1-1-2014

Using MIMIC Methods to Detect and Identify Sources of DIF among Multiple Groups

Seokjoon Chun
University of South Florida, seokjoon@mail.usf.edu

Using MIMIC Methods to Detect and Identify Sources of DIF among Multiple Groups

by

Seokjoon Chun

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science
Department of Psychology
College of Arts and Sciences
University of South Florida

Major Professor: Stephen Stark, Ph.D.
Eun Sook Kim, Ph.D.
Oleksandr S. Chernyshenko, Ph.D.
Michael Coovert, Ph.D.

Date of Approval:
September 24, 2014

Keywords: Multiple Indicator Multiple Cause Method, Differential Item Functioning, Constrained, Free, and Sequential-Free baseline approaches

DEDICATION

To the woman who gave me the unconditional love, my wife Yuna Chae.

Acknowledgements

My most sincere thanks go to my advisor, Dr. Stephen Stark. I am also deeply indebted to my committee members: Dr. Oleksander Chernyshenko, Dr. Eun Sook Kim, and Dr. Michael Coovert.

I want to express how grateful I am to my family, my father-in-law, mother-in-law, father, and my brother.

Mom, I want to say love you. And I hope you would enjoy my work with God.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This study investigated the efficacy of multiple indicators, multiple causes (MIMIC) methods in detecting uniform and nonuniform differential item functioning (DIF) among multiple groups, in which the underlying sources of DIF were manipulated. A sequential-free baseline procedure for running MIMIC models was developed and compared to free and constrained baseline model comparison methods. The sequential-free baseline procedure used the most discriminating non-DIF item identified in constrained baseline tests as a referent for subsequent free baseline model comparisons. The robustness of MIMIC DIF methods to violations of the equal factor variance assumption was also examined. Overall, a simulation study revealed that the practical sequential-free baseline method provided Type I error and power rates similar to the idealized free baseline method involving a designated non-DIF anchor, and much better Type I error and power rates than the constrained baseline method. However, when the MIMIC equal factor variance assumption was violated, Type I error was inflated. Although MIMIC methods were found to be effective in detecting uniform and nonuniform DIF, further methodological developments are needed to improve identification of the underlying sources.

INTRODUCTION

In industrial and organizational psychology, meta-analyses have shown that general mental ability (GMA) tests are among the best predictors of job performance (Schmidt & Hunter, 1998). Using GMA tests for entry level personnel screening therefore tends to improve overall productivity and organizational effectiveness, but because GMA tests often have adverse impact (Uniform Guidelines, 1978) against minority groups, it is incumbent that the measures are examined in local validation studies for predictive bias (Cleary, 1968) and differential item functioning (Drasgow, 1987) when sample sizes permit (American Psychological Association, 1999). *Differential item functioning* (DIF) is said to occur when the psychometric properties of an item, such as discrimination and difficulty, differ for individuals selected from subpopulations that have equal standing on the trait that is being measured. In other words, after accounting for true differences in ability, which are referred to as *impact*, the members of comparison groups still have different expected item scores (Drasgow, 1987; Stark, Chernyshenko, & Drasgow, 2004).

There is significant interest in identifying underlying "sources" of DIF in applied psychology and in educational measurement. Many authors have noted that DIF results from items measuring secondary factors or dimensions on which comparison groups systematically differ (Camilli, 1992; Lopez-Rivas Stark, & Chernyshenko, 2009; Shealy & Stout, 1993). For example, DIF on mathematical reasoning items might result from differences among comparison groups in English proficiency. Alternatively, in personality assessment, DIF on conscientiousness items might result from socially desirable responding that is more prevalent

among job applicants than nonapplicants or from cultural differences (Chernyshenko, Stark & Guenole, 2007; Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001).

A related and, perhaps, more perplexing practical problem in DIF analysis is how to parse examinee groups for comparisons and what to do if different items are flagged as problematic in separate analyses involving background variables that co-occur, such as gender and ethnicity or gender and language. For example, what remedies might be suggested if one subset of items in a math test exhibits DIF in comparisons of males and females, because they tap spatial abilities, and another subset exhibits DIF in comparisons of Whites and English-as-second-language (ESL) Hispanics because they are influenced by English proficiency? Preferably, one could compare item properties using both background variables simultaneously to determine their relative contributions to DIF and inform revisions that would minimize the overall untoward effects of bias.

The purpose of this research is to propose and evaluate a method that is suitable for detecting DIF due to two or more background variables and their potential interactions. More specifically, a Monte Carlo simulation will be conducted to explore the efficacy of DIF detection using three implementations of Multiple Indicator Multiple Cause (MIMIC; Jöreskog & Goldberger, 1975) methodology, which is rooted in the confirmatory factor analysis (CFA) tradition. The next section provides an overview of MIMIC methodology, its advantages over other multi-group CFA methods such as mean and covariance structure analysis (MACS), and a new approach to choosing a baseline model for DIF detection that is expected to provide better power and Type I error than conventional MIMIC implementations.

CFA Methods for DIF Detection

In the last decade alone, many papers have compared and contrasted CFA and IRT approaches to DIF detection (e.g., Kim & Yoon, 2011; Stark, Chernyshenko, & Drasgow, 2006). Some have suggested that IRT methods are advantageous because they fit nonlinear models directly to item responses, rather than linear models to inter-item correlation or covariance matrices, which is especially problematic with dichotomous variables. On the other hand, the unidimensionality assumption of many IRT DIF methods and software packages limit the scope of invariance testing relative to CFA software, which can easily accommodate multidimensionality and multiple groups. Moreover, with the advent of categorical CFA methods, there is no longer a fundamental distinction between the CFA and IRT item response models (McDonald, 1999), so the more general CFA approaches to DIF detection should be preferred.

There are at least two general ways to pursue DIF detection in the categorical CFA framework: Multi-group CFA (MGCFA) and MIMIC. MGCFA involves comparing the fit of increasingly constrained measurement models across two or more groups. Essentially, it compares a model with one or more parameters estimated freely across groups to a model in which those parameters are constrained to be equal. Although the efficacy of MGCFA methods has been well established (Kim, Yoon, & Lee, 2012), there are some noteworthy limitations. In MGCFA, only one categorical variable can be used to define the comparison groups, so comparisons involving more than one background variable are cumbersome and it is difficult to isolate the source of DIF when detected. In addition, because parameters are estimated for each group separately, each group must be large enough to adequately estimate model parameters,

which leads to large overall sample size requirements. In contrast, MIMIC methods estimate just one set of model parameters using the total sample of respondents and test for DIF by adding or deleting paths from variables associated with group membership to the items under investigation. Because the full sample is used for parameter estimation, the total sample size needed for effective MIMIC DIF detection can be considerably smaller than with MGCFA (Muthén,1989) and does not increase when more than two groups must be compared. Also, by allowing for the inclusion of more than one background variable and its interactions, MIMIC models can readily be used to explore why DIF occurs, which makes it an attractive alternative to MGCFA methods.

MIMIC DIF Detection

Like some other CFA and IRT DIF methods, MIMIC DIF analysis involves comparing the fit of a series of full and reduced models with the goal of determining whether items measuring a latent variable are equally discriminating and difficult across comparison groups. What makes MIMIC unique is the way this is accomplished. Rather than fixing and freeing parameters reflecting item discrimination (loadings) and difficulty (thresholds) across groups, MIMIC tests for DIF by adding or deleting direct paths to items emanating from the background variables associated with group membership, and impact is accounted for by paths from grouping variables to the common factor. Essentially, MIMIC tells us how grouping variables affect item properties and factor means.

The first step in building a baseline model for MIMIC DIF analysis is to select a categorical background variable *z* that defines group membership. If two-group characteristics are of interest, for example, then two background variables can be selected, such as gender and ethnicity, and called *z1* and *z2*. To test for a potential interaction of these variables, a third

background variable *z1z2* can be created. The next step in building a baseline model is to draw direct paths (γ) from each background variable to the common factor (θ) and paths (α) from the common factor to the items that will be tested for DIF, as shown in Figure 1.

The next step depends on whether one wishes to use a constrained baseline (all other) method or a free baseline (constant anchor) method to perform model comparisons (Stark, Chernyshenko, & Drasgow, 2006; Wang & Shih, 2010; Wang & Yeh, 2004). Most MIMIC DIF research has been conducted using the constrained baseline method in which items are tested for uniform DIF, associated with group differences in item thresholds (τ), by using the model shown in Figure 1 as a baseline and adding paths from each grouping variable (β) to individual items in a sequence of compact vs. augmented model comparisons. If an augmented model fits significantly better than the compact baseline, then the item under investigation is flagged as a DIF item (Kim et al., 2012). Figure 2a shows the full model that would be used to test for uniform DIF on Item 2 of a scale containing *k* items.

If one also wants to test for nonuniform DIF on Item 2, then additional variables can be created to reflect the moderating effect of background variables on common factors scores (Woods & Grimm, 2011). This is illustrated in Figure 2b, which shows three new variables $\theta z1$, $\theta z2$, and $\theta z1z2$ having direct paths (ω) to Item 2. Note that this augmented model performs an omnibus hypothesis test for DIF on item loadings (α) and thresholds(τ).

Although the constrained baseline approach to testing for DIF is convenient because it allows every item to be evaluated, it often leads to high Type I error rates because the baseline model is misspecified when DIF is present (Stark et al., 2006). To deal with this problem, researchers have explored using stricter p-values for DIF detection or corrections to chi-square statistics on which model comparisons may be based (Oort, 1998). However, evidence is

mounting that alternative free-baseline approaches are not only logically and statistically justified but more effective (Lopez-Rivas et al., 2009; Stark et al., 2006; Woods & Grimm, 2011)

Free-baseline approaches to DIF detection begin by forming a baseline model that has only the necessary constraints for identification. In MGCFA DIF analysis, this might involve constraining the loadings and thresholds for one item to be equal across comparison groups. Reduced models are then formed by constraining the loading and threshold parameters simultaneously for one additional item at a time and examining the change in goodness of fit for each reduced model relative to the baseline. If the fit worsens significantly, then the item under investigation is flagged as DIF. Stark et al. (2006) showed that this method yielded high power and low Type I error for IRT and mean and covariance structure (MACS; Sörbom, 1974) DIF detection, and Woods and Grimm (2011) showed that this general approach is effective with MIMIC.

Figure 3a shows a single-anchor free-baseline model for MIMIC DIF detection. Note that the model contains the same latent and observed variables as shown in Figure 2b, except that there are paths from the variables associated with group membership to all items except Item 1, which was conveniently chosen as an anchor for model identification. To perform an omnibus test for DIF due to group differences in loadings and thresholds on Item 2, the paths to Item 2 from the variables associated with group membership are deleted, as shown in Figure 3b. This process is repeated for the other items in the scale and, in each case, a statistically significant decrease in goodness of fit relative to the baseline model indicates DIF.

