V7=NRAND
NE V8=NRAND
NE V9=NRAND
NE V10=NRAND
NE V11=NRAND
NE V12=NRAND
NE V13=NRAND
NE ABILITY = V1
NE I1 = .4650*ABILITY + .8853*V2
NE I2 = .5550*ABILITY + .8319*V3
NE I3 = .6090*ABILITY + .7932*V4
NE I4 = .7400*ABILITY + .6726*V5
NE I5 = .6600*ABILITY + .7513*V6
NE I6 = .7760*ABILITY + .6307*V7
NE I7 = .6000*ABILITY + .8000*V8
NE I8 = .5480*ABILITY + .8365*V9
NE I9 = .5700*ABILITY + .8216*V10
NE I10 = .6280*ABILITY + .7782*V11
NE I11 = .4920*ABILITY + .8706*V12
NE I12 = .6080*ABILITY + .7939*V13
RE I1 OLD = -9.00--1.681,-1.681--0.665,-0.665--0.454,-0.454- 0.705, 0.705- 9.00 NEW
= 1,2,3,4,5
RE I2 OLD = -9.00--0.746,-0.746- 0.514, 0.514- 0.976, 0.976- 1.955, 1.955- 9.00 NEW
= 1,2,3,4,5
RE I3 OLD = -9.00--1.563,-1.563--0.625,-0.625--0.197,-0.197- 0.789, 0.789- 9.00 NEW
= 1,2,3,4,5
RE I4 OLD = -9.00--0.995,-0.995- 0.357, 0.357- 0.622, 0.622- 1.933, 1.933- 9.00 NEW
= 1,2,3,4,5
RE I5 OLD = -9.00--1.046,-1.046- 0.306, 0.306- 0.587, 0.587- 1.775, 1.775- 9.00 NEW
= 1,2,3,4,5
RE I6 OLD = -9.00--0.538,-0.538- 0.419, 0.419- 0.705, 0.705- 1.775, 1.775- 9.00 NEW
= 1,2,3,4,5
RE I7 OLD = -9.00--1.139,-1.139- 0.292, 0.292- 0.643, 0.643- 1.963, 1.963- 9.00 NEW
= 1,2,3,4,5
RE I8 OLD = -9.00--1.299,-1.299--0.016,-0.016- 0.419, 0.419- 1.274, 1.274- 9.00 NEW
= 1,2,3,4,5
RE I9 OLD = -9.00--1.123,-1.123- 0.090, 0.090- 0.408, 0.408- 1.376, 1.376- 9.00 NEW
```

```
= 1,2,3,4,5
RE I10 OLD = -9.00--0.995,-0.995- 0.367, 0.367- 0.773, 0.773- 1.963, 1.963- 9.00 NEW
= 1,2,3,4,5
RE I11 OLD = -9.00--0.454,-0.454- 0.665, 0.665- 1.274, 1.274- 1.888, 1.888- 9.00 NEW
= 1,2,3,4,5
RE I12 OLD = -9.00--0.732,-0.732- 0.154, 0.154- 0.454, 0.454- 1.563, 1.563- 9.00 NEW
= 1,2,3,4,5
NE GP = 1
SD V1 - V13 ABILITY
OU RA=bDIF500.DAT IX=123456
```

```
TITLE Generation Program: ab DIF for Focal Group for Sample Size of 500
DA NO=500
CO ALL
NE V1=NRAND
NE V2=NRAND
NE V3=NRAND
NE V4=NRAND
NE V5=NRAND
NE V6=NRAND
NE V7=NRAND
NE V8=NRAND
NE V9=NRAND
NE V10=NRAND
NE V11=NRAND
NE V12=NRAND
NE V13=NRAND
NE ABILITY  = V1
NE I1  = .4650*ABILITY + .8853*V2
NE I2  = .5550*ABILITY + .8319*V3
NE I3  = .6090*ABILITY + .7932*V4
NE I4  = .5630*ABILITY + .8265*V5
NE I5  = .6600*ABILITY + .7513*V6
NE I6  = .7760*ABILITY + .6307*V7
NE I7  = .4670*ABILITY + .8843*V8
NE I8  = .5480*ABILITY + .8365*V9
NE I9  = .5700*ABILITY + .8216*V10
NE I10 = .4860*ABILITY + .8740*V11
NE I11 = .4920*ABILITY + .8706*V12
NE I12 = .6080*ABILITY + .7939*V13
RE I1 OLD  = -9.00--1.681,-1.681--0.665,-0.665--0.454,-0.454- 0.705, 0.705- 9.00 NEW
= 1,2,3,4,5
RE I2 OLD  = -9.00--0.746,-0.746- 0.514, 0.514- 0.976, 0.976- 1.955, 1.955- 9.00 NEW
= 1,2,3,4,5
RE I3 OLD  = -9.00--1.563,-1.563--0.625,-0.625--0.197,-0.197- 0.789, 0.789- 9.00 NEW
= 1,2,3,4,5
RE I4 OLD  = -9.00--1.245,-1.245- 0.107, 0.107- 0.372, 0.372- 1.683, 1.683- 9.00 NEW
= 1,2,3,4,5
RE I5 OLD  = -9.00--1.046,-1.046- 0.306, 0.306- 0.587, 0.587- 1.775, 1.775- 9.00 NEW
= 1,2,3,4,5
RE I6 OLD  = -9.00--0.538,-0.538- 0.419, 0.419- 0.705, 0.705- 1.775, 1.775- 9.00 NEW
= 1,2,3,4,5
RE I7 OLD  = -9.00--1.389,-1.389- 0.042, 0.042- 0.393, 0.393- 1.713, 1.713- 9.00 NEW
= 1,2,3,4,5
RE I8 OLD  = -9.00--1.299,-1.299--0.016,-0.016- 0.419, 0.419- 1.274, 1.274- 9.00 NEW
= 1,2,3,4,5
RE I9 OLD  = -9.00--1.123,-1.123- 0.090, 0.090- 0.408, 0.408- 1.376, 1.376- 9.00 NEW
```

= 1,2,3,4,5
RE I10 OLD = -9.00--1.245,-1.245- 0.117, 0.117- 0.523, 0.523- 1.713, 1.713- 9.00 NEW
= 1,2,3,4,5
RE I11 OLD = -9.00--0.454,-0.454- 0.665, 0.665- 1.274, 1.274- 1.888, 1.888- 9.00 NEW
= 1,2,3,4,5
RE I12 OLD = -9.00--0.732,-0.732- 0.154, 0.154- 0.454, 0.454- 1.563, 1.563- 9.00 NEW
= 1,2,3,4,5
NE GP = 0
SD V1 - V13 ABILITY
