When can Multidimensional Item Response Theory (MIRT) Models be a Solution for Differential Item Functioning (DIF)? A Monte Carlo Simulation Study

Yuan-Ling Liaw

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Supervisory Committee:

Elizabeth A. Sanders, Chair

Min Li

Deborah McCutchen

Catherine Taylor

Dagmar Amtmann

Program Authorized to Offer Degree:

College of Education

Measurement and Statistics

University of Washington

Abstract

When can Multidimensional Item Response Theory (MIRT) Models be a Solution for Differential Item Functioning (DIF)? A Monte Carlo Simulation Study

Yuan-Ling Liaw

Chair of the Supervisory Committee:

Professor Elizabeth A. Sanders, Ph. D.

Measurement and Statistics

The present study was designed to examine whether *multidimensional* item response theory (MIRT) models might be useful in controlling for differential item functioning (DIF) when estimating primary ability, or whether traditional (and simpler) *unidimensional* item response theory (UIRT) models with DIF items removed are sufficient for accurately estimating primary ability.

Researchers have argued that the leading cause of DIF is the inclusion of "multidimensional" test items. That is, tests thought to be unidimensional—one latent (unobserved) construct or trait per item measured—are actually measuring at least one other latent trait besides the one of interest. Additionally, most "problem" DIF is likely due to items measuring multiple traits that are *noncompensatory* in nature: to get an item correct, an examinee

needs a sufficient amount of all relevant traits (one trait cannot compensate for another trait). However, few studies have conducted empirical research on MIRT models; of the few that have, none examined the use of MIRT models for the purpose of controlling for DIF, and none empirically compared the performance of compensatory and noncompensatory MIRT models. The present study contributes new information on the performance of these methodologies for multidimensional test items by addressing the following main research question: How accurately do UIRT and MIRT models calibrate the primary ability estimate (θ_1) for focal and reference groups?

The data in this simulation study were generated for a test with 40 items and 2,000 examinees, and assumed a 2-parameter logistic (2PL), 2-dimensional, noncompensatory case. Five conditions were manipulated, including: between-dimension correlation (0 and 0.3), reference-to-focal group size balance (1:1 and 9:1), primary dimension discrimination level (0.5 and 0.8), secondary dimension discrimination level (0.2 and 0.5), and percentage of DIF items (0%, 10%, 20%, and 30%; all DIF favored the reference group). Five model approaches were then applied for IRT calibration, with results saved and averaged for each condition:

- Approach 1 (UIRT*d*): UIRT, no items removed from analysis;
- Approach 2 (UIRT*nds*): UIRT, after removing DIF-detected items (using Mantel–Haenszel with standard criterion p-value ≤ 0.05);
- Approach 3 (UIRT*ndl*): UIRT, after removing DIF-detected items (using Mantel–Haenszel with standard criterion p-value ≤ 0.10);
- Approach 4 (MIRT*c*): compensatory MIRT, no items removed from analysis; and
- Approach 5 (MIRT*nc*): noncompensatory MIRT, no items removed from analysis.

The impact of these modeling approaches and manipulated conditions on the accuracy of primary ability estimates was the focus of the investigation. Accuracy was judged by bias, which was calculated using the typical definition of the mean difference across the 500 replications between the estimated $\hat{\theta}_1$ and the true primary θ_1 used to generate the data. Analyses of variance (ANOVAs) on model-derived mean ability estimates were then used to identify main effects and simple interactions among modeling approaches and conditions.

As was expected, for the focal group, the ANOVA results showed that the UIRT*d* model (no items removed from analysis) yielded the worst bias (the focal group's primary ability was consistently underestimated and reference group's primary ability was consistently over-estimated) compared to all other models. Using UIRT*nds* and UIRT*ndl* models (DIF-detected items removed from analyses, one with the standard alpha level and the other with a liberal alpha level) led to the smallest bias, and use of the two types of MIRT models (MIRT*c* and MIRT*nc*) led to slightly more bias than the two UIRT models with DIF removed, but these differences were not significant (i.e., the only model that differed from the others was the UIRT model that completely ignored DIF). In other words, the simple model UIRT approach works as well as the complex MIRT approaches, but only for researchers willing to remove items with DIF prior to calibration; for those with limited item pools, the MIRT approach works just as well without removing DIF items.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

There are many people I need to thank for their support while writing this dissertation and during my lengthy term in graduate school. First and foremost, I give my heartfelt thanks to Dr. Liz Sanders, my committee chair and academic advisor. She has dedicated an enormous amount of time to providing academic and professional guidance and support, including countless hours providing feedback on my ideas, reading and critiquing paper drafts, and general and ongoing support throughout this process.

Second, I would like to express deep appreciation to Dr. Cathy Taylor, who started me on this journey with her measurement courses, and who was the first to pique my interest in IRT and DIF. Her passion for rigorous and equitable test development continues to inspire my own passion for state of the art psychometrics research. Similarly, I would also like to thank Dr. Min Li, who went above and beyond the call of duty to support my learning from early morning to late into the night. Without her constant support, expertise, and knowledge, this dissertation certainly would not have been possible. I am also appreciative to my other two committee members, Drs. Deborah McCutchen and Dagmar Amtmann, for joining my committee despite their many other commitments and obligations. In addition, many thanks go out to Dr. Patricia Martinkova for donating her time and statistical expertise to critiquing my dissertation ideas.

Third, there are many others who should be thanked, from my many fellow graduate students, to the research scientists at Pearson's Psychometric Services office, to the friends all over the world. Drs. Kellie Wills, Julie Lorah, Fraser Bocell, and Pol Thummaphan, my M&S classmates, I thank you for the technical discussions and for helping me navigate through the program. I also thank Drs. Annie Kuo, Tiffany Katanyoutanant, and Wanda Liao, who are

qualitative researchers yet patiently listened to me talk about quantitative research for so long, as well as Alec Kennedy in the Evans School who never hesitated to offer his help in statistical programming. I cannot find the words to fully express my gratitude for their encouragement and help.

Last but not least, I am forever indebted to my family for their lifelong encouragement, patience, and sacrifice which have allowed me to study abroad in order to achieve my dreams. Without their unwavering love and support, I never would have made it to this point.

Chapter I: Introduction

Research has drawn attention to the importance of language demands in student performance on assessments in content-based areas such as mathematics (Abedi & Lord, 2001; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006). Most large-scale mathematics tests include word problems that not only measure mathematical ability, but also measure transfer of mathematics knowledge and skills to situations. Being able to complete these kinds of exercises is an integral part of mathematics learning because they assess students' ability to apply their mathematics knowledge to real-world situations. In order to correctly respond to these items, examinees must first read a problem presented in a written format, and then translate it into a mathematical equation, table, graph, or symbolic representation. Alternatively, the examinees might be provided with information given in a table, graph, or image, and then asked to translate what they see into a solvable manipulation form. Indeed, the language (literacy) demands of such items can be so great that some students do not respond accurately even though they have the mathematical knowledge to do so. In other words, under certain circumstances, examinees can fail to answer math problems correctly solely due to their language limitations — not their mathematics ability.

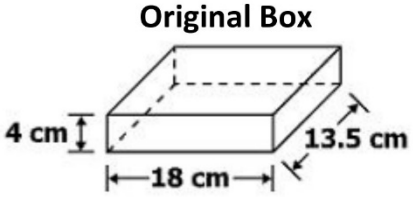
Although the details involved in constructing a given test item are complex, items do contain two distinct parts—stimulus material and a form for answering (Osterlind & Everson, 2009; Reckase, 2009). As a demonstration, consider the following two multiple-choice items with four response alternatives: Item 1 (Figure 1) is designed to measure algebraic ability, and Item 2 (Figure 2) is designed to measure geometrical ability. Item 1 appears to measure only algebraic symbol manipulation with little language demand. However, Item 2 appears to be measuring both geometry and language abilities.

What value of x will make the equation $3(x + 15) - 6x = -6(x - 3)$ true?

A -9
B -6
C 2
D 3

Figure 1. Algebra Item with Little Language Comprehension Demand

A cell phone box in the shape of a rectangular prism is shown. The height of the box is 4 cm.



The height of the original box will be increased by 3.5 centimeters so a new instruction manual and an extra battery can be included. Which is closest to the total surface area of the new box?

A 479 cm^2
B 707 cm^2
C 738 cm^2
D 959 cm^2

Figure 2. Geometry Item with Increased Language Comprehension Demand

Selecting a correct response for any item is the result of the interaction between the capabilities of the examinee and the characteristics of the test item. In order to answer Item 1 correctly, which is an abstract algebra item, the examinee only has to perform the computation and solve *x*. However, to correctly respond to Item 2, the examinee must be able to:

- read and comprehend English in order to achieve the correct answer,
- know that the task is not asking for the surface area of the box presented in the graph format, but the surface area of the new box described in the text format,
- know the relation between old height and new height through the preposition "by",
- know that the question is asking for the "closest" calculating result and the correct answer out of the response alternatives is the result rounded to the nearest whole number.

Furthermore, a number of factors can create different ways of organizing pieces of information given (Taylor & Lee, 2011; Walker, 2011). Zumbo (1999) drew upon the work of Camilli and Shepard (1994) as well as Clauser and Mazor (1998) to develop a taxonomy of different phenomena that can affect examinees' performance on a test item. According to Zumbo's definition (p. 12), *item impact* occurs when examinees from different groups show different probabilities of success on an item because there are true differences between the groups in the underlying ability being measured by the item. For example, when low-performing students are assigned to basic geometry classes and high-performing students are enrolled in advanced geometry classes, there will be true differences in their geometry ability. *Differential item functioning (DIF)* occurs when examinees from different groups have different probabilities of responding correctly to an item even when they have the same level of the underlying ability that the item is intended to measure (i.e., geometry ability). *Item bias* occurs when examinees of

one group are less likely to answer an item correctly than examinees of another group because of some characteristics of the test item or testing situation that are construct-irrelevant to the test purpose. For example, language comprehension (e.g., use of propositions, idioms, relative clauses, specialized vocabulary, and complex syntax; Abedi, Lord, & Plummer, 1997; Abedi & Lord, 2001) is a source of construct-irrelevant variance in a geometry test. Therefore, two groups of examinees with equal geometry ability may show different probabilities of success on an item due to their language comprehension ability. If there are items flagged as having DIF, then the research must focus on whether the potential cause of DIF is relevant or irrelevant to the construct.

For over two decades, researchers (Ackerman, 1992; Shealy & Stout, 1993; Roussos & Stout, 1996; Russell, 2005) have argued that the leading cause of DIF is the inclusion of "multidimensional" test items. That is, tests thought to be unidimensional—one latent (unobserved) construct or trait per item—are actually measuring at least one other latent trait besides the one of interest. In other words, items that are supposed to be measuring geometry ability are in fact also measuring language comprehension ability. If the secondary ability (i.e., language comprehension) is deemed to be irrelevant to the primary ability (i.e., geometry ability) intended to be measured, the use of *unidimensional* item response theory (UIRT) models with *multidimensional* items violates the UIRT's unidimensionality assumption and may pose threats to item and examinee parameter estimation, posing a threat to the accurate representation of a student's primary ability or trait being measured. (Notably, there exist tests in which a secondary ability, such as language comprehension, would be considered construct-*relevant*, such as when "problem solving" is measured in the context of algebra ability; importantly, this type of construct-relevant multidimensionality is not the focus of the present study.) Furthermore, these

inaccurate examinee parameter estimates (such as pure geometry ability) can subsequently lead to flawed test scores from and can lead to biased decision making, ranging from which individuals should be deemed to "pass" the test (proficiency), to which individuals will be afforded specific instructional intervention.

Overview of Item Response Theory Models (IRT)

IRT models have been developed predominantly in the education and psychology fields since the late 1960s. Focused on observed responses on binary or ordinal items, these models define a nonlinear relationship between a latent trait (θ) and the observed performance on the test that is designed to measure that trait (Hambleton, Swaminathan, & Rogers, 1991).The latent trait is assumed to be continuous and the location of person *j* on θ (denoted θ_j) is usually referred to as the person's ability or proficiency. The underlying assumption of IRT is that all of the items in a test are measuring a single ability dimension, which is referred to as the *unidimensionality* assumption; for example, pure mathematical proficiency may be considered a single dimension. However, the items are pairwise uncorrelated if ability level is held constant, which is referred to as the *local independence* assumption (de Ayala, 2009).

With IRT, each item is described by a set of item parameters that can be used to graphically depict the relationship between an item and a latent trait through use of an *item characteristic curve (ICC)* or *item response function (IRF)*. When dealing with items that have been scored dichotomously, three related IRT models are frequently used in the psychometric literature: the one parameter logistic (1PL) model, the two parameter logistic (2PL) model, and the three parameter logistic (3PL) model.

The most complex of these models is the three-parameter (3PL) IRT model, as follows

$$P(X_{ij} = 1 \mid \theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}. \quad (1)$$

In the model above, $P(X_{ij}=1|\ \theta_j)$ is the probability that a randomly chosen examinee *j* with ability θ answers item *i* correctly; *a* is item discrimination parameter; *b* is item difficulty parameter; and *c* is lower asymptote or pseudo-guessing parameter (termed "pseudo" because it is usually lower than would be predicted by a random guessing model (de Ayala, 2009, p. 126)).

An alternative way to express this model is in the slope-intercept form. Equation (1) can be represented as

$$P(X_{ij} = 1 \mid \theta_j, a_i, d_i, c_i) = c_i + (1 - c_i)\frac{e^{a_i\theta_j + d_i}}{1 + e^{a_i\theta_j + d_i}}, \quad (2)$$

where d_i is an intercept parameter equal to $-a_ib_i$ from Equation (1).

On the plot of IRF, the *b* parameter determines the location of the item on the underlying scale (θ) where the IRFs has maximum slope; the higher the *b-parameter* value, the more difficult it is to answer the item correctly. The discrimination *a* parameter determines the slope of the IRF; the higher the *a* parameter value, the stronger the relationship is between ability level and response on the item. Therefore, an item with a substantial *a* parameter value can powerfully differentiate examinees with different ability scores for a narrow range of the ability scale. The *c* parameter represents the lower intercept and corresponds to the probability that a person lacking in proficiency will answer the item correctly (Figure 3).

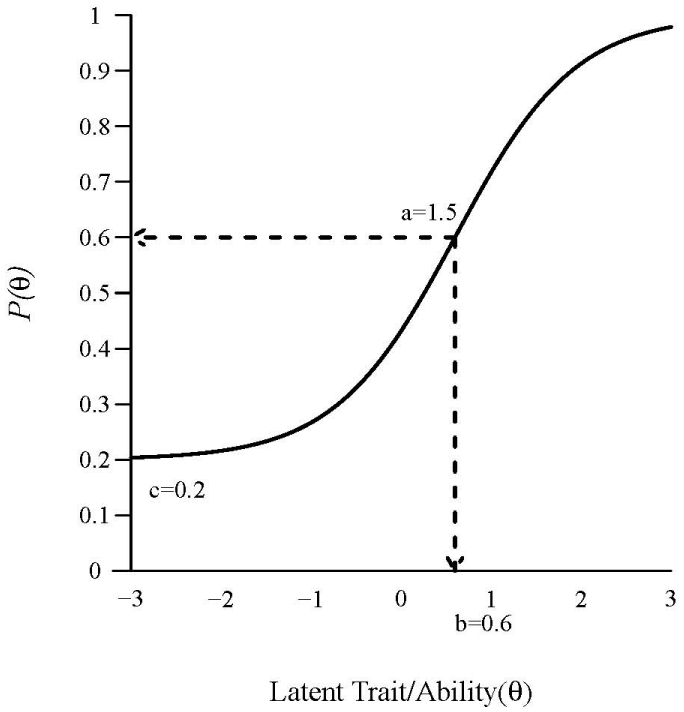


Figure 3. Three-parameter Logistic Model Item Characteristic Curve

The 3PL IRT model can be constrained to form the simpler two-parameter (2PL) IRT model by removing the item pseudo-guessing parameter c_i. The lower asymptote of each item's IRF is assumed to be zero. The reduced model, therefore, contains only estimates of item difficulty and item discrimination and has the following form

$$P(X_{ij} = 1 \mid \theta_j, a_i, b_i) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}. \tag{3}$$

A further restriction can be imposed to create the one-parameter (1PL) IRT model (or Rasch model). The item discrimination parameter a_i is constrained to be equal across all items and fixed as a. Therefore, the only parameter being estimated is the item difficulty b_i. The 1PL IRT model defines the probability of correctly responding to an item as follows

$$P(X_{ij} = 1 \mid \theta_j, a, b_i) = \frac{e^{a(\theta_j - b_i)}}{1 + e^{a(\theta_j - b_i)}}. \tag{4}$$

Differential Item Functioning (DIF)

According to Dorans and Holland (1993), *differential item functioning* (DIF) refers to, "a difference in item performance between two comparable groups of examinees, that is, groups that are matched with respect to the construct being measured by the test. The comparison of matched or comparable groups is critical because it is important to distinguish between differences in item functioning from differences between groups" (p. 35). Consider, for instance, groups defined by gender, ethnicity, culture, and language, with one assigned to be the *reference group* (i.e., majority or normative) and the other assigned to be the *focal group* (i.e., minority or marginal). The presence of DIF would indicate that the item is not performing the same across groups; that is, there is a lack of item-level measurement invariance and the item has lower construct validity for one of the groups (Steinberg & Thissen, 2006). Logically, if the context of some test items were unfamiliar to focal group examinees, then these examinees would be more likely to answer incorrectly compared to the reference group examinees, even if the focal group has the same level of the measured trait as the reference group. A test containing items flagged as having DIF, when the secondary abilities are deemed to be irrelevant to the primary ability intended to be measured, could in turn create inaccurate observed scores, resulting in inaccurate reflections of one group's true construct skills or traits.

Mathematically, the absence of DIF is characterized by a conditional probability distribution of X_{ij} that is independent of group membership and has the following form (Equation (5)). This suggests that the probability of correct response is identical for individuals belonging to different groups, but sharing the same value of θ.

$$P(X_{ij} = 1 \mid \theta, G = R) = P(X_{ij} = 1 \mid \theta, G = F), \tag{5}$$

where G corresponds to the grouping variable, R corresponds to the reference group, and F corresponds to the focal group.

In contrast to the above situation, suppose that the conditional probability of X_{ij} is not identical for the reference and focal groups. In this case, individuals having the same level of proficiency, but belonging to different groups, have a different probability distribution of Y. That is, there exists a dependency between group membership and item performance after controlling for θ. All DIF detection statistics, in one way or another, are concerned with either testing the null hypothesis of no DIF, as described by Equation (5), or providing a measure of the extent to which the situation described by Equation (5) is false.

Types of DIF: Uniform vs. Non-Unifrom. DIF can be either uniform or non-uniform (Camilli & Shepard, 1994; Mellenbergh, 1989), depending on the item parameter that differs across groups. *Uniform DIF* is present when only the *b* parameter differs across groups. The item response curves or functions (IRFs) for the two groups will be different but do not cross; more specifically, one group will be more or less likely to earn a higher score over the entire range of θ compared to the other group (see Figure 4).

Non-uniform DIF, on the other hand, exists when the *a* parameter differs across groups, regardless of whether or not the *b* parameter is different. In this context, the shapes of IRFs differ between the two groups, and often the graphs of the IRFs will cross (see Figure 5).