As discussed by Stark et al. (2006), this single-anchor omnibus free-baseline approach to DIF detection has some desirable features. For example, by keeping the baseline model consistent across comparisons, error does not propagate as it would if items found to be free of

DIF were subsequently added to the baseline to increase the size of the anchor group in the hope of increasing power, as is sometimes done in MGCFA invariance testing. In addition, by using just a single anchor item, there is less chance of contaminating the baseline model by including a DIF item. On the other hand, as noted by Lopez-Rivas et al. (2009), the effectiveness of this method relies on choosing a satisfactory anchor – one that is unbiased and adequately discriminating. Otherwise, performance could be as bad as or worse than the constrained baseline approach, which works reasonably well when contamination due to DIF is not severe.

To choose a suitable anchor, Lopez-Rivas et al. (2009) suggested a sequential-free baseline approach similar to Thissen and Steinberg (1988). Specifically, they suggested performing DIF analysis in two steps. First, conduct constrained baseline tests to identify items that appear to be free of DIF. Then, choose the most discriminating non-DIF item as the anchor for subsequent free baseline tests of the other items in the scale. A study by Meade and Wright (2012) found that this two-step method, sequential-free baseline method for DIF analysis was the most effective of several that were considered. But, there has been no research exploring its efficacy with MIMIC, so studies are needed to see whether that finding generalizes.

Summary

In field settings, there is often interest in testing for DIF in several groups simultaneously and identifying sources of DIF to help develop interventions and refine test construction practices. Although traditional IRT and MGCFA methods can be used toward that end, MIMIC methods provide a flexible alternative for multi-group DIF analysis that can more easily handle multiple background variables and their potential interactions without exceedingly large total samples (Kim et al., 2012; Woods, 2009). Moreover, with advances in structural equations

modeling software that allow for interactions between latent and observed variables (Muthén & Muthén, 1998-2013), it is now possible to construct MIMIC models for detecting nonuniform DIF as well as uniform DIF (Woods & Grimm, 2011). And, the use of free baseline methods for MIMIC DIF analysis (Woods & Grimm, 2011) can provide high power and effective Type I error control without statistical corrections for contamination due to DIF in constrained baseline applications. Nonetheless, research is still needed on several fronts, because most studies have addressed just a few key issues, rather than the array of possibilities that MIMIC methodology offers.

For example, no published studies have examined the efficacy of MIMIC DIF methods with more than two grouping variables and interactions. Second, there have been no large-scale evaluations of free and constrained baseline tests for nonuniform DIF when an unbiased anchor item isn't chosen *a priori*; choosing a problematic anchor item could completely undermine the benefits of the free baseline process. Third, very little research has examined the robustness of MIMIC methods to violations of the equal variance assumption for the common factor (θ) across groups. Violations of that assumption could inflate Type I error regardless of how the baseline model is specified. These unexplored issues motivated the following large-scale simulation study.

METHOD

Study Design

Unlike most DIF simulations which have focused on pairwise group comparisons, MIMIC DIF detection was investigated here using *four* comparison groups, resulting from the co-occurrence of gender (male, female) and ethnic group status (majority, minority). "Male-Majority (MMA)" was served as a reference group. "Male-Minority (MMI)," "Female-Majority (FMA)," and "Female-Minority (FMI)" were served as focal groups whose item parameters were manipulated to create DIF associated with gender (G), ethnicity (E), and gender by ethnicity interactions (GxE).

For comparability with previous research, and because many variables were examined, scale length was fixed at 15 items, and generating item parameters were based on the non-DIF and small DIF conditions of Stark et al. (2006). In non-DIF conditions, reference group item parameters were used to generate both reference and focal group item responses, based on a categorical MGCFA model (Muthén & Asparouhov, 2002), via Mplus 7.11 (Muthén & Muthén, 2013) scripts run using SAS PROC IML (SAS Institute, 2010). In all cases, responses to 15 items were generated based on Equations A1 and A2 of Appendix A, which describe the data generation process. In DIF conditions, DIF was simulated on items 3, 8, 11, and 15 by decreasing focal group loadings by 0.15 (nonuniform DIF) or by increasing focal group thresholds by 0.25 (uniform DIF).

Tables A1-A6 of Appendix A present the reference and focal group generating item parameters for this study. These parameters were used to create the main, marginal, and interactive effects associated with the grouping variables shown in Figure A1 and A2 of Appendix A.

Independent Variables

176 independent experimental conditions were created by manipulating seven independent variables:

1. Response categories: 2, 5.
2. Sample size per group (MMA = MMI = FMA = FMI): 125, 250.
3. Impact: none ($\mu_{Male} = \mu_{Female} = 0$), 0.5 SD ($\mu_{Male} = 0, \mu_{Female} = -0.5$).
4. Factor variance: equal ($\psi_{Male} = \psi_{Female} = 1$), unequal ($\psi_{Male} = 1, \psi_{Female} = 0.7$).
5. Type of DIF: none, nonuniform (λ), uniform (τ).
6. Source of DIF: G, G E, GxE, G GxE, G E GxE.
7. Baseline model: constrained, free, sequential-free.

Variable 6 was nested within the last two levels of variable 5. Variable 7 was completely nested. That is, the same data sets were used for the constrained, free, and sequential-free baseline conditions. In each condition, 100 data sets were generated for the comparison groups using the item parameters in Appendix A and, assuming normality, the factor means and variances associated with independent variables 3 and 4, respectively.

Dependent Variables

Type I error and power were the primary dependent variables. *Type I error* is defined as the proportion of items erroneously flagged as DIF (i.e., false positives) averaged over replications. P*ower* is defined as the proportion of items correctly identified as DIF (i.e., hits) averaged over replications.

In addition, to see whether MIMIC methodology could accurately identify the source of DIF, outcomes known as Type III errors (Mosteller, 1948) were recorded. Here, *Type III error* is defined as the number of times the source of DIF was misidentified, divided by the number of hits, averaged over replications.

Analysis Details

MIMIC analyses were performed using Mplus 7.11 (Muthén & Muthén, 2013). The robust maximum likelihood estimation (MLR) option was chosen so that the "XWITH" command could be used to model interactions between latent and grouping variables for nonuniform DIF tests. Model identification and standardization were accomplished by fixing the intercept of the common factor (θ) to 0 and fixing the variance of the common factor to 1 in the reference and focal groups. Because all items other than the one under investigation can be used to anchor the metric in constrained baseline analyses, every item was tested for DIF. However, free-baseline tests required an explicit referent for the model comparisons, so one item had to be left out of the DIF analyses.

For comparability with previous research by Stark et al. (2006), and to explore a recommendation by Lopez-Rivas et al. (2009) concerning the choice of anchor items for free baseline DIF analyses, this simulation examined MIMIC DIF detection using the following procedure. First, *constrained* baseline DIF analyses were performed on Items 1-15. In the *free* baseline conditions, Item 1, a discriminating non-DIF item, served as the designated referent for exploring DIF detection under the desirable, uncontaminated-one-item-anchor scenario. In the *sequential-free baseline* conditions, the most discriminating item identified as non-DIF in the constrained baseline analyses was chosen as an anchor for subsequent free baseline tests on each replication, and the 14 remaining items were analyzed. The sequential-free baseline conditions thus provided a more realistic picture of MIMIC performance than the free baseline conditions, because a contaminated anchor could have been chosen.

The constrained, free, and sequential-free baseline analyses were performed in accordance with the procedures described in connection with Figure 1 and Figure 2b. Omnibus likelihood ratio (LR) tests were conducted for uniform and nonuniform DIF based on the Satorra and Bentler (2001) method for nested model testing with scaled chi-squares, which adjusts for potential bias due to multivariate nonnormality (Bryant & Satorra, 2012):

$$\chi^2_{DIFF} = \frac{-2 \times (loglikelihood_{reduced} - loglikelihood_{full})}{C_{LR}}, \text{where}$$

$$C_{LR} = \frac{(q_{reduced})(c_{reduced}) - (q_{full})(c_{full})}{(q_{reduced}) - (q_{full})}, \quad (1)$$

q is the number of model parameters, c is a scaling correction factor for MLR chi-squares reported in the Mplus output, C_{LR} is the scaling factor for the chi-square difference statistic, and

subscripts *reduced* and *full* refer to the reduced (compact) and full (augmented) models. For every studied item, the observed χ^2_{DIFF} was compared to a critical chi-square (12.59) corresponding to p = .05 and 6 degrees of freedom (df), because there were three covariates (G, E, GxE) and three interactions with the common factor (θ*G, θ*E, θ*GxE). If the observed χ^2_{DIFF} exceeded the critical chi-square, the item was flagged as DIF.

The source of DIF was investigated by performing additional nested model comparisons following the omnibus DIF test. Beginning with a baseline model that contained paths from all grouping variables to a studied item, three reduced models were formed by successively deleting paths from the GxE, G, and E grouping variables in that order, and evaluating the χ^2_{DIFF} statistics with respect to a critical chi-square of 5.99 based on 2 df and critical p=.05. If a statistically significant result occurred for any grouping variable other than the true source of DIF, a Type III error was recorded.

Hypotheses (H) and Research Questions (Q)

Based on theoretical assumptions and previous research, the following hypotheses were formulated:

H1: Higher power and lower Type I error will be observed in the free baseline conditions than in the sequential-free and constrained baseline conditions.

- Rationale: The free baseline method has shown excellent performance in previous studies using a DIF free one-item anchor. The possibility of contaminated one-item anchors in the sequential-free baseline conditions may reduce power and increase Type I error. Because Type I error is typically high with the constrained baseline method, power results may be spurious and must be interpreted with caution.

H2: Power will be higher in the N = 500 conditions than in the N = 250 conditions.

- Rationale: Parameter estimation generally improves with sample size, so power to detect DIF across comparison groups should increase accordingly.

H3: Higher power will be observed when DIF is due to differences in item thresholds than item loadings.

- Rationale: Stark et al. (2006) found higher power for omnibus DIF detection with MACS when DIF was due to differences in item thresholds rather than loadings.

H4: Power will be higher for detecting DIF with polytomous data than with dichotomous data.

- Rationale: Kim and Yoon (2010) found higher power for MACS DIF detection with polytomous data than with dichotomous data.

H5: Type I error will be higher in the unequal factor variance conditions than in the equal variance conditions.

- Rationale: Unequal factor variance is a violation of MIMIC assumptions (Woods & Grimm, 2011), which causes model misspecification that will inflate Type I error.

No differences in power or Type I error were expected across impact and no-impact conditions, because MIMIC methods account for latent mean differences explicitly in the baseline model, and impact is rarely a significant factor in CFA or IRT DIF studies. In addition, no hypotheses were proposed concerning Type III error for sources of DIF, because it appears this study is the first involving CFA DIF methods to examine that question.

