Analytic Selection of a Valid Subtest for DIF Analysis when DIF has Multiple Potential Causes among Multiple Groups

by

Lietta Scott

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved July 2014 by the
Graduate Supervisory Committee:

Roy Levy, Co-Chair
Samuel Green, Co-Chair
Joanna Gorin
Leila Williams

ARIZONA STATE UNIVERSITY

August 2014

ABSTRACT

The study examined how ATFIND, Mantel-Haenszel, SIBTEST, and Crossing SIBTEST function when items in the dataset are modelled to differentially advantage a lower ability focal group over a higher ability reference group. The primary purpose of the study was to examine ATFIND's usefulness as a valid subtest selection tool, but it also explored the influence of DIF items, item difficulty, and presence of multiple examinee populations with different ability distributions on both its selection of the assessment test (AT) and partitioning test (PT) lists and on all three differential item functioning (DIF) analysis procedures. The results of SIBTEST were also combined with those of Crossing SIBTEST, as might be done in practice.

ATFIND was found to be a less-than-effective matching subtest selection tool with DIF items that are modelled unidimensionally. If an item was modelled with uniform DIF or if it had a referent difficulty parameter in the Medium range, it was found to be selected slightly more often for the AT List than the PT List. These trends were seen to increase as sample size increased. All three DIF analyses, and the combined SIBTEST and Crossing SIBTEST, generally were found to perform less well as DIF contaminated the matching subtest, as well as when DIF was modelled less severely or when the focal group ability was skewed. While the combined SIBTEST and Crossing SIBTEST was found to have the highest power among the DIF analyses, it also was found to have Type I error rates that were sometimes extremely high.

DEDICATION

This work is dedicated to my children, Thomas, Sarina, and Timothy, who have inspired me to reach farther than I thought I could ever go.

ACKNOWLEDGMENTS

First, I would like to acknowledge all of the help and encouragement the professors in the Measurement, Statistics, and Methodology Studies program, Roy Levy, Samuel Green, Marilyn Thompson, and Joanna Gorin, have given me throughout the past years. I especially want to thank Dr. Green for helping me get into the program, without him none of this would have been possible. I also want to especially thank Dr. Levy for his continued support and guidance through the whole dissertation process.

There are also others that need acknowledgement. Dr. Leila Williams at the Arizona Department of Education has been on my side for years, encouraging and consoling as I went through the ups and downs that doctoral programs entail. Thank you for your shoulder and ready smile. I also received support and encouragement from the other students in the program, especially from AAron Crawford, Derek Fay, and Dubravka Svetina. Your friendship and companionship helped ease the transition from teacher to scientist. As for technical support, I could not have completed this work without the software expertise of my brilliant son, Thomas Scott, both he and my brother in Scouting, Richard Cannon, helped me understand the software needed for the simulation studies herein. I also want to thank my extended family, those at work and those in my "tea group" - you know who you are! Without your love and support I could not have made it. Thank you! I will be forever grateful. And finally, I need to thank Bill, my husband of 42 years, my education has been a long and twisted road, but I *believe* we have finally reached the end.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Since the reauthorization of the Elementary and Secondary Education Act with the No Child Left Behind Act of 2001 (NCLB), state content assessments in mathematics, reading, writing, and science have increasingly become increasingly important. They are the basis for not only school and district evaluations but also gateways for student promotion and graduation as well as teacher evaluations, influencing decisions about pay and even possible job loss. Each year, every public school in our nation administers these high stakes tests to students enrolled in Grades 3 through 8 and one state designated grade in high school. All students, regardless of ability (or disability) must be assessed. Schools and districts are penalized if they do not assess at least 95% of each student subgroup, which is calculated by grade level and includes students with disabilities (SWD), English language learners (ELLs), students from low income families, and the major ethnicities within the state (United States Department of Education, 2000).

Because of the inclusion of students with disabilities and those just learning English, states have increasingly allowed students to use accommodations on the assessments that are specifically aligned with their needs so that they can better demonstrate their knowledge and skills (Thurlow, & Bolt, 2001; Johnstone, Altman, Thurlow, & Thompson, 2006, Bolt, & Thurlow, 2007). Of the students with disabilities in our nation's schools, over half have been found to be designated with learning disabilities (Torgesen, 2004). Learning disabilities, which affect students' ability to processes information, directly impact students' ability to learn and demonstrate knowledge in specific content areas such as reading, writing, and mathematics (Silver, 2004). Accommodations, such as having the writing prompt read out

loud or using a white board to work a mathematics question, are assumed to have negligible effect on the construct being assessed so that the resultant student scores are both valid and can be aggregated for teacher, school, district, and state evaluations (Arizona Department of Education (ADE), 2011c). The investigation and assurance of the validity of these accommodations for various groups of examinees falls on the assessment community for the states. To this end, multiple studies have been performed to investigate the lack of measurement invariance, including differential item functioning (DIF) between students who took an assessment with an accommodation (or group of accommodations) and those who did not (e.g. Cohen, Gregg, & Deng, 2005; Elliot, & Marquart, 2004; Middleton & Cahalan Laitusis, 2007). The work, however, is not complete.

With multiple groups defined by both various needs and the use of various accommodations, researchers are struggling with the task of teasing apart the data to determine whether items function similarly or differently in the many groups defined by the combinations of those needs and accommodations. For example, one state has 19 different need classifications which allow access to 22 specific accommodations. These students' needs include various education classifications such as Emotionally Disabled, Specific Learning Disabled, and Mild Mental Retardation, as well as physical disabilities such as students who have Visual Impairment, Hearing Impairment, or Traumatic Brain Injury. Additionally, ELLs are allowed to access a separate list of accommodations which contains some of the same accommodations that students with learning or physical disabilities may access. If a researcher restricts the students included in the DIF analyses to those who used only one accommodation, 418 analyses would be needed to compare the item functions for

each group/accommodation combination to the referent group of general education students who took the assessment with no accommodations.

Complicating these analyses is that the DIF process matches the two comparison groups on some measure of ability prior to determining whether the items function differently for these two groups. The groups are often identified as the focal, generally consisting of examinees in the subgroup of interest (in this case, the students with an identified need who took the test with an accommodation) and referent, the examinees in the majority group, generally (Millsap, 2011) but not necessarily, assumed to be advantaged by DIF. When students are matched on their total score for the assessment, the score for the focal group would include any items that had been impacted by their accommodation and cause the matching score to be increasingly unreliable as the number of DIF items that are contained in the assessment increases. In an ideal world, we would prefer to match examinees on a purified subtest that contains no DIF items to optimize the reliability of the analysis (Linn, 1993). Additionally, because of the large number of group/ accommodation combinations, it would be preferable to be able to have the same matching set of items for all DIF analyses to aid in comparability of the results across combinations.

The goal of this study is to investigate a method of analytically selecting a purified matching subtest for use when multiple focal groups have multiple causes for potential DIF as compared to a single referent group.

CHAPTER 2

BACKGROUND LITERATURE

Measurement Non-invariance

Definition of Differential Item Functioning. DIF is defined to exist when individuals having the same ability level but belonging to different groups are observed to have different probabilities of correctly answering an item (Hambleton, Swaminathan, & Rogers, 1991). That is, systematic differences in performance, between different groups of examinees, is observed after they are matched on the ability that the test was intended to measure (French & Finch, 2010). Using Hanson's (1998) notation, this definition can be written with the item response variable for the one item (V) under investigation (so no item subscript is used) with possible discrete values $v_1, v_2, \ldots v_i$. The group variable in this notation is G with possible discrete values $g_1, g_2, \ldots g_j$, and the matching variable is W which can be either discrete or continuous. It can be formulated as the conditional dependence of V on both G and W as below

$$P(V = v_i \mid W = w, G = g_j) \neq P(V = v_i \mid W = w) \quad (1)$$

for $1 \leq i \leq I$, $1 \leq j \leq J$, and for all w.

Dichotomous data, generally examined for one or both of two types of DIF. *Uniform* DIF tends to advantage one group over the other across the whole ability range. Non-uniform DIF however, occurs when there is an interaction between membership within a group and ability level (Narayanan & Swaminathan, 1996). Within item response theory

(IRT), dichotomous item responses are often modeled using some variant of the general curvilinear three parameter logistic (3-PL) model shown.

$$P(V_{is} = 1|\theta_s, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{e^{a_i(\theta_s - b_i)}}{1+e^{a_i(\theta_s - b_i)}} \quad (2)$$

In this model, a_i is the parameter for the item V_i's discrimination, b_i is the parameter for the item's difficulty, and c_i is the item's pseudo-guessing parameter where θ_s is the parameter for examinee's ability. Here, the probability of an examinee's correct response to the item is dependent solely on the three item parameters and the one person parameter and, disregarding error, should be the same regardless of how the examinees are grouped (Embretson & Reise, 2000). Within this framework, which will be utilized in the current study, uniform DIF is observed when two or more groups vary on item difficulty parameter after matching on ability. However, if the groups vary, either instead or additionally, on the item discrimination parameter, then non-uniform DIF is evident (de Ayala, 2009). A graphical representation of uniform and non-uniform DIF can be obtained using separate item characteristic functions for each of the two groups. Figure 1 displays an example of uniform DIF. In this figure, the two item characteristic functions (ICF) values differ only in difficulty parameter. The difficulty parameter in this example has a value of -0.5 for Group 2 (the reference group) and a value of 0.5 for Group 1 (the focal group), indicating that this item is harder (or takes more of the trait to answer correctly) for Group 1 than Group 2. Since this is the only parameter that varies between the groups for this item, Group 2 has a

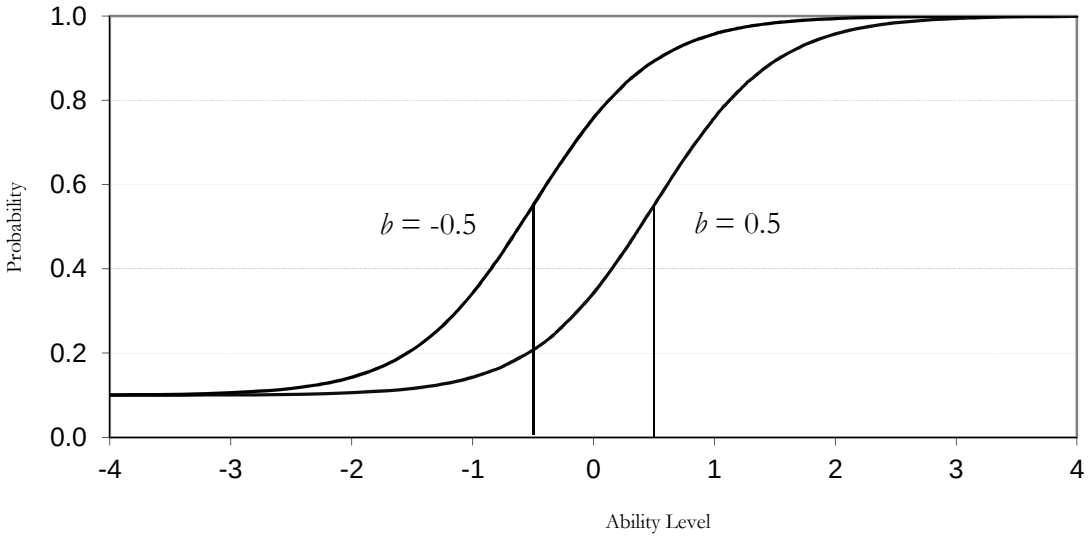


Figure 1. Probability curves for two groups on an item that displays uniform DIF.

higher probability of correctly answering the item than Group 1 across every level of ability ( ). The item characteristic curves (ICC) displayed will not meet except that they asymptote to the same values (in this example at .10 and 1).