OU RA=abDIF500.DAT IX=123456

APPENDIX B

SAMPLE PROGRAMS FOR MEAN AND COVARIANCE STRUCTURE ANALYSIS METHOD: ITERATIVE AND NONITERATIVE PROCEDURES

```
TITLE Iterative MACS Procedure for Sample Size of 500:  a DIF Baseline Model
!REFERENCE GROUP
DA NI=13 NG=2 NO=500
RA=Ref500.DAT
LA
I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 GP
SE
I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 /
MO NX=12 NK=1 TX=FR KA=FI
LK
ABILITY
VA 1.0 LX 6 1
FR LX 1 1 LX 2 1 LX 3 1 LX 4 1 LX 5 1 LX 7 1 LX 8 1
FR LX 9 1 LX 10 1 LX 11 1 LX 12 1
FR TX 1 TX 2 TX 3 TX 4 TX 5 TX 6 TX 7 TX 8
FR TX 9 TX 10 TX 11 TX 12
OU MI AD=OFF IT=300

Mean and Covariance Structure Model
!FOCAL GROUP
DA NI=13 NO=500
RA=aDIF500.DAT
LA
I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 GP
SE
I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 /
MO NX=12 NK=1 TX=FR KA=FI
LK
ABILITY
VA 1.0 LX 6 1
FR LX 1 1 LX 2 1 LX 3 1 LX 4 1 LX 5 1 LX 7 1 LX 8 1
FR LX 9 1 LX 10 1 LX 11 1 LX 12 1
FR TX 1 TX 2 TX 3 TX 4 TX 5 TX 6 TX 7 TX 8
FR TX 9 TX 10 TX 11 TX 12
!EQ LX 1 1 1 LX 2 1 1
!EQ LX 1 2 1 LX 2 2 1
!EQ LX 1 3 1 LX 2 3 1
!EQ LX 1 4 1 LX 2 4 1
!EQ LX 1 5 1 LX 2 5 1
EQ LX 1 6 1 LX 2 6 1
!EQ LX 1 7 1 LX 2 7 1
!EQ LX 1 8 1 LX 2 8 1
!EQ LX 1 9 1 LX 2 9 1
!EQ LX 1 10 1 LX 2 10 1
!EQ LX 1 11 1 LX 2 11 1
!EQ LX 1 12 1 LX 2 12 1
```

```
!EQ TX 1 1 TX 2 1
!EQ TX 1 2 TX 2 2
!EQ TX 1 3 TX 2 3
!EQ TX 1 4 TX 2 4
!EQ TX 1 5 TX 2 5
EQ TX 1 6 TX 2 6
!EQ TX 1 7 TX 2 7
!EQ TX 1 8 TX 2 8
!EQ TX 1 9 TX 2 9
!EQ TX 1 10 TX 2 10
!EQ TX 1 11 TX 2 11
!EQ TX 1 12 TX 2 12
OU MI AD=OFF IT=300
```

```
TITLE Iterative MACS Procedure for Sample Size of 500:  a DIF Examine Item 1
!REFERENCE GROUP
DA NI=13 NG=2 NO=500
RA=Ref500.DAT
LA
I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 GP
SE
I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 /
MO NX=12 NK=1 TX=FR KA=FI
LK
ABILITY
VA 1.0 LX 6 1
FR LX 1 1 LX 2 1 LX 3 1 LX 4 1 LX 5 1 LX 7 1 LX 8 1
FR LX 9 1 LX 10 1 LX 11 1 LX 12 1
FR TX 1 TX 2 TX 3 TX 4 TX 5 TX 6 TX 7 TX 8
FR TX 9 TX 10 TX 11 TX 12
OU MI AD=OFF IT=300

Mean and Covariance Structure Model
!FOCAL GROUP
DA NI=13 NO=500
RA=aDIF500.DAT
LA
I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 GP
SE
I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 /
MO NX=12 NK=1 TX=FR KA=FI
LK
ABILITY
VA 1.0 LX 6 1
FR LX 1 1 LX 2 1 LX 3 1 LX 4 1 LX 5 1 LX 7 1 LX 8 1
FR LX 9 1 LX 10 1 LX 11 1 LX 12 1
FR TX 1 TX 2 TX 3 TX 4 TX 5 TX 6 TX 7 TX 8
FR TX 9 TX 10 TX 11 TX 12
EQ LX 1 1 1 LX 2 1 1
!EQ LX 1 2 1 LX 2 2 1
!EQ LX 1 3 1 LX 2 3 1
!EQ LX 1 4 1 LX 2 4 1
!EQ LX 1 5 1 LX 2 5 1
EQ LX 1 6 1 LX 2 6 1
!EQ LX 1 7 1 LX 2 7 1
!EQ LX 1 8 1 LX 2 8 1
!EQ LX 1 9 1 LX 2 9 1
!EQ LX 1 10 1 LX 2 10 1
!EQ LX 1 11 1 LX 2 11 1
!EQ LX 1 12 1 LX 2 12 1
```

```
EQ TX 1 1 TX 2 1
!EQ TX 1 2 TX 2 2
!EQ TX 1 3 TX 2 3
!EQ TX 1 4 TX 2 4
!EQ TX 1 5 TX 2 5
EQ TX 1 6 TX 2 6
!EQ TX 1 7 TX 2 7
!EQ TX 1 8 TX 2 8
!EQ TX 1 9 TX 2 9
!EQ TX 1 10 TX 2 10
!EQ TX 1 11 TX 2 11
!EQ TX 1 12 TX 2 12
OU MI AD=OFF IT=300
```

```
TITLE Noniterative Mean & Covariance Structure Procedure: a DIF, Sample Size of 500
!REFERENCE GROUP
DA NI=13 NG=2 NO=500
RA=Ref500.DAT
LA
I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 GP
SE
I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 /
MO NX=12 NK=1 TX=FR KA=FI
LK
ABILITY
VA 1.0 LX 6 1
FR LX 1 1 LX 2 1 LX 3 1 LX 4 1 LX 5 1 LX 7 1 LX 8 1
FR LX 9 1 LX 10 1 LX 11 1 LX 12 1
FR TX 1 TX 2 TX 3 TX 4 TX 5 TX 6 TX 7 TX 8
FR TX 9 TX 10 TX 11 TX 12
OU MI AD=OFF IT=300

Mean and Covariance Structure Model
!FOCAL GROUP
DA NI=13 NO=500
RA=aDIF500.DAT
LA
I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 GP
SE