Analyses

ANOVA was used to test for main effects and interactions involving up to three independent variables and omega-square (ω^2) effect size statistics were reported, where .01, .06, and .14 represent small, medium, and large effects, respectively (Cohen, 1998). The specific hypotheses, H1-H5, were tested using planned comparisons with p = .05 for statistical significance.

RESULTS

Table 1 shows the Type I error rates for all no-DIF conditions. As expected, Type I error rates were near the nominal level (.05) in the equal factor variance conditions, with none exceeding .06. On the contrary, in the unequal factor variance conditions, which violated MIMIC assumptions, Type I error was often above .05, with many values exceeding .09 and one reaching .14. Interestingly, better results were observed here in the constrained baseline conditions, but this finding did not hold when DIF was simulated (see Tables 2 through 5).

Tables 2 through 5 present results for the remaining 160 DIF conditions. Tables 2 and 3 show the findings for the equal factor variance dichotomous and polytomous conditions, and Tables 4 and 5 show the results for the respective unequal factor variance conditions.

Overall, it can be seen that Type I error was lower in the equal variance conditions, Type I error was highest in the constrained baseline conditions, power improved as sample size increased, and power was higher for detecting DIF on thresholds than loadings. All of these findings were consistent with expectations. Importantly, similar power was observed in the sequential-free and free baseline conditions, indicating the viability of the sequential-free method in the absence of *a priori* information for choosing a referent. In fact, a detailed examination of individual simulation runs indicated that a DIF item was inappropriately chosen as an anchor only 1% of the time.

Further inspection of the results in Tables 2 and 3 revealed that, in accordance with expectations, Type I error rates in the equal variance conditions were markedly lower for the free

and sequential-free baseline methods. Whereas only a few values reached .08 in these conditions, values were frequently high in constrained baseline conditions and reached a maximum of .36 in the polytomous *N*=250 conditions with DIF due to main effects and interactions. Also consistent with expectations, power was higher in polytomous conditions than in dichotomous conditions, with power to detect DIF on loadings and thresholds using the sequential-free method averaging .50 and .85, respectively, in polytomous and .53 and .62 in dichotomous.

Detailed inspection of the results in Tables 4 and 5 reveals a similar pattern. The most noteworthy finding is the substantially higher Type I error that occurred in all unequal factor variance conditions. Clearly, violating the equal variance assumption of MIMIC methodology can lead to spurious DIF detection, so a stricter statistical criterion for flagging DIF items is warranted when model violations are suspected.

To support these results interpretations and to address the specific hypotheses of this study, separate ANOVAs and planned comparisons were conducted on the Type I error and power results. Tables 6 and 7 show the outcomes of the F-tests and ω^2 effect size statistics for the independent variables and interactions that accounted for at least 1% of the variance in the dependent variables.

All of the manipulated factors were statistically significant ($p < .05$). The baseline model accounted for 13% of the variance in Type I error and 1% of the variance in power. As expected, Type I error was much higher for the constrained baseline method than the free- and sequential-free baseline methods ($\omega^2 = .13$). Power was highest overall for the free baseline method, followed by the sequential-free, and constrained baseline methods ($\omega^2 = .01$), which supports Hypothesis 1.

Hypotheses 2 and 4 were also supported. As seen in Table 7, sample size had the largest effect on power ($\omega^2 = .32$), followed by the type of DIF ($\omega^2 = .23$). As expected, power was higher in the *N* = 250 conditions than in the *N* = 125 conditions. At the same time, power was higher for detecting DIF on thresholds than DIF on loadings.

Hypothesis 3 proposed that power would be higher for detecting DIF in polytomous conditions than in in dichotomous conditions. This hypothesis was supported, with the number of response categories accounting for 4% of the variance in power. Also, the ANOVA results in Table 7 showed a significant interaction ($\omega^2 = .07$) between response categories and type of DIF. The highest power was observed in the polytomous conditions when DIF was simulated on thresholds, and the lowest power was found in the polytomous conditions when DIF was simulated on loadings.

Hypothesis 5 was also supported. As expected, Type I error rates were higher in the unequal factor variance conditions than in the equal variance conditions ($\omega^2 = .11$). In addition, there was a statistically significant interaction between factor variance and baseline model ($\omega^2 = .05$). (Incidentally, note also that impact had no statistically significant effect on Type I error or power.)

In addition to testing the proposed hypotheses, the efficacy of MIMIC methods for identifying the source(s) of DIF was examined by performing ANOVA tests on the Type III errors. It is important to note that Type III error rates were not calculated for the "G E GxE" conditions, because DIF was simulated based on all of the sources (i.e., gender, ethnicity, and their interaction)..

Table 8 presents the Type III error ANOVA results for main effects and interactions. As before, only factors that accounted for at least 1% of the variance in the dependent variable are shown. This analysis revealed a remarkably large effect for the baseline model ($\omega^2 = .68$) and significant but smaller effects for the source of DIF and the interaction of the source with the baseline model ($\omega^2 = .15$ and .13, respectively).

To help shed light on these findings, Type III error results for the combinations of baseline models and sources of DIF are shown in Table 9. Overall the Type III error rates were extremely high, suggesting that although MIMIC methods are effective in detecting DIF, they are not effective in identifying the underlying source. Moreover, the free and sequential-free baseline methods, which were more effective for DIF detection, performed much worse than the constrained baseline method in this comparison. Clearly, more methodological work is needed if one wishes to develop and use MIMIC DIF detection methods for this purpose.

DISCUSSION

In recent years, MIMIC methods have been suggested for DIF detection in situations involving multiple groups (Kim et al., 2012; Woods, 2009). However, until now, there has been no hard evidence to support that practice because the simulation studies showing MIMIC efficacy have used two-group designs (Finch, 2005; Kim et al., 2012; Wang & Shih, 2010; Woods, 2009). MIMIC methods have also been discussed as a way of examining the effects on item responses due to background variables and interactions, but little, if any, research has examined MIMIC accuracy for detecting DIF due to multiple sources, such as the co-occurrence of gender and ethnicity in samples, where the underlying causes of gender and ethnic DIF may be different. Furthermore, since MIMIC methods were extended for nonuniform DIF detection by Woods and Grimm (2012), there have been no published follow-up studies, and there have been no studies examining CFA DIF evaluations of the sequential-free baseline strategy outlined by Lopez-Rivas et al. (2009).

This study aimed to address needs in all of these areas. MIMIC DIF detection was explored using four groups, defined by combinations of gender and ethnicity. Uniform and nonuniform DIF was simulated based on combinations of background variables and interactions. The robustness of MIMIC DIF detection to violations of the equal variance assumption was examined. And the efficacy of the more practical sequential-free baseline method was compared to the idealized free-baseline method with a DIF-free anchor and the often-used constrained baseline method.

As expected, the free and sequential-free baseline methods provided much better Type I error and power rates than the constrained baseline method, and importantly there was only a minor diminution in efficacy for the sequential-free baseline, relative to the free baseline, because a DIF item was rarely selected as an anchor. Also, as expected, MIMIC performed markedly better when the equal factor variance assumption was satisfied. When the assumption was violated, Type I error was above .05 regardless of the baseline model specification. Consequently, a correction or stricter criterion for flagging items for DIF is needed when unequal variances are suspected. Interestingly, despite its intuitive appeal for identifying sources of DIF, MIMIC was highly ineffective in this regard. Although DIF items were identified correctly in many conditions, the Type III error results clearly showed that drawing substantive conclusions about underlying source(s) of DIF was problematic.

It is hoped that future research will build on the design and results of this investigation. It would be worthwhile, for examine, to examine how MIMIC DIF detection efficacy varies as a function of the difference in factor variances across comparison groups. This study explored only one possibility (0.3SD difference) due to the large number of conditions, but smaller variance differences may be more realistic and have a smaller adverse effect on performance. It may also be worthwhile to examine DIF detection with varying degrees of DIF across items, rather than using a magnitude chosen in previous studies for comparability. Future research should also examine why MIMIC methods could not reliably detect the source of DIF when a DIF item was correctly identified. Perhaps the omnibus tests used here were problematic, but more effective processes can be developed.

Altogether, this research conclusively shows the MIMIC DIF detection methods provide a powerful alternative to multi-group factor analysis and traditional item response theory DIF

approaches. The sequential-free baseline method is not only easy to implement, but also effective in detecting nonuniform and uniform DIF across multiple examinee groups. And it appears fairly straightforward to extend the methods to tests that may involve more than one dimension by design. Although Mplus is currently the only commercial software that allows for interactions between factors underlying test performance and background variables, other software packages are likely to incorporate those capabilities over time and thus expand the array of tools available to practitioners for item screening and test revision.

TABLE 1. Type I Error in No DIF Conditions

Factor Variance	Response Categories	*N*	Impact	Baseline		
				Constrained	Free	Sequential-Free
Equal	Dichotomous	125	none	.05	.05	.05
			0.5 SD	.06	.06	.06
		250	none	.04	.04	.04
			0.5 SD	.06	.05	.04
	Polytomous	125	none	.06	.06	.05
			0.5 SD	.05	.05	.04
		250	none	.05	.05	.05
			0.5 SD	.05	.04	.04
Unequal	Dichotomous	125	none	.05	.07	.06
			0.5 SD	.06	.08	.08
		250	none	.06	.09	.09
			0.5 SD	.05	.09	.10
	Polytomous	125	none	.05	.09	.09
			0.5 SD	.06	.10	.09
		250	none	.06	.13	.14
			0.5 SD	.09	.05	.13

**Note.* *N* = sample size per group; DIF = differential item functioning.

TABLE 2. Power And Type I Error In Dichotomous, Equal Variance Conditions

Source of DIF	*N*	Type of DIF	Impact	Baseline					
				Constrained		Free		Sequential-Free	
				Power	Type I	Power	Type I	Power	Type I
G	125	Threshold	none	.36	.08	.35	.05	.31	.06
			0.5 SD	.35	.10	.38	.07	.30	.08
		Loading	none	.29	.06	.31	.05	.29	.04
			0.5 SD	.28	.07	.32	.06	.31	.06
	250	Threshold	none	.66	.10	.67	.05	.65	.04
			0.5 SD	.62	.11	.62	.04	.58	.05
		Loading	none	.48	.05	.55	.05	.55	.04
			0.5 SD	.50	.07	.61	.05	.60	.04
GE	125	Threshold	none	.69	.12	.71	.05	.69	.07
			0.5 SD	.66	.14	.69	.06	.61	.06
		Loading	none	.39	.06	.44	.05	.43	.04
			0.5 SD	.43	.07	.52	.05	.53	.05
	250	Threshold	none	.96	.18	.97	.05	.93	.04
			0.5 SD	.91	.17	.92	.05	.88	.05
		Loading	none	.69	.06	.77	.05	.77	.04
			0.5 SD	.75	.07	.81	.05	.80	.05
GxE	125	Threshold	none	.40	.09	.39	.05	.37	.05
			0.5 SD	.44	.10	.42	.06	.40	.07
		Loading	none	.23	.05	.26	.05	.26	.05
			0.5 SD	.28	.07	.28	.06	.26	.06
	250	Threshold	none	.68	.10	.68	.05	.63	.04
			0.5 SD	.64	.12	.69	.05	.61	.06
		Loading	none	.51	.06	.58	.05	.57	.05
			0.5 SD	.52	.07	.59	.05	.59	.05
G GxE	125	Threshold	none	.47	.10	.43	.05	.42	.06
			0.5 SD	.49	.12	.46	.07	.38	.08
		Loading	none	.28	.05	.33	.05	.31	.04
			0.5 SD	.31	.07	.34	.05	.35	.05
	250	Threshold	none	.74	.12	.76	.05	.71	.04
			0.5 SD	.72	.13	.72	.05	.68	.08
		Loading	none	.53	.06	.59	.05	.59	.04
			0.5 SD	.55	.06	.64	.05	.64	.04
GE GxE	125	Threshold	none	.73	.13	.77	.05	.70	.05
			0.5 SD	.71	.14	.75	.06	.64	.07
		Loading	none	.46	.06	.49	.05	.50	.05
			0.5 SD	.48	.07	.55	.05	.55	.05
	250	Threshold	none	.97	.21	.99	.05	.97	.04
			0.5 SD	.94	.19	.96	.05	.95	.04
		Loading	none	.74	.06	.82	.05	.82	.04
			0.5 SD	.78	.08	.85	.04	.85	.04