Figure 2 is an example of non-uniform DIF where not only is there a difference between groups in probability in correctly answering the item but the group having the advantage changes at some point within the ability range. Here, one group has a higher probability of correctly answering the question at end of the scale and the other group has a higher probability at the other (Walker, 2001). In this example, the ICF values differ in terms of discrimination and pseudo-guessing parameters, as well as the difficulty parameter. The ICCs in this example show that, at the lower levels of ability, examinees in Group 1 have a higher probability of correctly answering the item than those in Group 2, where this trend is reversed at the higher end of the ability scale. Some researchers have termed non-uniform

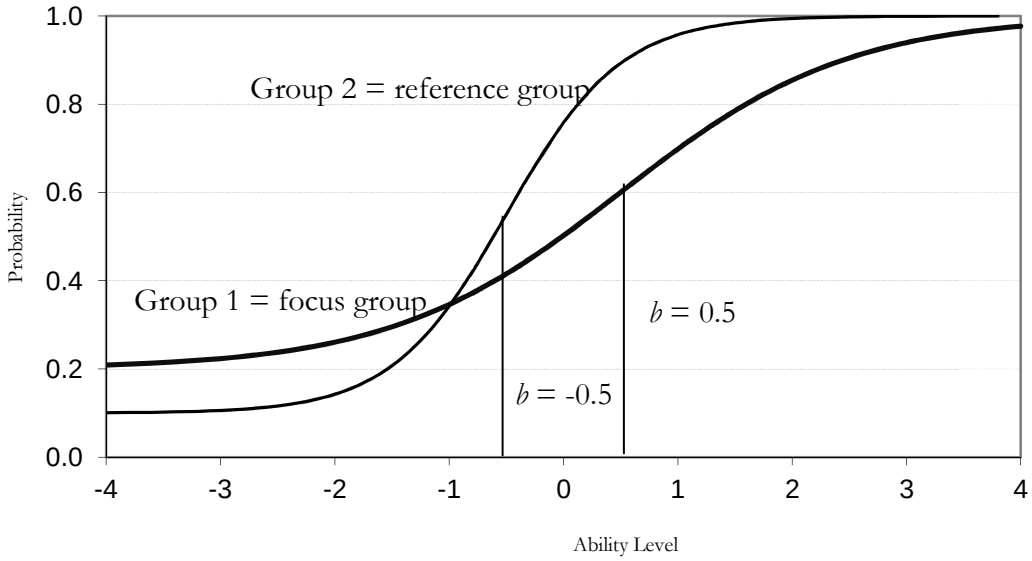


Figure 2. Probability curves for two groups on an item that displays non-uniform DIF.

DIF as "crossing" due to the visual juncture of the ICFs sometimes observed (Bolt & Stout, 1996, p. 88). It should be noted that it is entirely possible for an item to have invariant difficulty values for two groups and only have discrimination and/or pseudo-guessing values that vary. In this case, if the analysis is insensitive to non-uniform DIF, DIF might not be identified (Camilli, 2006). It is also possible that the ICCs do not intersect, except at the asymptotes, but the magnitude of DIF varies across the ability scale. With this "non-crossing-nonuniform" (Güler & Penfield, 2009) or "unidirectional" DIF, one group remains advantaged over the other across the scale but the amount of advantage is not consistent (Li & Stout, 1996).While DIF within the IRT framework, as explained above and used within this study, can be conceived as a lack of parameter invariance between groups matched on some estimate of ability, researchers have conceptualized it from other perspectives also.

Parameter and Test Level Non-invariance. Within parametric approaches to DIF, parameter non-invariance has been defined as equivalent population parameters across different populations (Teresi, 2006b). These population parameters can be either factor loadings and variable intercepts or item discrimination, difficulty and pseudo-guessing parameters (Teresi, 2006b). There are two general frameworks within which parametric approaches to parameter non-invariance (or DIF) studies are based: item response theory (IRT) and structural equation modeling (SEM). Following Byrne (2006), this statement subsumes confirmatory factor analysis (CFA) as a special case of SEM where the relationships between latent factors are not causally modeled. IRT and SEM are related in that some models used in IRT (normal ogive and logistic ogive) have been shown to be linked to one form of an SEM model (multiple indicator-multiple cause, MIMIC) that is used to investigate parameter non-invariance. Additionally, Takane and de Leeuw (1987) demonstrated that the models used for factor analysis and IRT are equivalent (Teresi, 2006b).

Multidimensional Framework for DIF. Assessments can be designed to (or incidentally) address one or more skills or traits. When the ability to correctly respond to all of the items on an assessment depends solely on one skill, or one set of composite skills (Reckase, Ackerman, & Carlson, 1998), the test is unidimensional. When the modeling of assessment performance is more complex (e.g. accounting for either multiple abilities concurrently or various mixtures of abilities assessed within the various items), then a multidimensional model might be more appropriate (Hartig, & H□hler, 2009). The differences in the observed probabilities, as illustrated above, could be conceived as due to some ability, outside those for which the test was designed to measure, that varies between

groups. If the probability of the correct score was not based solely on the ability level of the primary trait (θ) but also that of the secondary trait (η, also know as an auxiliary or nuisance trait), then DIF could occur if the probability density functions of η conditioned on the primary trait were not equivalent for the two groups (Roussos & Stout, 1996a). Using a multidimensional framework, there can be multiple primary traits ($\Theta = \theta_1, \theta_2, \theta_3, \ldots \theta_n$) that the assessment is intended to measure and multiple unintended (or nuisance) traits ($\mathrm{H} = \eta_1, \eta_2, \eta_3 \ldots \eta_k$) that it also measures. Each of these traits would have one or more items that function as an indicator. Figure 3 presents an example, using a factor structure of an assessment, which measures two intended traits and one nuisance trait.

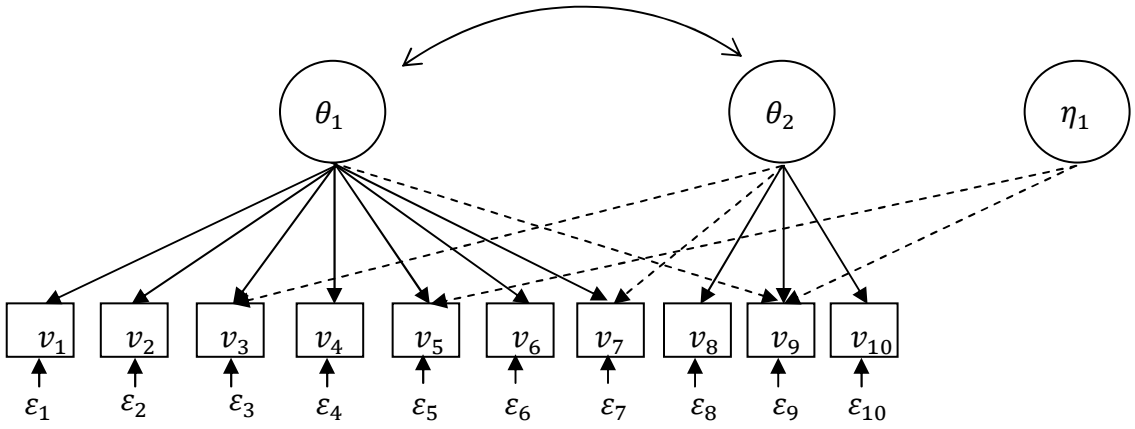


Figure 3. Example of a multi-trait factor structure.

In this example, significant loadings and correlations are indicated by dashed and solid arrows where non-existent and non-significant loadings are excluded for clarity. Here, as represented by the solid arrows, items v_1 through v_7 are intended to measure latent trait θ_1, items represented by the solid arrows, items v_1 through v_7 are intended to measure latent trait θ_1, items v_8 through v_{10} are intended to measure latent trait θ_2, ε_1 through ε_{10}

represent the impact of random and unique measurement error on the items, and the curved double arrow line indicates that θ_1 is correlated with θ_2. Some of these items, however, also load on other traits as displayed by the connecting dashed arrows. As indicated item v_9 loads not only on θ_2 as intended, but also on θ_1 while v_3 and v_7 load not only on θ_1, but also on θ_2. The third trait, the nuisance trait η_1, has two items that load on it, v_5 and v_9. In this example, the remaining items' non-significantly loading on η_1 and the non-significant correlations between the nuisance trait and the intended traits are not displayed for simplification of illustration. There is no supposition that this must always be the case, rather researchers might expect that there would always be at least some correlation between traits.

A hypothetical example of this type of structure might appear in a high school algebra quiz, where θ_1 would be general algebra concepts including some calculation, θ_2 would be logic problems using algebra concepts, and η_1 would be computational skill. In this scenario, v_5 and v_9, since they additionally load on η_1, might require relatively high levels of computation to answer correctly while the rest of the items require low levels or no computation to correctly answer. Since the intent for this quiz is to assess students on algebra and logic concepts and not on their computational skill (arithmetic), this additional trait might be considered a nuisance trait. If it was determined that there were two or more groups of students who varied systematically on this trait, DIF might be observed.