I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 /
MO NX=12 NK=1 TX=FR KA=FI
LK
ABILITY
VA 1.0 LX 6 1
FR LX 1 1 LX 2 1 LX 3 1 LX 4 1 LX 5 1 LX 7 1 LX 8 1
FR LX 9 1 LX 10 1 LX 11 1 LX 12 1
FR TX 1 TX 2 TX 3 TX 4 TX 5 TX 6 TX 7 TX 8
FR TX 9 TX 10 TX 11 TX 12
EQ LX 1 1 1 LX 2 1 1
EQ LX 1 2 1 LX 2 2 1
EQ LX 1 3 1 LX 2 3 1
EQ LX 1 4 1 LX 2 4 1
EQ LX 1 5 1 LX 2 5 1
EQ LX 1 6 1 LX 2 6 1
EQ LX 1 7 1 LX 2 7 1
EQ LX 1 8 1 LX 2 8 1
EQ LX 1 9 1 LX 2 9 1
EQ LX 1 10 1 LX 2 10 1
EQ LX 1 11 1 LX 2 11 1
EQ LX 1 12 1 LX 2 12 1
```

```
EQ TX 1 1 TX 2 1
EQ TX 1 2 TX 2 2
EQ TX 1 3 TX 2 3
EQ TX 1 4 TX 2 4
EQ TX 1 5 TX 2 5
EQ TX 1 6 TX 2 6
EQ TX 1 7 TX 2 7
EQ TX 1 8 TX 2 8
EQ TX 1 9 TX 2 9
EQ TX 1 10 TX 2 10
EQ TX 1 11 TX 2 11
EQ TX 1 12 TX 2 12
OU MI AD=OFF IT=300
```

APPENDIX C

SAMPLE PROGRAMS FOR MODIFICATION INDEX METHOD AND MODIFICATION INDEX-DIVIDED SAMPLE METHOD

```
TITLE Modification Index Method: a DIF, Sample Size of 500
DA NI=13 NO=1000
RA=ALL500_ADIF_ANALYSIS.DAT
LA
I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 GP
SE
I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 GP/
MO NX=13 NK=2 PH=DI
LK
ITEMS GROUP
FR LX 1 1 LX 2 1 LX 3 1 LX 4 1 LX 5 1 LX 7 1 LX 8 1 LX 9 1 LX 10 1
FR LX 11 1 LX 12 1
FI LX 6 1 LX 13 2
VA 1.0 LX 6 1 LX 13 2
FI TD 13 13
VA 0.0 TD 13 13
OU MI AD=OFF
```

```
TITLE Modification Index Divided:  a DIF, Sample Size of 500
DA NI=13 NO=500
RA=MID500_A_11.DAT
LA
I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 GP
SE
I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 GP/
MO NX=13 NK=2 PH=DI
LK
ITEMS GROUP
FR LX 1 1 LX 2 1 LX 3 1 LX 4 1 LX 5 1 LX 7 1 LX 8 1 LX 9 1 LX 10 1
FR LX 11 1 LX 12 1
FI LX 6 1 LX 13 2
VA 1.0 LX 6 1 LX 13 2
FI TD 13 13
VA 0.0 TD 13 13
OU MI AD=OFF
```

APPENDIX D

DESCRIPTIVE STATISTICS FOR

MODIFICATION INDEXES AND CHI-SQUARE DIFFERENCE TEST RESULTS

Table D.1

Noniterative MACS Method: Descriptive Statistics for Modification Indexes

N	Type of DIF	ITEM	GROUP	Data Sets λ	Data Sets τ	Minimum λ	Minimum τ	Maximum λ	Maximum τ	Mean λ	Mean τ	Std. Dev. λ	Std. Dev. τ
250	Loadings	1	Foc.	50	50	0	0	0	0	0.11	0.00	0.09	0.00
			Ref.	50	50	0	0	1	0	0.13	0.00	0.10	0.00
		2	Foc.	50	50	0	0	1	0	0.19	0.00	0.09	0.00
			Ref.	50	50	0	0	1	0	0.22	0.00	0.11	0.00
		3	Foc.	50	50	0	0	1	0	0.24	0.00	0.14	0.00
			Ref.	50	50	0	0	1	0	0.27	0.00	0.17	0.00
		4	Foc.	50	50	1	0	6	1	3.29	0.11	1.23	0.12
			Ref.	50	50	1	0	7	1	3.69	0.11	1.40	0.12
		5	Foc.	50	50	0	0	1	0	0.30	0.00	0.11	0.00
			Ref.	50	50	0	0	1	0	0.34	0.00	0.13	0.00
		7	Foc.	50	50	0	0	5	0	1.79	0.09	0.89	0.10
			Ref.	50	50	0	0	6	0	2.02	0.09	1.02	0.10
		8	Foc.	50	50	0	0	1	0	0.19	0.00	0.10	0.00
			Ref.	50	50	0	0	1	0	0.22	0.00	0.11	0.00
		9	Foc.	50	50	0	0	0	0	0.22	0.00	0.10	0.00
			Ref.	50	50	0	0	1	0	0.25	0.00	0.12	0.00
		10	Foc.	50	50	1	0	5	1	1.96	0.10	0.88	0.14
			Ref.	50	50	1	0	5	1	2.21	0.10	0.99	0.14
		11	Foc.	50	50	0	0	0	0	0.14	0.00	0.07	0.00
			Ref.	50	50	0	0	0	0	0.16	0.00	0.08	0.00
		12	Foc.	50	50	0	0	1	0	0.24	0.00	0.11	0.00
			Ref.	50	50	0	0	1	0	0.27	0.00	0.12	0.00
	Thresholds	1	Foc.	50	50	0	0	0	2	0.01	1.04	0.02	0.43
			Ref.	50	50	0	0	0	2	0.01	1.04	0.02	0.43
		2	Foc.	50	50	0	1	0	3	0.03	1.80	0.04	0.51
			Ref.	50	50	0	1	0	3	0.03	1.80	0.05	0.51
		3	Foc.	50	50	0	1	0	4	0.03	2.37	0.04	0.67
			Ref.	50	50	0	1	0	4	0.04	2.37	0.04	0.67
		4	Foc.	50	50	0	14	2	38	0.30	27.13	0.35	5.79
			Ref.	50	50	0	14	2	38	0.31	27.13	0.37	5.79
		5	Foc.	50	50	0	2	0	5	0.04	2.85	0.06	0.66
			Ref.	50	50	0	2	0	5	0.04	2.85	0.06	0.66
		7	Foc.	50	50	0	13	2	34	0.32	22.97	0.49	4.94
			Ref.	50	50	0	13	3	34	0.34	22.97	0.52	4.94
		8	Foc.	50	50	0	1	0	3	0.03	1.84	0.04	0.54
			Ref.	50	50	0	1	0	3	0.03	1.84	0.04	0.54
		9	Foc.	50	50	0	1	0	3	0.02	1.94	0.03	0.60