TABLE 3. Power and Type I Error in Polytomous, Equal Variance Conditions

Source of DIF	*N*	Type of DIF	Impact	Baseline					
				Constrained		Free		Sequential-Free	
				Power	Type I	Power	Type I	Power	Type I
G	125	Threshold	none	.61	.12	.63	.06	.59	.05
			0.5 SD	.56	.12	.63	.05	.57	.06
		Loading	none	.22	.07	.25	.06	.23	.06
			0.5 SD	.25	.07	.28	.06	.26	.05
	250	Threshold	none	.90	.18	.95	.05	.93	.04
			0.5 SD	.88	.16	.93	.04	.90	.04
		Loading	none	.46	.06	.49	.05	.49	.05
			0.5 SD	.47	.06	.49	.05	.48	.04
GE	125	Threshold	none	.91	.18	.94	.06	.89	.05
			0.5 SD	.90	.20	.94	.06	.90	.05
		Loading	none	.41	.08	.45	.06	.44	.05
			0.5 SD	.44	.07	.46	.06	.46	.05
	250	Threshold	none	1.00	.32	1.00	.05	.99	.04
			0.5 SD	.99	.32	1.00	.04	1.00	.05
		Loading	none	.76	.07	.79	.05	.77	.05
			0.5 SD	.77	.08	.81	.04	.80	.04
GxE	125	Threshold	none	.67	.13	.69	.06	.63	.06
			0.5 SD	.68	.13	.71	.05	.65	.06
		Loading	none	.20	.07	.24	.06	.23	.05
			0.5 SD	.26	.07	.27	.06	.26	.05
	250	Threshold	none	.93	.17	.97	.05	.94	.04
			0.5 SD	.93	.18	.96	.04	.91	.03
		Loading	none	.45	.06	.50	.05	.49	.04
			0.5 SD	.53	.06	.57	.05	.57	.04
G GxE	125	Threshold	none	.74	.14	.77	.05	.70	.06
			0.5 SD	.73	.15	.78	.05	.69	.06
		Loading	none	.28	.07	.29	.06	.29	.05
			0.5 SD	.66	.07	.32	.05	.31	.05
	250	Threshold	none	.96	.22	.97	.05	.96	.04
			0.5 SD	.94	.21	.97	.04	.95	.04
		Loading	none	.52	.06	.57	.06	.56	.05
			0.5 SD	.56	.06	.58	.04	.58	.04
GE GxE	125	Threshold	none	.93	.20	.97	.05	.93	.05
			0.5 SD	.94	.21	.96	.05	.93	.04
		Loading	none	.44	.08	.51	.06	.48	.05
			0.5 SD	.49	.08	.53	.06	.52	.05
	250	Threshold	none	1.00	.36	1.00	.05	1.00	.04
			0.5 SD	1.00	.37	1.00	.04	1.00	.04
		Loading	none	.79	.08	.84	.05	.84	.05
			0.5 SD	.83	.10	.88	.04	.88	.04

TABLE 4. Power and Type I Error in Dichotomous, Unequal Variance Conditions

Source of DIF	N	Type of DIF	Impact	Baseline					
				Constrained		Free		Sequential-Free	
				Power	Type I	Power	Type I	Power	Type I
G	125	Threshold	none	.41	.09	.42	.06	.37	.07
			0.5 SD	.36	.10	.39	.08	.31	.10
		Loading	none	.34	.05	.48	.07	.48	.06
			0.5 SD	.34	.07	.53	.09	.54	.08
	250	Threshold	none	.68	.12	.76	.10	.70	.10
			0.5 SD	.63	.13	.69	.09	.64	.11
		Loading	none	.60	.04	.80	.09	.80	.09
			0.5 SD	.63	.05	.83	.09	.83	.09
GE	125	Threshold	none	.71	.13	.79	.07	.70	.07
			0.5 SD	.66	.14	.71	.08	.66	.10
		Loading	none	.47	.06	.60	.08	.60	.07
			0.5 SD	.47	.07	.64	.08	.65	.08
	250	Threshold	none	.96	.19	.98	.10	.95	.10
			0.5 SD	.93	.18	.95	.09	.90	.11
		Loading	none	.76	.05	.86	.09	.86	.09
			0.5 SD	.81	.06	.89	.09	.89	.08
GxE	125	Threshold	none	.45	.10	.43	.07	.40	.08
			0.5 SD	.43	.10	.19	.20	.36	.09
		Loading	none	.22	.06	.30	.08	.31	.07
			0.5 SD	.26	.08	.34	.08	.33	.08
	250	Threshold	none	.70	.11	.72	.10	.71	.10
			0.5 SD	.71	.11	.74	.09	.69	.10
		Loading	none	.48	.07	.58	.10	.57	.10
			0.5 SD	.54	.07	.64	.09	.64	.09
G GxE	125	Threshold	none	.52	.10	.55	.07	.49	.09
			0.5 SD	.49	.11	.50	.08	.43	.09
		Loading	none	.34	.05	.48	.07	.46	.07
			0.5 SD	.38	.06	.53	.08	.55	.08
	250	Threshold	none	.79	.14	.87	.10	.81	.10
			0.5 SD	.73	.13	.78	.09	.73	.10
		Loading	none	.61	.05	.79	.09	.79	.09
			0.5 SD	.67	.05	.82	.09	.82	.09
GE GxE	125	Threshold	none	.42	.14	.83	.07	.74	.08
			0.5 SD	.72	.14	.77	.08	.68	.10
		Loading	none	.50	.06	.60	.08	.60	.07
			0.5 SD	.49	.06	.66	.08	.66	.08
	250	Threshold	none	.98	.22	.99	.10	.97	.11
			0.5 SD	.96	.21	.98	.09	.95	.10
		Loading	none	.76	.05	.88	.10	.88	.10
			0.5 SD	.80	.06	.91	.09	.91	.08

TABLE 5. Power and Type I Error in Polytomous, Unequal Variance Conditions

Source of DIF	N	Type of DIF	Impact	Baseline					
				Constrained		Free		Sequential-Free	
				Power	Type I	Power	Type I	Power	Type I
G	125	Threshold	none	.62	.12	.70	.11	.67	.11
			0.5 SD	.58	.13	.68	.11	.62	.12
		Loading	none	.26	.06	.47	.11	.45	.10
			0.5 SD	.26	.06	.46	.10	.45	.10
	250	Threshold	none	.93	.17	.97	.13	.96	.12
			0.5 SD	.88	.18	.94	.12	.89	.14
		Loading	none	.47	.04	.75	.14	.74	.13
			0.5 SD	.52	.05	.82	.13	.82	.13
GE	125	Threshold	none	.91	.20	.95	.10	.91	.09
			0.5 SD	.89	.20	.94	.11	.90	.11
		Loading	none	.38	.07	.56	.10	.55	.09
			0.5 SD	.45	.06	.56	.11	.56	.11
	250	Threshold	none	1.00	.33	1.00	.13	1.00	.13
			0.5 SD	.99	.34	1.00	.12	.99	.14
		Loading	none	.74	.06	.89	.13	.88	.12
			0.5 SD	.74	.06	.90	.13	.91	.12
GxE	125	Threshold	none	.68	.14	.74	.10	.67	.10
			0.5 SD	.68	.15	.75	.11	.70	.11
		Loading	none	.19	.06	.24	.10	.23	.09
			0.5 SD	.21	.06	.27	.11	.27	.10
	250	Threshold	none	.93	.17	.97	.13	.96	.13
			0.5 SD	.92	.18	.97	.12	.96	.11
		Loading	none	.41	.06	.53	.13	.52	.13
			0.5 SD	.49	.06	.60	.12	.36	.08
G GxE	125	Threshold	none	.75	.15	.82	.11	.77	.11
			0.5 SD	.73	.15	.81	.11	.76	.12
		Loading	none	.27	.06	.45	.11	.45	.10
			0.5 SD	.28	.05	.49	.11	.48	.10
	250	Threshold	none	.96	.21	.99	.13	.99	.13
			0.5 SD	.93	.23	.98	.12	.95	.13
		Loading	none	.53	.05	.77	.14	.76	.14
			0.5 SD	.58	.06	.81	.13	.81	.13
GE GxE	125	Threshold	none	.94	.21	.97	.10	.95	.09
			0.5 SD	.93	.22	.96	.11	.93	.11
		Loading	none	.37	.06	.56	.10	.56	.09
			0.5 SD	.47	.06	.63	.10	.62	.10
	250	Threshold	none	1.00	.37	1.00	.13	1.00	.14
			0.5 SD	1.00	.39	1.00	.13	1.00	.14
		Loading	none	.71	.07	.89	.13	.89	.12
			0.5 SD	.78	.07	.92	.12	.92	.12

TABLE 6. ANOVA Results for Type I Error

Source	df_B	F	ω²
Type of DIF (D)	1	1586.08	.13
Baseline model (B)	2	761.19	.13
Factor variance (V)	1	1297.63	.11
Response categories (R)	1	516.56	.04
Source of DIF (S)	4	56.23	.02
Sample size per group (N)	1	176.39	.02
D*B	2	1242.56	.21
V*B	2	320.26	.05
S*B	8	68.90	.05
D*S	4	45.87	.02
R*B	2	90.76	.02
N*B	2	75.78	.01
D*S*B	8	50.53	.03
D*R*B	2	187.86	.03
N*D*B	2	149.89	.03

**Note.* The listed independent variables are ones that accounted for at least 1% of the variance in the dependent variables. All effects shown were significant at $p < .05$. ω² = proportion of variance accounted for by the independent variables. df_B = degree of freedom between; for all effects, degrees of freedom within = 479.