Using this scenario and applying Ackerman's (1992, 1994) multidimensional perspective, θ_1 and θ_2 could be considered a plane of valid traits with each on an orthogonally-placed axis. When only the intended traits are used, all items are represented by

vectors that lie either close to one of the two axes or, in the case of v_3, v_7, and v_9 (since they load on both intended traits), lie between them. One possible example of this plane is presented in Figure 4. Please note that the following figures only include the first quadrant and were conceptualized as scaled from zero at the origin to positive 4. The scale values, however, have been left off to simplify the artwork. In a more realistic scenario, all two-dimensional (2D) drawings would be on a regular coordinate plane ranging from - ∞ to + ∞ and all three-dimensional (3D) drawings would be placed on a three dimensional coordinate grid with each dimension ranging from - ∞ to + ∞.

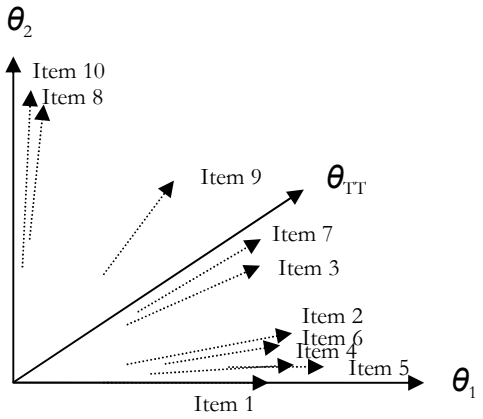


Figure 4. Example of an algebra quiz, item vectors in relation to the intended traits.

In Figure 4, each item's location between the two axes represents a composite of the amount of each trait it is measuring, while the distance away from the origin represents the difficulty of the item, and the length of the vector represents its level of multidimensional discrimination. The unidimensional vector θ_{TT} in this example, following Froelich and Habing (2008), is depicted as an approximation of the weighted average of the item vectors and can be thought of as the ability composite that is intended to be measured by the whole quiz. The direction of the items in this intended plane would be changed with the addition of

a third trait η_1, orthogonal to the intended plane. Figure 5 presents one possible example of this change.

Here, Items 5 and 9 have been pulled off the intended plane (depicted as a shaded rectangle) by the additional trait they measure, η_1. While it is hard to display, both of these items retain their relationship between the intended traits. Their maintained relationship is illustrated by the inclusion of both items' reflections (Item 5' and Item 9') onto the intended plane. While the remaining items' possible non-significant loadings on η_1 are not included in this figure for clarity.

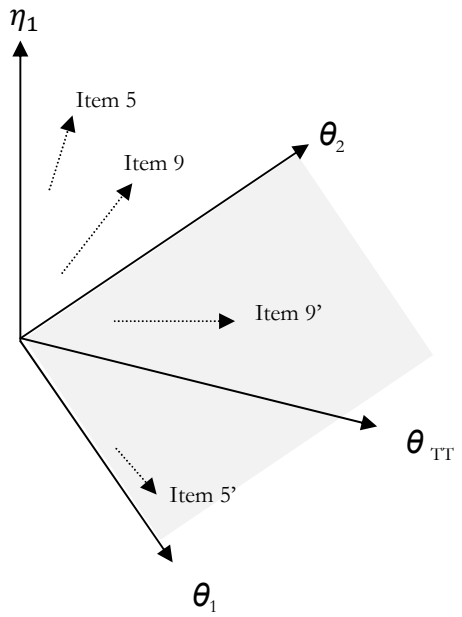


Figure 5. An example of items relation to the 2 intended traits and 1 unintended trait.

If instead of this 3D representation, the θ_{TT} vector was used to represent the intended construct, the items could be viewed once again in a 2D graph as in Figure 6. In this depiction most of the items mainly measure the intended construct and non-significantly load on the nuisance trait. However, two items (Items 5 and 9) more heavily and

significantly, measure the unintended construct in the example scenario, computation or arithmetic. Ackerman (1992) proposed that the items most closely measuring the intended construct of an assessment could be distinguished from other items that might be measuring other constructs, those suspected of containing DIF, through the creation of a "Validity Sector." This sector (also illustrated in Figure 6), and its reflection through the origin, contains the items that would constitute a valid subtest. Ackerman (1992) recommends that practitioners establish this sector and its width as they are developing each assessment. To this end, he presents an index of item validity that is a function only of the direction *a* that the item measures and the angle of the reference composite for the items chosen as most representative of measuring the intended trait. Since this index is population-dependent, he cautions that groups should be combined for its determination. He additionally recommends that for DIF analyses the estimate of ability (in whatever form) used as the matching criterion be based on items within this sector.

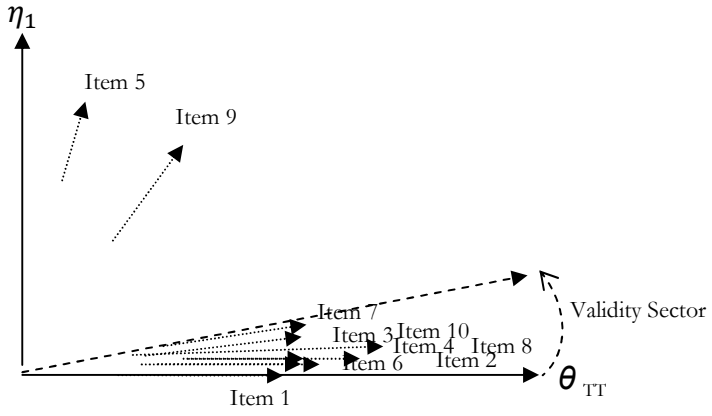


Figure 6. A 2D example of quiz items relation to intended and unintended trait vectors.

The mere existence of an unintended dimension, however, does not mean that DIF will be observed. For DIF to be present within this framework, not only must there be an additional dimension, but there also needs to be differences between at least two groups' conditional ability distributions on that additional dimension. Ackerman (1992) provided a case-wise analysis of the effects differences in mean and variances of a reference and a focal group would have on the detection of DIF. His conclusions are summarized in Table 1.

Table 1

The effect of distribution of group ability on detection of DIF.

Case#/ Effect	μ_{F_θ} μ_{R_θ}	μ_{F_η} μ_{R_η}	σ_{F_θ} σ_{R_θ}	σ_{F_η} σ_{R_η}
1. No DIF/ no impact	=	=	=	=
2. Impact[1]	≠	=	=	=
3. Uniform DIF	=	≠	=	=
4. Nonuniform DIF	=	=	=	≠
5. If $\rho_{F_{\theta,\eta}} \neq \rho_{R_{\theta,\eta}}$ then nonuniform DIF	=	=	=	=

DIF versus Impact. DIF, as described above, exists when there are differences in the probability of correctly answering a question between groups, after matching on (or conditioning on) ability. This term was introduced by Holland and Thayer (1988) as a more "neutral" term for what was originally called "item bias." However, some authors still prefer "item bias" (Ackerman, 1992) or even the more general term of "measurement bias" (Millsap & Everson, 1993), which may be used when discussing not only item level, but also

testlet and factor level bias. For these authors, these terms deal with the statistical evidence of differences between group performance after matching on some estimation of ability or abilities without judgment with regard to whether cause for the difference is a threat to construct validity.

This is similar to Camilli's (2006) explanation of DIF as synonymous to "statistical bias," which he defines as a systematic (as opposed to random) over- or under-estimation of item or person parameters or when two parameters that should be the same are systematically different. He, however, also uses the term "unfairness" to describe DIF or statistical bias that is due to measurement differences in traits that are irrelevant to the constructs the test is intended to measure. The term "unfairness" used by Camilli (2006) in this context, seems to be a replacement for the term "item bias," which Camilli and Shepard (1994) used to describe items that displayed DIF and also were found to be caused by some source irrelevant to the construct being measured by the test.

Impact has been defined as the existence of differences in the average performance on a test or test item between groups, without regard to matching examinees on ability (Camilli, 2006). This can be a measure of true differences in ability between groups as evidenced by the results of their test scores (Camilli, 2006). He gives the example of impact and statistical bias using running a race where two groups' (A and B) times are determined by two different stop watches. An example of impact would be if group A's time is, on average, better than group B's and if both watches are working accurately, then group A might be said to have higher ability for running, on average, than group B. However, if the reason for group A having better times than group B was that the watch used to measure group A's time ran slow, recording faster times for this group, then the results would contain

bias (Camilli, 2006). He notes that in this scenario, since all of the runners in each group were measured by a single watch, comparisons within groups may be accurate, but due to the inaccuracy of group A's watch, comparisons between the two groups would be confounded.

Ackerman (1992) describes impact from a validity standpoint by defining it in terms of valid skills or constructs. In his definition, impact occurs when there are between-group differences in test performance that are caused by between-group differences in a valid skill. His multi-trait perspective has an emphasis on valid and invalid constructs that might be considered similar to Camilli's (2006) fair and unfair perspective when conceptualized within the construct validity framework presented by Camilli and Shepard (1994). All of these researchers seem to agree that impact is an unbiased difference between two groups where differences that are unfair or invalid are biased.

As outlined above, it seems that authors have come close to consensus regarding the usage of impact as defined above. Given the historical usage of the term "item bias," however, there still seems to be some differing opinions about its usage. The original "item bias" was 1) sometimes (e.g. Ackerman, 1992) intended to be a statistical bias based only on difference in performance after matching on ability (this is the definition that is associated with DIF) and 2) sometimes (e.g. Camilli & Shepard, 1994) intended to be based both on a statistical bias in group performance conditioned on ability and on an evaluation of the item over and above differences in group performance (Angoff, 1993). The term "item bias" continues to be variously used. In a recent book, Millsap (2011) used item bias in preference to DIF to emphasize the importance of multivariate target-latent variables that the test is intended to measure. This usage would have an evaluative component and would, therefore, place the usage with those whose intent is based both on statistical bias and evaluation.

Camilli (2006, p. 234) however, used both DIF and "item bias" for statistical bias and the term "fairness" for statistical bias and evaluation.