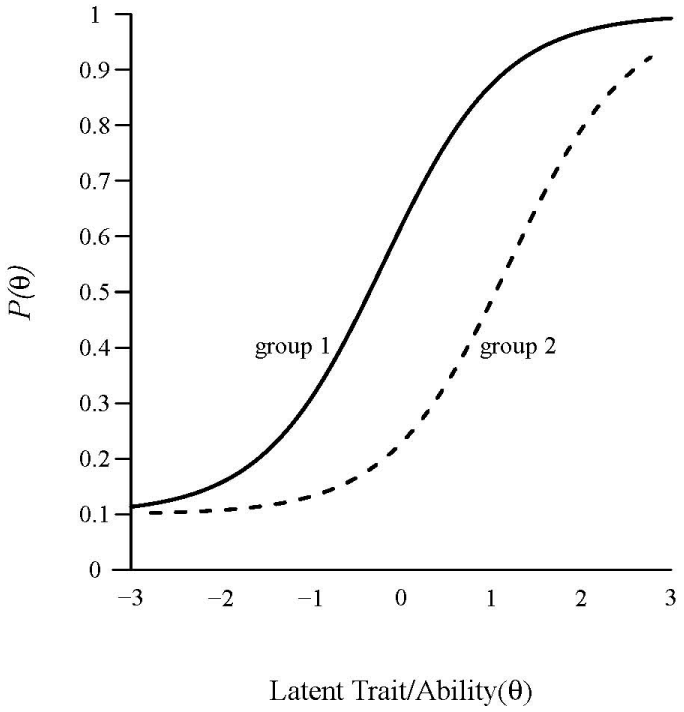


Figure 4. Example of Uniform DIF

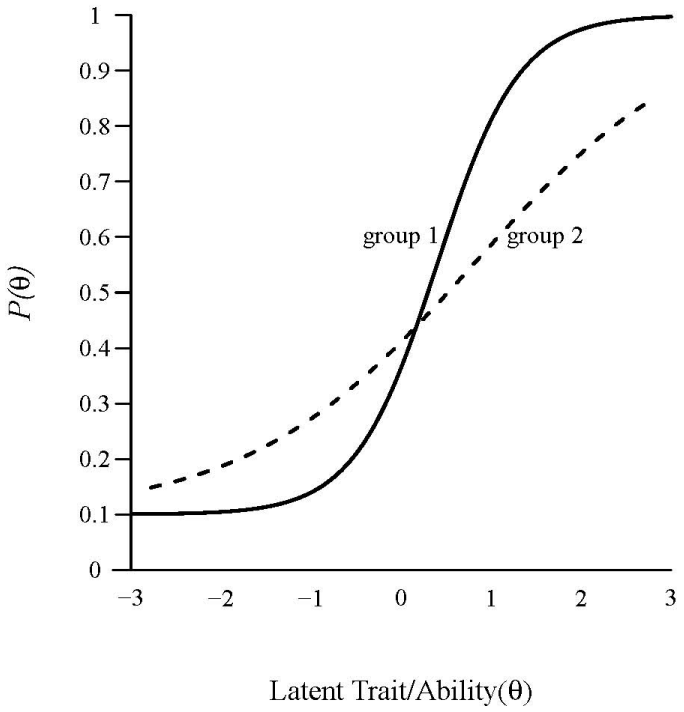


Figure 5. Example of Non-uniform DIF

IRT-based Methods of DIF Detection. Several IRT techniques have been proposed for detecting differential item functioning. Lord (1977, 1980) used a chi-square test to compare the item parameters between the focal and reference groups, with the null hypothesis being that there is no difference between groups. Subsequently, Thissen and colleagues (1994) employed a chi-square-based likelihood ratio test (TSW-ΔG^2) to compare the overall fit of an IRT model with and without separate group parameter estimates, where if the more complex model (separate group parameter estimates) fit significantly better, the null model (same group parameter estimates) was rejected. One issue with Thissen's method of DIF evaluation is that the entire model, rather than the individual items in the models, is compared to competing models. As such, no evaluation can be made regarding individual items. Finally, a third method of assessing DIF was proposed by Raju (1988, 1990), which estimates the *area* between the focal and reference groups' IRFs, similar to the original conception by Lord and colleagues but with a more fine-grained approach. Graphically, the IRFs are completely determined by their corresponding item parameters, and thus DIF can be identified by comparing the item parameters to determine whether two IRFs are different or not.

In addition, the Simultaneous Item Bias (SIB) implements a nonparametric IRT estimation and hypothesis testing statistical method of assessing DIF, with versions available for detecting both the uniform (Shealy & Stout, 1993) and non-uniform (Li & Stout, 1996) cases. The presence of impact between groups can inflate the Type I error rates of DIF detection methods because the expected value of the reference group's primary ability will tend to differ from the corresponding expected value for the focal group. SIBTEST uses a nonlinear regression correction to correct for this inflated Type-I error rate (Jiang & Stout, 1998). The conditioning variable used in SIBTEST is a matching subtest that consists of a set of items hypothesized to be

unidimensional and therefore only measuring the primary ability. The studied item is hypothesized to be multidimensional, measuring a secondary dimension on which the two matched groups are believed to differ, in addition to the primary construct of interest.

Non-IRT-based Methods of DIF Detection. Other non-IRT-based techniques exist as an alternative to IRT-based procedures for detecting DIF, particularly when sample sizes are small or when strong assumptions are not tenable. These include delta plot (Angoff, 1972, 1993), standardization (Dorans & Kulick, 1986; Dorans & Holland, 1993), Mantel–Haenszel (MH; (Holland & Thayer, 1988; Mantel & Haenszel, 1959), and logistic regression (Swaminathan & Rogers, 1990).

The MH procedure is widely employed by many large-scale assessments including the NAEP and some statewide achievement assessments (e.g., National Center for Education Statistics [NCES], 2009; Educational Testing Service, 2015), which is an extension of the chi-square test. Specifically, MH allows for the comparison of item responses between the focal and reference groups, controlling for matched subtest scores (scores serve as a proxy for the latent trait being measured by the instrument). The first step in using MH for assessing DIF is the calculation of a total score on the instrument, which is used to create matching groups for the examinees (see Holland & Thayer, 1988). From there, k 2x2 contingency tables are created, where k is the number of score categories on the matching criterion. The data can be summarized as N_{R1k} *and* N_{F1k}, which denote the number of examinees in the reference and focal groups, respectively, who answered correctly for the kth score level. In contrast, N_{R0k} *and* N_{F0k} are the numbers of examinees in the reference and focal groups who answered incorrectly, respectively. The computational formula for the MH common-odds ratio α_{MH} (Penfield & Camilli, 2007; Dorans, 1989; Holland & Thayer, 1988) is

$$\alpha_{MH} = \frac{\sum_k N_{R1k} N_{F0k} \,/\, N_k}{\sum_k N_{F1k} N_{R0k} \,/\, N_k}, \tag{6}$$

where N_k is the total number of examinees in score group k. The statistical significance of α_{MH} can be test with a χ^2 test with 1 degree of freedom as follows

$$\chi^2_{MH} = \frac{\left(\left| \sum_k N_{R1k} - \sum_k E(N_{R1k}) \right| - \frac{1}{2} \right)^2}{\sum_k Var(N_{R1k})}, \tag{7}$$

where $E(N_{R1k}) = n_{Rk} m_{1k} / N_k$, $Var(N_{R1k}) = \frac{n_{Rk} n_{Fk} m_{1k} m_{0k}}{N_k^2 (N_k - 1)}$, n_{Rk} and n_{Fk} denote the numbers of examines in the reference and focal groups, respectively, m_{1k} represents the number of examinees who answered the item correctly, and m_{0k} is the number who answered incorrectly, N_k is the total number of examinees in score group k. A statistically significant result for this test indicates the presence of uniform DIF for the target item. Items that are not statistically significantly different based on the MH $\chi 2$ are considered to have similar performance between the two studied groups; these items are considered to be functioning appropriately. Otherwise, items display statistically significant DIF. For polytomous items, k r x2 contingency tables are created, where k is the number of score categories on the matching criterion and r is the number of score levels for the item.

Multidimensional Item Response Theory (MIRT)

In the recent development of IRT models, unidimensional (UIRT) models may be generalized to multidimensional (MIRT) models, which allows the researcher to relax the assumption of UIRT by simultaneously estimating multiple abilities. In other words, each item in the dataset may be analytically specified to have a relationship with both traits.

There are two major types of MIRT models (Reckase, 2009; Ackerman, 1994; Sympson, 1977). The first type, called *compensatory MIRT*, is based on a weighted *linear* combination of abilities and assumes that deficits in one trait can be compensated by over-abundance in another trait, such that the probability of a correct response on a given item is the weighted sum of an individual's probabilities of correct response given each trait (Ackerman, Gierl, & Walker, 2003). For example, if the ability on one of the dimensions is low, such algebraic symbol manipulation ability, the individual may still have a good chance of getting a correct answer on an item if they have a sufficiently high ability in another dimension, such as arithmetic symbol manipulation ability.

In contrast, the second type of MIRT, known as the *noncompensatory MIRT*, assumes that proficiency in all the dimensions (i.e., multiple latent traits) of a given item must be sufficient to have a good chance of getting the correct answer. More precisely, the probability of correct response for the item depends on the weighted *product* of the individual's probabilities of correct response for each latent trait (which is why this class of models is nonlinear in nature) (Sympson, 1977; Reckase & McKinley, 1991). Using the example given earlier, an individual must have sufficiently high ability in both geometry ability and language comprehension for a high probability of correct response to an item that demands both of these characteristics.

Compensatory MIRT Models for Dichotomous Items. Compensatory multidimensional two-parameter model (M2PL) for dichotomous data is a generalization of the UIRT model in the slope-intercept form given by

$$P(X_{ij} = 1 \mid \boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i) = \frac{e^{\sum_{l=1}^{m} a_{il}\theta_{lj} + d_i}}{1 + e^{\sum_{l=1}^{m} a_{il}\theta_{lj} + d_i}}. \quad (8)$$

In the UIRT models, the exponent of the form $a_i(\theta_j - b_i)$ can be represented as the slope-intercept form, $a_i\theta_j + d_i$. When there are multiple elements in the $\boldsymbol{\theta}$-vector, the simple slope-intercept form is replaced by the expression $\boldsymbol{a}_i\boldsymbol{\theta}' + d_i$, where $\boldsymbol{a}$ is a $1 \times m$ vector of item discrimination parameters and $\boldsymbol{\theta}$ is a $m \times 1$ vector of person coordinates with m indicating the number of dimensions in the coordinate space. Each person has a vector of θ values, and each items has a vector of a values with only one d value. The direction of a indicates the composite of the skills best measured by the item and its length indicates the amount of multidimensional item discrimination (Ackerman, 1994; Reckase & McKinley, 1991). The d_i parameter in the model is related to the difficulty of the test item. However, the value of this parameter cannot be interpreted in the same way as the b-parameter of UIRT models because the model is in its slope/intercept form. In general, compensatory models involve summing a series of latent trait (θ) values, with each multiplied by a different discrimination value, and then this sum is added to an intercept (a) value (note that if there is only one non-zero a value, the model simplifies to the unidimensional case).

Figure 6 shows the form of the compensatory model for the two-dimensional case in two ways: the surface plot, which is three-dimensional and illustrates the probability of a correct item response as a function of the item response surface (IRS) (Panel A), and the contour plot, which shows the probability of a correct item response by selected segments of the IRS contours (Panel B). Note that the contours of equal probability form straight lines. Both plots clearly show that a high θ on one dimension can compensate for a low θ on the other dimension. For instance, a person with $\theta = -2.0$ on the second dimension of the item in Figure 6 might still have a probability of endorsing the item as high as 0.9 if the person's θ value on the first dimension is close to 3.0.

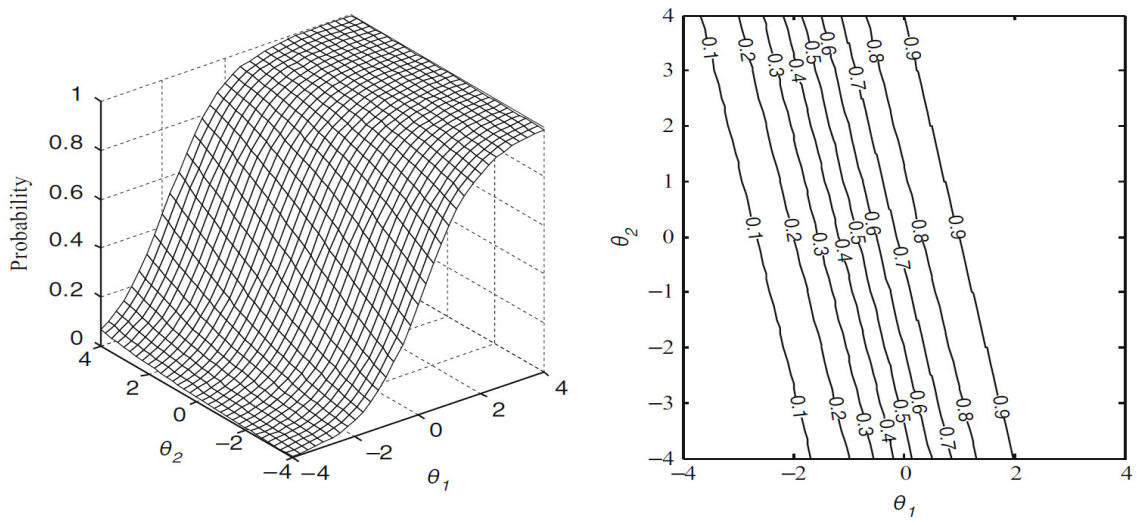


Figure 6. Surface Plot (Panel A, Left) and Contour Plot (Panel B, Right) for the Probability of Correct Response for a Two-Dimensional Compensatory Item with $a_1 = 1.2$, $a_2 = 0.3$, $d = 1.0$ (Cited from Reckase (2009, p. 88))

Nonompensatory MIRT Models for Dichotomous Items. The form of the noncompensatory MIRT for dichotomous data is given by

$$P(X_{ij} = 1 \mid \boldsymbol{\theta}_j, \boldsymbol{a}_i, \boldsymbol{b}_i) = \prod_{l=1}^{m} \frac{e^{a_{il}(\theta_{lj} - b_{il})}}{1 + e^{a_{il}(\theta_{lj} - b_{il})}}, \tag{9}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ is a $1 \times m$ vector of item discrimination and difficulty parameters, respectively, and $\boldsymbol{\theta}$ is a $m \times 1$ vector of person coordinates with ℓ indicating the number of dimensions in the coordinate space. This model involves calculating a series of individual probabilities and then multiplying these probabilities together to obtain the overall probability of response. Unlike compensatory models, noncompensatory models require both a discrimination and a difficulty parameter on each dimension, for each item (note that if the probability values of all but one dimension are equal to 1, then the model collapses into the unidimensional form).

Figure 7 contains IRS (Panel A) and contour (Panel B) plots of a two-dimensional noncompensatory item. A person with one low θ will not have a high probability of endorsing

the item no matter how much higher the other θ value is. The probability of correct response for low values of either θ_1 or θ_2, or both, is close to zero. Only when both θ values are high does the model yield a high probability of correct response.

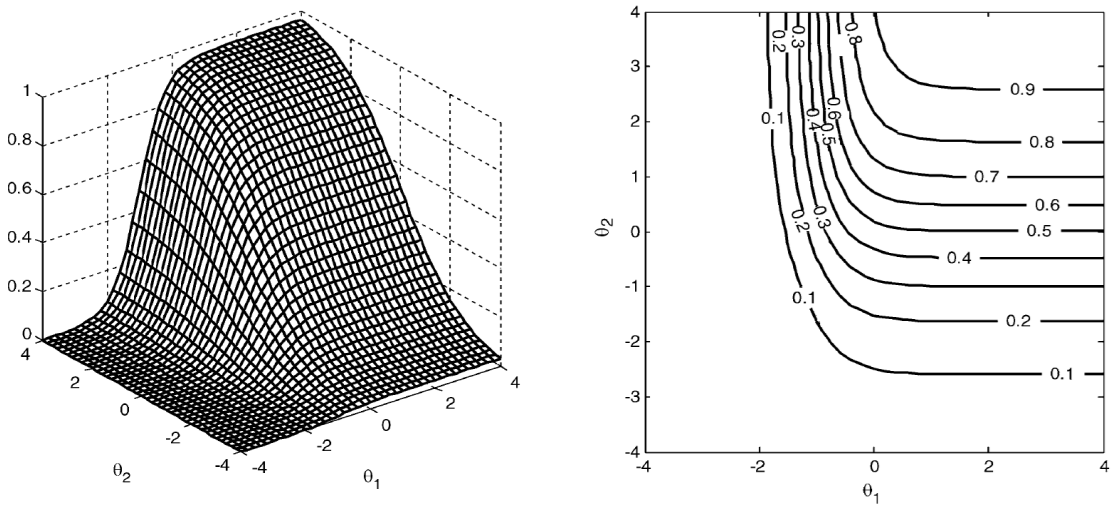


Figure 7. Surface Plot (Panel A, Left) and Contour Plot (Panel B, Right) for the Probability of Correct Response for a Two-Dimensional Noncompensatory Item with a_1=1.2, a_2=0.5, b_1=1.0, b_2=0.0 (Cited from Reckase (2007, p. 616))

The noncompensatory MIRT, on the other hand, separates the cognitive tasks in a test item into parts (latent traits) and uses a unidimensional model for each part. The probability of correct response for the item is the product of the probabilities for each part. As a result, in this model, even if an individual is relatively high on one ability, one single trait cannot compensate for being low on another ability/trait essential to responding to the item. For example, a person with very low language comprehension skills would not be able to comprehend the geometry problem that needed to be solved because even if such a person had very high geometry skills, he/she would not be able to determine the correct answer because of a lack of understanding the question in the first place.

Unfortunately, the multiplicative structure of noncompensatory MIRT presents severe estimation challenges. Unlike compensatory MIRT, noncompensatory MIRT often necessitate separate difficulty parameters for each item on each dimension. As Bolt and Lall (2003) noticed, estimating these parameters "requires sufficient variability in the relative difficulties of components across items to identify the dimensions" (p. 396). Therefore, despite a noncompensatory structure potentially modeling cognitive processes better than a compensatory structure, the proposed methods and software programs for estimating noncompensatory MIRT parameters are less commonly used.

In summary, the noncompensatory (product) model is consistent with the hypothesis that test items have different parts related to different skills or knowledge, and that overall, success requires success on each part. On the other hand, the compensatory (linear) model is consistent with a more holistic view of the interaction of persons and test items. Individuals bring all of their skills and knowledge to bear on all aspects of the items, and hence it is conceivable that strengths in one trait could compensate for deficiencies in another trait. Ultimately, the usefulness of each model will be determined by how accurately they represent the responses from actual test items (Bolt & Lall, 2003). However, researchers have done relatively little work with noncompensatory models because of inherent difficulties in model estimation (Chalmers & Flora, 2014; Babcock, 2011).