TABLE 7. ANOVA Results for Power

Source	df_B	F	ω²
Sample size per group (N)	1	5458.65	.32
Type of DIF (D)	1	4013.43	.23
Source of DIF (S)	4	921.94	.21
Response categories (R)	1	678.33	.04
Factor variance (V)	1	229.45	.01
Baseline model (B)	2	112.82	.01
D*R	1	1219.38	.07

**Note.* The listed independent variables are ones that accounted for at least 1% of the variance in the dependent variables. All effects shown were significant at $p < .05$. ω² = proportion of variance accounted for by the independent variables. df_B = degree of freedom between; for all effects, degrees of freedom within = 479.

TABLE 8. ANOVA Results for Type III Error

Source	df_B	F	ω^2
Baseline model (B)	2	4842.32	.68
Source of DIF (S)	3	706.19	.15
S*B	6	315.55	.13

**Note.* All effects shown were significant at $p < .05$. ω^2 = proportion of variance accounted for by the independent variables. df_B = degree of freedom between; for all effects, degrees of freedom within = 383.

TABLE 9. Type III Error Rates for Sources of DIF and Baseline Model

	Baseline model		
Source of DIF	Constrained	Free	Sequential-Free
GxE	.85	1.00	1.00
G	.16	.98	.98
G GxE	.21	.99	.99
GE	.08	.82	.81

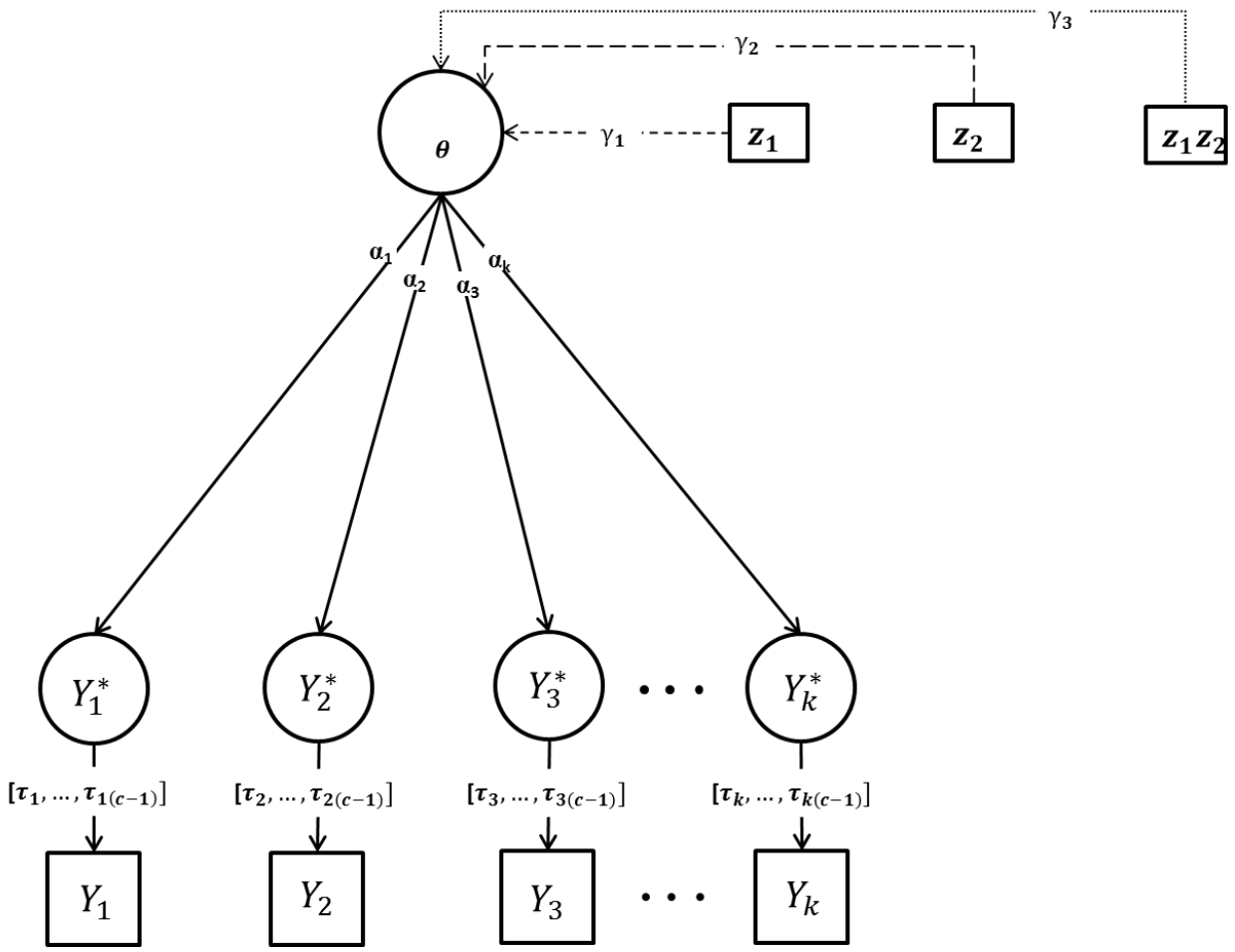

FIGURE 1. A Baseline Model for Constrained Baseline Approach

**Note.* γ_j = Mean difference by group membership (z_j) on the latent variable θ; items i = 1, 2, … , k; α_i = discrimination; τ_i = threshold; c is the number of categories in each item

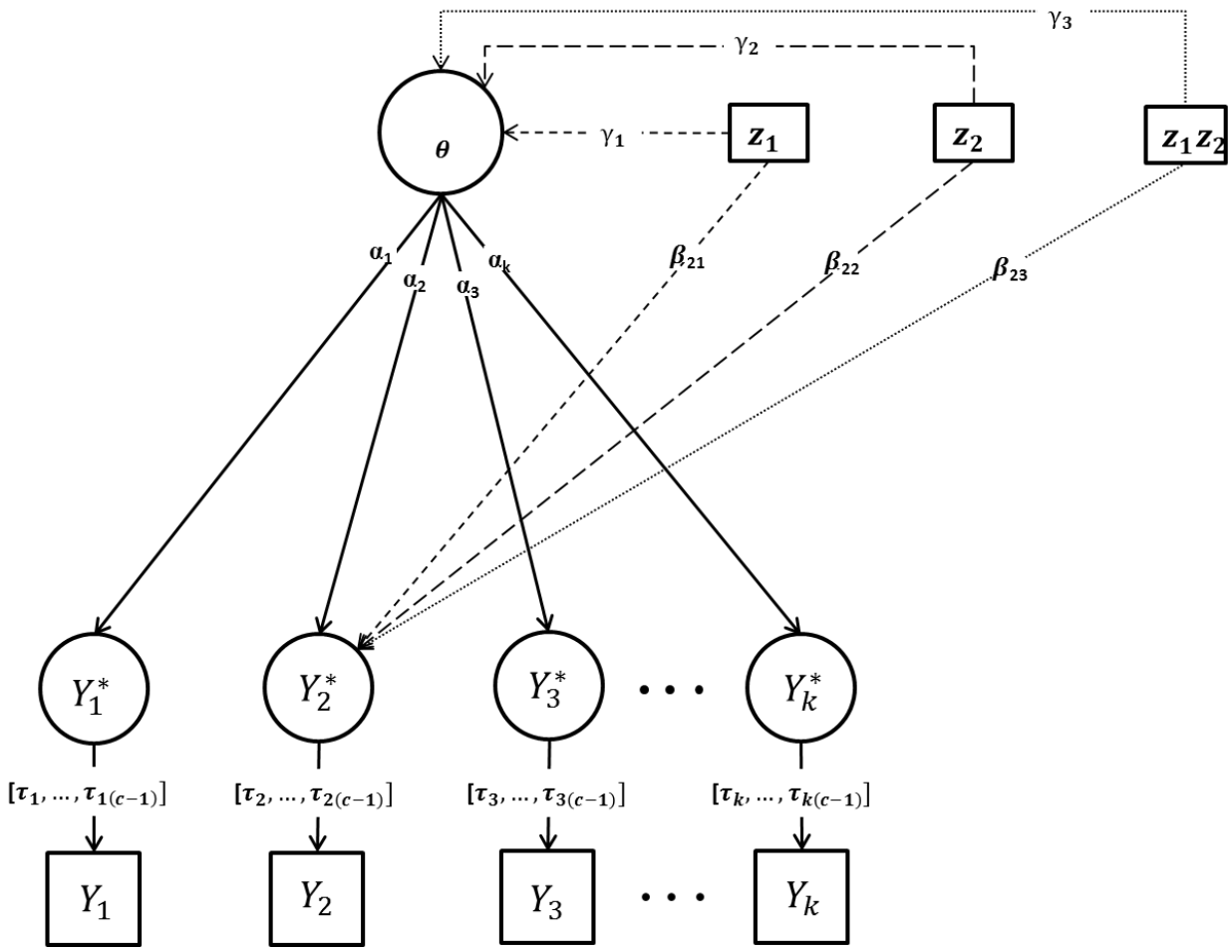

FIGURE 2a. The Full Model of Constrained Baseline Approach for Testing Uniform DIF on Item 2 of a Scale Containing *k* Items

**Note.* γ_j = Mean difference by group membership (z_j) on the latent variable θ; items i = 1, 2, … , k; α_i = discrimination; τ_i = threshold; c is the number of categories in each item; β_{ij} = group difference by z_j in the threshold of item i