Importance of DIF. An item is determined to be biased if it is unfairly harder for one group than it is for another (Clauser & Mazor, 1998). DIF and bias are related in that for an item to be biased, it must display DIF. However, the presence of DIF alone is not sufficient a sufficient condition for the determination of bias. When differences between groups are found after matching individuals within those groups by ability, the item could be measuring not only the primary ability of interest but also an additional ability. This item could be considered bias or not based on the relevancy of the additional ability it is measuring. The item could be considered unbiased if the additional ability it is measuring is relevant to the purpose of the test (Clauser & Mazor).

Contemporary validity theorists such as Cronbach (1988) and Messick (1988) have argued that test validation efforts should be focused on its application and the interpretation of its results. Further, DIF analysis is just one piece of the whole validity argument for a test. It is a separate piece of evidence from the others used within the overall test development process but is used in conjunction with all other validity evidence to support the validity argument (Clauser & Mazor, 1998).

One possibility for an item exhibiting DIF is that it doesn't accurately measure the ability of interest for members in one of the groups (Camilli & Shepard, 1994). Items found to exhibit DIF, could reviewed by a committee of content experts for additional dimension(s) not integral to the intended construct of the test. If the committee is able to identify the additional dimension(s) that is causing the DIF, and if the dimension is not integral to the construct that the test is intended to measured, then the item could be

considered for exclusion from future test forms or from interpretations of test scores. If however, the additional dimension causing DIF is found to be construct relevant, then the item could be retained (Camilli & Shepard). The ultimate decisions as to the deposition of DIF items, generally is a matter of policy (Clauser & Mazor, 1998). These decisions are typically informed both by practicality and the intended use of the test (Clauser & Mazor). Additionally, if the reason an item exhibits DIF can be identified, then perhaps it can be re-written to eliminate the cause, or future items can be written insuring that the identified DIF-producing characteristics are not present.

Causes of DIF. Many, but not all, of the causes of local dependence within performance assessments delineated by Yen (1993) can cause DIF to be observed within dichotomously scored multiple-choice items. Her list, with examples of cases of DIF provided by respondent, includes

a) External assistance or interference – Special education students (the focal group), especially at the lower end of the ability scale, who were allowed the use of computational tools on a high school mathematics test, scaled based on responses of students who were not allowed computational tools, were found to have a higher probability of correctly answering computational items than either general education or other special education students not allowed this accommodation (Scott, 2009). It could be that the additional use of computational tools by the focal group acted as an external assistance rather than simply allowing the student to show what they knew or were able to do.
b) Speededness – An effect of differential speededness at the end of tests was found to disadvantage Blacks and Hispanics in two editions of Verbal and

Mathematical portions of the SAT (Schmitt, et all, 1990). It could be that the items in the early portions of the test took more time for these examinees to answer correctly or that they had less skill in managing their time to insure accurate completion of the test.

c) Item or response format – In both mathematics and in social studies, if an item contains a graph or figure (e.g. for social studies: map or chart), White examinees have been found to be advantaged over Black examinees, even when some written material is included (O'Neill & Mc Peek, 1993). It could be that differences in an auxiliary ability may be contributing to these findings, but it could also be an artifact of item difficulty since there was generally a moderate to strong correlation between difficulty and DIF.

d) Content – The inclusion of words that have multiple meanings (homographs) such as the word "hide," which can mean either the skin of an animal or the act of seeking concealment, was found to disadvantage Black and Hispanic examinees (O'Neill & Mc Peek, 1993). It could be that, when paired with multiple-choice options that played off the alternative meaning (e.g. hunt or shelter), these examinees got confused (O'Neill & Mc Peek, 1993).

e) Knowledge – Males were found to have a higher probability than female examinees, across the ability scale, of correctly answering verbal analogy SAT questions that are related to typical male pursuits (O'Neill & Mc Peek, 1993). These words, dealing with such topics as hunting, fishing, or ice hockey, could advantage males simply because of their additional knowledge of these topics (Camilli & Shepard, 1994).

f) Abilities – General education students, across the ability scale, were found to have a higher probability of correctly answering items that required logic to answer correctly than special education students who were allowed to use computational tools as an accommodation (Scott, 2009). It could be that the special education students not only had a handicap with computation but also an additional handicap associated with the use of logic. It could also be that the teachers of these students' had focused on computational or other low-level mathematical skills while choosing not to address the higher-order thinking skills required for the application of logic in mathematics problems. This lack of opportunity to learn is also evident for other subgroups such as English language learners on state science content assessments (e.g., Ilich, 2013) and between students from different countries in tests like the Second International Mathematics Study (McDonnell, 1995).

Some of the other causes of local dependence between items which Yen (1993) identifies, that may cause lack of unidimensionality, might not lead to differential item functioning between groups because they cause the lack of unidimensionality equally for both groups. However, DIF can also be observed (or hidden) when multidimensional data is analyzed and scaled using a unidimensional IRT model. For example, when two components of the Law School Admission Test were analyzed using testlets, rather than individual items that function autonomously, DIF was found between various groups that had previously been undetected (Wainer, 1995).

Populations of Interest. Since the institution of NCLB in 2001, there has been increased concern about differences between groups of students on standardized

assessments. States routinely perform DIF analyses on items on their annual achievement tests, comparing groups by gender and by ethnicity: White as compared to each of the other major ethnicities of at least Black, Hispanic, Native American, and Asian (e.g. ADE, 2011b). However, the list of populations of interest continues to grow. For example, linguistic complexity within items has been a concern for students who are still learning the English language. To investigate this issue, Martiniello (2009) compared the performance of ELL students on a mathematics assessment to students who were native speakers of English (non-ELL).

With the increase in the inclusion of various subgroups, such as ELLs and SWDs, who had previously been excluded from taking standardized assessments, much more concern about appropriate accommodations for those students has been evident. In addition to multiple studies of a specific accommodation for a specific subgroup (e.g. Cohen, Gregg, & Deng, 2005; Elliot, & Marquart, 2004; Middleton & Cahalan Laitusis, 2007), one recent study by a state examined the science performance of 19 different need categories of students who used one of 22 different accommodations on their annual state assessment (ADE, 2011a). This study attempted to tease apart the performance information for the various need groups rather than lumping all of the need categories for students into one overall variable, such as SWD. While not containing a DIF analysis, it did compare the students within each need group who took the test with accommodations to both the non-need students and students with the same need who took the assessment without standard accommodations. The concern for validating accommodations for examinees in specific subgroups has recently been highlighted within the Journal of Applied Testing Technology (JATT) when they chose to present this as a topic for a special issue (Camara, 2009).

The various studies sometimes have divergent results. The results presented within the special issue of JATT showed that while consistent, large differences in performance between groups was evident, there was no evidence that the accommodations changed the underlying constructs that the tests were measuring (Camara, 2009). Other studies (e.g. Scheuneman, Camara, Cascallar, Wendler, & Lawrence, 2002; Scott, 2009) have found a number of items that were influenced by the use of accommodations.

In a study that investigated the causes of DIF when accommodations were used on a state high school mathematics assessment, the students who were identified as having disabilities had a very different response distribution than the students in the regular population, regardless of whether they had used an accommodation or not (Scott, 2009). This study grouped students not only based on whether they had an identified need, but also by the type of accommodation they used: non-standard such as calculators or manipulatives, or standard accommodation such as the use of a white board or an alternative setting. Due to federal requirements, all tests of students using non-standard accommodations since 2007 have been invalidated and not included within aggregations of student performance or for accountability purposes (Franciosi, 2008). However, the distributions of both the mainstream students and those with disabilities who took the assessment with either standard or no accommodation (SWD) are suitable to inform conditions for the current study. In that state 2006 data, the SWD group had a population mean proportion correct of around .438 (σ= .164), where the regular education students (REG) had a mean proportion correct that was much higher at .688 (σ=.183). An associated difference was that these two disparate groups' population distributions had vastly different skewness with the SWD group having an extreme positive skewness of .898 (S_E=.036), and the REG group having a

moderately negative skewness of -.411 (S_E=.010). Another population of interest in current research concerns students who are ELLs that take assessments with standard accommodations geared specifically to address their needs. In the same state 2006 data, these students had a response distribution similar but slightly higher and less skewed than the students with disabilities with a mean proportion correct of .448 (σ=.150) and skewness of .772 (S_E=.034). The distributions of these diverse subgroups are important in that they inform the conditions that researchers should investigate prior to proposing new methodologies for use with groups such as these.

Methods of Examining Measurement Non-Invariance.

There are multiple methods for examining non-invariance in IRT and in SEM (non-IRT parametric) frameworks. For example within the IRT framework, DIF can be explored using various likelihood ratio tests (e.g. General IRT-LR, Bock & Aitkin, 1981; Loglinear IRT-LR, Thissen & Mooney, 1989), ratios of parameters (e.g. IRT-D^2, Bock, Muraki & Pfeiffenberger, 1988), or based on the area between ICCs (e.g. Raju, 1988) defined by the parameters (Wainer, 1993). Within the IRT framework, the parameters that are explored for non-invariance are the difficulty parameter (*b*), the discrimination parameter (*a*), and the pseudo-guessing parameter (*c*), usually in that order of priority, as modeled with the 3-PL as shown in Equation 2 above. Within SEM, investigations of DIF have been performed using multi-group CFA and MIMIC models (Teresi, 2006b).

IRT Methods. Within IRT, DIF can be investigated either parametrically or non-parametrically. While there are a great number of different parametric methods to investigate DIF, all of these require distributional assumptions. Historically, they assumed normal distributions for both focal and reference groups (Woods, 2011). Non-parametric

methodologies, however, place no restrictions on examinee distributions (Douglas, Roussos, Stout, 1996). While non-IRT parametric methods to investigate measurement non-invariance will be discussed because of the usefulness in test level analyses, parametric IRT methods might not be appropriate with the extreme performance distributions found within the subgroups of interest discussed above because of their assumption of normal distributions. The discussion here, therefore, is limited to two often-used non-parametric methods, namely Mantel-Haenszel and SIBTEST (Clauser & Mazor, 1998; Narayanan & Swaminathan, 1996). Both procedures begin by computing the number of correct and incorrect responses to items after matching examinees from each group, focal and referent. For these computations, they both can use the observed total score. SIBTEST, however, has the additional option of computing these using weighting by either the reference or the combined-group target ability distribution (Shealy & Stout, 1993b). From that point, the procedures differ.