Earlier in this chapter, I provided two example items: Item 1 (Figure 1), measuring only algebraic manipulation ability with little reading demand, and Item 2 (Figure 2), measuring both geometric ability and language comprehension ability. The fact is actually that both items require capabilities on two latent abilities/traits to achieve a correct response, and this in turn suggests that a trait space with two coordinate axes is needed to fully describe the variation in examinee

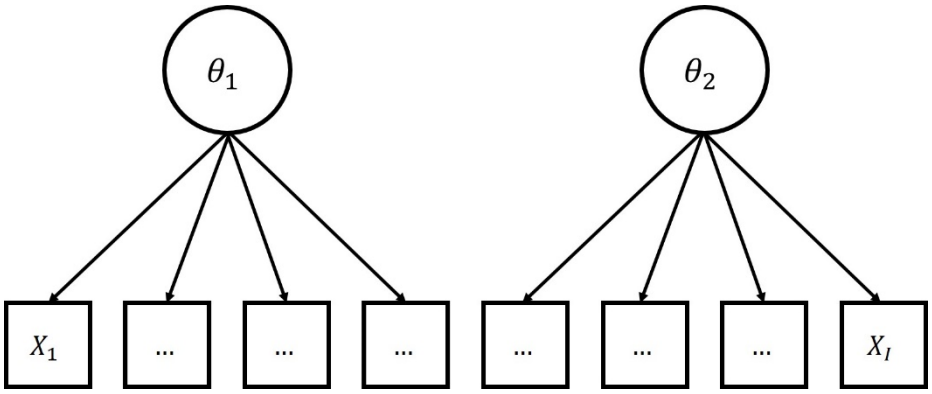
responses. The coordinates for a specific examinee *j* for this space are indicated by θ_{j1} and $\theta_{j2.}$ In terms of Item1, θ_{j1} could be an estimate of the examinee's level on algebraic symbol manipulation and θ_{j2} could be an estimate of the examinee's level on arithmetic symbol manipulation. As for Item 2, on the other hand, θ_{j1} could be an estimate of the examinee's level on geometric ability and θ_{j2} could be an estimate of the examinee's level on language comprehension ability in Item 2. Further, the function $P(\theta_{j1,}\theta_{j2})$ is used to indicate the probability of correct response to an item given the location of examine *j* in the space (Reckase, 2009).

What is interesting about these two items is that they would likely require different MIRT models. Item 1 captures the essence of a compensatory (linear) MIRT model: if the examinee is lacking in algebraic problem solving for example, his/her ability in arithmetic manipulation may compensate for the algebraic deficit, and there may still be a fair chance that he/she responds to the item correctly. However, for Item 2, we would likely need a noncompensatory MIRT model: the two latent abilities involved (e.g., geometry and language comprehension) are both pre-requisite to correctly solving the problem – no amount of either ability/trait could compensate for a deficit in the other.

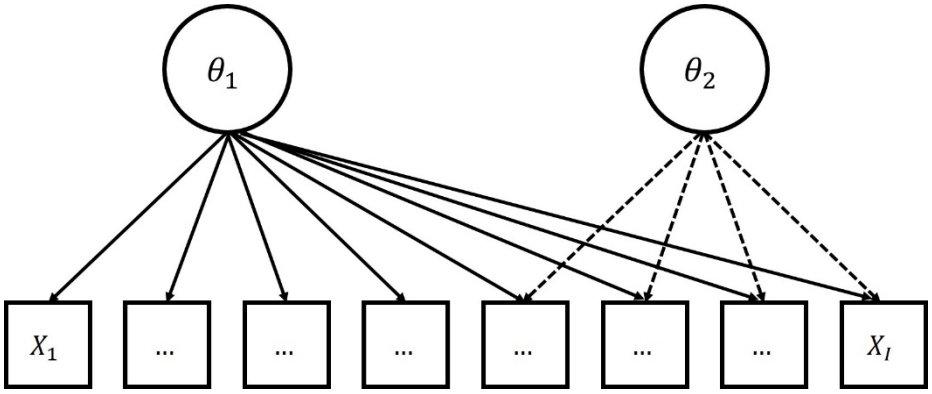
MIRT Simple vs. Complex Structures. In addition to whether an MIRT is considered compensatory or noncompensatory, MIRTs are also distinguished by whether their structure is considered simple or complex at the item level. *Complex* structure items are associated with multiple latent abilities/traits, whereas the *simple* structure items are associated with only one latent ability/trait (Zhang, 2012; Wang & Chen, 2004). Figure 6 illustrates the various dimensionality types. Note that the figure has no specific factors.

More specifically, simple structure refers to situations in which any given item is primarily associated with only one dimension (or ability/trait), although nonzero coefficients in

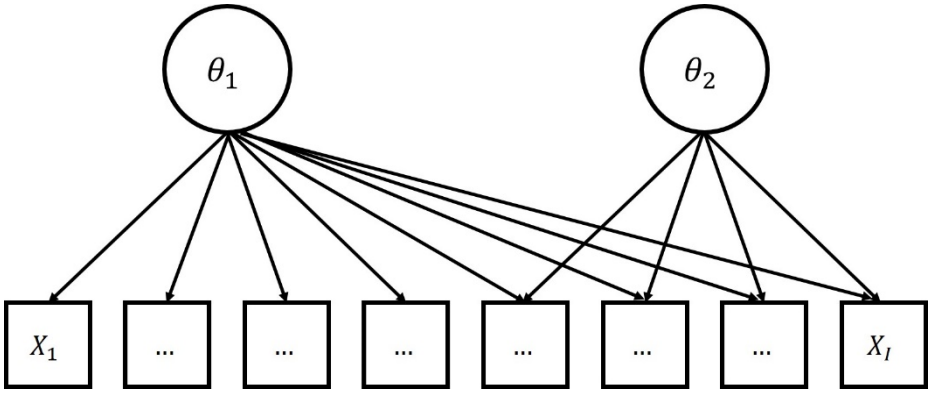
the vector of item discriminations still may allow items to be associated with multiple latent variables (but these associations are considered trivial). A complex structure, however, extends any given item's association with multiple latent variables; however, those associations are considered important to capturing the multidimensional nature of the item(s).



Panel A: Exact Simple Structure



Panel B: Approximate Simple Structure



Panel C: Complex Structure

Figure 8. Geometric Representation of Multidimensional IRT Models with Exact Simple Structure (A), Approximate Simple Structure (B), and Complex Structure (C)

Can MIRT Solve the Problem of DIF?

In the foregoing discussion of MIRT models, it is clear that items exhibiting bias (i.e., differential item functioning that leads to biased estimates of a focal group's true ability/trait) are likely to be multidimensional. How might we apply this in educational test development? Is it possible for DIF to be accounted for by MIRT models such that examinees' estimated ability/trait scores on a particular dimension of interest (such as geometry proficiency) are untainted by DIF due to a secondary dimension (such as language comprehension ability)?

As already discussed, in the multidimensional interpretation, items displaying DIF are in fact measuring at least one dimension in addition to the primary dimension. The dimension that a test is designed to measure is the primary dimension of the test; the additional potential DIF-causing dimension is referred to as secondary dimensions. The secondary dimensions can be further classified into *auxiliary* if the items are placed intentionally to measure the construct; or *nuisance* (if the second dimension is not intended). The former is defined as *benign* DIF, and the latter as *adverse* DIF (Roussos & Stout, 1996).

The magnitude of DIF could be further treated as the difference between the conditional expectation of η, given θ, for the reference group and focal group (Shealy & Stout, 1993),

$$E[\eta_R \mid \theta] - E[\eta_F \mid \theta] = d_\eta - \rho d_\theta, \tag{10}$$

where ρ is the correlation between the two dimensions and d_θ is the difference between the means of the focal and reference group on θ (impact), and d_η is the mean of the focal group on η subtracted from the mean of the reference group on η. Positive values of d_η indicate DIF against the focal group, and negative values indicate DIF against the reference group.

However, few studies have examined the impact of the primacy of the primary dimension (e.g., whether the primary dimension has a higher discrimination than that of the secondary

dimension) on the detection of DIF (Furlow, Ross, & Gagne, 2009). Therefore, in the current study, different levels of loading on the primary and secondary dimensions will be considered in order to explore whether the potential cause of DIF is from a relatively high discrimination value for the secondary dimension than for the primary dimension.

Understanding how to appropriately deal with items flagged as having DIF is crucial to modern test development. Follow the steps to conducting DIF analyses described by Penfield and Camilli (2007), if the secondary dimension is deemed to be irrelevant to the constructs intended to be measured, then the DIF items should be considered for revision or removal. In practice, during test construction, classifications of DIF, from prior test administrations, are available for most items chosen for test forms. When items previously flagged for DIF are chosen for operational test forms, content specialists review these items to determine whether or not the item content lends itself to DIF (Office of Assessment and Accountability [OAA], 2014). However, after operational testing, DIF analyses are typically conducted after equating and scaling has already been done. In other words, the parameter estimates are assumed to be robust to violation of unidimensionality assumption, and UIRT models are ubiquitously employed to create tests that likely include DIF items.

The potential source of DIF that causes biased test scores is arguably due to the item being multidimensional in nature but yet treated as unidimensional analytically. Further, most "problem" DIF is likely due to items with multidimensional traits that are noncompensatory in nature: to get the item correct, an examinee needs a sufficient amount of both (one trait cannot compensate for the other).

Hence, the current study simulated samples of scores of uni- and multi-dimensional items (i.e., items with and without DIF), assuming one primary and one secondary dimension (with a

number of varying item parameter conditions) that were then analyzed using five estimation model approaches to calibrate each set of items, as follows:

- Approach 1 (UIRT*d*): UIRT, no items removed from analysis;
- Approach 2 (UIRT*nds*): UIRT, after removing DIF-detected items (DIF items detected using standard criterion p-value ≤ 0.05);
- Approach 3 (UIRT*ndl*): UIRT, after removing DIF items (DIF items detected using liberal criterion p-value of 0.10);
- Approach 4 (MIRT*c*): compensatory MIRT, no items removed from analysis; and
- Approach 5 (MIRT*nc*): noncompensatory MIRT, no items removed from analysis.

Research Questions

The research questions for the current study were specifically as follows (noting that the second research question is the primary focus).

Q1: For item response data were simulated as 2PL two-dimensional noncompensatory items in which there was a low to modest a-parameter (a_2) on the secondary dimension (θ_2), how accurately do UIRT models detect DIF items?

Q2: For item response data were simulated as 2PL two-dimensional noncompensatory items in which there was a low to modest a-parameter (a_2) on the secondary dimension (θ_2), how accurately do UIRT and MIRT models calibrate the primary ability estimate for focal and reference groups?

Chapter II: Methods

The research questions regarding the accuracy and precision of primary ability estimation in the two-dimensional structure case were addressed using a Monte Carlo simulation study. All of the simulated conditions described below were completely crossed, with 500 replications per combination or "cell". The manipulated variables under investigation were estimation model approach, group size balance level, factor correlation level, percentage of DIF items, primary discrimination level, and secondary discrimination level. The impact of these variables on the accuracy of primary ability estimates was the focus of the investigation. Analyses of variance (ANOVAs) on model-derived mean ability estimates were used to identify main effects and simple interactions among conditions.

Monte Carlo Simulation

In order to investigate the effects of several complex conditions associated with the accuracy of primary ability estimation, the present study employed a simulation study using a Monte Carlo (MC) approach. Unlike real datasets, the "true" ability and "true" dimensionality structure are known a priori and thus estimates be compared to known parameters. MC simulation studies have been used widely to study IRT models (Harwell, Stone, Hsu, & Kirisci, 1996; Harwell, 1997). This technique can model realistic data conditions to compare competing statistics or methodologies in ways not possible with empirical data in which population parameters are unknown.

Experimental Conditions

Estimation Model Approach. The person and item parameters for all conditions, regardless of the model used to generate the item parameters, were estimated using five IRT models: UIRT, no items removed from analysis, which is a UIRT with DIF items (UIRT*d*);

UIRT, after removing DIF-detected items (DIF items detected using standard criterion p-value ≤ 0.05) (UIRT*nds*); UIRT, after removing DIF items (DIF items detected using liberal criterion p-value of 0.10) (UIRT*ndl);* compensatory MIRT, with no items removed from analysis (MIRT*c*); and noncompensatory MIRT, with no items removed from analysis (MIRT*nc*).

Group Size Balance Level. Two different sample size ratios of reference group to focal group members (1:1 and 9:1) were used. Although sample size is one of the most commonly studied conditions simulated in DIF research, it is an important factor to include as a control for absolute sample size when sample size ratios are being manipulated (Furlow et al., 2009; Willse & Goodman, 2008). Instances in which the ratio of reference to focal group members in the population might be expected to be approximately equal would include studies of gender DIF in which there are relatively equal numbers of boys and girls (who constitute the reference and focal group, respectively). An example of the second instance (unequal proportions) might be when a large number of non-ELL students take an exam compared with a smaller number of ELL students.

Factor Correlation Level. Two levels of factor correlation level were used, none ($\rho = 0$) and small ($\rho = 0.3$). The intent was to use no correlation to lower correlation. Prior research has used a wide variety of values for the correlation between dimensions from no or low correlation (0 and 0.3) to a fairly large correlation (0.8). Bateley and Boss (1993) included values of 0, 0.25, and 0.5 in their study, whereas Miller (1991) used 0 or 0.5 and Flora and Curran (2004) used 0.3. Other researchers have included higher inter-factor correlations, such as Gosz and Walker (2002) with 0.5, 0.75, and 0.9; Tate (2003) with 0.6; and Finch (2010) with 0, 0.3, 0.5, or 0.8.

Percentage of DIF Items. The percentage of DIF items was set at 10%, 20%, or 30%; these levels were selected based on previous research showing performance differences by these

levels (Finch & French, 2007). In addition, a no-DIF condition (0%) was also specified as a check on baseline model estimates when no item bias is present (this condition was not fully crossed with all other conditions; discussed shortly).

Discrimination Parameters. Two levels of *primary* discrimination (a_1) were used, modest (0.5) and high (0.8), and two levels of *secondary* discrimination (a_2) were also used, including low (0.2) and modest (0.5). The choice of level of each discrimination, individually, was based on low ($\delta = 0.2$), modest ($\delta = 0.5$), and high ($\delta = 0.8$) effect size levels from Cohen (1988); the selection of the specific pairs of discrimination values was based on personal experience with a range of potential values that would be seen in real data. Specifically, for DIF-free items, a_2 does not exist. However, for an item with DIF, it is possible to have relatively higher discrimination on the primary ability (i.e., a stronger relationship) than on the secondary ability (a less strong relationship), yielding the combination of $a_1 = 0.8$, and $a_2 = 0.2$. A second possibility is that the discrimination on the primary ability is strong ($a_1 = 0.8$) and additionally, that the secondary ability is increased to modest, at $a_2 = 0.5$. The third and fourth possibilities, respectively, are that items could have only modest and weak relationships with the primary and secondary abilities, respectively ($a_1 = 0.5$, and $a_2 = 0.2$). And finally, the fourth possibility could be that both discrimination parameters are modest ($a_1 = 0.5$, and $a_2 = 0.5$).

To summarize, the present study included 240 fully crossed conditions (5 Estimation Model Approaches × 2 Group Size Balance Levels × 2 Factor Correlation Levels × 3 Percentages of DIF Items × 2 Primary Discrimination Levels × 2 Secondary Discrimination Levels), along with 20 no-DIF cells (5 Estimation Model Approaches × 2 Group Size Balance Levels × 2 Primary Discrimination Levels; for these conditions, there was no factor correlation

or secondary discrimination to be concerned with). To recap, specific main effects under investigation were as follows.

- Estimation Model Approach: UIRT*d*, UIRT*nds*, UIRT*ndl*, MIRT*c*, and MIRT*nc*.
- Group Size Balance Level: Equal (1:1) and Unequal (9:1).
- Factor Correlation Level: None ($\rho = 0$) and Low ($\rho = 0.3$).
- Percentage of DIF Items: 0% (not fully crossed), 10%, 20%, and 30%.
- Primary Discrimination (a_1): High (0.8) and Modest (0.5).
- Secondary Discrimination (a_2): Modest (0.5) and Low (0.2).

Note that the 20 cells involving no DIF were incorporated in the design purposefully so that bias in baseline ability estimates across the five estimation model approaches could be established separately from the full set of conditions.

Data Generation

The data was generated according to the conditions describes above, and 500 replications of each set were simulated using *R* (R Core Team, 2015). For each of the conditions, 2,000 simulated response patterns (i.e., subjects' responses) on 40 items on a single test were generated. This sample size was chosen to calibrate IRT models with reasonable minimum sample size (Hulin, Lissak, & Drasgow, 1982; Reckase, 2009), and the test length of 40 items was used to reflect realistic testing scenarios (Kim & Oshima, 2012); this said, note that DeMars and Lau (2011) found that test length did not seriously impact recovery of person and item parameters, nor did DIF effect size.

Shealy and Stout's (1993) model for DIF magnitude was employed and given by

$$E[\theta_{2R} \mid \theta_1] - E[\theta_{2F} \mid \theta_1] = d_{\theta_2} - \rho d_{\theta_1}, \tag{11}$$

where ρ is the correlation between the θ_1 and θ_2, $d_{\theta 1}$ is the difference between the means of the focal and reference group on θ_1 (impact), and $d_{\theta 2}$ is the mean of the focal group on θ_2 subtracted from the mean of the reference group on θ_2. The mean on the primary ability held constant for both groups (indicating no impact), resulting in a $d_{\theta 1}$ value of 0. On the other hand, the difference between the reference group and focal group on the secondary ability was fixed to 1, resulting in a $d_{\theta 2}$ value of 1.

For unidimensional data (i.e., DIF-free items), items measuring only the primary ability (θ), dichotomous data were generated using the one-parameter logistic model (1PL),

$$P(X_{ij} = 1 \mid \theta_{j1}, a_1, b_{i1}) = \frac{e^{a1(\theta_{j1} - b_{i1})}}{1 + e^{a1(\theta_{j1} - b_{i1})}}, \tag{12}$$

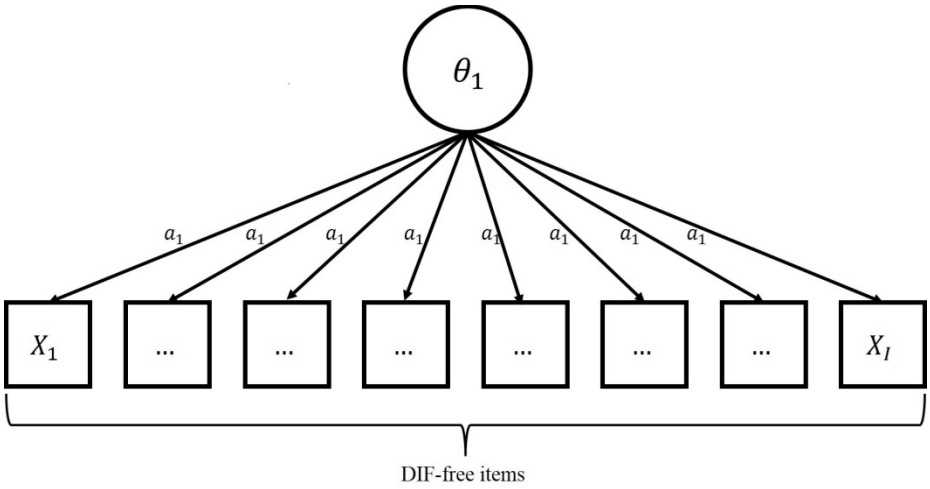
where $P(X_{ij}=1 \mid \theta_{j1})$ is the probability that an examinee correctly answers item i, b_{i1} is the item difficulty, a_1 is the item discrimination. That is, the value of a_1 was fixed to either 0.5 or 0.8 on the primary dimension.