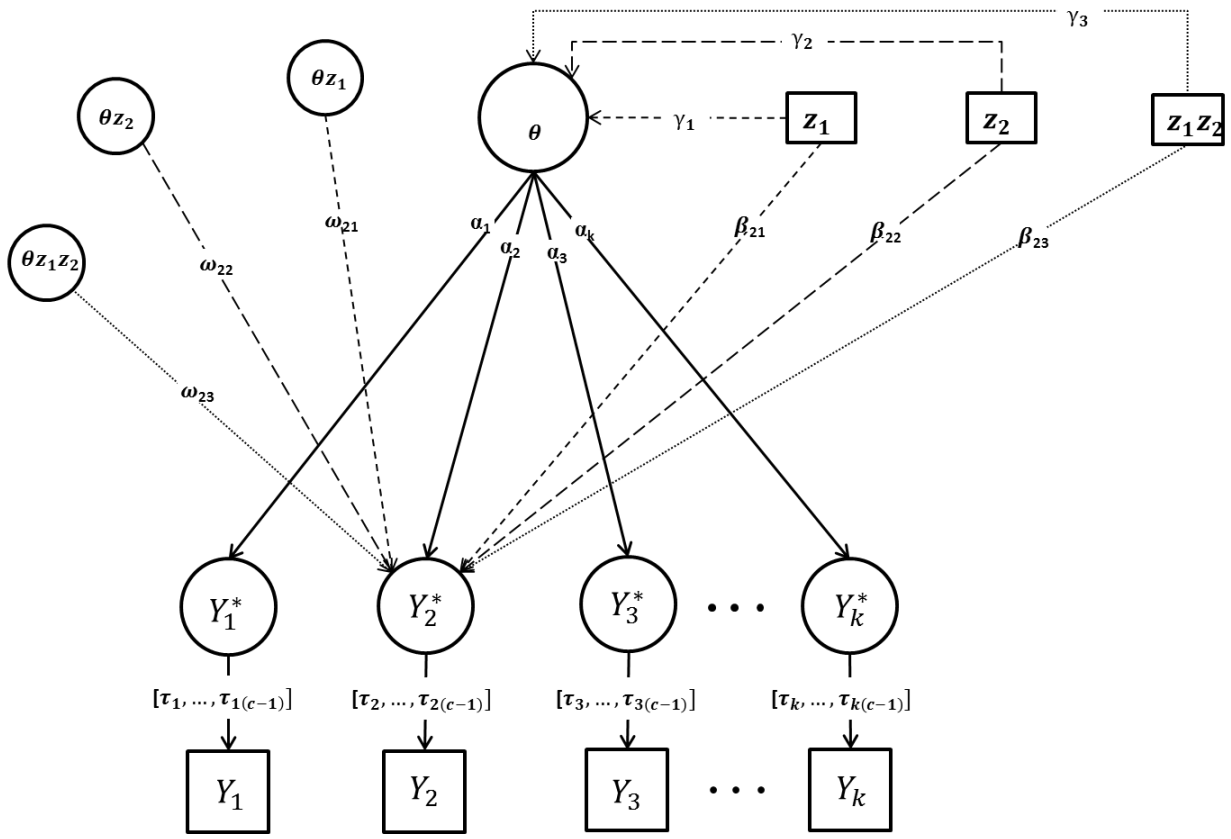

FIGURE 2b. The Full Model of Constrained Baseline Approach with Interaction between Grouping Variables and θ for Testing Uniform and Nonuniform DIF on Item 2 of a Scale Containing *k* Items

**Note.* γ_j = Mean difference by group membership (z_j) on the latent variable θ; items i = 1, 2, … , k; α_i = discrimination; ω_{ij} = nonuniform DIF effect on item i by z_j; τ_i = threshold; c is the number of categories in each item; β_{ij} = group difference by z_j in the threshold of item i

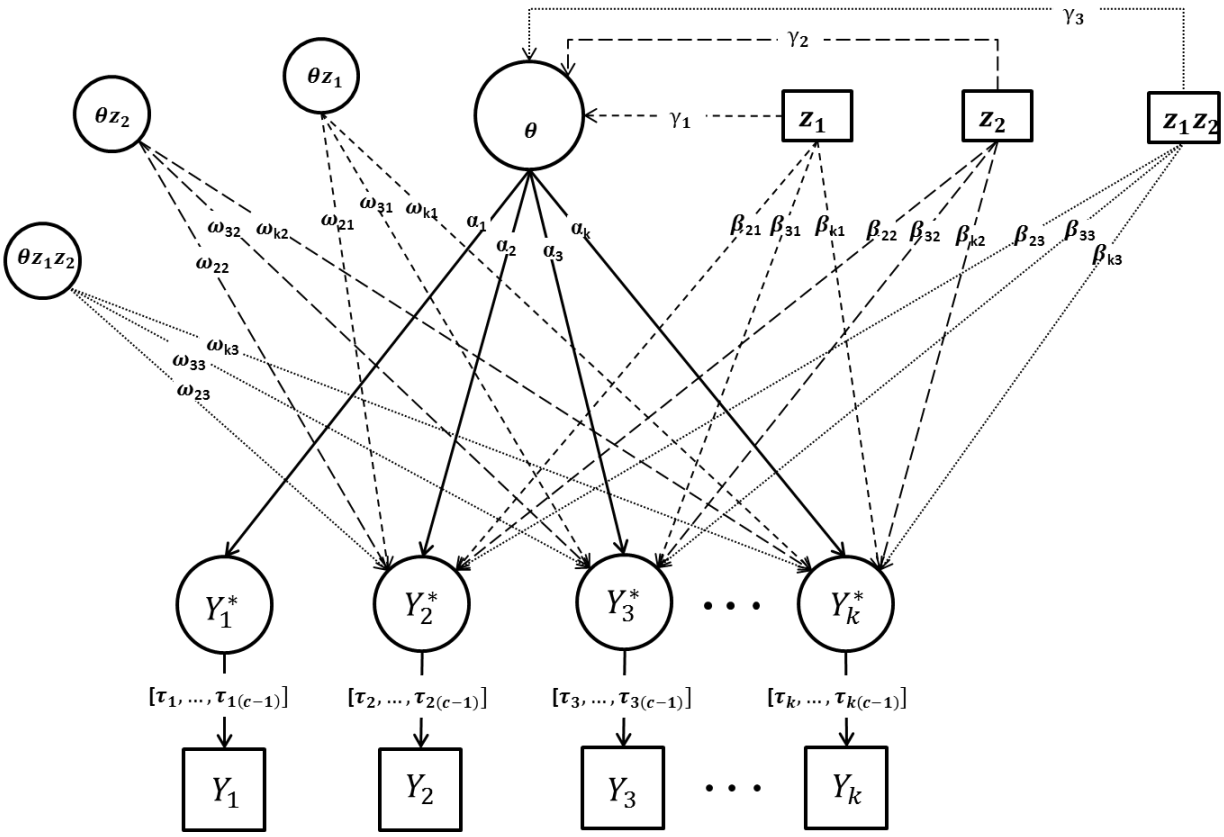

FIGURE 3a. A Baseline Model of a Single-Anchor Free Baseline Approach for Testing Uniform and Nonuniform DIF in which Item 1 is Used as the Anchor Item

**Note.* γ_j = Mean difference by group membership (z_j) on the latent variable θ; items i = 1, 2, … , k; α_i = discrimination; ω_{ij} = nonuniform DIF effect on item i by z_j; τ_i = threshold; c is the number of categories in each item; β_{ij} = group difference by z_j in the threshold of item i

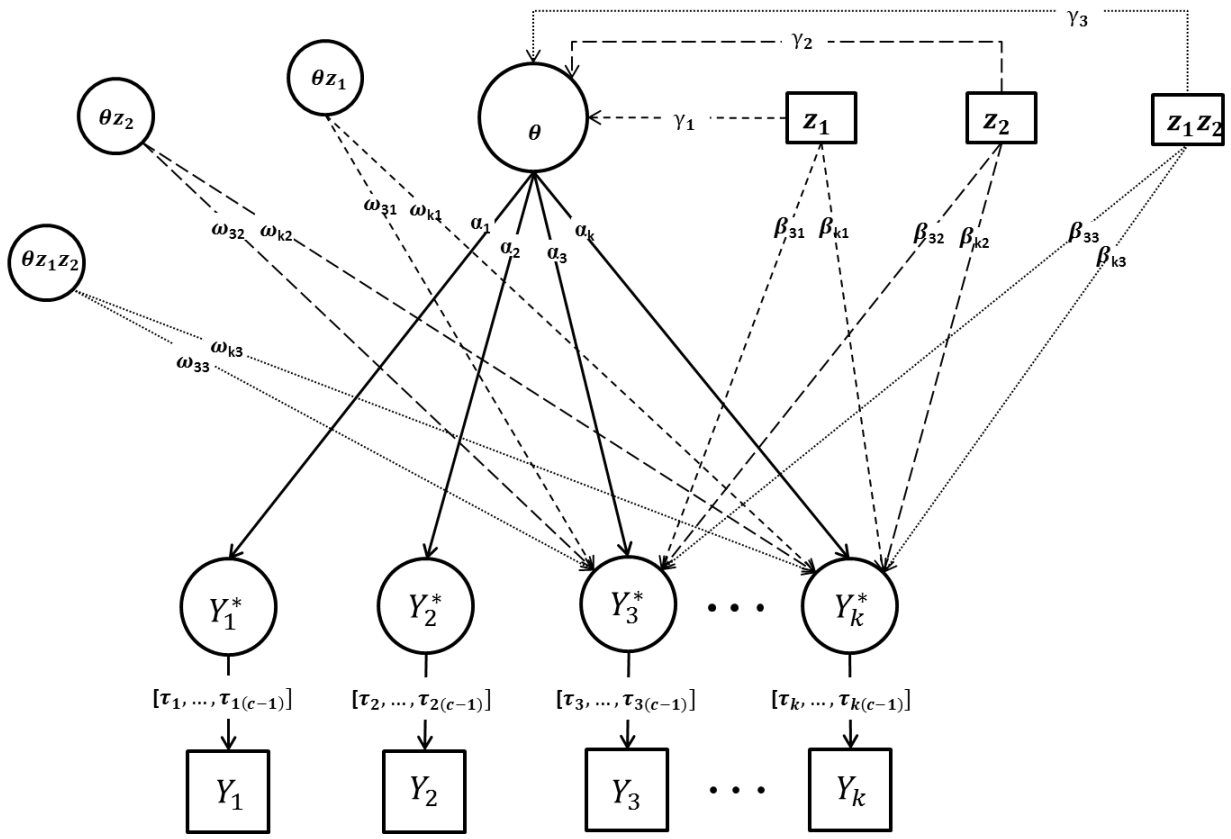

FIGURE 3b. The Compact Model of a Single-Anchor Free Baseline Approach for Testing Uniform and Nonuniform DIF on Item 2 in Which Item 1 is Used as the Anchor Item

**Note.* γ_j = Mean difference by group membership (z_j) on the latent variable θ; items i = 1, 2, … , k; α_i = discrimination; ω_{ij} = nonuniform DIF effect on item i by z_j; τ_i = threshold; c is the number of categories in each item; β_{ij} = group difference by z_j in the threshold of item i

REFERENCES

American Psychological Association (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bryant, F. B., & Satorra, A. (2012) Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*, 372-398.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16*, 129-147.

Chernyshenko, O. S., Stark, S., & Guenole, N. (2007). Can the discretionary nature of certain criteria lead to differential prediction across cultural groups? *International Journal of Selection and Assessment, 15*, 175 – 184.

Cleary, T. A. (1968). Test bias: prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5*, 115-124.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, *72*, 19-29.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70*, 631-639.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: a comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling, 18*, 212-228.

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC: Likelihood ratio test with a critical value adjustment. *Educational and Psychological measurement, 72*, 469-492.

Lopez-Rivas, G.E., Stark, S., & Chernyshenko, O.S. (2009). The effects of referent item parameters upon DIF detection using the free-baseline likelihood ratio test. *Applied Psychological Measurement, 33*, 251 – 265.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology, 97*, 1016-1031.