Mantel-Haenszel Method. The Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988) detects DIF by first clustering both the focal and reference groups by each examinee's total test score. It then creates a 2 x 2 matrix of the number correct and incorrect responses for each score level (Clauser & Mazor, 1998). The null hypothesis for MH is that the focal and reference groups have the same odds of answering the item correctly at a given score level, for all score levels (Narayanan & Swaminathan, 1996). MH's null and alternative hypotheses can be written using Dorans' and Holland's (1993) notation as

$$H_o: [R_{rw}/M_{rw}]/[R_{fw}/M_{fw}] = 1 \quad w = 1, 2, \ldots W \tag{3}$$

$$H_a: [R_{rw}/M_{rw}] = \alpha\,[R_{fw}/M_{fw}] \qquad w = 1, 2, \ W \text{ and } a \neq 1. \tag{4}$$

In these equations, *R* stands for the number of correct (right) responses and *M* stands for the number of incorrect (missed) responses. The subscripts *r* and *f* refer to group, reference and focal, respectively, and the subscript *w* refers is the score level. The number of score levels is denoted as *W*.

Again using Dorans' and Holland's (1993) notation, the estimate of effect size for MH DIF as given by Clauser and Mazor (1998) can be written as

$$\alpha_{MH} = \frac{\sum_m (R_{rw} M_{fw})/\, T_w}{\sum_m (M_{rw} R_{fw})/\, T_w}. \tag{5}$$

In this equation T_w refers to the total number of responses from both groups at score level *w*. Dorans and Holland give The *MH* X^2 statistic, given by Dorans and Holland, which is revised here to use the same notation as equations 3 through 5, is

$$MH_X2 = \left[\left|\sum_w R_{rw} - \sum_w \mathrm{E}(R_{rw})\right| - 0.5\right]^2 \Big/ \sum_w Var(R_{rw}). \tag{6}$$

They note that in this equation, the purpose of subtracting 0.5 from the absolute value is to improve the statistic's accuracy. The $\mathrm{E}(R_{rw})$ term within this statistic can be defined as

$$\mathrm{E}(R_{rw}) = T_{rw}\, R_{tw} / T_w. \tag{7}$$

In this definition, T_{rw} is the total number of reference group responses and R_{tw} is the total number of correct responses at that score level. Additionally, the $Var(R_{rw})$ term can be defined as

$$Var(R_{rw}) = \frac{T_{rw} R_{tw} T_{fw} M_{tw}}{[T_w^2 (T_w - 1)]} . \quad (8)$$

Here, T_{fw} is the total number of focal group responses and M_{tw} is the total number of incorrect responses at that score level.

SIBTEST Method. The simultaneous item bias test (SIBTEST, Jiang & Stout, 1998; Shealy & Stout, 1993; Nandakumar & Stout, 1993) is also a non-parametric DIF procedure. While it is based on a multidimensional model for DIF, conceptually it can be thought of as very similar to the MH procedure (Clauser & Mazor, 1998). SIBTEST however incorporates some major innovations including a new matching procedure and the ability to test bundles of suspect items (Clauser & Mazor, 1998). While SIBTEST's matching process is operationalized as using a regression-corrected matched-score, it is conceptualized as matching on a latent score. The purpose of the addition of the model-based regression correction was to reduce Type I error (Shealy & Stout, 1993). It is based on the observed score for a group of items believed to be free of DIF. This is contrasted with MH's matching procedure which is strictly based on observed score. The other innovation for SIBTEST is the ability to "bundle" items. The purpose for this functionality is to evaluate the amplification or cancellation of DIF across items. In a confirmatory mode bundling can be used to group items in a predefined subtest. In an exploratory mode it can be used to

investigate whether individual items with low levels of DIF will exhibit significant DIF as a group (Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001).

For SIBTEST, DIF occurs at an ability level if the probability of correctly answering an item is not the same for the reference group ($P_R(\theta)$) as it is for the focal group ($P_F(\theta)$) at that ability level (Bolt & Stout, 1996). In SIBTEST's formulation, ability level is generalized as a corrected matching score, termed as the one target ability θ (Bolt & Stout). In their review of improvements made to the original version of SIBTEST (Shealy & Stout, 1993), Bolt and Stout state that within the SIBTEST framework, abilities other than that primarily intended to be measured by the assessment are secondary abilities. These additional abilities are represented with the vector η.

Secondary abilities can be either integral or extraneous to the dominant construct being measured. They term these secondary abilities auxiliary and nuisance, respectively. The probability functions, $f_R(\eta \mid \theta)$ and $f_F(\eta \mid \theta)$, incorporate conditional distributions of η for the respective groups at a fixed θ. Bolt and Stout explain that since all of the abilities that influence examinee performance are included within the latent space of the model, item response function invariance holds across the two groups (Bolt & Stout). The index they give for the amount of DIF that is attributable to differences in the distributions of the nuisance ability between the two groups at θ is

$$B(\theta) = E_R(Y \mid \theta) - E_F(Y \mid \theta)$$
$$= \int E(Y \mid \theta, \eta)\,[f_R(\eta \mid \theta) - f_F(\eta \mid \theta)]\,d\eta. \quad (9)$$

The suspect item score *Y* in this index can be defined as either the score of one dichotomous item or one polytomous item. It could, however be defined as the sum of the item scores for a group of either dichotomous or polytomous items (Bolt & Stout, 1996). Because the index assesses DIF at a specific ability level, it is a measure of "local DIF." The global index of DIF, β_{uni}, is computed by integrating $B(\theta)$ over the whole target ability range θ. In this integral, $f_F(\theta)$ is the density function of the focal group's target ability

$$\beta_{uni} = \int B(\theta) f_F(\theta) \, d\theta \, . \tag{10}$$

The subscript *UNI* in this index refers to the type of DIF being investigated (unidirectional). The $f_F(\theta)$ indicates that with this formulation the DIF found would be against the focal group (Bolt & Stout). This unidirectional formulation is for DIF that occurs for one group over the other across the entire ability range. Therefore it specifically examines the item(s) for uniform DIF. Sometimes however, in conditions such as when accommodations are used (Scott, 2009), the difference between the two groups' probabilities of correctly answering the item changes sign within the target ability scale. The more general non-uniform DIF function is calculated using a piece-wise integral, with the first integral going from the lower bound to the estimated crossing point ability and the second going from that estimate to the upper bound. Bolt and Stout give this as

$$\beta_{CRO} = \int_{\theta < \theta c} [E_F\,(Y \mid \theta) - E_R\,(Y \mid \theta)]\, f_F(\theta) d\theta$$

$$+ \int_{\theta > \theta c} [E_R\,(Y \mid \theta) - E_F\,(Y \mid \theta)]\, f_F(\theta) d\theta\,. \qquad (11)$$

In this formulation, θ_c is the estimated ability level for the location where the functions cross (Bolt & Stout). A positive value for this index indicates that the focal group has the advantage below the crossing point and the reference group has the advantage above that point where a negative value indicates that the reverse is true (Bolt & Stout).

All of these integrals in practice are estimated by dividing the curve into segments, estimation of the area under the curve between segments and then adding the areas together. This is a standard mathematical technique for estimating the area under a curve. For the crossing point used in the β_{CRO} computation, a linear regression is used. The crossing point is the estimate of the ability level where the curve heights are equal (Bolt & Stout, 1996).

Non-IRT Parametric Methods. SEM can be thought of as a marriage of measurement models and structural models so that parameter non-invariance investigations address both models together (Byrne, 2006). Investigations of the measurement model portion within which the item or variable loadings and intercepts lie can be variously performed using single- or multiple-group multiple indicator-multiple causes (MIMIC) or multiple-group CFA analyses (Teresi, 2006b). Investigations of non-invariance within SEM using CFA procedures can examine differences in dimension, configuration, metric, factor intercepts, and item-specific errors with some authors recommending the investigation proceed in that order (Teresi, 2006b).

The MIMIC model. The MIMIC model, when based on full information, can be thought of variously as an ordinal dependent variable CFA model or as a generalization of the IRT 2-parameter normal ogive function (Jones, 2006). It is a special case of Muthén's (2002) general model

$$y_i^* = \tau + \Lambda\Theta_i + \mathrm{K}g_i + \varepsilon_i \,, \tag{12}$$

for individual *i*, where y_i^* is (usually) a continuous latent response variable, τ is the vector of measurement intercepts, Λ is the matrix of factor loadings that relate the latent response variable to the latent trait (or measurement slopes), Θ is the vector of latent factors, g are the studied group categorical or continuous variables, the matrix of regression slopes (*K*) measures the direct effects of the studied variables on the response to the item, and ε is the vector of measurement errors (or residuals) that are uncorrelated with the other variables in the model. In the MIMIC model when dichotomous items are used, they are formulated as realizations of the latent underlying continuous variables (Teresi, 2006b).

With the MIMIC model, uniform DIF is assumed when parameter estimates within *K* are significantly different than 0 (Teresi, 2006b). The estimate of the magnitude of DIF, conditioned on the indirect effect of the other covariates on the item by the path through the latent variable, is the value of the coefficient of the path from the studied background (or grouping) variable to the item (Teresi, 2006a). The single group MIMIC model has expanded the basic SEM model developed by Jöreskog (1979) to incorporate terms that not only model the intercepts of the studied variables (comparable to the difficulty parameter in IRT), and the direct effects of the variables (Teresi, 2006b), but also explicitly recognize the

discrete and ordinal nature of dichotomous data (Millsap, 2006). With a single group MIMIC model, only uniform DIF can be detected while nonuniform DIF can be investigated if a multiple-group MIMIC model is used (Jones, 2006).

Both the MIMIC model and multiple-group SEM (or CFA) have distinct advantages for use in the investigation of DIF. Advantages for MIMIC are that (especially for discrete data) sample size requirements are reduced (Millsap, 2006) and that it allows the examination of the direct effect of background variables on factors as well as the use of continuous rather than only categorical background (or grouping) variables (Teresi, 2006a).

Advantages for multiple-group SEM include the ability for the investigation of the estimation of different loadings for each item (Teresi, 2006a). With single-group MIMIC, these loadings are treated as the same across groups which makes it impossible to detect differences in the discrimination parameter (Teresi, 2006a). While the use of the multiple-group MIMIC model can investigate differences in discrimination parameters, large sample sizes are required and the categorical group variables must be examined separately, one at a time (Teresi, 2006a).