For multidimensional data (i.e., DIF items), a noncompensatory two-dimensional model was used to simulate items measuring the primary and secondary dimensions,

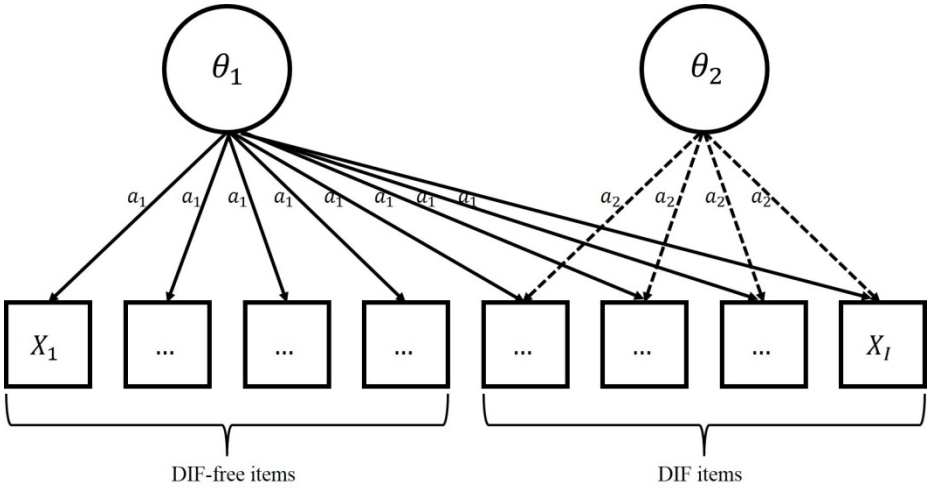
$$P(X_{ij} = 1 \mid \theta_{j1}, \theta_{j2}, a_1, a_2, b_{i1}, b_{i2}) = \frac{e^{a1(\theta_{j1} - b_{i1})}}{1 + e^{a1(\theta_{j1} - b_{i1})}} \times \frac{e^{a2(\theta_{j2} - b_{i2})}}{1 + e^{a2(\theta_{j2} - b_{i2})}}, \tag{13}$$

where $P(X_{ij}=1 \mid \theta_{j1})$ is the probability that an examinee correctly answers item i given θ_1 and θ_2, a_{i1} and a_{i2} are the discrimination parameters corresponding to the primary and secondary dimension for item i, and b_{i1} and b_{i2} are the difficulty parameters corresponding to the primary and secondary dimension for item i. The value of a_1 was fixed to either 0.5 or 0.8 on the primary dimension, additionally, the values of a_2 was either 0.2 or 0.5 on the secondary dimension (see Figure 7). The dimension-based item difficulty parameters, b_{i1} and b_{i2} were randomly generated once from a unit normal distribution $N(0, 1)$ at the beginning of the simulation process.

As for person parameters, examinees' (true) ability scores for θ_1 and $\theta_2$were drawn from a bivariate normal distribution. For reference group, the distribution of θ_1 and θ_2 were fixed to have a mean of 0 and a standard deviation of 1. For focal group, the distribution of θ_1 was fixed to have a mean of 0 and a standard deviation of 1, whereas the distribution of θ_2 was fixed to have a mean of –1 and a standard deviation of 1. In this way, the distributions of θ_1 and θ_2 for the reference and focal groups resulted in a $d_{\theta 1}$ value of 0 as well as $d_{\theta 2}$ value of 1. Finally, two different correlations between θ_1 and θ_2 were considered: 0 and 0.3. In short, θ_1 and θ_2 were simulated using the `rmvnorm` function in *R*.



Panel A: Conditions with no DIF items



Panel B: Conditions with DIF items

Figure 9. Geometric Representation of Conditions with (a) No DIF and (b) DIF Items

Finally, based on the person and item parameters, the probability of a correct response was calculated for each examinee on each item using the `plogis` function in *R*. Probabilities were then converted to item responses by comparing each probability to a random number from a uniform distribution using the `runif` function in *R*. The examinee was considered to have answered an item correctly (i.e., a response of 1) if $P(X_{ij}=1|\theta_{j1}) = 1$ was greater than the random number; otherwise, the examinee's response was scored as incorrect (i.e., 0). (Note that

this simulation process follows De Ayala's recommendations (2013, p. 417).) Figure 10 shows an example of generating one dataset under one specific condition; Figure 11 illustrates the complete process for data generation, model estimation, and results analyses (the full *R* code is given in the Appendix).

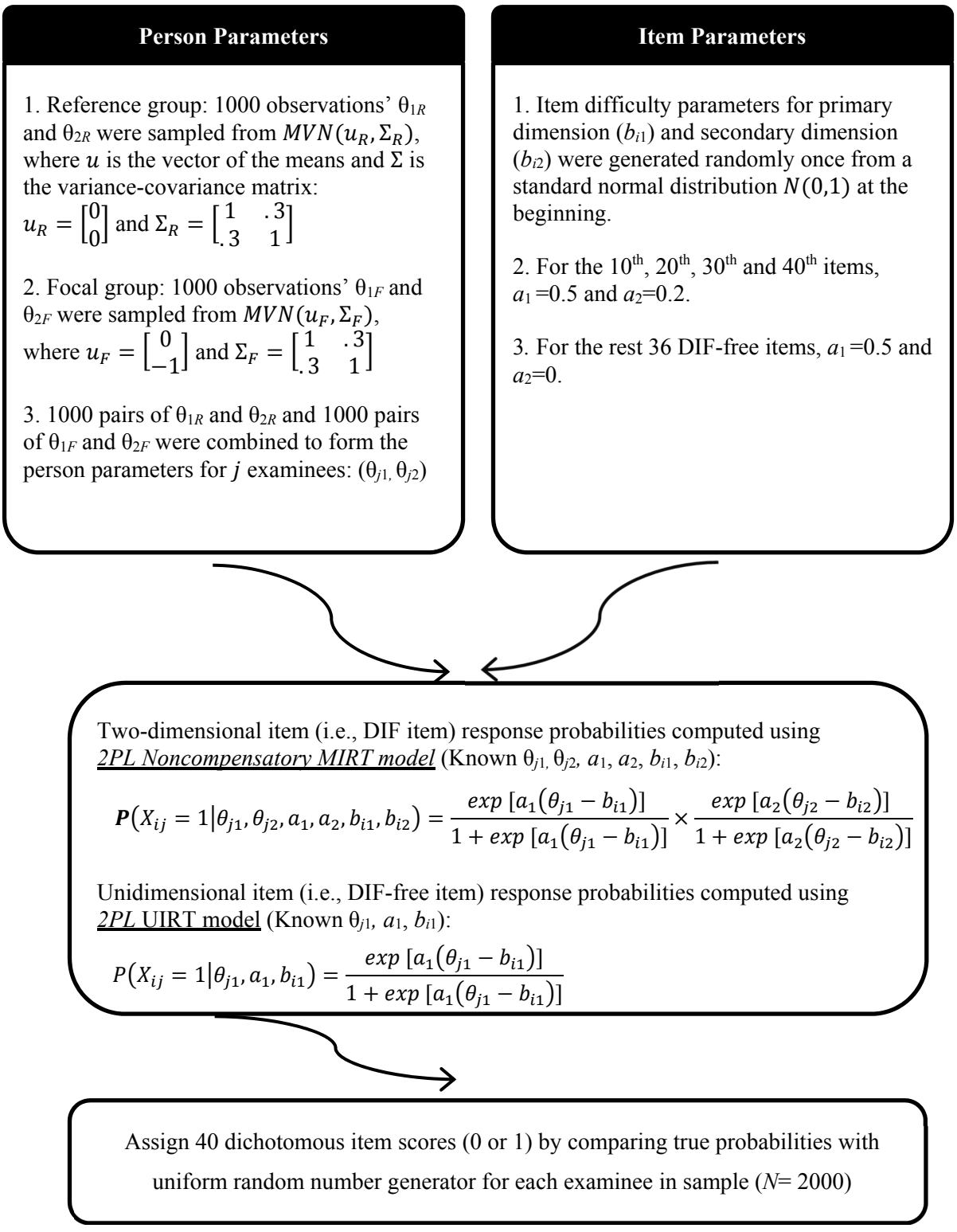


Figure 10. Data Generation of Condition with Group Size Balance = Equal; Factor Correlation Level = 0.3; Percentage of DIF items = 10%; Primary Discrimination = Modest (a_1=0.5); and Secondary Discrimination = Low (a_2=0.2)

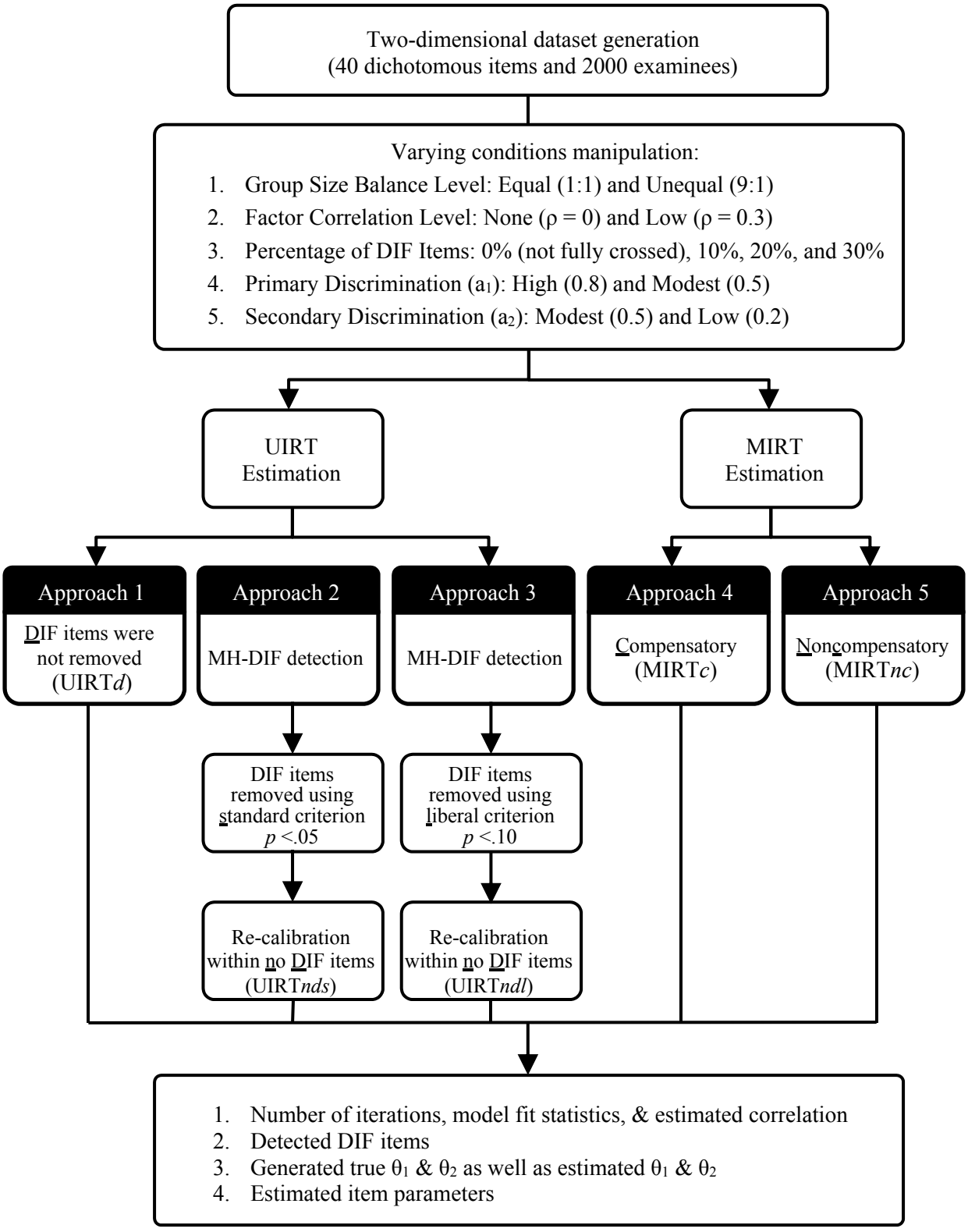


Figure 11. Flowchart for Data Generation, Calibration, and Data Storage Procedures

Model Estimation and Secondary Data Analysis

Parameter Estimation. Models were estimated in *R* using the `smirt` function within the *sirt: Supplementary Item Response Theory Models* package (Robitzsch, 2015). This function estimates the noncompensatory and compensatory MIRT models as well as the UIRT models for dichotomous data. These models were estimated using an Expectation-Maximization (EM; Dempster, Laird, & Rubin, 1977; Bock & Aitkin, 1981) algorithm employing marginal maximum likelihood (MML). A value of 1.01 was used indicating the extent of the decrease of the maximum increment in each iteration required to reach convergence for some non-converging analyses. In terms of person parameters, three IRT ability estimate methods have been reported including Expected a posteriori (EAP), Maximum (mode) a posteriori (MAP) and Maximum likelihood estimation (MLE). Although Bayesian estimators generally lead to biased estimates, their overall errors tend to be relatively small due to shrinking to the mean (EAP) or mode (MAP) (Kim, Moses, & Yoo, 2015). Therefore, EAP was utilized for estimating person parameters.

The first analytic approach involved ignoring the potential multidimensionality in the data, and all the items were fit using a 2PL UIRT model (i.e., assuming a single ability). This approach would be analogous to a researcher being unaware of the possibility that two dimensions underlie their data (i.e., unaware of potential DIF) and thus fit a standard UIRT model to the items. In the second approach, the detected DIF items were removed from analyses (using the MH chi-square procedure with either a standard (0.05) or liberal (0.10) *p*-value criterion), and the remaining items were fit again using a 2PL UIRT model (this approach would be analogous to a researcher being aware of the potential for DIF).

Finally, two types of MIRT models were estimated (compensatory and noncompensatory) using target loading matrices. For both MIRT models, all items were free to load on the primary ability but only the DIF items were free to load on the additional secondary ability.

Number of iterations and model fit information (including likelihood ratio test (LRT), Akaike's information criterion (AIC), corrected AIC (cAIC), Bayesian information criterion (BIC), and consistent AIC (CAIC)) were each saved for each simulated dataset's five estimation model approaches. For UIRT models, true ability parameter (θ_1) and ability parameter estimates ($\hat{\theta}_1$) were saved. For MIRT models, primary and secondary true ability parameters (θ_1 and θ_2) and ability estimates ($\hat{\theta}_1$ and $\hat{\theta}_2$) and correlation between ability parameter estimates ($r(\hat{\theta}_1 , \hat{\theta}_2)$) were saved. This said, in the present study, only the estimation of the first ability (i.e., the primary dimension) was the focus. In other words, only true primary ability (θ_1) and primary ability estimate ($\hat{\theta}_1$) were used for the further analyses. All bias values, the difference between θ_1 and $\hat{\theta}_1$, were then averaged for each of the conditions for use in subsequent analyses of variance. Fit indices and estimates of the secondary ability will be examined in future work.

DIF Detection. DIF detection using the Mantel-Haenszel (MH) method employed the `difMH` function within the *difR*: *Collection of Methods to Detect Dichotomous Differential Item Functioning* package (Magis, Beland, & Raiche, 2015). The resulting MH statistic is distributed approximately as chi-square with one degree of freedom (Penfield & Camille, 2007). An item was detected as DIF when its chi-square statistic was significant either at the 5% or 10% level. Power (percentage of correct identification of the known DIF items) and Type I error (percentage of non-DIF items incorrectly identified) were examined separately for all conditions.

Accuracy of Primary Ability Estimate. For each sample, with each condition, bias was calculated using the typical definition of the mean difference across the 500 replications between the estimated primary ability estimated and the true primary θ_1 used to generate the data.

Evaluation of Analysis Approach Results. In order to determine which conditions contributed to variation in the estimation of primary ability estimates, ANOVAs with two-way interactions were conducted on cell means (recall that there were 500 replicates per crossed condition or "cell") of bias, for the reference and focal groups, separately. Due to the high precision of estimates for these models, only main effects or interactions with substantive (nonzero) effect sizes were examined for follow-up pairwise comparisons (Tukey's HSD used). Specifically, the statistic omega-squared (ω^2) was used to define effect size, which is as follows:

$$\omega^2 = \frac{SSeffect - (J-1)MSwithin}{SStotal + MSwithin}, \quad (14)$$

where SS_{effect} is the sum of squared deviations for main effect or interaction, SS_{total} is the sum of squares for the model, and MS_{within} is the mean square within groups/cells. Effect size values of .01, .059, and .138 have been suggested to constitute small, medium, and large effects, respectively (Cohen, 1988; Kirk, 1996).

Analysis of DIF detection results. DIF detection Type I error (falsely identifying an item as DIF when it is not) and power (correctly identifying an item as DIF when it is) from the MH chi-square test (during the UIRT*nds* model estimates) were analyzed using a 5-factor ANOVA testing main effects and all possible 2-way interactions (note that model approach as a factor is irrelevant for this analysis). Type I error rates were specifically calculated as number of false alarms divided by the number of items with DIF in the condition; for the 0% DIF items condition, this is the number of items falsely detected divided by 40 possible items; for the 10% condition, the divisor was 36; for the 20% and 30% conditions, the divisors were 32 and 28,

respectively. The same computation was completed for power, except that the 0% DIF condition was removed from this particular analysis since there were no items to be identified. Finally, mean Type I error and power for each cell, across replicates, were computed and used in the secondary analyses of variance.

Analysis of ability estimate results from null conditions (no DIF items). The null conditions for the present study refer to the 20 conditions in which there was no DIF (i.e., items in the dataset only measured a single primary dimension such that the additional secondary dimension does not exist. Primary ability estimates for each of the 20 null conditions were analyzed using a 3-factor main effects-only ANOVA, with main effects including model approach, group size balance, and primary discrimination level. These analyses were specifically used to examine baseline data to ensure that any effects observed in the non-null conditions were not simply a reflection of baseline estimation problems.

Analysis of ability estimate results from non-null conditions (DIF items present). After establishing baseline DIF Type I error and Ability estimate bias results, both DIF and Ability Estimate data were subjected to 6-factor ANOVAs with 2-way interactions with model approach (note that it is not possible to test all interactions for a simulation study due to the cell counts exactly equal to the number of fully crossed conditions, leaving no *df* for estimation of a factorial model; further, the focal interest of the study was in moderators of approach effects on primary ability estimates). Recall that the main effects across all non-null conditions included model approach, group size balance, factor correlation, percentage of DIF items, primary discrimination, and secondary discrimination.

Chapter III: Results

The present study was designed to examine whether multidimensional item response theory (MIRT) models might be useful in controlling for differential item functioning (DIF) when estimating primary ability, or whether traditional (and simpler) unidimensional item response theory (UIRT) models with DIF items removed are sufficient for accurately estimating primary ability. Recall that the data in this study were generated for a test with 40 items and 2,000 examinees, and assumed a 2-dimensional, noncompensatory case (e.g., testing a primary dimension of geometry skills but involving also a secondary dimension of language comprehension proficiency). Between-dimension correlation, reference-to-focal group size balance, and primary and secondary dimension discrimination levels were varied, as was percentage of DIF items. Most conditions were fully crossed, and the levels selected for use in each condition were grounded in prior applied and theoretical research. For each of the 500 simulated datasets per condition, five estimation model approaches were applied, with results saved and averaged for each condition. Recall that the five approaches were as follows.