			Ref.	50	50	0	1	0	3	0.03	1.94	0.03	0.60
		10	Foc.	50	50	0	14	2	38	0.40	24.30	0.50	5.00
			Ref.	50	50	0	14	2	38	0.42	24.30	0.52	5.00

Table D.1 (*continued*)

N	Type of DIF	ITEM	GROUP	Data Sets λ	Data Sets τ	Minimum λ	Minimum τ	Maximum λ	Maximum τ	Mean λ	Mean τ	Std. Dev. λ	Std. Dev. τ
		11	Foc.	50	50	0	0	0	2	0.04	1.18	0.06	0.42
			Ref.	50	50	0	0	0	2	0.04	1.18	0.07	0.42
		12	Foc.	50	50	0	1	0	4	0.04	2.36	0.05	0.55
			Ref.	50	50	0	1	0	4	0.04	2.36	0.05	0.55
	Loadings and Thresholds	1	Foc.	50	50	0	0	0	0	0.10	0.21	0.07	0.10
			Ref.	50	50	0	0	0	0	0.11	0.21	0.08	0.10
		2	Foc.	50	50	0	0	0	1	0.18	0.35	0.09	0.11
			Ref.	50	50	0	0	0	1	0.20	0.35	0.11	0.11
		3	Foc.	50	50	0	0	1	1	0.22	0.47	0.14	0.13
			Ref.	50	50	0	0	1	1	0.25	0.47	0.16	0.13
		4	Foc.	50	50	1	4	6	13	3.00	6.95	1.28	1.85
			Ref.	50	50	1	4	7	13	3.36	6.95	1.45	1.85
		5	Foc.	50	50	0	0	1	1	0.29	0.57	0.13	0.16
			Ref.	50	50	0	0	1	1	0.34	0.57	0.15	0.16
		7	Foc.	50	50	0	2	4	10	1.39	6.31	0.71	2.11
			Ref.	50	50	0	2	4	10	1.57	6.31	0.81	2.11
		8	Foc.	50	50	0	0	0	1	0.18	0.36	0.09	0.12
			Ref.	50	50	0	0	1	1	0.20	0.36	0.10	0.12
		9	Foc.	50	50	0	0	0	1	0.19	0.38	0.10	0.14
			Ref.	50	50	0	0	1	1	0.21	0.38	0.11	0.14
		10	Foc.	50	50	1	3	5	12	2.04	6.83	1.02	2.18
			Ref.	50	50	1	3	6	12	2.30	6.83	1.16	2.18
		11	Foc.	50	50	0	0	0	1	0.14	0.23	0.11	0.09
			Ref.	50	50	0	0	1	1	0.16	0.23	0.12	0.09
		12	Foc.	50	50	0	0	0	1	0.23	0.46	0.10	0.12
			Ref.	50	50	0	0	1	1	0.26	0.46	0.12	0.12
500	Loadings	1	Foc.	50	50	0	0	0	0	0.21	0.00	0.09	0.00
			Ref.	50	50	0	0	1	0	0.23	0.00	0.10	0.00
		2	Foc.	50	50	0	0	1	0	0.36	0.00	0.12	0.00
			Ref.	50	50	0	0	1	0	0.40	0.00	0.14	0.00
		3	Foc.	50	50	0	0	1	0	0.40	0.00	0.14	0.00
			Ref.	50	50	0	0	1	0	0.46	0.00	0.16	0.00
		4	Foc.	50	50	4	0	12	1	7.26	0.14	1.75	0.19
			Ref.	50	50	4	0	14	1	8.11	0.14	1.95	0.19
		5	Foc.	50	50	0	0	1	0	0.61	0.00	0.19	0.00
			Ref.	50	50	0	0	1	0	0.69	0.00	0.22	0.00
		7	Foc.	50	50	1	0	5	1	3.20	0.07	1.15	0.12
			Ref.	50	50	1	0	6	1	3.60	0.07	1.30	0.12
		8	Foc.	50	50	0	0	1	0	0.36	0.00	0.11	0.00
			Ref.	50	50	0	0	1	0	0.41	0.00	0.13	0.00
		9	Foc.	50	50	0	0	1	0	0.43	0.00	0.15	0.00

Table D.1 (*continued*)

N	Type of DIF	ITEM	GROUP	Data Sets λ	Data Sets τ	Minimum λ	Minimum τ	Maximum λ	Maximum τ	Mean λ	Mean τ	Std. Dev. λ	Std. Dev. τ
		9	Ref.	50	50	0	0	1	0	0.49	0.00	0.17	0.00
		10	Foc.	50	50	1	0	7	0	3.81	0.10	1.34	0.11
			Ref.	50	50	1	0	7	0	4.28	0.10	1.50	0.11
		11	Foc.	50	50	0	0	1	0	0.26	0.00	0.13	0.00
			Ref.	50	50	0	0	1	0	0.29	0.00	0.15	0.00
		12	Foc.	50	50	0	0	1	0	0.49	0.00	0.15	0.00
			Ref.	50	50	0	0	1	0	0.56	0.00	0.18	0.00
	Thresholds	1	Foc.	50	50	0	1	0	4	0.02	2.07	0.03	0.57
			Ref.	50	50	0	1	0	4	0.02	2.07	0.03	0.57
		2	Foc.	50	50	0	2	0	5	0.05	3.46	0.05	0.73
			Ref.	50	50	0	2	0	5	0.05	3.46	0.06	0.73
		3	Foc.	50	50	0	3	0	6	0.03	4.31	0.04	0.70
			Ref.	50	50	0	3	0	6	0.03	4.31	0.04	0.70
		4	Foc.	50	50	0	38	3	80	0.40	54.01	0.50	7.91
			Ref.	50	50	0	38	3	80	0.42	54.01	0.52	7.91
		5	Foc.	50	50	0	3	0	9	0.07	5.94	0.07	1.16
			Ref.	50	50	0	3	0	9	0.07	5.94	0.07	1.16
		7	Foc.	50	50	0	36	2	68	0.29	48.66	0.37	6.83
			Ref.	50	50	0	36	2	68	0.30	48.66	0.39	6.83
		8	Foc.	50	50	0	2	0	5	0.03	3.44	0.03	0.72
			Ref.	50	50	0	2	0	5	0.03	3.44	0.03	0.72
		9	Foc.	50	50	0	2	0	6	0.04	3.96	0.05	0.83
			Ref.	50	50	0	2	0	6	0.04	3.96	0.05	0.83
		10	Foc.	50	50	0	35	3	67	0.56	48.43	0.57	7.07
			Ref.	50	50	0	35	3	67	0.58	48.43	0.59	7.07
		11	Foc.	50	50	0	1	0	4	0.03	2.40	0.03	0.63
			Ref.	50	50	0	1	0	4	0.03	2.40	0.03	0.63
		12	Foc.	50	50	0	3	0	7	0.03	4.74	0.03	0.95
			Ref.	50	50	0	3	0	7	0.03	4.74	0.03	0.95
	Loadings and Thresholds	1	Foc.	50	50	0	0	0	1	0.19	0.40	0.10	0.11
			Ref.	50	50	0	0	1	1	0.21	0.40	0.11	0.11