Multi-group CFA within SEM. In the discussion of CFA non-invariance, testing a conceptual example using seven items intended to measure one factor as depicted in Figure 7, following Camilli's (2006) diagram, will be used. In Figure 7 v_1 through v_7 are test items, $\theta_1$is the intended latent construct to be assessed, $\eta_1$is an unintended secondary construct affecting item performance, G is a dichotomized grouping variable for the referent and focus groups and includes a predictive variable, X. The predictive variable X is derived by the simple sum of scored item responses. This variable is often the defacto estimate of examinee ability used in DIF analyses (e.g., MH and logistic regression; Clauser & Mazor, 1998) and is

indicated by the fixed unitaries specified on the paths connecting the test items to X. The solid lines indicate the original (or intended) relationships within the system where the dashed lines indicate the final (or actual) model in our example. Also, curved double-arrowed lines imply correlations where single-headed lines imply causality.

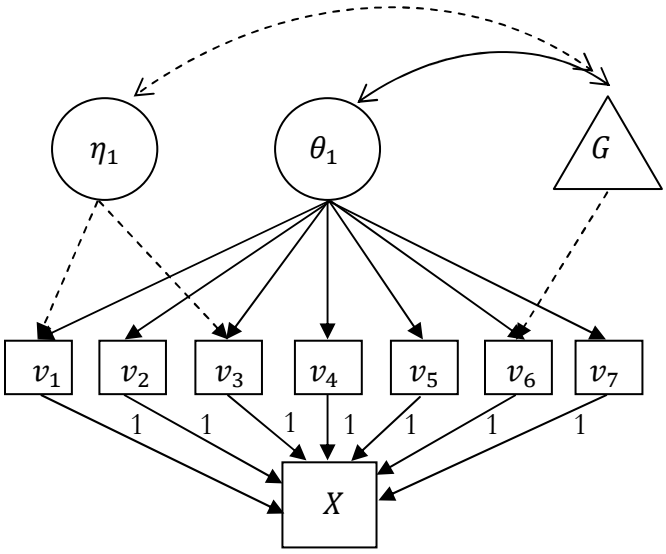


Figure 7. Multi-group structural model depicting both item and test level non-invariance.

Camilli (2006) explains that DIF within this diagram is depicted directly in v_6 by the path from the grouping variable to this item which indicates that it is affected by group membership beyond the two latent proficiency factors that influence item performance. He additionally contends that DIF may also be observed indirectly in items v_1 and v_3 due to the correlation between the grouping variable and the secondary latent trait. He argues that the utility of this type of diagram is that, in addition to modeling DIF, the impact of DIF on test usage can be modeled (Camilli, 2006). While these additional considerations are not the focus of this study, it is important to keep in mind that investigations of measurement non-

invariance or DIF are not an end in themselves, but only one step in the effort to ensure unbiased assessments and assessment usage.

As was mentioned earlier, investigations of non-invariance within SEM using CFA procedures have been recommended to proceed from examining differences in dimension, to configuration, to metric, to factor intercepts, and finally to item-specific errors (Teresi, 2006b). Within SEM, the factor loadings and the factor model intercepts can be considered comparable to the discrimination and difficulty parameters within IRT (Teresi, 2006b). The information of factor loadings and model intercepts are contained within the measurement model portion of SEM. The measurement model portion is equivalent to a factor analysis model of the data (Byrne, 2006), making confirmatory factor analysis procedures used for investigations of DIF within factor analysis models appropriate for this portion of non-invariance investigations within SEM (Teresi, 2006b).

Authors vary in the recommended number of steps or order in which the steps for investigation of invariance (or non-invariance) should proceed (see Gregorich, 2006; Mulaik & Millsap, 2000; and Thompson & Green, 2006): however, all agree that the model should approximately fit for all groups prior to investigating for DIF. A diagram of the original measurement model shown in Figure 7 is used in Figure 8 to display a unidimensional factor model in two groups that follows that presented by Gregorich (2006). This representation excludes factor means which are important considerations for factorial invariance as a clue that differential additive response bias might be existent (Gregorich, 2006). These means, however, are not considered when investigating DIF, since any difference between them for the groups would amount to impact rather than bias. Included are the single-headed arrows that represent the regression parameters with common factor loading values noted as λ and

common factor intercept values noted as τ for each item for each group. As previously stated, these are conceptually the same as item discriminations and difficulty parameters, respectively, investigated for DIF in the IRT framework and will be used to identify item parameter invariance within this CFA example.

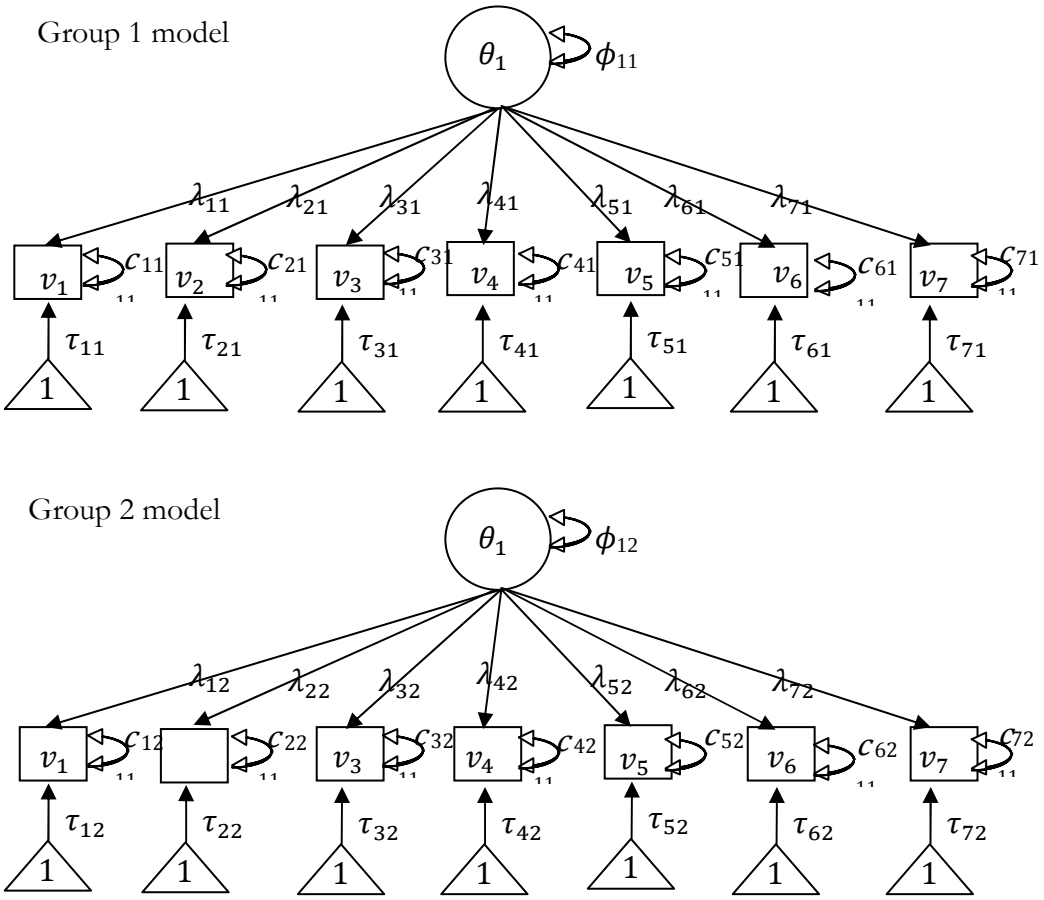


Figure 8. Example of a unidimensional factor model in 2 groups.

For one group the general common factor model has been offered as

$$v_{ij} = \sum_{k=1}^{q} \lambda_{ik}\theta_{jk} + \tau_{ij} + \alpha_i \ , \qquad (13)$$

where i runs from 1 to *p* items, *j* runs from 1 to *N* examinees, *k* runs from 1 to *q* factors and $N > p > q$. In this equation, v_{ij} are the scores for each of the examinees on each of the items, and as in Figure 8, λ_{ik} represent the common factor loadings (or regression coefficients) of each item on each of the common factors. Additionally, θ_{jk} represents the value of each examinee's factor score, and τ_{ij} and α_i represent the random errors of measurement and the intercepts from the regression of each of the items onto the factor, respectively. This model also aligns nicely to the IRT framework used within this study, where the IRT difficulty and discrimination parameters, *b* and *a* are non-linear transformations of τ and λ, respectively (Mc Donald, 1999: Takane & de Leeuw, 1987). The transformation equations given by Mc Donald (1999) are

$$b_i = {}^{\lambda_i}\Big/\sqrt{1-{\lambda_i}^2} \tag{14}$$

and

$$a_i = {}^{-\tau_i}\Big/\sqrt{1-{\lambda_i}^2}\,. \tag{15}$$

The investigation of measurement invariance generally relates to "full" invariance; however, this restriction can be relaxed at any step to obtain partial factorial invariance. This is particularly important in the investigation of DIF which is a search for non-invariance

between groups, at the item level. If partial invariance of the measurement model exists, then some but not all of the items' intercepts are equivalent across groups (Teresi, 2006b). Also, if an item (or factor) is found to be non-invariant between groups, it can be removed from further investigations of invariance, therefore purifying the comparison or conditioning estimate (Teresi, 2006b). This is comparable to processes within IRT where items that have been determined to contain DIF are removed from the estimate of θ in an iterative purification process (Teresi, 2006b). Meredith and Teresi (2006) cite the main causes for failure of measurement invariance (or DIF) as either the presence of multiple common factors and/or differences between groups on the means and/or variances of specific factors. They warn that applying models that are intended for continuous data with data that is ordered categorical might adversely affect the estimates of factor structure obtained (Meredith & Teresi, 2006).

Model Considerations. There are similarities between CFA and IRT approaches to invariance testing. The aim for both in this context is an investigation of the probability of observed patterns of item responses (Teresi, 2006b). Both rely on one or more latent variable, use model-based goodness-of-fit tests, and estimate a parameter similar to the corrected item total correlation from classic test theory (Teresi, 2006b). As indicated above, for IRT this is the discrimination parameter (a_i) and for CFA this is the factor loading (λ) where the difficulty (b_i) parameter is comparable to the item or variable intercepts (Teresi, 2006b). However, IRT processes conditions the probability of observing item response patterns on an estimate of ability, while CFA uses the marginal probability of a response pattern (Teresi, 2006b).