- Approach 1 (UIRT*d*): UIRT, no items removed from analysis;
- Approach 2 (UIRT*nds*): UIRT, after removing DIF-detected items (DIF items detected using standard criterion p-value ≤ 0.05);
- Approach 3 (UIRT*ndl*): UIRT, after removing DIF items (DIF items detected using liberal criterion p-value of 0.10);
- Approach 4 (MIRT*c*): compensatory MIRT, no items removed from analysis; and
- Approach 5 (MIRT*nc*): noncompensatory MIRT, no items removed from analysis.

Table 1 provides the mean number of iterations required for model convergence for each level of each main effect as well as the four combinations of a-parameter conditions.

Table 1. *Mean Number of Iterations by Model Approach and Condition Levels*

Factor Correlation Level	$\rho = 0$	$\rho = 0.3$		
UIRT*d*	15.87 (1.66)	15.49 (1.69)		
UIRT*nds*	14.95 (1.93)	14.47 (1.80)		
UIRT*ndl*	14.42 (1.92)	13.92 (1.78)		
MIRT*c*	54.85 (52.77)	80.49 (61.72)		
MIRT*nc*	292.96 (203.71)	463.00 (102.85)		
Reference:Focal Group Size Ratio	*Equal (1:1)*	*Unequal (9:1)*		
UIRT*d*	15.75 (1.70)	15.66 (1.67)		
UIRT*nds*	14.55 (2.05)	14.94 (1.69)		
UIRT*ndl*	13.96 (2.05)	14.45 (1.65)		
MIRT*c*	57.00 (52.23)	74.68 (62.35)		
MIRT*nc*	342.65 (179.94)	389.01 (194.55)		
Percentage of DIF	*0%*	*10%*	*20%*	*30%*
UIRT*d*	17.27 (1.73)	16.11 (1.88)	15.24 (1.43)	14.98 (0.99)
UIRT*nds*	16.83 (1.61)	15.73 (1.47)	14.43 (1.02)	13.03 (1.34)
UIRT*ndl*	16.35 (1.50)	15.28 (1.31)	13.90 (0.88)	12.37 (1.33)
MIRT*c*	17.27 (1.73)	107.76 (61.16)	65.46 (56.58)	48.58 (42.47)
MIRT*nc*	17.27 (1.73)	418.09 (102.83)	417.79 (136.48)	435.89 (152.93)
Primary Discrimination	$a_1 = 0.5$	$a_1 = 0.8$		
UIRT*d*	17.06 (1.19)	14.35 (0.67)		
UIRT*nds*	15.76 (1.89)	13.72 (1.18)		
UIRT*ndl*	15.08 (1.97)	13.33 (1.27)		
MIRT*c*	52.95 (45.33)	78.72 (66.16)		
MIRT*nc*	354.50 (188.66)	377.16 (188.37)		
Secondary Discrimination	$a_2 = 0.2$	$a_2 = 0.5$		
UIRT*d*	15.91 (1.75)	15.50 (1.59)		
UIRT*nds*	15.15 (1.73)	14.34 (1.95)		
UIRT*ndl*	14.64 (1.66)	13.77 (1.97)		
MIRT*c*	103.96 (58.37)	27.71 (18.86)		
MIRT*nc*	462.14 (189.86)	269.52 (126.72)		
Discrimination Combinations	$a_1 = 0.5, a_2 = 0.2$	$a_1 = 0.5, a_2 = 0.5$	$a_1 = 0.8, a_2 = 0.2$	$a_1 = 0.8, a_2 = 0.5$
UIRT*d*	17.37 (1.05)	16.76 (1.28)	14.44 (0.82)	14.25 (0.49)
UIRT*nds*	16.33 (1.40)	15.20 (2.19)	13.97 (1.11)	13.48 (1.24)
UIRT*ndl*	15.67 (1.46)	14.48 (2.27)	13.61 (1.16)	13.06 (1.36)
MIRT*c*	82.89 (47.36)	23.02 (9.74)	125.04 (62.25)	32.40 (24.42)
MIRT*nc*	457.66 (190.24)	251.35 (121.71)	466.62 (196.55)	287.7 (133.52)

Note. Numbers in parentheses represent standard deviations.

The stopping criterion for model convergence was set at a maximum change of < 0.001. The UIRT models only needed a few iterations to converge at 0.001: UIRT*d* M = 15.70 (SD = 1.67), UIRT*nds* M = 14.74 (SD = 1.87), and UIRT*ndl* M = 14.21 (SD = 1.86). The MIRT models, on the other hand, needed more iterations: MIRT*c* M = 65.84 (SD = 57.68), and MIRT*nc* M = 365.83 (SD = 187.14). As for three UIRT approaches, UIRT*ndl* and UIRT*nds* averaged the least number of iterations, respectively, since the data were close to one dimension after removing the detected DIF items. However, using UIRT*d* to calibrate data involving two-dimensional items, the relatively more iterations were needed. For the primary and secondary discrimination combination of a_1 = 0.8 or 0.5 with a_2 = 0.2 (DIF items with stronger relationships with the primary ability, which would be similar to unidimensional data), both MIRT approaches needed more iterations to converge. In contrast, for the combination of a_1 = 0.5 and a_2 = 0.5 (equal relationships with each dimension), MIRT approaches needed relatively fewer iterations to converge.

Research Question 1. For simulated item data in which there is a low to modest a-parameter on the secondary dimension, how accurately do UIRT models detect DIF items?

The descriptive and inferential results from the Mantel-Haenszel (MH) DIF tests are given in Tables 2 and 3 for Type I error rates, respectively. In the present context, Type I error is the percentage of *non-DIF items* that were incorrectly identified as DIF (using a nominal alpha level of 0.05 would indicate that 5% of items should be incorrectly identified if the test is working properly; and by extension, using a nominal alpha level of 0.10 would indicate that 10% of items should be incorrectly identified). Power is simply the percentage of correctly identified DIF items.

MH DIF Detection Type I Error Rates. The Type I error rates for the null conditions (no DIF items) served as a baseline condition to assist with evaluating the effectiveness in the non-null conditions (DIF items present).The empirical Type I error rates reported here is the standard nominal alpha of 0.05 (i.e., the UIRT*nds* model, Approach 1). Type I error for the alpha level of 0.10 yielded nearly identical patterns; these results are available from the author upon request). Error rates represent the mean across 40 items within each condition.

Across the 0% DIF conditions, the mean Type I error rate was slightly lower than the nominal 5% value, averaging 0.04 (*SD* = 0.003) (this rate was consistent across all subconditions).

However across the 10-30% DIF conditions, the mean Type I error rate for incorrectly detecting DIF (for the remaining 90-70% of items, respectively) was inflated, averaging *M* = 0.10 (*SD* = 0.08) and increasing with the percentage of real DIF items present (up to 16% of non-DIF items falsely identified for the conditions in which 30% of the other items did have DIF).

Table 2. *MH DIF Detection Type I Error Rates Using UIRTnds (Alpha = 0.05)*

Factor Correlation Level	*ρ = 0*	*ρ = 0.3*		
	0.095 (0.080)	0.101 (0.074)		
Group Size Ratio	*Equal (1:1)*	*Unequal (9:1)*		
	0.127 (0.094)	0.069 (0.037)		
Percentage of DIF items	*0%*	*10%*	*20%*	*30%*
	0.042 (0.003)	0.054 (0.013)	0.092 (0.044)	0.162 (0.104)
Primary Discrimination	$a_1 = 0.5$	$a_1 = 0.8$		
	0.110 (0.092)	0.086 (0.058)		
Secondary Discrimination	$a_2 = 0$	$a_2 = 0.2$	$a_2 = 0.5$	
	0.042 (0.003)	0.069 (0.029)	0.137 (0.096)	

Note. Numbers in parentheses represent standard deviations

As shown in Table 3, an analysis of variance (ANOVA) with main effects and all two-way interactions showed that MH DIF detection Type I error rates were affected by group size

balance, percentage of DIF items, primary discrimination, and secondary discrimination. Specifically, the inflation was worse for equal group sizes, a greater number of DIF items, and when the primary discrimination was lower, and the secondary discrimination was higher. Although there was a significant difference in the Type I error rate between the factor correlation levels, the effect size was negligible; further, there were no interactions between factor correlation and other conditions. Follow-up comparisons using Tukey's HSD pairwise *q*-tests showed that, while the 0% and 10% DIF conditions did not significantly differ, each of the other pairs of DIF levels did significantly differ (*ps* < 0.001); further, the pattern in the data is a near-perfect quadratic function in predicting Type I error by DIF percentage level. Follow-up tests also showed that the three levels of secondary discrimination significantly differed from one another, with the smallest Type I error rate found in the $a_2 = 0$ condition (when no secondary dimension existed), and the highest Type I error for the modest condition ($a_2 = 0.5$).

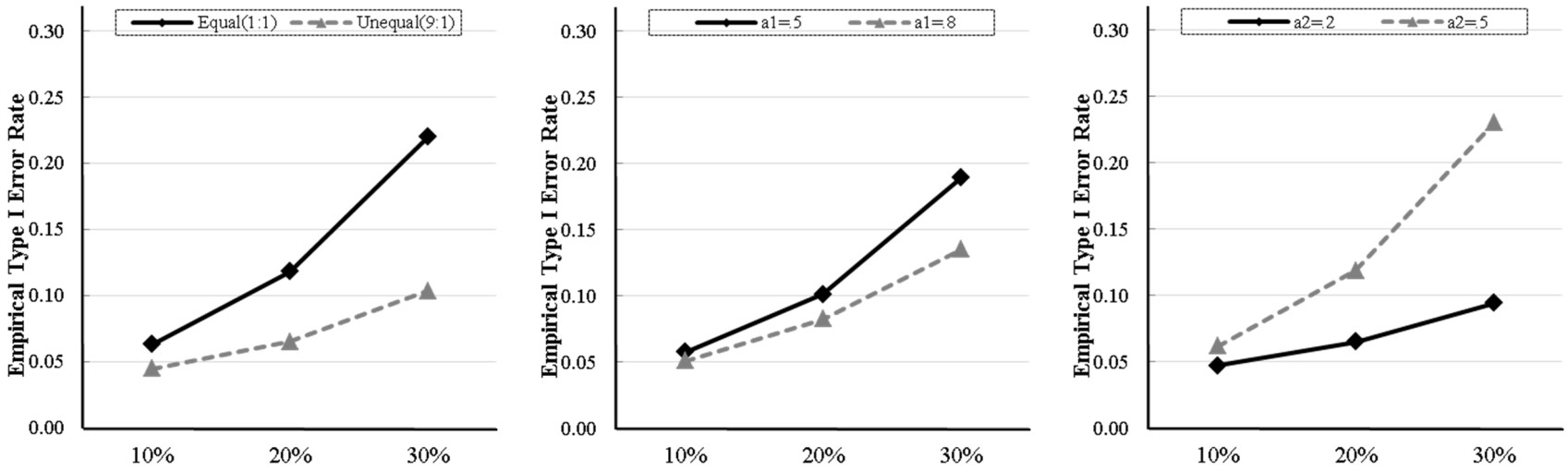
Table 3. *ANOVA Results for MH DIF Detection Type I Error Using UIRT*nds *(Alpha = 0.05)*

Source	*SS*	*df*	*MS*	*F*	ω^2
Correlation Level	0.00	1	0.00	4.48*	0.00
Group Size Balance	0.04	1	0.04	113.32***	0.13
% of DIF Items	0.11	3	0.04	115.55***	0.40
Primary Discrim a_1	0.01	1	0.01	20.27***	0.02
Secondary Discrim a_2	0.06	1	0.06	145.94***	0.17
Correlation × Group	0.00	1	0.00	0.16	0.00
Correlation × % of DIF	0.00	2	0.00	0.91	0.00
Correlation × Primary a_1	0.00	1	0.00	0.32	0.00
Correlation × Secondary a_2	0.00	1	0.00	0.07	0.00
Group × % of DIF	0.02	2	0.01	25.77***	0.07
Group × Primary a_1	0.00	1	0.00	4.02	0.00
Group × Secondary a_2	0.01	2	0.01	18.25***	0.03
% of DIF × Primary a_1	0.01	2	0.00	6.36**	0.01
% of DIF × Secondary a_2	0.03	2	0.02	40.15***	0.09
Primary a_1 × Secondary a_2	0.00	2	0.00	4.79*	0.01
Error	0.01	28	0.00		

Note. $R^2 = 0.97$ (Adjusted $R^2 = 0.94$).
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

As also shown in the ANOVA results (Table 3), there were significant interaction effects on Type I error rates among all main effect conditions except factor correlation level. Combinations of these condition means were plotted in Figure 12 to determine the nature of these joint effects. The figure shows that Type I error inflation was greater for conditions in which there was a relatively high percentage of DIF items combined with equal group sizes (over 20% of items falsely detected), lower primary discrimination (approximately 20% falsely detected), and higher secondary discrimination (over 20% Type I error rate) (Panels A-C). There was also increased Type I error inflation for conditions in which there was both equal group sizes and lower primary or higher secondary discriminations (Panels D-E), as well as for the combination of lower primary and higher secondary discriminations (Panel F). Although it is likely that there are 3- or 4-way interactions present, it was not possible to test all interactions given the number of cells available. Nevertheless, it seems noteworthy that the cell mean for the combination of 30% DIF items, equal group sizes, and the combination of both low primary and high secondary discrimination was extraordinarily high, at 38% (SD = 0.04).

(A) Group Size Balance by % of DIF Items (B) Primary Discrim a_1 by % of DIF Items (C) Secondary Discrim a_2 by % of DIF Items



(D) Group Size by Primary Discrim a_1 (E) Group Size by Secondary Discrim a_2 (F) Primary a_1 by Secondary Discrim a_2

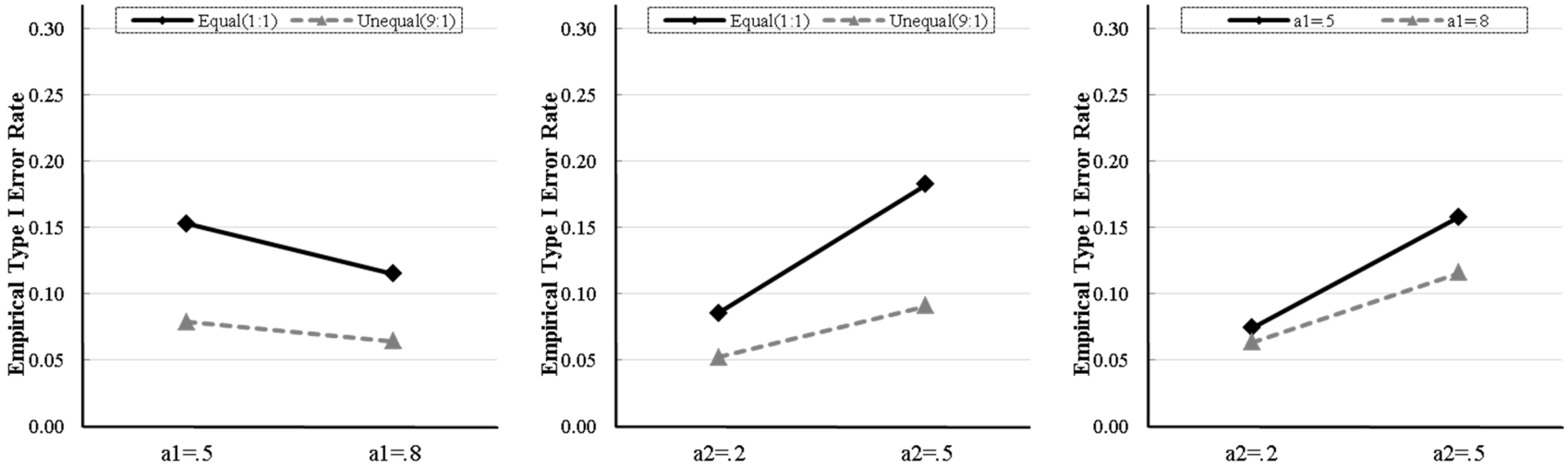


Figure 12. Interactions among Conditions on MH DIF Detection Type I Error Rates (UIRT*nds*, Alpha = 0.05)

MH DIF Detection Power. Power rates for MH DIF detection across conditions that had DIF present, with a nominal alpha of 0.05 (UIRT*nds*), averaged 58% (*SD* = 0.27); levels of each condition are summarized in Table 4.

Table 4. *MH DIF Detection Power for UIRT*nds *(Alpha = 0.05)*

Factor Correlation Level	*ρ = 0*	*ρ = 0.3*	
	0.586 (0.270)	0.590 (0.281)	
Group Size Ratio	*Equal (1:1)*	*Unequal (9:1)*	
	0.747 (0.220)	0.430 (0.225)	
Percentage of DIF	*10%*	*20%*	*30%*
	0.654 (0.277)	0.584 (0.269)	0.526 (0.274)
Primary Discrimination	$a_1 = 0.5$	$a_1 = 0.8$	
	0.646 (0.264)	0.530 (0.274)	
Secondary Discrimination	$a_2 = 0.2$	$a_2 = 0.5$	
	0.390 (0.188)	0.786 (0.186)	

Note. Numbers in parentheses represent standard deviations.

As Table 5 shows, the ANOVA results revealed that power was affected by group size balance, percentage of DIF items, primary discrimination, and secondary discrimination, but not factor correlation level. More specifically, highest power was achieved when the group size ratio was equal, when percentage of DIF items decreased, and when the primary and secondary discriminations were modest. None of the interaction effects were statistically significant.

Table 5. *ANOVA Results for MH DIF Detection Power Using UIRT*nds *(Alpha = 0.05)*

Source	*SS*	*df*	*MS*	*F*	ω^2
Correlation Level	0.00	1	0.00	0.08	0.00
Group Size Balance	1.21	1	1.21	412.50***	0.34
% of DIF Items	0.13	2	0.07	22.42***	0.04
Primary Discrim a_1	0.16	1	0.16	55.30***	0.05
Secondary Discrim a_2	1.88	1	1.88	643.45***	0.54
Correlation × Group	0.00	1	0.00	0.02	0.00
Correlation × % of DIF	0.01	2	0.01	1.84	0.00
Correlation × Primary a_1	0.00	1	0.00	0.32	0.00
Correlation × Secondary a_2	0.00	1	0.00	0.13	0.00
Group × % of DIF	0.01	2	0.00	0.83	0.00
Group × Primary a_1	0.01	1	0.01	2.49	0.00
Group × Secondary a_2	0.00	1	0.00	0.00	0.00
% of DIF × Primary a_1	0.00	2	0.00	0.50	0.00
% of DIF × Secondary a_2	0.00	2	0.00	0.72	0.00
Primary a_1 × Secondary a_2	0.00	1	0.00	0.01	0.00
Error	0.08	27	0.00		

Note. $R^2 = 0.98$ (Adjusted $R^2 = 0.96$).
$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Research Question 2: For simulated item data in which there is a low to modest a-parameter on the secondary dimension, how accurately do UIRT and MIRT models calibrate the primary ability estimate for focal and reference groups?

For each primary ability estimate, bias was calculated using the typical definition of the mean differences between the estimated primary ability ($\hat{\theta}_1$) and the true primary θ_1 used to generate theta; the focal group results are presented first, followed by the reference group results.