		2	Foc.	50	50	0	0	1	1	0.36	0.66	0.16	0.13
			Ref.	50	50	0	0	1	1	0.41	0.66	0.18	0.13
		3	Foc.	50	50	0	1	1	1	0.37	0.82	0.17	0.14
			Ref.	50	50	0	1	1	1	0.42	0.82	0.19	0.14
		4	Foc.	50	50	1	8	10	21	6.19	13.70	1.91	2.94
			Ref.	50	50	2	8	11	21	6.91	13.70	2.16	2.94
		5	Foc.	50	50	0	1	1	2	0.58	1.14	0.20	0.25
			Ref.	50	50	0	1	1	2	0.66	1.14	0.23	0.25
		7	Foc.	50	50	1	9	6	20	2.98	13.10	1.20	2.40
			Ref.	50	50	1	9	7	20	3.35	13.10	1.37	2.41

Table D.1 (*continued*)

N	Type of DIF	ITEM	GROUP	Data Sets λ	Data Sets τ	Minimum λ	Minimum τ	Maximum λ	Maximum τ	Mean λ	Mean τ	Std. Dev. λ	Std. Dev. τ
		8	Foc.	50	50	0	0	1	1	0.34	0.66	0.14	0.14
			Ref.	50	50	0	0	1	1	0.39	0.66	0.15	0.14
		9	Foc.	50	50	0	0	1	1	0.40	0.75	0.15	0.16
			Ref.	50	50	0	0	1	1	0.45	0.75	0.18	0.16
		10	Foc.	50	50	1	9	9	20	4.00	13.27	1.53	2.55
			Ref.	50	50	1	9	10	20	4.49	13.27	1.73	2.55
		11	Foc.	50	50	0	0	1	1	0.25	0.46	0.11	0.13
			Ref.	50	50	0	0	1	1	0.29	0.46	0.13	0.13
		12	Foc.	50	50	0	1	1	1	0.45	0.91	0.14	0.19
			Ref.	50	50	0	1	1	1	0.51	0.91	0.17	0.19

Note. Modification index results for λ refer to loadings data. Modification index results for τ refer to intercepts data.

Table D.2

Iterative MACS Method: Descriptive Statistics for Chi-Square Differences

N	Type of DIF	Item	Data Sets	Minimum	Maximum	Mean	Std. Dev.
250	λ	1	50	0.00	0.06	0.01	0.01
		2	50	0.00	0.06	0.01	0.01
		3	50	0.00	0.12	0.01	0.02
		4	50	1.01	5.53	3.23	1.11
		5	50	0.00	0.07	0.01	0.01
		7	50	0.57	5.04	2.18	0.85
		8	50	0.00	0.06	0.01	0.01
		9	50	0.00	0.07	0.01	0.01
		10	50	0.91	4.65	2.38	0.88
		11	50	0.00	0.06	0.01	0.01
		12	50	0.00	0.05	0.01	0.01
	γ	1	50	0.00	0.10	0.02	0.02
		2	50	0.00	0.13	0.02	0.03
		3	50	0.00	0.27	0.02	0.05
		4	50	20.77	49.71	34.51	6.44
		5	50	0.00	0.09	0.02	0.02
		7	50	20.04	45.33	31.38	5.44
		8	50	0.00	0.10	0.02	0.02
		9	50	0.00	0.08	0.01	0.02
		10	50	21.64	44.49	32.91	5.26
		11	50	0.00	0.07	0.02	0.02
		12	50	0.00	0.10	0.02	0.02
	λ and γ	1	50	0.00	0.08	0.01	0.01
		2	50	0.00	0.05	0.01	0.01
		3	50	0.00	0.09	0.01	0.02
		4	50	5.48	17.27	11.17	2.66
		5	50	0.00	0.06	0.01	0.02
		7	50	4.11	13.53	9.50	2.38
		8	50	0.00	0.04	0.01	0.01
		9	50	0.00	0.06	0.01	0.02
		10	50	6.28	19.44	10.67	2.85
		11	50	0.00	0.12	0.01	0.02
		12	50	0.00	0.07	0.01	0.02
500	λ	1	50	0.00	0.04	0.01	0.01
		2	50	0.00	0.04	0.01	0.01
		3	50	0.00	0.06	0.01	0.01
		4	50	3.91	10.82	7.04	1.56
		5	50	0.00	0.05	0.01	0.01

Table D.2 (*continued*)

N	Type of DIF	Item	Data Sets	Minimum	Maximum	Mean	Std. Dev.
		7	50	1.99	6.60	4.02	1.17
		8	50	0.00	0.07	0.01	0.01
		9	50	0.00	0.06	0.01	0.01
		10	50	1.49	7.54	4.55	1.38
		11	50	0.00	0.04	0.01	0.01
		12	50	0.00	0.05	0.01	0.01
	γ	1	50	0.00	0.09	0.02	0.02
		2	50	0.00	0.07	0.02	0.02
		3	50	0.00	0.14	0.03	0.03
		4	50	51.71	93.24	68.34	8.32
		5	50	0.00	0.16	0.02	0.03
		7	50	50.26	89.03	65.01	7.82
		8	50	0.00	0.10	0.02	0.02
		9	50	0.00	0.10	0.02	0.03
		10	50	49.05	84.79	64.96	7.89
		11	50	0.00	0.15	0.01	0.02
		12	50	0.00	0.12	0.02	0.02
	λ and γ	1	50	0.00	0.07	0.01	0.01
		2	50	0.00	0.10	0.01	0.02
		3	50	0.00	0.10	0.02	0.02
		4	50	11.71	31.07	22.25	4.04
		5	50	0.00	0.06	0.01	0.01
		7	50	14.81	28.15	19.64	3.19
		8	50	0.00	0.13	0.02	0.02
		9	50	0.00	0.08	0.01	0.02
		10	50	12.94	30.72	20.69	3.56
		11	50	0.00	0.03	0.01	0.01
		12	50	0.00	0.14	0.01	0.02

Note. λ=loading values; γ=threshold values. Chi-square differences were computed by subtracting the chi-square goodness-of-fit value of the compact baseline model from the chi-square goodness-of-fit value from an augmented model.

Table D.3

Modification Index Method: Descriptive Statistics for Modification Indexes

N	Type of DIF	Item	Data Sets	Minimum	Maximum	Mean	Std. Dev.