Mosteller, F. (1948). A k-sample slippage test for an extreme population. *The Annals of Mathematical Statistics, 19*, 58-65.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557-585.

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus*. Unpublished manuscript (Available at http://www.statmodel.com/mplus/examples/webnote.html#web4).

Muthén, L. K., & Muthén, B. O. (1998-2013). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén.

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

SAS Institute (2010). *SAS 9.3 user's guide*. SAS Institute, Cary, NC.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moments structure analysis. *Psychometrika, 66*, 507-514.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Shealy, R. & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects bias/DTF as well as item bias/DIF. *Psychometrika, 59*, 159-194.

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

Stark, S., Chernyshenko, O.S., Chan, K.Y., Lee, W.C., & Drasgow, F. (2001). Effects of the testing situation on item responding: cause for concern. *Journal of Applied psychology, 86*, 943-953.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: when are statistically significant effects practically important? *Journal of Applied Psychology, 89*, 497-508.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with CFA and IRT: Toward a unified strategy. *Journal of Applied Psychology, 19,* 1292-1306.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Erlbaum.

Uniform Guidelines on Employee selection procedures. 43 *Federal Register* 38290-38315 (1978).

Wang, W.-C. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement, 34,* 166-180.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44,* 1-27.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied psychological measurement, 35,* 339-361.

APPENDIX A

Response Data Generation

This section describes how to generate ordered-categorical response data for the 15 items in the proposed simulation. Tables A1-A6 show the reference and focal group generating item parameters.

Two- and five- category response data will be generated using the linear common factor model with an embedded threshold structure. In this framework, the ordered categorical observed scores (Y_{ij}) are construed as the manifestation of the underlying latent response variates (Y_{ij}^{*}). The latent response variates (Y_{ij}^{*}) are assumed to be continuous and linearly related with factor score (θ_i):

$$Y_{ij}^{*} = \lambda_i \theta_j + \varepsilon_{ij}, \tag{A1}$$

where *i* represents items, indexed *i* = 1, 2, …, 15; *j* represents simulated respondents (simulees); θ_j represents the factor score for respondent *j*; λ_i represents the loading of item *i* on the common factor θ; and ε_{ij} represents the unique factor score for simulee *j* on item *i*, which is obtained by 1 minus λ_i^2.

For each item response, Y_{ij} will be determined by comparing the latent response variates (Y_{ij}^{*}) with the item threshold (τ_i) as follows:

$$Y_{ij} = c, \text{if } \tau_{ic} < Y_{ij}^{*} \leq \tau_{ic+1}, \quad \text{(A2)}$$

where τ_{ic} indicates the *C* ordered categorical responses of item *i* ($\nu_{i0} = -\infty$; $\nu_{i(c+1)} = \infty$; and *c* = 0, 1, …, *C*-1). The number of thresholds of an item equals the number of categories of that item minus one (*C*-1). Any latent response variate (Y_i^*) that falls between the threshold for a response category (*c*) and the threshold for the next higher response category (*c*+1) is manifested as a response category, *c*.

TABLE A1. Generating Item Parameters for Reference Group in Dichotomous and Polytomous Conditions

	Response Categories					
	Both	Binary	Polytomous			
Item no.	λ	τ	τ_1	τ_2	τ_3	τ_4
1	.90	-.26	-1.21	-.57	-.26	.42
2	.66	-.06	-1.42	-.50	-.06	1.11
3	.83	-.42	-1.69	-.98	-.42	.64
4	.71	-.14	-1.28	-.46	-.14	.61
5	.77	-.37	-1.56	-.82	-.37	.54
6	.68	-.34	-1.72	-.84	-.34	.70
7	.58	-.48	-1.73	-.94	-.48	.36
8	.80	-.07	-1.18	-.42	-.07	.77
9	.85	-.30	-1.57	-.78	-.30	.71
10	.85	-.48	-1.54	-.91	-.48	.33
11	.82	-.27	-1.52	-.73	-.27	.72
12	.80	-.26	-1.51	-.63	-.26	.56
13	.85	-.03	-1.19	-.38	-.03	.86
14	.84	-.14	-1.35	-.54	-.14	.80
15	.86	-.27	-1.32	-.62	-.27	.48

**Note*. In the "none" (no-DIF) conditions, these same parameters will be used to generate item responses for focal groups. The reference group is "Male-Majority".

TABLE A2. Generating Parameters for DIF Items in Gender (G) Conditions

DIF Item no.	Groups	Response Categories					
		Both	Binary	Polytomous			
		λ	τ	τ_1	τ_2	τ_3	τ_4
3	MMA	.83	-.42	-1.69	-.98	-.42	.64
	MMI	.83	-.42	-1.69	-.98	-.42	.64
	FMA	.68	-.17	-1.44	-.73	-.17	.89
	FMI	.68	-.17	-1.44	-.73	-.17	.89
7	MMA	.58	-.48	-1.73	-.94	-.48	.36
	MMI	.58	-.48	-1.73	-.94	-.48	.36
	FMA	.43	-.23	-1.48	-.69	-.23	.61
	FMI	.43	-.23	-1.48	-.69	-.23	.61
11	MMA	.82	-.27	-1.52	-.73	-.27	.72
	MMI	.82	-.27	-1.52	-.73	-.27	.72
	FMA	.67	-.02	-1.27	-.48	-.02	.97
	FMI	.67	-.02	-1.27	-.48	-.02	.97
15	MMA	.86	-.27	-1.32	-.62	-.27	.48
	MMI	.86	-.27	-1.32	-.62	-.27	.48
	FMA	.71	-.02	-1.07	-.37	-.02	.73
	FMI	.71	-.02	-1.07	-.37	-.02	.73

**Note*. MMA (Male-Majority); MMI (Male-Minority); FMA (Female-Majority); FMI (Female-Minority); Male groups (MMA and MMI); Female groups (FMA and FMI); Majority groups (MMA and FMA); and Minority groups (MMI and FMI).
In "*G*" conditions, DIF is caused only by gender differences in item parameters. To create uniform DIF for gender, the average item threshold for females was obtained by adding 0.25 to the average threshold for males. To create nonuniform DIF for gender, the average loading for females was obtained by subtracting 0.15 from the average for males.

TABLE A3. Generating Parameters for DIF Items in Gender and Ethnicity (G E) Conditions

		Response Categories					
		Both	Binary	Polytomous			
Item no.	Groups	λ	τ	τ_1	τ_2	τ_3	τ_4
3	MMA	.83	-.42	-1.69	-.98	-.42	.64
	MMI	.68	-.17	-1.44	-.73	-.17	.89
	FMA	.68	-.17	-1.44	-.73	-.17	.89
	FMI	.53	.08	-1.19	-.48	.08	1.14
7	MMA	.58	-.48	-1.73	-.94	-.48	.36
	MMI	.43	-.23	-1.48	-.69	-.23	.61
	FMA	.43	-.23	-1.48	-.69	-.23	.61
	FMI	.28	.02	-1.23	-.44	.02	.86
11	MMA	.82	-.27	-1.52	-.73	-.27	.72
	MMI	.67	-.02	-1.27	-.48	-.02	.97
	FMA	.67	-.02	-1.27	-.48	-.02	.97
	FMI	.52	.23	-1.02	-.23	.23	1.22
15	MMA	.86	-.27	-1.32	-.62	-.27	.48
	MMI	.71	-.02	-1.07	-.37	-.02	.73
	FMA	.71	-.02	-1.07	-.37	-.02	.73
	FMI	.56	.23	-.82	-.12	.23	.98

**Note*. MMA (Male-Majority); MMI (Male-Minority); FMA (Female-Majority); FMI (Female-Minority); Male groups (MMA and MMI); Female groups (FMA and FMI); Majority groups (MMA and FMA); and Minority groups (MMI and FMI).
In "*G E*" conditions, DIF is caused by both gender and ethnicity differences in item parameters. To create uniform DIF for gender, the average item threshold for females was obtained by adding 0.25 to the average threshold for males. To create nonuniform DIF for gender, the average loading for females was obtained by subtracting 0.15 from the average for males. Similarly, to create uniform DIF for ethnicity, the average item threshold for the minority groups was obtained by adding 0.25 to the average threshold for the majority groups. To create nonuniform DIF for ethnicity, the average loading for the minority groups was obtained by subtracting 0.15 from the average loading for the majority groups.

Table A4. Generating Parameters for DIF Items in Gender by Ethnicity (GxE) Conditions

		Response categories					
		Both	Binary	Polytomous			
Item no.	Groups	λ	τ	τ_1	τ_2	τ_3	τ_4
3	MMA	.83	-.42	-1.69	-.98	-.42	.64
	MMI	.68	-.17	-1.44	-.73	-.17	.89
	FMA	.68	-.17	-1.44	-.73	-.17	.89
	FMI	.83	-.42	-1.69	-.98	-.42	.64
7	MMA	.58	-.48	-1.73	-.94	-.48	.36
	MMI	.43	-.23	-1.48	-.69	-.23	.61
	FMA	.43	-.23	-1.48	-.69	-.23	.61
	FMI	.58	-.48	-1.73	-.94	-.48	.36
11	MMA	.82	-.27	-1.52	-.73	-.27	.72
	MMI	.67	-.02	-1.27	-.48	-.02	.97
	FMA	.67	-.02	-1.27	-.48	-.02	.97
	FMI	.82	-.27	-1.52	-.73	-.27	.72
15	MMA	.86	-.27	-1.32	-.62	-.27	.48
	MMI	.71	-.02	-1.07	-.37	-.02	.73
	FMA	.71	-.02	-1.07	-.37	-.02	.73
	FMI	.86	-.27	-1.32	-.62	-.27	.48

**Note*. MMA (Male-Majority); MMI (Male-Minority); FMA (Female-Majority); FMI (Female-Minority); Male groups (MMA and MMI); Female groups (FMA and FMI); Majority groups (MMA and FMA); and Minority groups (MMI and FMI).
In "*G*x*E*" conditions, DIF is due solely to the interaction of gender with ethnicity. Parameters were manipulated to create uniform and nonuniform DIF consistent with the corresponding panels in Figure A1, such that gender differences depended on ethnicity.