Focal Group Analyses

Null Conditions. Recall that the null conditions for the present study refer to the 20 conditions in which there was no DIF generated. Means for the focal group when no DIF was present showed that $\hat{\theta}_1$s averaged close to zero across model approaches, with –0.044 (UIRT*d*), –0.026 (UIRT*nds*), –0.028 (UIRT*ndl*), –0.036 (MIRT*c*), and –0.034 (MIRT*nc*). (Note that the standard errors for all conditions were extremely close to zero.) Table 6 given below gives the

mean bias for each level of each main effect, which clearly demonstrates tiny, if any, bias (overall mean of 0.001).

Table 6. *Focal Group Mean Bias in θ_1 by Condition (Null Conditions Only)*

Model Approach	*UIRTd*	*UIRTnds*	*UIRTndl*	*MIRTc*	*MIRTnc*
	0.001	0.001	0.001	0.001	0.001
Group Size Ratio	*Equal (1:1)*	*Unequal(9:1)*			
	0.008	–0.006			
Primary Discrimination a_1	*$a_1 = 0.5$*	*$a_1 = 0.8$*			
	0.005	–0.003			

Table 7 reports the 3-factor main effects-only ANOVA results on the null conditions, with main effects including model approach, group size balance, and primary discrimination level. Results indicated that group size balance and primary discrimination were significant predictors of bias; however, there was not a significant difference in bias among the five estimation model approaches. Clearly, although the overall bias was negligible across conditions in terms of analytic approaches, there was some bias in estimates present for different group size ratios (equally sized groups had a positive bias for the focal group and unequal groups had a negative bias for the focal group) and primary discrimination levels (positive bias for the focal group with modest discrimination whereas slightly negative bias for strong discrimination).

Table 7. *ANOVA Results for Focal Group Mean Bias in θ_1 (Null Conditions Only)*

Source	*SS*	*df*	*MS*	*F*	ω^2
Model Approach	0.00	4	0.00	0.00	0.00
Group Size Balance	0.00	1	0.00	17.93**	0.46
Primary Discrim a_1	0.00	1	0.00	5.04*	0.11
Error	0.00	13	0.00		

Note. $R^2 = 0.64$ (Adjusted $R^2 = 0.45$).
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

DIF Conditions. Descriptive statistics for bias on the primary ability estimate across conditions with DIF items are given in Table 8 below for each of the levels of the main effects as

well as the four combinations of the *a*-parameter levels. Based on visual inspection, mean bias was greatest for the UIRT*d* model and least for the UIRT*nds* model, but on average across all model approaches was greater than was seen in the null conditions.

Table 8. *Focal Group Mean Bias in θ_1 by Condition (DIF Conditions)*

Model Approach	*UIRTd*	*UIRTnds*	*UIRTndl*	*MIRTc*	*MIRTnc*
	–0.048	–0.028	–0.030	–0.039	–0.037
Factor Correlation Level	*ρ = 0*	*ρ = 0.3*			
	–0.033	–0.040			
Group Size Ratio	*Equal (1:1)*	*Unequal(9:1)*			
	–0.025	–0.048			
% of DIF Items	*10%*	*20%*	*30%*		
	–0.015	–0.038	–0.056		
Primary Discrimination a_1	*a_1 = 0.5*	*a_1 = 0.8*			
	–0.043	–0.030			
secondary Discrimination a_2	*a_2 = 0.2*	*a_2 = 0.5*			
	–0.029	–0.044			
Discrimination Combinations	*a_1 = 0.5, a_2 = 0.2*	*a_1 = 0.5, a_2 = 0.5*	*a_1 = 0.8, a_2 = 0.2*	*a_1 = 0.8, a_2 = 0.5*	
	–0.037	–0.049	–0.022	–0.039	

After establishing baseline results (recall that there were no differences between model approaches on bias for null conditions), a 6-factor ANOVA with all 2-way interactions was conducted on cell means from all of the fully crossed non-null conditions (i.e., data with DIF items); these results are reported in Table 9. Results indicated that all main effects were significant. Follow-up post hoc comparisons using Tukey's HSD pairwise *q*-tests showed that the only significant difference among the five estimation model approaches was between UIRT*d* and UIRT*nds* and UIRTnd*l* (*p*s < 0.01; removal of DIF items yielded significantly less underestimate of primary ability for the focal group). The follow-up tests also showed that the three DIF percentages each significantly differed from one another (*q*-test *p*s < 0.001), with the least underestimation of primary ability found in the 10% DIF condition, and the most in the

30% DIF condition (see again Table 8). The other main effects each contain only two levels; hence, the direction of differences can be found by inspecting means in Table 8: underestimation bias was greatest for correlated factors (compared to uncorrelated), unequal group sizes (compared to equal), moderate primary ability discrimination (compared to higher), and moderate secondary discrimination (compared to lower). In fact, the combination of moderate primary and secondary discriminations together yield the highest bias compared to the other combinations of the a-parameters.

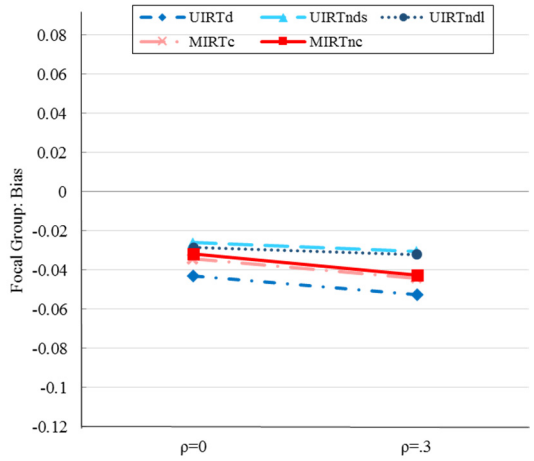
Although none of the interactions were statistically significant, because this study was interested in the mean bias of primary ability estimated by different model approach under various condition levels, plots were created (see Figure 13) to illustrate the general lack of interaction effects between model approaches and other main effects. As can be seen, the greatest bias was found with the UIRT*d* model for all conditions. The least bias was found with the UIRT*nds* and UIRT*ndl* models except in the 30% DIF items condition, where UIRT*nds*, UIRT*ndl*, MIRT*c*, and MIRT*nc* all demonstrate the same amount of bias.

Table 9. *ANOVA Results for Focal Group Mean Bias in θ_1 by Condition (DIF Conditions)*

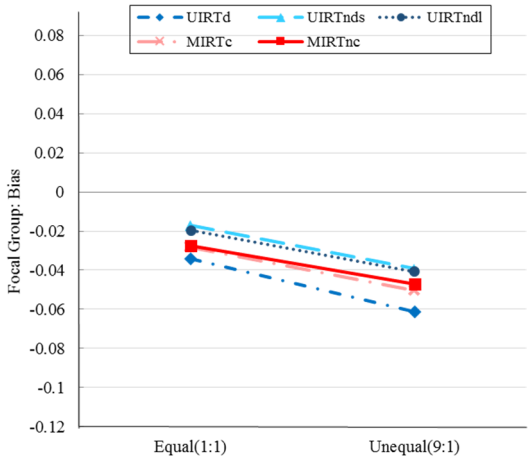
Source	*SS*	*df*	*MS*	*F*	ω^2
Model Approach	0.01	4	0.00	4.51**	0.03
Correlation Level	0.00	1	0.00	5.33*	0.01
Group Size Balance	0.03	1	0.03	46.20***	0.11
% of DIF Items	0.07	2	0.03	52.23***	0.24
Primary Discrim a_1	0.01	1	0.01	14.58***	0.03
Secondary Discrim a_2	0.01	1	0.01	19.99***	0.04
Model × Correlation	0.00	4	0.00	0.20	0.00
Model × Group	0.00	4	0.00	0.15	0.00
Model × % of DIF	0.00	8	0.00	0.58	0.00
Model × Primary a_1	0.00	4	0.00	0.34	0.00
Model × Secondary a_2	0.00	4	0.00	1.42	0.00
Error	0.13	205	0.00		

Note. $R^2 = .78$ (Adjusted $R^2 = .74$).
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

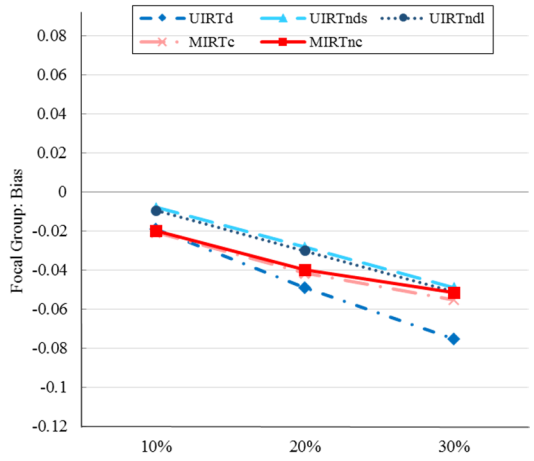
Panel A: Factor Correlation



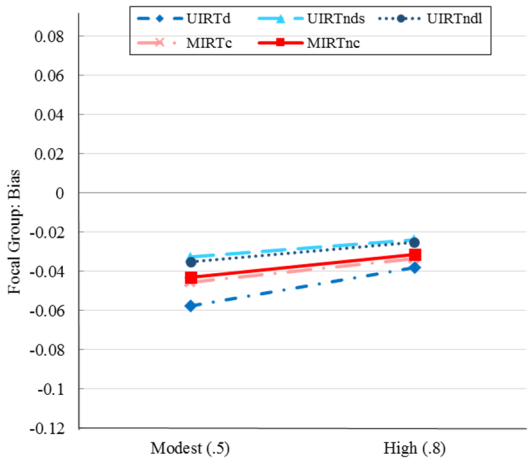
Panel B: Group Size Ratio



Panel C: Percentage of DIF items



Panel D: Primary Discrimination



Panel E: Secondary Discrimination

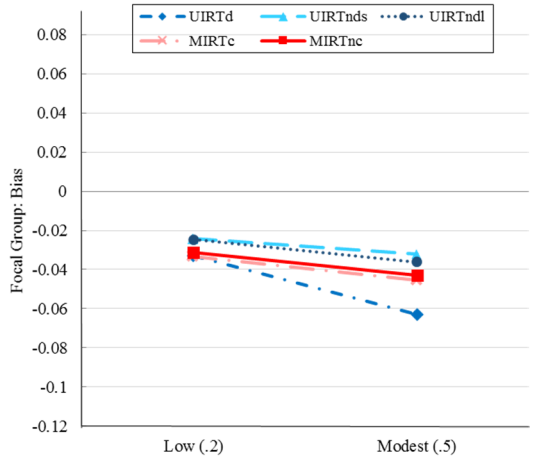


Figure 13. Focal Group Mean Bias by Model Approach and Main Condition Levels for Non-Null Conditions (DIF items present)

Reference Group Analyses

Although estimates for the focal group are indeed the focus of this study, reference group estimates should also be examined. As was done in the foregoing, results are presented for null and non-null conditions,

Null Conditions. Again, recall that the null conditions for the present study refer to the 20 conditions in which there was no DIF generated. Means for the reference group when no DIF was present showed that $\hat{\theta}_1$s averaged close to zero across model approaches, with –0.001 (UIRT*d*), –0.002 (UIRT*nds*), –0.002 (UIRT*ndl*), –0.001 (MIRT*c*), and –0.001 (MIRT*nc*). (Note that the standard errors for all conditions were extremely close to zero.) Table 10 given below gives the mean bias for each level of each main effect, which clearly demonstrates tiny, if any, bias (overall mean of –0.002).

Table 10. *Reference Group Mean Bias in θ_1 by Condition (Null Conditions Only)*

Model Approach	*UIRTd*	*UIRTnds*	*UIRTndl*	*MIRTc*	*MIRTnc*
	–0.001	–0.002	–0.002	–0.001	–0.001
Group Size Ratio	*Equal (1:1)*	*Unequal(9:1)*			
	0.005	–0.008			
Primary Discrimination a_1	$a_1 = 0.5$	$a_1 = 0.8$			
	–0.002	–0.001			

Table 11 reports the 3-factor main effects-only ANOVA results on the null conditions, with main effects including model approach, group size balance, and primary discrimination level. Results indicated that group size balance was significant predictor of bias; however, there were no significant differences in bias among the five estimation model approaches or between two primary discrimination levels. Clearly, there was some bias in estimates present for different group size ratios (equally sized groups had a positive bias for the reference group and unequal groups had a negative bias for the reference group).

Table 11. *ANOVA Results for Reference Group Mean Bias in θ_1 (Null Conditions Only)*

Source	*SS*	*Df*	*MS*	*F*	ω^2
Model Approach	0.00	4	0.00	0.00	0.00
Group Size Balance	0.00	1	0.00	14.87**	0.48
Primary Discrim a_1	0.00	1	0.00	0.30	0.00
Error	0.00	13	0.00		

Note. $R^2 = 0.55$ (Adjusted $R^2 = 0.33$).
**$p < 0.01$

DIF Conditions. Descriptive statistics for bias on the primary ability estimate across conditions with DIF items are given in Table 12 below for each of the levels of the main effects as well as the four combinations of the *a*-parameter levels. Based on visual inspection, mean bias was greatest for the UIRT*d* model and least for the UIRT*nds* and *MIRTnc* models, but on average across all model approaches was greater than was seen in the null conditions. It is important to note that, while the bias for the focal group was negative, bias for the reference group was positive. This suggests that, under the DIF conditions, ability for the focal group is underestimated and ability for the reference group is overestimated.

Table 12. *Reference Group Mean Bias in θ_1 by Condition (DIF Conditions)*

Model Approach	*UIRTd*	*UIRTnds*	*UIRTndl*	*MIRTc*	*MIRTnc*
	0.022	0.012	0.013	0.014	0.012
Factor Correlation Level	*ρ=0*	*ρ=.3*			
	0.015	0.014			
Group Size Ratio	*Equal(1:1)*	*Unequal(9:1)*			
	0.021	0.009			
% of DIF Items	*10%*	*20%*	*30%*		
	0.008	0.013	0.024		
Primary Discrimination a_1	*a_1 = 0.5*	*a_1 = 0.8*			
	0.014	0.015			
Secondary Discrimination a_2	*a_2 = 0.2*	*a_2 = 0.5*			
	0.015	0.014			
Discrimination Combinations	*a_1 = 0.5, a_2 = 0.2*	*a_1 = 0.5, a_2 = 0.5*	*a_1 = 0.8, a_2 = 0.2*	*a_1 = 0.8, a_2 = 0.5*	
	0.013	0.016	0.017	0.013	

After establishing baseline results (recalling that there were no significant differences between model approaches on bias), a 6-factor ANOVA (main effects with model approach 2-way interactions) was conducted on cell means from all of the fully crossed non-null conditions (i.e., data with DIF items); these results are reported in Table 13. Results indicated that only group size balance and percentage of DIF items were significant.

Table 13. *ANOVA Results for Reference Group Mean Bias in θ_I by Condition (DIF Conditions)*

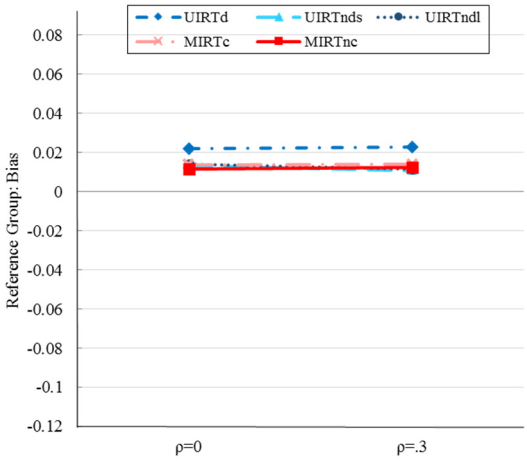
Source	*SS*	*df*	*MS*	*F*	ω^2
Model Approach	0.00	4	0.00	1.60	0.01
Correlation Level	0.00	1	0.00	0.04	0.00
Group Size Balance	0.01	1	0.01	14.95***	0.06
% of DIF Items	0.01	2	0.01	9.44***	0.07
Primary Discrim a_1	0.00	1	0.00	0.05	0.00
Secondary Discrim a_2	0.00	1	0.00	0.02	0.00
Model × Correlation	0.00	4	0.00	0.06	0.00
Model × Group	0.00	4	0.00	0.73	0.00
Model × % of DIF	0.00	8	0.00	0.10	0.00
Model × Primary	0.00	4	0.00	0.11	0.00
Model × Secondary	0.00	4	0.00	0.25	0.00
Error	0.12	205	0.00		

Note. $R^2 = 0.39$ (Adjusted $R^2 = 0.29$).
***$p < 0.001$.

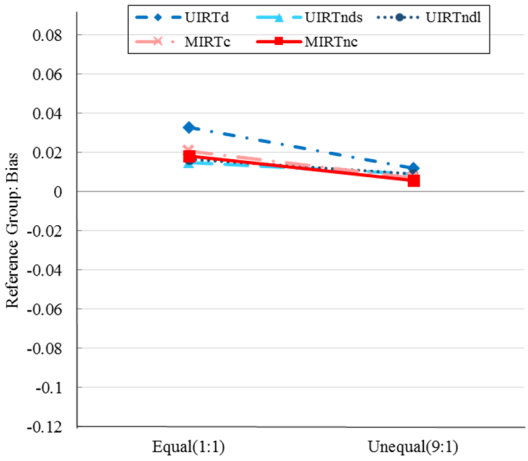
Inspection of the means in Table 12 show that balanced groups yields more overestimation of primary ability for the reference group than does imbalanced groups. To determine the location of the differences for the %DIF items effect, follow-up comparisons using Tukey's HSD pairwise *q*-tests showed that 30% of DIF items differed significantly from 10% of DIF items (*q*-test $p < 0.001$) and 20% of DIF items ($p < 0.05$), respectively; there was no significant difference between the 10% and 20% DIF conditions. In other words, having a high percentage of DIF items yielded significantly more overestimation of primary ability for the reference group (which is directly opposite of the focal group results).

Finally, although none of the interactions were statistically significant, we were interested in examining mean bias of primary ability estimated by different model approach under each of the other main effect conditions (see Figure 14). As can be seen, under all conditions, the bias for the UIRT*d* model showed the greatest bias and there was little to no difference in bias for the other four models.