250	λ	1	50	0.00	0.01	0.00	0.00
		2	50	0.00	0.01	0.00	0.00
		3	50	0.00	0.02	0.00	0.00
		4	50	0.00	0.51	0.11	0.12
		5	50	0.00	0.01	0.00	0.00
		7	50	0.00	0.37	0.09	0.10
		8	50	0.00	0.01	0.00	0.00
		9	50	0.00	0.01	0.00	0.00
		10	50	0.00	0.71	0.11	0.14
		11	50	0.00	0.01	0.00	0.00
		12	50	0.00	0.01	0.00	0.00
	γ	1	50	0.25	2.48	1.05	0.44
		2	50	0.96	2.81	1.82	0.52
		3	50	1.39	4.09	2.39	0.67
		4	50	14.46	38.23	27.27	5.83
		5	50	1.76	4.56	2.87	0.66
		7	50	13.04	34.29	23.09	4.96
		8	50	0.85	3.19	1.85	0.54
		9	50	1.15	3.43	1.95	0.60
		10	50	14.42	37.74	24.42	5.03
		11	50	0.50	2.18	1.18	0.42
		12	50	1.04	3.58	2.37	0.55
	λ and γ	1	50	0.05	0.48	0.21	0.10
		2	50	0.15	0.62	0.35	0.11
		3	50	0.19	0.79	0.46	0.13
		4	50	3.56	12.93	7.04	1.87
		5	50	0.25	1.03	0.56	0.16
		7	50	2.42	10.49	6.39	2.13
		8	50	0.18	0.61	0.36	0.12
		9	50	0.18	0.76	0.38	0.14
		10	50	3.53	12.52	6.90	2.20
		11	50	0.09	0.52	0.23	0.09
		12	50	0.16	0.70	0.46	0.12
500	λ	1	50	0.00	0.01	0.00	0.00
		2	50	0.00	0.01	0.00	0.00
		3	50	0.00	0.02	0.00	0.00
		4	50	0.00	0.72	0.14	0.19
		5	50	0.00	0.02	0.00	0.00

Table D.3 (*continued*)

N	Type of DIF	Item	Data Sets	Minimum	Maximum	Mean	Std. Dev.
		7	50	0.00	0.64	0.07	0.12
		8	50	0.00	0.02	0.00	0.00
		9	50	0.00	0.02	0.00	0.00
		10	50	0.00	0.47	0.10	0.11
		11	50	0.00	0.01	0.00	0.00
		12	50	0.00	0.02	0.00	0.00
	γ	1	50	1.07	3.55	2.08	0.57
		2	50	2.03	5.33	3.47	0.73
		3	50	2.93	5.96	4.33	0.70
		4	50	38.31	80.43	54.13	7.93
		5	50	3.43	9.08	5.96	1.16
		7	50	36.57	68.62	48.77	6.85
		8	50	1.72	5.22	3.46	0.73
		9	50	2.00	6.08	3.97	0.83
		10	50	34.90	67.14	48.53	7.09
		11	50	1.20	3.81	2.41	0.63
		12	50	3.01	7.41	4.76	0.95
	λ and γ	1	50	0.18	0.63	0.39	0.11
		2	50	0.39	0.93	0.65	0.13
		3	50	0.53	1.18	0.81	0.14
		4	50	7.83	20.96	13.84	2.96
		5	50	0.63	1.86	1.12	0.24
		7	50	9.54	19.94	13.21	2.42
		8	50	0.37	0.99	0.65	0.14
		9	50	0.44	1.15	0.74	0.15
		10	50	8.69	20.33	13.38	2.57
		11	50	0.21	0.70	0.45	0.12
		12	50	0.52	1.32	0.90	0.18

Note. λ=loading values; γ=threshold values.

Table D.4
Modification Index-Divided Method Low Scoring Group: Descriptive Statistics for Modification Indexes

N	Type of DIF	Item	Data Sets	Minimum	Maximum	Mean	Std. Dev.
250	λ	1	49	0.00	0.13	0.02	0.03
		2	49	0.00	0.27	0.05	0.07
		3	49	0.00	0.12	0.03	0.03
		4	49	0.19	3.20	1.42	0.74
		5	49	0.00	0.38	0.05	0.08
		7	49	0.00	2.95	0.95	0.73
		8	49	0.00	0.18	0.03	0.04
		9	49	0.00	0.45	0.05	0.09
		10	49	0.02	2.68	0.83	0.60
		11	49	0.00	0.18	0.03	0.04
		12	49	0.00	0.21	0.04	0.06
	γ	1	50	0.00	1.99	0.60	0.48
		2	50	0.01	2.82	0.92	0.74
		3	50	0.07	4.39	1.36	0.99
		4	50	4.08	21.48	11.77	3.98
		5	50	0.06	4.36	1.41	0.96
		7	50	3.53	21.06	9.74	4.00
		8	50	0.01	2.93	0.76	0.72
		9	50	0.01	4.47	0.86	0.77
		10	50	2.46	19.90	10.24	3.59
		11	50	0.00	2.17	0.46	0.48
		12	50	0.01	3.49	0.85	0.63
	λ and γ	1	47	0.00	0.43	0.11	0.12
		2	47	0.00	1.09	0.13	0.20
		3	47	0.00	0.50	0.15	0.13
		4	47	0.00	2.12	0.51	0.41
		5	47	0.00	0.65	0.16	0.15
		7	47	0.02	4.50	1.19	0.90
		8	47	0.00	1.47	0.14	0.24
		9	47	0.00	0.64	0.13	0.14
		10	47	0.08	3.16	1.04	0.70
		11	47	0.00	0.44	0.10	0.11
		12	47	0.00	0.49	0.10	0.13
500	λ	1	50	0.00	0.44	0.04	0.07
		2	50	0.00	0.25	0.05	0.05
		3	50	0.00	0.29	0.06	0.08
		4	50	0.80	5.72	3.09	1.25

Table D.4 (*continued*)

N	Type of DIF	Item	Data Sets	Minimum	Maximum	Mean	Std. Dev.
		5	50	0.00	0.26	0.06	0.06
		7	50	0.19	4.81	1.49	0.85
		8	50	0.00	0.38	0.06	0.08
		9	50	0.00	0.27	0.05	0.06
		10	50	0.31	3.39	1.59	0.81
		11	50	0.00	0.35	0.05	0.07
		12	50	0.00	0.20	0.05	0.05
	γ	1	50	0.01	2.84	0.98	0.64
		2	50	0.01	3.70	1.43	0.88
		3	50	0.33	6.24	2.32	1.08
		4	50	12.47	38.73	23.59	6.00
		5	50	0.25	5.80	2.21	1.24
		7	50	12.45	36.64	20.43	5.15
		8	50	0.20	3.59	1.70	0.80
		9	50	0.21	3.98	1.63	1.03
		10	50	10.72	36.56	20.59	5.98
		11	50	0.04	2.45	0.92	0.65
		12	50	0.20	6.51	2.27	1.38
	λ and γ	1	50	0.00	0.95	0.16	0.17
		2	50	0.00	0.87	0.22	0.22
		3	50	0.01	0.85	0.28	0.20
		4	50	0.06	3.78	1.09	0.80
		5	50	0.00	1.14	0.30	0.26
		7	50	0.48	5.39	2.12	0.90
		8	50	0.00	1.20	0.18	0.23
		9	50	0.00	0.79	0.18	0.19
		10	50	0.15	3.69	1.85	0.80
		11	50	0.00	0.94	0.16	0.22
		12	50	0.00	1.29	0.32	0.27

Note. λ=loading values; γ=threshold values.