TABLE A5. Generating Parameters for DIF Items in G GxE Conditions

		Response categories					
		Both	Binary	Polytomous			
Item no.	Groups	λ	τ	τ_1	τ_2	τ_3	τ_4
3	MMA	.83	-.42	-1.69	-.98	-.42	.64
	MMI	.76	-.30	-1.57	-.86	-.30	.77
	FMA	.61	-.05	-1.32	-.61	-.05	1.02
	FMI	.68	-.17	-1.44	-.73	-.17	.89
7	MMA	.58	-.48	-1.73	-.94	-.48	.36
	MMI	.51	-.36	-1.61	-.82	-.36	.49
	FMA	.36	-.11	-1.36	-.57	-.11	.74
	FMI	.43	.23	-1.48	-.69	.-.23	.61
11	MMA	.82	-.27	-1.52	-.73	-.27	.72
	MMI	.75	-.15	-1.40	-.61	-.15	.85
	FMA	.60	.11	-1.15	-.36	.11	1.10
	FMI	.67	-.02	-1.27	-.48	-.02	.97
15	MMA	.86	-.27	-1.32	-.62	-.27	.48
	MMI	.79	-.15	-1.20	-.50	-.15	.61
	FMA	.64	.11	-.95	-.25	.11	.86
	FMI	.71	-.02	-1.07	-.37	-.02	.73

**Note*. MMA (Male-Majority); MMI (Male-Minority); FMA (Female-Majority); FMI (Female-Minority); Male groups (MMA and MMI); Female groups (FMA and FMI); Majority groups (MMA and FMA); and Minority groups (MMI and FMI).
In the "*G GxE*" conditions, DIF is due to gender and the interaction of gender with ethnicity. Parameters were manipulated to create uniform and nonuniform DIF consistent with the corresponding panels in Figure A2.

TABLE A6. Generating Parameters for DIF Items in G E GxE Conditions

		Response categories					
		Both	Binary	Polytomous			
Item no.	Groups	λ	τ	τ_1	τ_2	τ_3	τ_4
3	MMA	.83	-.42	-1.69	-.98	-.42	.64
	MMI	.76	-.30	-1.57	-.86	-.30	.77
	FMA	.76	-.30	-1.57	-.86	-.30	.77
	FMI	.53	.08	-1.19	-.48	.08	1.14
7	MMA	.58	-.48	-1.73	-.94	-.48	.36
	MMI	.51	-.36	-1.61	-.82	-.36	.49
	FMA	.51	-.36	-1.61	-.82	-.36	.49
	FMI	.28	.02	-1.23	-.44	.02	.86
11	MMA	.82	-.27	-1.52	-.73	-.27	.72
	MMI	.75	-.15	-1.40	-.61	-.15	.85
	FMA	.75	-.15	-1.40	-.61	-.15	.85
	FMI	.52	.23	-1.02	-.23	.23	1.22
15	MMA	.86	-.27	-1.32	-.62	-.27	.48
	MMI	.79	-.15	-1.20	-.50	-.15	.61
	FMA	.79	-.15	-1.20	-.50	-.15	.61
	FMI	.56	.23	-.82	-.12	.23	.98

**Note*. MMA (Male-Majority); MMI (Male-Minority); FMA (Female-Majority); FMI (Female-Minority); Male groups (MMA and MMI); Female groups (FMA and FMI); Majority groups (MMA and FMA); and Minority groups (MMI and FMI).
In "*G E G*x*E*" conditions, DIF is caused by gender, ethnicity, and their interaction. Parameters were manipulated to create uniform and nonuniform DIF consistent with the corresponding panels in Figure A2.

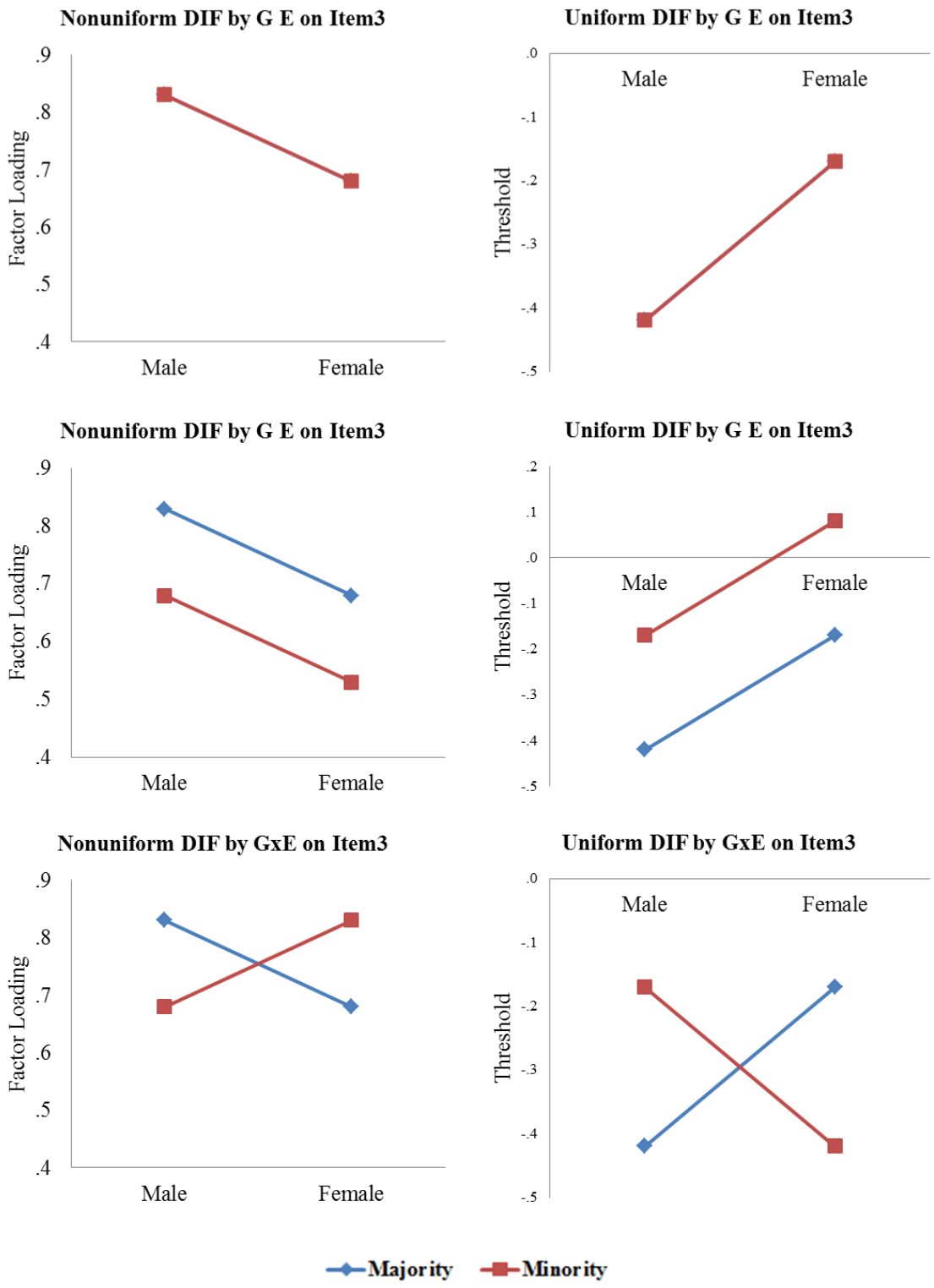

FIGURE A1. Examples of Nonuniform and Uniform DIF on Dichotomous Item 3 in Gender (G), Gender and Ethnicity (G E), and Gender by Ethnicity (GxE) Conditions

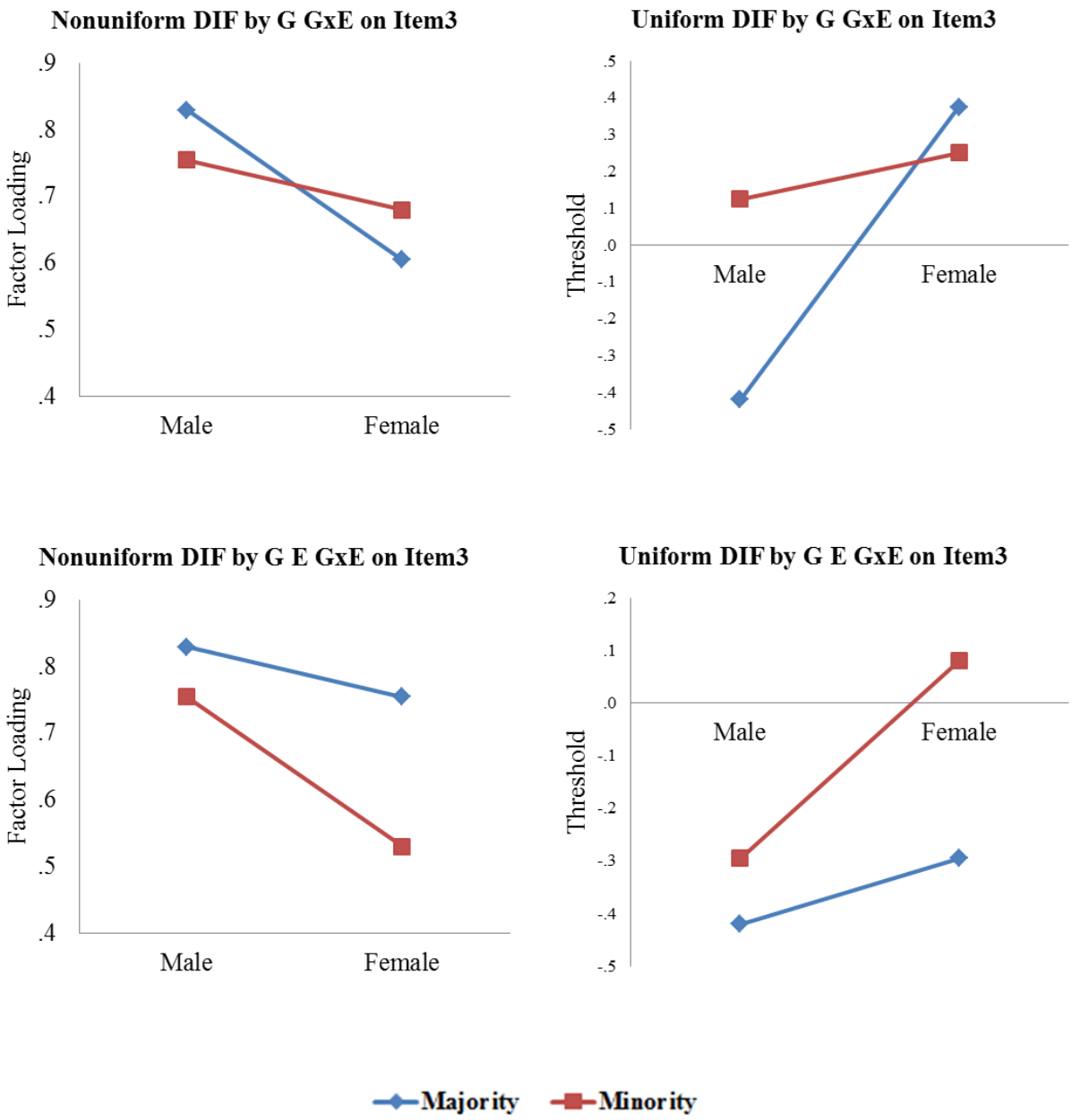

FIGURE A2. Examples of Nonuniform and Uniform DIF on Dichotomous Item 3 in Gender and Gender by Ethnicity (G GxE), and Gender and Ethnicity and Gender by Ethnicity (G E GxE) Conditions