Both CFA and IRT have their strengths and weaknesses. One of the strengths of IRT has been its focus on individual ability estimates which are critical in some applications such as computerized adaptive testing (Teresi, 2006b). Also, IRT was specifically created to be used with dichotomous items where the SEM (including CFA) models are generally designed for continuous, or at least ordinal, variables where special processes are needed to accommodate categorical or dichotomous variables (Byrne, 2006). While this difference has started to blur as models are created within each framework to handle all types of variables (e.g. Jones, 2006; Muraki, 1999), some authors suggest that decisions for the choice of model might rightfully involve consideration of this difference (Teresi, 2006a, Hambleton, 2006).

IRT methods, unlike other parametric approaches, are comprehensive methods in that they can investigate differences between groups in all three parameters, difficulty, discrimination power, and pseudo-guessing (Angoff, 1993). While SEM methods can investigate parameter differences between groups in difficulty (item or variable intercept) and discrimination (item or variable loading on the factor), they are unable to address the issue of differences due to guessing (Angoff, 1993). However, IRT methods are similar to the SEM methods in that for some of the models (e.g. Rasch, 2-PL), guessing can be assumed to be nonexistent. For example, with the Rasch model, which assumes that all items are equally discriminating, an investigation of DIF would be limited to uniform DIF because the only difference allowed between items and item curves is difficulty (Angoff, 1993). With IRT, a mismatch between model and data can lead to misidentification of DIF; in fact some researchers insist that without a good model fit, DIF analyses are "pointless" (Hambleton, 2006, p. S186). A major strength of the use of SEM models (both MIMIC and CFA) is the ability to focus and explicitly model the relationship between the factors as well as the

relationships between factors and various variables of interest (Teresi, 2006b). A major weakness of both is that there needs to be a purification process to filter out the items that contain DIF from the analysis; in DIF, this resultant subset of items is called the matching subtest (Teresi, 2006b; Roussos & Stout, 1996a).

Current Subtest Selection Methods

Two-Step Process. One method of selection proposed by Holland and Thayer (1988) and studied by Clauser, Mazor, and Hambleton (1993), consists of a two-step process. In this process, DIF analyses is first run matching examinees on their total test score, then after the items that are found to contain DIF are removed, DIF analysis is rerun matching them on their "purified" score. Clauser, et al. (1993) found that the two-step process was superior to matching on total test score alone when using MH analysis for DIF detection. This was particularly true when the ability distributions of the focal and referent group were equal and the percentage of DIF items in the data was relatively large (20%). However, even in the least optimal cases (unequal ability distributions, small test length, and a small percentage of DIF items in the data) the improvement in both the Type I error rate and the percentage of DIF items identified by MH analysis was generally observed. While this two-step process is an improvement over simply matching examinees on their total test score, the DIF results are informed by matching on some items that might contain DIF and are therefore less than reliable.

External Criteria. Other researchers (Mazor, Kanjee, & Clauser, 1993; Williams, 1997) have suggested that the addition of one or more multivariate external criteria, such as student grades, might both "enhance the validity of subgroup comparisons and inferences of possible item bias" (Williams, 1997, p. 253) and allow for one designator to be imbedded as

an anchor across multiple test forms. There are multiple issues with this proposal, a few of which are delineated here. First, some testing situations are devoid of external criteria for this use. Second, the validity of teacher assignment of student grades can be called into question. Additionally, the consistency of grading procedures across teachers is non-existent (Williams, 1997). While the addition of one or more valid external matching variables to the matching subtest would be useful, the issues associated with this addition make it a non-viable option in many cases.

Dimensionality Based. Roussos and Stout (1996a) used the multidimensional model for DIF put forth by Shealy and Stout (1993a) to describe a paradigm that, in part, is designed to reduce the Type I error inflation of DIF analyses through the selection of a matching criterion based on understanding of the dimensionality of the dataset. The first stage of their unified DIF analysis process requires that the items be inspected for characteristics that are known or suspected to cause DIF. They suggest that, among other resources, these characteristics could be based upon the results of previously published DIF analyses, substantive content considerations, and dimensionality of archival test data. They state that the need for a substantive hypothesis base for DIF analysis is driven by the "overwhelming failure" (Roussos & Stout, 1996a, p. 360) to identify underlying causes of DIF. The second step of their process is to run confirmatory DIF analysis with the items suspected of measuring the same additional dimension bundled and assessed together. It is this additional dimension that, in the Shealy and Stout framework (1993a), is thought to vary between the focus and referent group and cause an item to exhibit DIF.

Group differences in ability on one or more secondary dimensions can cause DIF to be evident or obscured if the groups are matched on items that contain those dimensions

(O'Neill & McPeek, 1993; Roussos & Stout, 1996a). Roussos and Stout contend that the optimal matching subtest in the multidimensional framework would use only the items that "most purely measure only the primary dimension(s) of the test" (p. 367). They go on to say that, if this optimal matching subtest cannot be found, that the second best alternative would to be to use items that have secondary dimensions that are not contaminated by DIF. As examples of two methods to identify groups of suspect items, they briefly outline the processes presented in Douglas, Roussos, and Stout (1996). The first of these examples is based on expert review of the items to bundle possible suspect items. The second, using a process put forward by Stout, Douglas, Habing, Kim, Roussos, & Zhang (1996) combines the dimensional analysis programs hierarchical cluster analysis (HCA) and DIMTEST (Nandakumar & Stout, 1993; Stout, 1987) to identify bundles that corresponded to various secondary dimensions. In both examples, the bundles are then analyzed for DIF in a confirmatory mode using SIBTEST (Shealy & Stout, 1993b).

The intent of Roussos & Stout's (1996a) paradigm is to identify characteristics in items that cause DIF so that these characteristics could be consciously addressed during future item and assessment development. One might deal with identified causes of DIF by either modifying items to eliminate the DIF-causing characteristics or by balancing the benefit/disadvantage for groups within each test form. While the elimination of the characteristics identified as causing DIF within items is admirable and a prudent process for assessment developers, it is not applicable to cases, such as those described above, where the cause for DIF is not in how the item is written but rather in the interaction of examinees, who have access to non-standard resources (accommodations) with those items.

Using DIMTEST. Nandakumar (1994) also suggested that DIMTEST, which assesses essential unidimensionality of a dataset, might be useful in selecting an analytically valid subtest for use in DIF analyses. She investigated the first version of DIMTEST, which contained factor analysis as the mechanism for the selection of the comparison subtest and a non-bootstrap adjustment for bias based on large scale statistics. Working within the Shealy and Stout (1993a, 1993b) framework of multidimensionality for DIF items, she ran DIMTEST repeatedly with each analysis using two randomly assigned groups, one for the exploratory factor analysis and the other for the confirmatory dimensionality analysis. She then eliminated from the matching subtest, items that consistently were chosen by the factor analysis and reran DIMTEST to see if the resultant set of items was essentially unidimensional. If the result of this analysis was that the remaining items still did not compose an essentially unidimensional set, the DIMTEST analyses - removal of items process was repeated, until the final set of items was confirmed as essentially unidimensional. She then used this essentially unidimensional set of items as the matching set for DIF analysis (Nandakumar, 1994).

Her preliminary results showed promise; she found that she was able to identify the multidimensional items associated with DIF in most of the conditions. She cautioned that several conditions could cause this process to not be successful. She delineated these conditions as when 1) the number of items influenced by multidimensionality is pervasive, 2) multiple major dimensions affect the function of many items, or 3) one major auxiliary dimension influences many items. In these instances, she recommended that intensive item analysis by content experts be sought (Nandakumar, 1994). If DIF is caused by an accommodation used throughout the assessment, one would expect that many, if not most,

of the items might be influenced by the accommodation. Nandakumar warned that this is one of the cases where her process might not be successful so using it for the scenario outlined above would be inadvisable. This possibility, which is not being investigated within the current study, is one of this study's limitations.

Proposed Subtest Selection Method

While the above methods have proven useful for selecting an empirically valid subtest, it may be possible to improve and perhaps simplify this process for DIF analysis in the presence of multiple groups and items with multiple dimensions of various strengths. In this study, a method of selecting matching subtests for DIF analysis when there are multiple potential causes for DIF among multiple groups is proposed.

The mechanism for this new method is grounded in the work of Roussos and Stout (1996a), Stout et al. (1996), and Nandakumar (1994) as described above but also incorporates the improvements to the subtest selection process, called ATFIND, within DIMTEST introduced by Froelich and Habing (2008). ATFIND uses a combination of the non-parametric conditional covariance programs, HCA/CCPROX (Roussos, Stout, & Marden, 1998) and DETECT (Kim, 1994; Zhang & Stout, 1999b), to empirically select an assessment subtest (AT). Like DIMTEST, both HCA/CCPROX and DETECT were created to assess dimensionality. DIMTEST then compares AT to the remaining items (the partitioning subtest, PT). DIMTEST bases its dimensionality statistic on the examinee responses to those two sets of items. DIMTEST with AT selected by ATFIND was found to both maintain a Type I error rate around the nominal rate of α=.05 and have similar or significantly more power than when factor analysis was used to select AT when the data was modeled using a compensatory model (Froelich & Habing, 2008).

This study proposes the use of ATFIND to select a comparison subtest for use within DIF analysis. This proposal is based on this program's ability to select a dimensionally homogeneous set of items that is dimensionally distinct from the remaining set of items (Froelich & Habing, 2008) and is an extension of Nandakumar's (1994) work using DIMTEST's selection of AT to purify a valid subtest for DIF. With the program ATFIND, the dimensionally homogeneous set of items is designated as AT and the remaining heterogeneous and dimensionally distinct set of items is designated at the PT. It is the items in PT that are proposed as the matching subtest for DIF analysis and the items in the AT that are proposed to be suspect of DIF.

DIMTEST and ATFIND

DIMTEST's Purpose. The DIMTEST (Froelich & Habing, 2008; Nandakumar & Stout, 1993; Stout, 1987; Stout, Froelich, & Gao, 2001) procedure was created as a test of essential unidimensionality within assessments for at least three of the reasons discussed in section (b) above: 1) the requirement that the assessment is in reality measuring the trait which it purports to measure rather than being significantly contaminated by other traits, 2) that an assessment that is designed to measure differences between examinees is actually measuring a unified trait, and 3) if unidimensionality is violated, then the validity of the use of the resultant scores would be called into question (Stout, 1987).