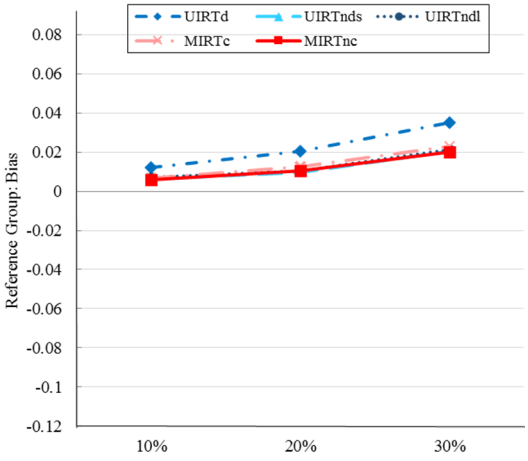
Panel A: Factor Correlation



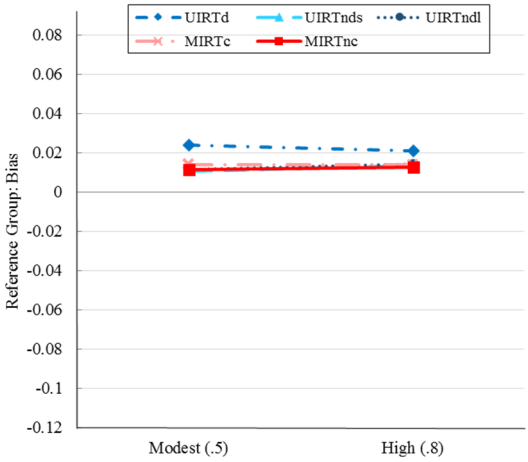
Panel B: Group Size Ratio



Panel C: Percentage of DIF items



Panel D: Primary Discrimination



Panel E: Secondary Discrimination

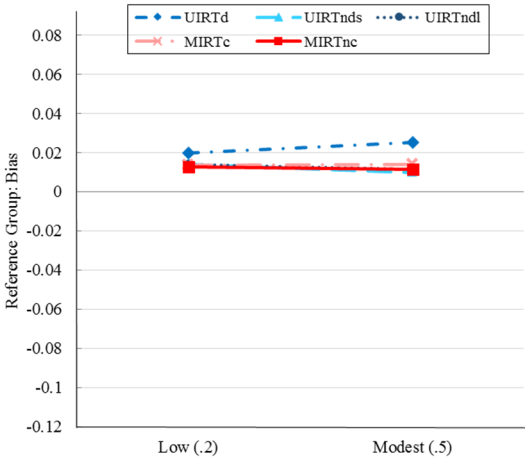


Figure 14. Reference Group Mean Bias by Model Approach and Main Condition Levels for Non-Null Conditions (DIF items present)

Chapter IV: Discussion

The present Monte Carlo simulation study compared five estimation model approaches to items that were generated with and without differential item functioning (DIF) due to low scores on a second (secondary) ability. An example of this issue occurring in practice would easily occur when an English language learner (ELL) must respond to a geometry item that has a high demand on language comprehension: the ELL student would have a difficult time responding to the question correctly due to the language comprehension– not geometry– demand; whereas a student who is a native English speaker with equal geometry ability would have a much higher probability of answering the question correctly. Hence, use of an item response theory (IRT) model, which assumes only one true underlying dimension and which assumes no DIF, would inaccurately estimate the ELL child's true geometry ability (i.e., the primary ability of interest) for any items that require a construct-irrelevant secondary (yet not estimated) ability. In other words, primary ability (geometry) estimates for the focal, or non-normative, group (ELL individuals in this case) would be biased (underestimated).

To avoid this issue of bias, particularly if the secondary dimension is deemed to be irrelevant to the constructs intended to be measured, one can either remove items that exhibit DIF and then estimate primary ability using the remaining items (which requires detecting DIF in the first place – one may use the standard chi-square test statistic alpha level of .05, or a more liberal alpha level of .10), or the researcher might use a more complex multidimensional item response theory (MIRT) model, which is an extension of unidimensional item response theory (UIRT) and relaxes the assumption of unidimensionality by estimating multiple abilities simultaneously and allowing for the inclusion of items that measure multiple abilities or traits. This said, there are two kinds of MIRT modeling approaches: noncompensatory and

compensatory. Noncompensatory MIRT modeling (MIRT*nc*) has not been studied as much as compensatory MIRT (MIRT*c*) due to its complexity. Yet, ability in one dimension (i.e., a secondary ability) will not necessarily compensate for low ability in another dimension (i.e., primary ability).

As a result, the present study simulated data (40-item test with 2,000 examinees) under a number of realistic conditions (with 500 replicates per condition) which included both no-DIF and DIF items, balanced and imbalanced group sizes (reference:focal), differing levels of discriminations, and with and without correlations between the two dimensions. These data were then subjected the data to five estimation model approaches to determine the conditions under which UIRT and MIRT models might perform well, as follows.

- Approach 1 (UIRT*d*): UIRT, no items removed from analysis;
- Approach 2 (UIRT*nds*): UIRT, after removing DIF-detected items (DIF items detected using *standard* criterion p-value ≤ 0.05);
- Approach 3 (UIRT*ndl*): UIRT, after removing DIF items (DIF items detected using *liberal* criterion p-value of 0.10);
- Approach 4 (MIRT*c*): compensatory MIRT, no items removed from analysis; and
- Approach 5 (MIRT*nc*): noncompensatory MIRT, no items removed from analysis.

All DIF items were simulated to favor the reference group in the present study; in other words, the primary ability estimate of focal group was contaminated. In terms of the accuracy of primary ability estimate (i.e., bias), for the focal group, the ANOVA results showed that the UIRT*d* model yielded the worst bias (the focal group's primary ability was consistently underestimated and reference group's primary ability was consistently over-estimated) compared to all other models. Using UIRT*nds* and UIRT*ndl* models led to the smallest bias, and use of the

two types of MIRT models (MIRT*c* and MIRT*nc*) led to slightly more bias than the two UIRT models with DIF removed, but these differences were not significant (i.e., the only model that differed from the others was the UIRT model that completely ignored DIF). It is worth noting that none of the 2-way interaction effects between model approaches and other main effects were statistically significant. That is, the performance of five estimation model approaches were consistent under various condition levels.

In summary, the simple model UIRT approach works as well as the complex MIRT approaches, but only for researchers willing to remove items with DIF prior to calibration; for those with limited item pools, the MIRT approach works just as well without removing DIF items. Below the practical implications and study limitations, as well as future research directions are discussed in more detail.

Practical implications

First, the results suggested that the simpler UIRT models, after removing DIF-detected items, are best for minimizing bias. It is worth noting that none of the interactions were statistically significant. That is, the performance of five estimation model approaches were consistent under various condition levels. Second, if the secondary dimension is deemed to be irrelevant to the constructs intended to be measured, it is suggested that the DIF items should be considered for revision or removal (Penfield & Camilli, 2007). In practice, however, removing DIF items is actually not an ideal solution, particularly when there is a large volume of DIF items and/or the test item pool is relatively small. Also, under some test construction situations, test developers selected items so that DIF is balanced across focal and reference groups. As such, MIRT models should be used in order to factor out DIF effects, particularly for tests with relatively high amounts of DIF. Third, as with many DIF studies involving sample size, equally

balanced group sizes generally resulted in less bias. However, there is usually little control researchers have over reference and focal group sizes. Researchers should therefore keep both group sizes as large as possible rather than relying on one very large group. Finally, when only primary ability estimate is considered, it is preferable that the primary dimension has a higher discrimination than that of the secondary dimension.

Future Research and Limitations

Data Generation. The data were generated to follow Sympson's (1977) noncompensatory two-parameter MIRT model (i.e., MIRT*nc* was the true model). The numbers and types of conditions were limited, particularly, test length and sample size were held constant based on the principle of avoiding confounding effects. In this study, items were generated for a medium test length (40 items) and a sufficiently large sample size (2000 examinees), both of which are typical in practice. However, differences in test length can have an impact on internal consistency reliability and can influence the precision of ability estimation (Narayanan & Swaminathan, 1994), and sample size has had an impact on DIF detection Type I error and power rates (e.g., Mazor, Clauser, & Hambleton, 1992; Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). Also, all data conditions with DIF items were simulated to favor the reference group over the focal group. Furthermore, this study only considered manipulating levels of discrimination parameters, but not the levels of the difficulty parameter (i.e., difficulty parameters were sampled from the normal distribution once before data generation and were fixed through all the conditions and replications).

IRT Calibrations. Across the five estimation model approaches used to analyze the data, the calibrations were estimated using an Expectation-Maximization (EM; Dempster, Laird, & Rubin, 1977; Bock & Aitkin, 1981) algorithm employing marginal maximum likelihood (MML).

It is important to note that there are also other popular techniques for estimating the parameters including approximating a factor analytic model using least squares and tetrachoric correlations, as well as Bayesian Markov chain Monte Carlo (MCMC) methods. This said, Knol and Berger (1991) stated that "[t]he disadvantage of noncompensatory models is that no efficient algorithms for estimation of the item parameters are available", and to date, this statement partially remains true.

It is also noteworthy that the noncompensatory MIRT model calibration requires heavy computational demands. Recently, Babcock (2011) and Chalmers (2015) argued that the Bayesian MCMC method offers several benefits for fitting noncompensatory MIRT models over frequentist ML approaches. However, the extension and application of Bayesian algorithms requires much more data if the parameters are to be adequately estimated or the models to converge at all. Indeed, in the present study, convergence rates were slowest for MIRT noncompensatory modeling. As such, the convergence properties of the noncompensatory model could be investigated using larger sample sizes in future work.

DIF Detection. Mantel-Haenszel (MH) DIF detection method was selected in the current study because of its familiarity and acceptance in the measurement community (OSPI, 2014; OAA, 2014). The results of the current study indicated that the MH method can maintain reasonably low Type I error rates when there were unequal sample size ratio of reference group to focal group (9:1), less percentage of DIF items (10%), high primary discrimination ($a_1 = 0.8$), and low secondary discrimination ($a_2 = 0.2$). In terms of power rates, however, the results indicated that when there were equal sample size ratio of reference group to focal group (1:1), less percentage of DIF items (10%), low primary discrimination ($a_1 = 0.5$), and modest secondary discrimination ($a_2 = 0.5$) provided relatively high power rates.

There are certainly a few factors that are sometimes included in DIF analysis but which were not manipulated in the current study, thus limiting the generalizability of the results of the current study. For example, the distribution of mean ability differences is considered a significant factor that can influence DIF detection (Jodoin & Gierl, 2001). Mean standardized ability differences between focal and reference groups could be manipulated from 0.0 to 1.0 in increments of a fixed section (Li, Brooks, & Johanson, 2012; Willse, & Goodman, 2008). Additionally, the effect of unequal variances in reference group and focal group ability distributions could be considered (Monahan & Ankenmann, 2005).

Another factor that impacts DIF detection that could be further evaluated in future research includes the item parameter the levels of item parameters. This study only considered manipulating levels of discrimination parameters, but not the levels of the difficulty parameter. However, a constraint for multidimensional items is that the difficulty level is often held constant across dimensions and to establish baseline results for the present investigation, this parameter was not manipulated. Nevertheless, prior research has investigated the effect of different levels of difficulty on DIF detection, especially when the DIF detection relied on the accuracy of item parameter estimation (e.g., Donoghue & Allen, 1993; Rogers & Swaminathan, 1993).

Test length and sample size were held constant based on the principle of avoiding confounding effects, however, sample size has had an impact on DIF detection Type I error and power rates. More specifically, 200 to 250 members in each group was recommended as the minimum sample size for adequate DIF detection power, yet there were only 200 examinees in the focal group under the present study's unequal group ratio balance condition. As such, the MH DIF detection Type I error rate inflation and power rate deflation we observed is consistent with the prior research.

Finally, it is important to note that, for the MH procedure, effect sizes are typically used in concert with the chi-square test to determine item removal (Zieky, 1993; Zwick & Ercikan, 1989). However, DIF effect sizes were not considered in the present study, and as such, future research could consider both for item removal (and then subsequent model estimation). Further, iterative "purification" has been suggested to control inaccuracies in DIF detection in lieu of MH DIF detection (Ackerman, 1992; Candell & Drasgow, 1988; Cohen & Kim, 1993). This method attempts to identify a set of non-DIF items from the instrument under evaluation to be used as the matching criterion in DIF detection. Holland and Thayer (1988) suggested a two-step purification process with the MH procedure, where the total score is used as the matching criterion for initial DIF screening and a new score comprised of only items not displaying DIF is used for final analysis. This process can result in higher rates of correct DIF identification. For the present study, MH DIF detection was selected because it is so widely employed by many large-scale assessments. However, the use of purification in practice is uncertain (Clauser, Mazor & Hambleton, 1993), since detailed descriptions of DIF analyses often are not released (French & Maller, 2007). Future research is certainly warranted on this topic.

Furthermore, it cannot be emphasized enough that many DIF experts believe that more than one DIF methods have to be conducted in practice (Finch, 2005; Taylor & Lee, 2011; Taylor & Lee, 2012). Particularly, the multidimensional approach to DIF, as implemented in Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993), allows for a variety of scenarios that comprise differential dimensionality as the source for DIF. With similar sample-size requirements, SIBTEST performed similarly to MH in Type I error control when there was no impact (i.e., means of ability distributions are the same between groups) but better than MH when there was impact (Roussos & Stout, 1996). This was because a regression correction was

used in SIBTEST to adjust for bias due to possible difference in ability distributions (Jiang & Stout, 1998). Future research should employ additional SIBTEST for detection of DIF or other methods.

Other Limitations and Directions. Overall, few studies have conducted empirical research on MIRT models; of the studies of MIRT, none have examined their use for the purpose of controlling for DIF, and none have empirically compared the compensatory and noncompensatory models. Hence, the present study contributes new information on these methodologies. Nevertheless, the findings are limited in several aspects. In addition to the aforementioned limitations, a major challenge of the application of the MIRT procedure in dealing with DIF items includes the fact that the number of additional dimension(s) causing DIF is actually unknown. Identifying the "correct" number of dimensions is not always simple and easy in real data analysis (Reckase, 2009), and investigating the number of dimensions on performance is important for future work. In the present study, only two-dimensional data were generated. The mean and standard deviation of the secondary ability was fixed at 0 and 1 for reference group, and was fixed at –1 and 1 for focal group, respectively. Future research is needed to examine whether the approaches considered in this study can be applied when the distributions of additional dimensions are different.

The present study did not focus on the secondary dimension ability estimates, correlations between true and observed estimates, or overall model fit criteria. The focus of the current research was truly on whether MIRT model approaches could be used to control for DIF, assuming that the primary dimension is the one that researchers would be interested in estimating. Nevertheless, future research could examine the secondary dimension as well as model fit criteria.

Only dichotomous items were studied here, however, the inclusion of polytomous items, which have more than two scored outcomes, in large-scale tests is a relatively new phenomenon (Penfield, 2014). The extended polytomous IRT models, such as Graded Response Model or Partial Credit Model, are more complex than their dichotomous counterparts, having more parameters and a more sophisticated mathematical form. As assessment practices continue to advance and technology allows for increasing use of innovative item formats, future research should consider the impact of polytomous item scores on DIF statistics and the accuracy of the ability estimated.

Last but not least, the relationship between the items and the dimensionality of the simulated test was established within a confirmatory analysis framework in this study. This assumption, however, may not work in practice since the structure of the relationships between items and primary and secondary dimensions is actually unknown. In such cases, it may be wise to use an exploratory analysis be carried out on items prior to a confirmatory measurement model. It is left for a future study to examine how differently and accurately the different approaches can detect differential functioning and estimate accurate parameters.

References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234.

Abedi, J., Lord, C., & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance*. National Center for Research on Evaluation, Standards, and Student Testing.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67-91.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*(4), 255-78.

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*(3), 37-51.

Angoff, W. H. (1972). *A technique for the investigation of cultural differences*. Paper presented at the annual meeting of the American Psychological Association, Honolulu.

Angoff, W. H. (1993). Perspectives on differential item functioning methodologies. In P. W. Holland, & H. Wainer (Eds.). *Differential Item Functioning* (pp. 67–114). Hillsdale, NJ: Erlbaum.

Babcock, B. (2011). Estimating a noncompensatory IRT model using metropolis within Gibbs sampling. *Applied Psychological Measurement, 35*(4), 317-329.

Batley, R. M., & Boss, M. W. (1993). The effects on parameter estimation of correlated dimensions and a distribution-restricted trait in a multidimensional item response model. *Applied psychological measurement*, *17*(2), 131-141.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, *27*(6), 395-414.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied psychological measurement*, *12*(3), 253-260.

Chalmers, R. (2015). Extended mixed-effects item response models with the MH-RM algorithm. *Journal of Educational Measurement, 52*(2), 200-222.

Chalmers, R., & Flora, D. (2014). Maximum-likelihood estimation of noncompensatory IRT models with the MH-RM algorithm. *Applied Psychological Measurement, 38*(5), 339-358.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31-44.

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, *6*(4), 269-279.

Cohen, A. S., & Kim, S. H. (1993). A comparison of Lord's $\chi 2$ and Raju's area measures in detection of DIF. *Applied Psychological Measurement*, *17*(1), 39-52.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J.: L. Erlbaum Associates.

de Ayala, R. J. (2009). *The theory and practice of item response theory* (Methodology in the social sciences). New York: Guilford Press.

DeMars, C. (2015). Partially compensatory multidimensional item response theory models: Two alternate model forms. *Educational and Psychological Measurement,* 0013164415589595.

DeMars, C. E., & Lau, A. (2011). Differential item functioning detection with latent classes: how accurately can we detect who is responding differentially? *Educational and Psychological Measurement*, *71*(4) 597–616.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational and Behavioral Statistics*, *18*(2), 131-154.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale NJ: Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of educational measurement*, *23*(4), 355-368.

Educational Testing Service. (2015). *2015 grade 3-8 technical report for the Washington comprehensive assessment program*. Retrieved from

http://www.k12.wa.us/assessment/pubdocs/WCAP2014SpringAdministrationTechnicalReport.pdf

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278-295.

Finch, H. (2010). Item parameter estimation for the MIRT model bias and precision of confirmatory factor analysis—based models. *Applied Psychological Measurement*, *34*(1), 10-26.

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning a comparison of four methods. *Educational and Psychological Measurement*, *67*(4), 565-582.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods*, *9*(4), 466-491.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement*, *67*(3), 373-393.

Furlow, C., Raiford-Ross, T., & Gagne, P. (2009). The impact of multidimensionality on the detection of differential bundle functioning using simultaneous item bias test. *Applied Psychological Measurement*, *33*(6), 441-464.

Gosz, J., K. & Walker, C. M. (2002, April). *An empirical comparison of multidimensional item response data using TESTFACT and NOHARM*. Paper presented at the annual meeting of the National Council for Measurement in Education, New Orleans, LA.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, Calif: Sage Publications.

Harwell, M. R. (1997). Analyzing the results of Monte Carlo studies in item response theory. *Educational and Psychological Measurement*, *57*(2), 266-279.

Harwell, M. R., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied psychological measurement*, *20*(2), 101-125.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two-and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied psychological measurement*, *6*(3), 249-260.

Jiang, H., & Stout, W. (1998). Improved Type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics*, *23*(4), 291-322.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329-349.

Kim, J., & Oshima, T. C. (2012). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement*, *73*(3), 458-470.

Kim, S., Moses, T., & Yoo, H. H. (2015). A comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement*, *52*(1), 70-79.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and psychological measurement*, *56*(5), 746-759.

Knol, D. L., & Berger, M. P. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate behavioral research*, *26*(3), 457-477.

Li, H. H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, *61*(4), 647-677.

Li, Y., Brooks, G. P., & Johanson, G. A. (2012). Item discrimination and type I error in the detection of differential item functioning. *Educational and Psychological Measurement*, *72*(5), 847-861.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga, *Basic problems in cross-cultural psychology* (pp. 19-29). Amsterdam: Swets and Zeitlinger.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, N.J: L. Erlbaum Associates.

Magis, D., Beland, S., & Raiche, G. (2015). *Package 'difR"*. Retrieved from https://cran.r-project.org/web/packages/difR/difR.pdf

Mantel. N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.

Masor, K. M., Clauser, B. E., & Hambleton, R. K. (1993). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54*(2), 284-291.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*(2), 127-143.

Miller, T. R. (1991). *Empirical estimation of standard errors of compensatory MIRT model parameters obtained from the NOHARM estimation program*. ACT Research Report Series.

Monahan, P. O., & Ankenmann, R. D. (2005). Effect of unequal variances in proficiency distributions on Type-I error of the Mantel-Haenszel chi-square test for differential item functioning. *Journal of Educational Measurement*, *42*(2), 101-131.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*(4), 315-328.