Table D.5
Modification Index-Divided Method High Scoring Group: Descriptive Statistics for Modification Indexes

N	Type of DIF	Item	Data Sets	Minimum	Maximum	Mean	Std. Dev.
250	λ	1	50	0.00	0.09	0.02	0.02
		2	50	0.00	0.17	0.03	0.04
		3	50	0.00	0.15	0.02	0.03
		4	50	0.13	3.94	1.42	0.94
		5	50	0.00	0.18	0.03	0.04
		7	50	0.14	1.49	0.69	0.39
		8	50	0.00	0.11	0.02	0.03
		9	50	0.00	0.12	0.02	0.03
		10	50	0.02	1.95	0.77	0.52
		11	50	0.00	0.09	0.02	0.03
		12	50	0.00	0.25	0.03	0.04
	γ	1	50	0.00	1.97	0.47	0.44
		2	50	0.05	3.08	0.91	0.63
		3	50	0.09	3.05	1.05	0.71
		4	50	5.44	24.47	15.65	5.06
		5	50	0.15	4.09	1.36	0.82
		7	50	5.14	23.50	13.62	3.72
		8	50	0.10	3.30	1.10	0.66
		9	50	0.20	2.78	0.98	0.64
		10	50	7.36	22.87	14.10	3.47
		11	50	0.05	1.90	0.64	0.42
		12	50	0.27	3.54	1.35	0.80
	λ and γ	1	50	0.00	1.06	0.14	0.19
		2	50	0.00	0.85	0.27	0.23
		3	50	0.01	0.92	0.32	0.24
		4	50	4.91	14.51	8.95	2.59
		5	50	0.00	1.46	0.40	0.31
		7	50	1.55	11.17	6.30	2.22
		8	50	0.00	1.20	0.28	0.28
		9	50	0.00	0.87	0.24	0.20
		10	50	3.39	13.63	6.94	2.09
		11	50	0.00	0.76	0.18	0.18
		12	50	0.05	1.19	0.36	0.29
500	λ	1	50	0.00	0.62	0.04	0.09
		2	50	0.00	0.19	0.04	0.04
		3	50	0.00	0.32	0.03	0.05
		4	50	0.23	5.35	2.67	1.07

Table D.5 (*continued*)

N	Type of DIF	Item	Data Sets	Minimum	Maximum	Mean	Std. Dev.
		5	50	0.00	0.14	0.04	0.04
		7	50	0.15	2.66	1.19	0.53
		8	50	0.00	0.29	0.03	0.05
		9	50	0.00	0.16	0.03	0.04
		10	50	0.36	2.93	1.37	0.66
		11	50	0.00	0.29	0.04	0.06
		12	50	0.00	0.17	0.04	0.04
	γ	1	50	0.09	4.20	1.20	0.91
		2	50	0.29	4.46	2.10	0.95
		3	50	0.16	4.99	2.10	1.07
		4	50	14.40	41.11	29.90	6.18
		5	50	1.23	6.90	3.59	1.23
		7	50	18.94	40.68	27.96	5.17
		8	50	0.09	3.74	1.70	0.85
		9	50	0.53	5.22	2.25	1.02
		10	50	14.65	48.47	27.77	5.70
		11	50	0.10	3.77	1.49	0.83
		12	50	0.26	8.22	2.42	1.28
	λ and γ	1	50	0.00	1.86	0.31	0.33
		2	50	0.04	1.47	0.54	0.37
		3	50	0.02	1.56	0.50	0.36
		4	50	6.77	23.79	16.68	4.12
		5	50	0.12	2.23	0.88	0.49
		7	50	6.99	19.25	13.00	2.94
		8	50	0.02	1.82	0.54	0.37
		9	50	0.04	1.91	0.65	0.41
		10	50	7.55	24.59	13.58	3.17
		11	50	0.01	1.24	0.42	0.28
		12	50	0.02	1.68	0.60	0.35

Note. λ=loading values; γ=threshold values.

APPENDIX E

DIF DETECTION COUNTS

Table E.1
Noniterative MACS Method: Sample Size 250

Type of DIF	Item	DIF Detection Result		Total
		Non-DIF	DIF	
λ	1	50		50
	2	50		50
	3	50		50
	4	**30**	**20**	**50**
	5	50		50
	7	**49**	**1**	**50**
	8	50		50
	9	50		50
	10	**47**	**3**	**50**
	11	50		50
	12	50		50
λ Total		526	24	550
γ	1	50		50
	2	50		50
	3	48	2	50
	4		**50**	**50**
	5	44	6	50
	7		**50**	**50**
	8	50		50
	9	50		50
	10		**50**	**50**
	11	50		50
	12	50		50
γ Total		392	158	550
λ and γ	1	50		50
	2	50		50
	3	50		50
	4	**1**	**49**	**50**
	5	50		50
	7	**5**	**45**	**50**
	8	50		50
	9	50		50
	10	**4**	**46**	**50**
	11	50		50
	12	50		50
λ and γ Total		410	140	550
Grand Total		1328	322	1650

Note. λ=loading values; γ=threshold values.

Table E.2
Noniterative MACS Method: Sample Size 500

Type of DIF	Item	DIF Detection Result: Non-DIF	DIF	Total
λ	1	50		50
	2	50		50
	3	50		50
	4		**50**	**50**
	5	50		50
	7	**35**	**15**	**50**
	8	50		50
	9	50		50
	10	**25**	**25**	**50**
	11	50		50
	12	50		50
λ Total		460	90	550
γ	1	50		50
	2	34	16	50
	3	14	36	50
	4		**50**	**50**
	5	2	48	50
	7		**50**	**50**
	8	38	12	50
	9	27	23	50
	10		**50**	**50**
	11	50		50
	12	8	42	50
γ Total		223	327	550
λ and γ	1	50		50
	2	50		50
	3	50		50
	4		**50**	**50**
	5	50		50
	7		**50**	**50**
	8	50		50
	9	50		50
	10		**50**	**50**
	11	50		50
	12	50		50
λ and γ Total		400	150	550
Grand Total		1083	567	1650

Note. λ=loading values; γ=threshold values.

Table E.3
Iterative MACS Method: Sample Size 250

Type of DIF	Item	DIF Detection Result		Total
		Non-DIF	DIF	
λ	1	50		50
	2	50		50
	3	50		50
	4	**45**	**5**	**50**
	5	50		50
	7	**49**	**1**	**50**
	8	50		50
	9	50		50
	10	**49**	**1**	**50**
	11	50		50
	12	50		50
λ Total		543	7	550
γ	1	50		50
	2	50		50
	3	50		50
	4		**50**	**50**
	5	50		50
	7		**50**	**50**
	8	50		50
	9	50		50
	10		**50**	**50**
	11	50		50
	12	50		50
γ Total		400	150	550
λ and γ	1	50		50
	2	50		50
	3	50		50
	4		**50**	**50**
	5	50		50
	7	**1**	**49**	**50**
	8	50		50
	9	50		50
	10		**50**	**50**
	11	50		50
	12	50		50
λ and γ Total		401	149	550
Grand Total		1344	306	1650

Note. λ=loading values; γ=threshold values.

Table E.4
Iterative MACS Method: Sample Size 500

Type of DIF	Item	DIF Detection Result: Non-DIF	DIF	Total
λ	1	50		50
	2	50		50
	3	50		50
	4	**3**	**47**	**50**
	5	50		50
	7	**37**	**13**	**50**
	8	50		50
	9	50		50
	10	**27**	**23**	**50**
	11	50		50
	12	50		50
λ Total		467	83	550
γ	1	50		50
	2	50		50
	3	50		50
	4		**50**	**50**
	5	50		50
	7		**50**	**50**
	8	50		50
	9	50		50
	10		**50**	**50**
	11	50		50
	12	50		50
γ Total		400	150	550
λ and γ	1	50		50
	2	50		50
	3	50		50
	4		**50**	**50**
	5	50		50
	7		**50**	**50**
	8	50		50
	9	50		50
	10		**50**	**50**
	11	50		50
	12	50		50
λ and γ Total		400	150	550
Grand Total		1267	383	1650

Note. λ=loading values; γ=threshold values.

Table E.5
Modification Index Method: Sample Size 250

Type of DIF	Item	DIF Detection Result		Total
		Non-DIF	DIF	
λ	1	50		50
	2	50		50
	3	50		50
	4	**50**		**50**
	5	50		50
	7	**50**		**50**
	8	50		50
	9	50		50
	10	**50**		**50**
	11	50		50
	12	50		50
λ Total		550		550
γ	1	50		50
	2	50		50
	3	48	2	50
	4		**50**	**50**
	5	43	7	50
	7		**50**	**50**
	8	50		50
	9	50		50
	10		**50**	**50**
	11	50		50
	12	50		50
γ Total		391	159	550
λ and γ	1	50		50
	2	50		50
	3	50		50
	4	**1**	**49**	**50**
	5	50		50
	7	**5**	**45**	**50**
	8	50		50
	9	50		50
	10	**2**	**48**	**50**
	11	50		50
	12	50		50
λ and γ Total		408	142	550
Grand Total		1349	301	1650

Note. λ=loading values; γ=threshold values.

Table E.6
Modification Index Method: Sample Size 500

Type of DIF	Item	DIF Detection Result		Total
		Non-DIF	DIF	
λ	1	50		50
	2	50		50
	3	50		50
	4	**50**		**50**
	5	50		50
	7	**50**		**50**
	8	50		50
	9	50		50
	10	**50**		**50**
	11	50		50
	12	50		50
λ Total		550		550
γ	1	50		50
	2	34	16	50
	3	13	37	50
	4		**50**	**50**
	5	2	48	50
	7		**50**	**50**
	8	37	13	50
	9	27	23	50
	10		**50**	**50**
	11	50		50
	12	8	42	50
γ Total		221	329	550
λ and γ	1	50		50
	2	50		50
	3	50		50
	4		**50**	**50**
	5	50		50
	7		**50**	**50**
	8	50		50
	9	50		50
	10		**50**	**50**
	11	50		50
	12	50		50
λ and γ Total		400	150	550
Grand Total		1171	479	1650

Note. λ=loading values; γ=threshold values.

Table E.7

Modification Index-Divided Sample Method: Sample Size 250

Type of DIF	Item	DIF Detection Result		Total
		Non-DIF	DIF	
λ	1	49		49
	2	49		49
	3	49		49
	4	**48**	**1**	**49**
	5	49		49
	7	**49**		**49**
	8	49		49
	9	49		49
	10	**49**		**49**
	11	49		49
	12	49		49
λ Total		538	1	539
γ	1	50		50
	2	50		50
	3	46	4	50
	4		**50**	**50**
	5	43	7	50
	7		**50**	**50**
	8	50		50
	9	49	1	50
	10		**50**	**50**
	11	50		50
	12	50		50
γ Total		388	162	550
λ and γ	1	47		47
	2	47		47
	3	47		47
	4		**47**	**47**
	5	47		47
	7	**4**	**43**	**47**
	8	47		47
	9	47		47
	10		**47**	**47**
	11	47		47
	12	47		47
λ and γ Total		380	137	517
Grand Total		1306	300	1606

Note. λ=loading values; γ=threshold values.

Table E.8

Modification Index-Divided Sample Method: Sample Size 500

Type of DIF	Item	DIF Detection Result		Total
		Non-DIF	DIF	
λ	1	50		50
	2	50		50
	3	50		50
	4	**31**	**19**	**50**
	5	50		50
	7	**49**	**1**	**50**
	8	50		50
	9	50		50
	10	**50**		**50**
	11	50		50
	12	50		50
λ Total		530	20	550
γ	1	49	1	50
	2	33	17	50
	3	13	37	50
	4		**50**	**50**
	5	2	48	50
	7		**50**	**50**
	8	37	13	50
	9	26	24	50
	10		**50**	**50**
	11	50		50
	12	7	43	50
γ Total		217	333	550
λ and γ	1	50		50
	2	50		50
	3	50		50
	4		**50**	**50**
	5	50		50
	7		**50**	**50**
	8	50		50
	9	50		50
	10		**50**	**50**
	11	50		50
	12	50		50
λ and γ Total		400	150	550
Grand Total		1147	503	1650

Note. λ=loading values; γ=thresholds values.

VITA

Jenny Chia Yi Kuang received her B.A. from the University of California, Berkeley, in 1997. She graduated with a double major in psychology and anthropology. She received her M.S. in psychology from Old Dominion University in May of 2000. She received her Ph.D. in industrial and organizational psychology from Old Dominion University in May of 2007.

The address for the Department of Psychology at Old Dominion University is 250 Mills Godwin Building, Norfolk, VA, 23529.