National Center for Education Statistics. (2009, June 8). *NAEP technical documentation: The Mantel-Haenszel procedure*. Retrieved from http://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_checks_dif_proced_mh.aspx

Office of Assessment and Accountability (2014). *Kentucky performance rating for educational progress 2013–14 technical manual.* Retrieved from http://education.ky.gov/AA/KTS/Documents/2013-2014%20K-PREP%20Technical%20Manual%20v1.pdf

Office of the Superintendent of Public Instruction (2014). *Washington comprehensive assessment program, grades 3 – 8, high school, technical report for spring 2014*. Retrieved from http://www.k12.wa.us/assessment/pubdocs/WCAP2014SpringAdministrationTechnicalReport.pdf

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning.* Sage.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. *Handbook of statistics: vol. 26. Psychometrics* (pp. 125–167). Amsterdam: Elsevier.

Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice*, *33*(1), 36-48.

R Core Team (2015). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197-207.

Reckase, M. D. (2007). Multidimensional item response theory. *Handbook of statistics: vol. 26. Psychometrics* (pp. 607–642). Amsterdam: Elsevier

Reckase, M. (2009). *Multidimensional item response theory*. New York: Springer.

Reckase, M. D., & McKinley, R. L. (1991). The discrimination power of items that measure more than one dimension. *Applied Psychological Measurement, 14*(4), 361–373.

Robitzsch (2015). *Package 'sirt'*. Retrieved from https://cran.r-project.org/web/packages/sirt/sirt.pdf

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105-116.

Roussos, L. A., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*(4), 355–371.

Russell, S. S. (2005). Estimates of Type I error and power for indices of differential bundle and test functioning. *Dissertation Abstracts International*, 66 (5B), 2867. (UMI No. 3175804)

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105-126.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychological Methods*, *11*(4), 402-415.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*(4), 361-370.

Sympson, J. B. (1977). A model for testing multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference* (pp. 82-98). Minneapolis: University of Minnesota,Department of Psychology, Psychometric Methods Program.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, *27*(3), 159-203.

Taylor, C. S., & Lee, Y. (2011). Ethnic DIF in reading tests with mixed item formats. *Educational Assessment, 16*(1), 35-68.

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education*, *25*(3), 246-280.

Thissen, D., Wainer, H., & Wang, X-B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31*(2),

Walker, C. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, *29*(4), 364-376.

Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *28*(5), 295-316.

Willse, J. T., & Goodman, J. T. (2008). Comparison of multiple-indicators, multiple-causes–and item response theory–based analyses of subgroup differences. *Educational and Psychological Measurement*, *68*(4), 587-602.

Zhang, J. (2012). Calibration of response data using MIRT models with simple and mixed structures. *Applied Psychological Measurement*, *36*(5), 375-398.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, *26*(1), 55-66.

Appendix A

R Macro Code

```
rm(list=ls())
setwd("H:/")
install.packages("mvtnorm"); library(mvtnorm)
install.packages("ltm"); library(ltm)
install.packages("difR"); library(difR)
install.packages("mirt"); library(mirt)
install.packages("sirt"); library("sirt")

I<-40; N<-2000; reps<-500
ratio_list = c(.5, .9) #group size balance
corr_list = c(0, .3)  #factor correlation
nDIF_list = c(0, .1, .2, .3) #percentage of DIF items
a1_list = c(.8, .5) #primary discrimination
a2.diff_list = c(.2, .5) #nuisance discrimination

## b parameters
set.seed(85610)
b<-rnorm(I*2, mean = 0, sd = 1)
b<-ifelse( b > -2 & b < 2, b, NA); b<-na.omit(b)
b<-matrix(rep(b[1:(I)],2), nrow=I)

## Data generation function
## Simulate data from a two-dimensional noncompensatory IRT
dgf<-function(I, N, nDIF, a1, a2.diff, corr, ratio, k, seed1, seed2){

  ## set initial seed for theta generation
  set.seed(seed1)
  ## reference group
  N.ref<-N*ratio
  ref.theta<-rmvnorm(N.ref, mean=c(0, 0),
                     sigma=matrix(c(1, corr, corr, 1), 2))
  ## focal group
  N.foc<-N*(1-ratio)
  foc.theta<-rmvnorm(round(N.foc,0), mean=c(0,-1), #Impact,F_F2~N(-1,1)
                     sigma=matrix(c(1, corr, corr, 1), 2))
  theta<-rbind(ref.theta, foc.theta)
  theta1<-theta[, 1]; theta2<-theta[, 2]

  a2<-rep(0, I)
  diff<-I*nDIF
  diffree<-(I-diff)
  if(nDIF!=0) a2[seq(round(I/diff,0), round(I/diff,0)*diff,
      by=round(I/diff,0))]<-a2.diff
  a<-matrix(c(rep(a1, I), a2), nrow=I)

  # simulate data
```

```
  prob<-simdat<-matrix(0, nrow=N, ncol=I)
  for (ii in 1:I){
    prob[,ii]<-(plogis(theta1 - b[ii,1]))^a[ii,1]
    prob[,ii]<-prob[,ii]*(plogis(theta2 - b[ii,2]))^a[ii,2]
  }

  # set seed for data generation
  set.seed(seed2)
  simdat[prob > matrix(runif(N*I),N,I)]<-1

  ## add group information
  group<-c(rep(0, N.ref), rep(1,round(N.foc, 0))) #0=reference, 1=focal
  dataset<-cbind(group, theta, simdat)
  colnames(dataset)<-c("group", "theta1", "theta2",
      rep(paste("i",1:I,sep="")))
  return(dataset)
}

truetheta1.list<-truetheta2.list<-NULL
m1.ITER.list<-m2.ITER.list<-m3.ITER.list<-NULL
m4.ITER.list<-m5.ITER.list<-NULL
m1.modelFit.list<-m2.modelFit.list<-m3.modelFit.list<-NULL
m4.modelFit.list<-m5.modelFit.list<-NULL
m1.EAP.V1.list <-m2.EAP.V1.list<-m3.EAP.V1.list<-NULL
m4.EAP.V1.list<-m5.EAP.V1.list<-NULL

set.seed(85610)
SEEDS<-sample(1e5, 1000, replace = FALSE)
seed1 = SEEDS[(nDIF_list+.5)*(corr_list+1)*100]
seed2 = SEEDS[(nDIF_list+1)*(corr_list+2)*100]

for (ratio in ratio_list) {
  for (corr in corr_list) {
    for (nDIF in nDIF_list) {
      for (a1 in a1_list) {
        for (a2.diff in a2.diff_list) {
          for (k in 1:reps) {

       condition<-cbind(ratio, corr, nDIF, a1, a2.diff, k)
       CON<-c("ratio", "corr", "nDIF", "a1", "a2.diff", "reps")
       dataset<-dgf(I, N, nDIF, a1, a2.diff, corr, ratio, k, seed1, seed2)
       truetheta<-dataset[, c(2:3)]
       model<-0
       true.theta1<-cbind(model, condition, matrix(c(truetheta[,1]),nrow=1))
       true.theta2<-cbind(model, condition, matrix(c(truetheta[,2]),nrow=1))
       colnames(true.theta1)<-c("model", CON, paste("F1",1:N,sep=""))
       colnames(true.theta2)<-c("model", CON, paste("F2",1:N,sep=""))
       truetheta1.list<-rbind(truetheta1.list, true.theta1)
       truetheta2.list<-rbind(truetheta2.list, true.theta2)
```

```
dat<-dataset[,-c(2:3)]
     # define the data matrix
     QMIRT<-matrix(1, nrow=I, ncol=2)

     if (nDIF == .1) {
       QMIRT[ -seq(10,40,length=4), 2]<-0  #nDIF<-.1
     }
     else if (nDIF == .2) {
       QMIRT[ -seq(5,40,length=8), 2]<-0   #nDIF<-.2
     }
     else if (nDIF == .3) {
       QMIRT[ -seq(3,36,length=12), 2]<-0  #nDIF<-.3
     }
     else {
       QMIRT[ ,2]<-0  #nDIF<-0
     }

     # define the correlation between two factors
     if (corr ==0) {
       variance.fixed<-as.matrix( cbind( 1,2,0) )
     }
     else {
       variance.fixed = NULL
     }

     ###### model 1: Unidimensional 2PL model ######
     QUIRT<-matrix(1 , nrow = I , ncol=1 )

     model<-1
     m1.uirt<-smirt(dat[,-1], Qmatrix=QUIRT, est.a="2PL",
                    maxiter=1000, increment.factor=1.01)

     ## ITERATION
     m1.ITER<-m1.uirt$iter
     m1.ITER.fm<-cbind(model, condition, m1.ITER)
     row.names(m1.ITER.fm)<-NULL
     colnames(m1.ITER.fm)<-c("model", CON, "ITER")
     m1.ITER.list<-rbind(m1.ITER.list, m1.ITER.fm)

     ## model FIT
     m1.modelFit<-cbind(m1.uirt$deviance, m1.uirt$ic$AIC,
     m1.uirt$ic$BIC, m1.uirt$ic$AICc, m1.uirt$ic$CAIC)
     m1.modelFit.fm<-cbind(model, condition, m1.modelFit)
     colnames(m1.modelFit.fm)<-c("model", CON, "deviance", "AIC",
     "BIC", "AICc", "CAIC")
     m1.modelFit.list<-rbind(m1.modelFit.list, m1.modelFit.fm)

     ## PERSON
     m1.EAP.V1<-cbind(model, condition, matrix(m1.uirt$person$EAP.V1,
     nrow=1))
```

```
colnames(m1.EAP.V1)<-c("model",CON,paste("EAP.V1_",1:N,sep=""))
m1.EAP.V1.list<-rbind(m1.EAP.V1.list, m1.EAP.V1)

###### model 2: UIRT and Mantel-Haenszel, p.value<.05 ######
MH<-difMH(dat, group="group", focal.name=1, alpha = 0.05)

if (is.numeric(MH$DIFitems)==TRUE) {
  MH.item<-rep(0,I); MH.item[MH$DIFitems]<-1
  MH.item<-matrix(c(MH.item)[1:I],nrow=1)

  model<-2

  MH.item.fm<-cbind(model, condition, MH.item)
  colnames(MH.item.fm)<-c("model", CON, paste("I",1:I,sep=""))
  m2.MH.item.list<-rbind(m2.MH.item.list, MH.item.fm)

  MHdat<-dat[,-c(MH$DIFitems + 1)]

  QUIRT.DIF<-matrix(1 , nrow = I-length(MH$DIFitems) , ncol=1 )

  m2.uirt<-smirt(MHdat[,-1], Qmatrix=QUIRT.DIF, est.a="2PL",
                  maxiter=1000, increment.factor=1.01)
  ## ITERATION
  m2.ITER<-m2.uirt$iter
  m2.ITER.fm<-cbind(model, condition, m2.ITER)
  row.names(m2.ITER.fm)<-NULL
  colnames(m2.ITER.fm)<-c("model", CON, "ITER")
  m2.ITER.list<-rbind(m2.ITER.list, m2.ITER.fm)

  ## model FIT
  m2.modelFit<-cbind(m2.uirt$deviance, m2.uirt$ic$AIC,
  m2.uirt$ic$BIC, m2.uirt$ic$AICc, m2.uirt$ic$CAIC)
  m2.modelFit.fm<-cbind(model, condition, m2.modelFit)
  colnames(m2.modelFit.fm)<-c("model", CON, "deviance", "AIC",
  "BIC", "AICc", "CAIC")
  m2.modelFit.list<-rbind(m2.modelFit.list, m2.modelFit.fm)

  ## PERSON
  m2.EAP.V1<-cbind(model,condition,matrix(m2.uirt$person$EAP.V1,
  nrow=1))
  colnames(m2.EAP.V1)<-c("model",CON,paste("EAP.V1_",1:N,sep=""))
  m2.EAP.V1.list<-rbind(m2.EAP.V1.list, m2.EAP.V1)

} else {

  model<-2

  MH.item<-rep(0,I); MH.item<-matrix(c(MH.item)[1:I],nrow=1)
  MH.item.fm<-cbind(model, condition, MH.item)
  colnames(MH.item.fm)<-c("model", CON, paste("I",1:I,sep=""))
```

```
        m2.MH.item.list<-rbind(m2.MH.item.list, MH.item.fm)

        m2.ITER.list<-rbind(m2.ITER.list, m1.ITER.fm)
        m2.modelFit.list<-rbind(m2.modelFit.list, m1.modelFit.fm)
        m2.EAP.V1.list<-rbind(m2.EAP.V1.list, m1.EAP.V1)
      }

      ###### model 3: UIRT and Mantel-Haenszel, p.value<.10 ######

      MH<-difMH(dat, group="group", focal.name=1, alpha = 0.10)

      if (is.numeric(MH$DIFitems)==TRUE) {
        MH.item<-rep(0,I); MH.item[MH$DIFitems]<-1
        MH.item<-matrix(c(MH.item)[1:I],nrow=1)

        model<-3

        MH.item.fm<-cbind(model, condition, MH.item)
        colnames(MH.item.fm)<-c("model", CON, paste("I",1:I,sep=""))
        m3.MH.item.list<-rbind(m3.MH.item.list, MH.item.fm)

        MHdat<-dat[,-c(MH$DIFitems + 1)]

        QUIRT.DIF<-matrix(1 , nrow = I-length(MH$DIFitems) , ncol=1 )

        m3.uirt<-smirt(MHdat[,-1], Qmatrix=QUIRT.DIF, est.a="2PL",
                       maxiter=1000, increment.factor=1.01)

        ## ITERATION
        m3.ITER<-m3.uirt$iter
        m3.ITER.fm<-cbind(model, condition, m3.ITER)
        row.names(m3.ITER.fm)<-NULL
        colnames(m3.ITER.fm)<-c("model", CON, "ITER")
        m3.ITER.list<-rbind(m3.ITER.list, m3.ITER.fm)

        ## model FIT
        m3.modelFit<-cbind(m3.uirt$deviance, m3.uirt$ic$AIC,
        m3.uirt$ic$BIC, m3.uirt$ic$AICc, m3.uirt$ic$CAIC)
        m3.modelFit.fm<-cbind(model, condition, m3.modelFit)
        colnames(m3.modelFit.fm)<-c("model", CON, "deviance", "AIC",
        "BIC", "AICc", "CAIC")
        m3.modelFit.list<-rbind(m3.modelFit.list, m3.modelFit.fm)

        ## PERSON
        m3.EAP.V1<-cbind(model, condition,
matrix(m3.uirt$person$EAP.V1,
        nrow=1))
        colnames(m3.EAP.V1)<-c("model",CON,
paste("EAP.V1_",1:N,sep=""))
        m3.EAP.V1.list<-rbind(m3.EAP.V1.list, m3.EAP.V1)
```

```
      } else {

        model<-3

        MH.item<-rep(0,I); MH.item<-matrix(c(MH.item)[1:I],nrow=1)
        MH.item.fm<-cbind(model, condition, MH.item)
        colnames(MH.item.fm)<-c("model", CON, paste("I",1:I,sep=""))
        m3.MH.item.list<-rbind(m3.MH.item.list, MH.item.fm)

        m3.ITER.list<-rbind(m3.ITER.list, m1.ITER.fm)
        m3.modelFit.list<-rbind(m3.modelFit.list, m1.modelFit.fm)
        m3.EAP.V1.list<-rbind(m3.EAP.V1.list, m1.EAP.V1)
      }

      ###### model 4: Compensatory 2PL model ######
      model<-4
      m4.smirt<-smirt(dat[,-1], Qmatrix=QMIRT, irtmodel="comp",
est.a="2PL", variance.fixed=variance.fixed, maxiter=1000,
increment.factor=1.01)

      ## ITERATION
      m4.ITER<-m4.smirt$iter
      m4.ITER.fm<-cbind(model, condition, m4.ITER)
      row.names(m4.ITER.fm)<-NULL
      colnames(m4.ITER.fm)<-c("model", CON, "ITER")
      m4.ITER.list<-rbind(m4.ITER.list, m4.ITER.fm)

      ## model FIT
      m4.modelFit<-cbind(m4.smirt$deviance, m4.smirt$ic$AIC,
      m4.smirt$ic$BIC, m4.smirt$ic$AICc, m4.smirt$ic$CAIC)
      m4.modelFit.fm<-cbind(model, condition, m4.modelFit)
      colnames(m4.modelFit.fm)<-c("model", CON, "deviance", "AIC",
      "BIC", "AICc", "CAIC")
      m4.modelFit.list<-rbind(m4.modelFit.list, m4.modelFit.fm)

      ## PERSON
      m4.EAP.V1<-cbind(model, condition, matrix(m4.smirt$person$EAP.V1,
      nrow=1))
      colnames(m4.EAP.V1)<-c("model", CON, paste("EAP.V1_",1:N,sep=""))
      m4.EAP.V1.list<-rbind(m4.EAP.V1.list, m4.EAP.V1)

      ## Estimated correlation (COVARIANCE)
      m4.est.CORR<-(m4.smirt$cor.trait)[1,2]
      m4.est.CORR.fm<-cbind(model, condition, m4.est.CORR)
      colnames(m4.est.CORR.fm)<-c("model", CON, "CORR")
      m4.est.CORR.list<-rbind(m4.est.CORR.list, m4.est.CORR.fm)

      ###### model 5: Noncompensatory 2PL model ######
      ###### avoid convergence problems with increment.factor
```

```
            model<-5
            m5.smirt<-smirt(dat[,-1], Qmatrix=QMIRT , irtmodel="noncomp",
        est.a="2PL",variance.fixed=variance.fixed, maxiter=1000,
        increment.factor=1.01)

            ## ITERATION
            m5.ITER<-m5.smirt$iter
            m5.ITER.fm<-cbind(model, condition, m5.ITER)
            row.names(m5.ITER.fm)<-NULL
            colnames(m5.ITER.fm)<-c("model", CON, "ITER")
            m5.ITER.list<-rbind(m5.ITER.list, m5.ITER.fm)

            ## model FIT
            m5.modelFit<-cbind(m5.smirt$deviance, m5.smirt$ic$AIC,
            m5.smirt$ic$BIC, m5.smirt$ic$AICc, m5.smirt$ic$CAIC)
            m5.modelFit.fm<-cbind(model, condition, m5.modelFit)
            colnames(m5.modelFit.fm)<-c("model", CON, "deviance", "AIC",
            "BIC", "AICc", "CAIC")
            m5.modelFit.list<-rbind(m5.modelFit.list, m5.modelFit.fm)

            ## PERSON
            m5.EAP.V1<-cbind(model, condition, matrix(m5.smirt$person$EAP.V1,
            nrow=1))
            colnames(m5.EAP.V1)<-c("model", CON, paste("EAP.V1_",1:N,sep=""))
            m5.EAP.V1.list<-rbind(m5.EAP.V1.list, m5.EAP.V1)

            ## Estimated correlation (COVARIANCE)
            m5.est.CORR<-(m5.smirt$cor.trait)[1,2]
            m5.est.CORR.fm<-cbind(model, condition, m5.est.CORR)
            colnames(m5.est.CORR.fm)<-c("model", CON, "CORR")
            m5.est.CORR.list<-rbind(m5.est.CORR.list, m5.est.CORR.fm)

          }
          seed2 = seed2 + 1
        }
        seed1 = seed1 + 1
      }
    }
  }
}
}
```