Definition of Conditional Independence. The basic idea behind the conditional or local independence (LI) assumption is "that the trait value provides all relevant information about the examinee's performance and that once that trait value is taken into account, item responses are independent" (Yen & Fitzpatrick, 2006, p. 122). The assumption of conditional or local item independence, which Mokken (1996) distinguishes as

independence of responses within persons as opposed to sampling independence or independence of responses between persons, holds when for every single person for every fixed trait level, the response to any item in the test is independent of the responses to any other item in the test (Mokken, 1996). This implies that all of the systematic variation that exists within the item responses is only due to the variations of people over the trait range, the variation in any one trait value is random, and the residual, which is the only source of systematic variation, is kept constant (Mokken). A "strong" version of conditional independence can be formulated as

$$P(\{V_i = v_i\}|\theta) = \prod_{i=1}^{n} P(V_i = v_i|\theta) \quad (16)$$

where V_i is the score on item i (Yen & Fitzpatrick, 2006, p. 122). This definition requires that independence is conditioned on the trait (θ). This is contrasted to the case of correlations between the item scores of all the respondents who take the test when not accounting for θ (unconditioned). In the unconditioned case, the item scores are not expected to be, and generally should not be, independent (Yen & Fitzpatrick, 2006).

A "weak" version of conditional independence, as proposed by Mc Donald (1979), requires only the pair-wise conditional covariances among the items in the assessment to be equal to zero for every θ (Yen & Fitzpatrick, 2006). This "weak" version of conditional independence can be formulated as

$$P\big(V_i = v_i, V_j = v_j|\theta\big) = P(V_i = v_i|\theta)P\big(V_j = v_j|\theta\big) \quad (17)$$

where V_i is the score on the ith item and V_j is the score on item j (Yen & Fitzpatrick, 2006, p. 123). The concept behind these definitions is similar to that found within step-wise factor analysis where, analogous to conditioning on θ, the first factor is defined, and its variance accounted for, by the factor's removal in the first step (Yen & Fitzpatrick). The residual item correlations can then be inspected. If the item correlations are all zero, no more factors will be identified. However, if two or more items remain correlated, these items will contribute to additional factors being defined (Yen & Fitzpatrick).

Definition of Essential Unidimensionality. In 1990, Stout proposed essential independence as an even "weaker" definition of LI that was based on his 1987 definition of essential unidimensionality. In his definition of essential independence, it is only required that the average of the pair-wise correlations approach zero as the number of items in the assessment goes to infinity (Stout, 1990). This definition implies that only the dominant traits must be used by the model rather than requiring that the traits in the model completely explain the covariance between all of the items (Nandakumar, 1991). Nandakumar (1991, p. 102) gives this definition for every value of θ as

$$D_N(\theta) \equiv \frac{\sum_{1 \le i < j \le N} |Cov(V_i, V_j | \Theta = \theta)|}{\binom{N}{2}} \rightarrow 0 \; as \; N \rightarrow \infty \tag{18}$$

where V is the item pool, i and j are items within that pool, and Θ are, possibly, multiple latent traits. The data do not have to be unidimensional for LI to hold. If more than one trait is modeled, LI can still hold as long as each of those traits modeled completely describe the underlying responses within the items (Gorin & Embretson, 2008).

The unidimensionality assumption holds when only one ability or trait is measured by all of the items within the assessment (Hambleton, Swaminathan, & Rogers, 1991). In other words, the item responses are solely a function of the single, continuous, latent trait variable (de Ayala, 2009). The dimensionality of an assessment could be conceptualized as the number of traits that underlie the item responses (Gorin & Embretson, 2008) or alternatively, as the number of traits that must be modeled in order to achieve weak LI (Yen & Fitzpatrick, 2006). This second definition, put forth by Mc Donald (1981, 1982), has the added benefit of defining unidimensionality in operational terms, allowing it to be determined as a function of the conditional covariance between items (Yen & Fitzpatrick, 2006).

Rarely, if ever, does an assessment meet the strong assumption of unidimensionality (Hambleton, Swaminathan, & Rogers, 1991; Gorin & Embretson, 2008). A more appropriate criterion might be that the test data contains one dominant component, or trait, that influences performance on the test (Hambleton, Swaminathan, & Rogers, 1991). The idea of a single dominant trait being sufficient to explain the relationship between all of the item responses is the basis for Stout's (1987) essential unidimensionality (Gorin & Embretson, 2008). Stout defined essential dimensionality as the minimum number of dimensions required to attain essential independence (Yen & Fitzpatrick, 2006). Using Nandakumar's (1991) notation as above, formally Stout's (1987) definition of essential dimensionality is

$$d_E(\theta) \equiv \frac{\sum_{1 \le i \ne j \le N} \left| Cov(V_i, V_j | \Theta = \theta) \right|}{\binom{N}{2}} \approx 0 \qquad (19)$$

for all values of θ. Essential unidimensionality holds, when $d_E = 1$ (Nandakumar, 1991).

While the models discussed above have been based on the assumption of unidimensionality, it should be noted that researchers have been working to create models that can handle multidimensional assessments for both dichotomous and polytomous items (e.g. Mc Donald, 1996; Reckase, 1996), but the discussion of these models is outside the scope of the current study.

Rephrasing of the definition, Stout's essential unidimensionality states that on average, the pair-wise covariances between items, after accounting for the trait (or conditioning on the trait), will approach zero as the number of items in the test grows larger (Yen & Fitzpatrick, 2006). Stout (1987, p. 591) states that this definition is based on the "*fundamental principle* that LI should hold *approximately* when sampling from a subpopulation of examinees of *approximately equal ability*" (italics are author's). He goes on to say that if a test is multidimensional, then the fundamental principle would be violated because examinees with approximately the same test scores might differ widely in the items that they correctly answered to get their total score (Stout, 1987).

DIMTEST Hypotheses and Assumptions. Stout's (1987) DIMTEST is based on a nonparametric multidimensional latent trait model. He assumes that data contains the following characteristics: (a) that LI holds, (b) that a specific population, with a distribution of abilities, was randomly sampled for the examinees, (c) that different examinees had independent response patterns, (d) that a fixed set of items could possibly have been selected from a large set of items such as an item pool, and (e) that the data are adequately represented by monotonically increasing IRT. The DIMTEST methodology, however, only

assumes (a), (b), and (e) as stated above (Nandakumar & Yu, 1994). DIMTEST was originally created to assess the unidimensionality of dichotomous data which is the focus of the discussion here (Stout, 1987). It has since been expanded to assess the unidimensionality polytomous items with a separate program Poly-DIMTEST (Li, 1995), but since the current study is limited to dichotomous data, this extension will not be addressed here.

The hypothesis that is tested by DIMTEST is $H_0: d = 1$, which is compared to the alternative hypothesis of $H_0: d > 1$ (Stout, 1987). Stout (1987) noted that the conceptual raw data consists of $\{V_{ij}\}$ observations where i is the item index, which runs from 1 to $N = n + Q$ with n being the number of items in the "Partitioning" subtest (PT) and Q being the number of items in the "Assessment" subtest (AT), and j is the person index, which runs from 1 to J, the number of people in the sample. Additionally, each examinee's responses consist of a vector containing 0's and 1's with 0 indicating an incorrect response and 1 indicating a correct response (Stout, 1987). The basic algorithm was designed for tests administered to small samples ($J \leq 2000$) but includes a modification for larger sample sizes (Stout, 1987).

DIMTEST Procedure Logic. The general logic of DIMTEST, described in turn below, is: 1) the items within the test are divided into two subtests which are dimensionally distinct from each other (AT and PT), 2) the examinees are grouped based on their PT score, 3) the variance for each grouping of examinees is estimated, 4) an estimate of the variance for each grouping of examinees is computed using a formula that assumes unidimensionality, 5) normalize and then combine the different subgroup variances to form the statistic (T_1), 6) correct for statistical bias, and 7) test for unidimensionality using one of three comparisons derived for small ($J \leq 2{,}000$), moderate ($2{,}000 < J \leq 40{,}000$), or large ($J >$

40,000) administrations (Stout, 1987). Most of the steps within this logic remain virtually unchanged from the original launching of DIMTEST; however, changes in the process for performing the steps have changed and have been included in the newer DIMTEST 2.0 version. While Stout (1987) suggested empirically selecting the AT and PT item sets using factor analysis, the program ATFIND, which is a combination of two other programs (HCA/CCPROX and DETECT, described below), has recently been recommended and included within the program.

ATFIND Procedure. When originally proposed, Stout (1987) suggested that the two subtests required for the DIMTEST procedure be either 1) selected using expert opinion or 2) based on the results of principle axis factoring analysis. Later researchers (e.g., Blais & Laurier, 1995; Deng & Ansley, 2000; Froelich, 2000) found that the second option was less than ideal, especially when the underlying model was noncompensatory or when the examinee ability distribution was different from the item difficulty parameter distribution. Froelich and Habing (2008) proposed a new process (now called ATFIND within the program) using the conditional covariance-based theory (Stout, et al., 1996; Zhang and Stout, 1999a) as implemented with Roussos, Stout, and Marden's (1998) agglomerative hierarchical cluster analysis (HCA/CCPROX) and the DETECT procedure and index (Kim, 1994; Zhang & Stout,1999b).

In ATFIND, first HCA/CCPROX is run to cluster items into potential ATs. These clusters contain between four items and one-half of the items with the corresponding potential PTs identified as the rest of the items in the test. Once the clusters are formed, DETECT is run and its index is calculated for each of the AT/PT cluster pairs identified by HCA/CCPROX. The pair with the greatest DETECT index is then chosen as the AT and

the PT. It is this pair of clusters, that are the most dimensionally distinct, that are used for calculating the DIMTEST statistic as explained in steps 2 through 7.

Conditional Covariance-Based Theory. The conditional covariance-based theory that underlies HCA/CCPROX, DETECT, and DIMTEST is based on each item's discrimination vector as illustrated in Figure 9 for the 2D case. In this diagram, θ_1 and θ_2 are orthogonal traits being measured by the test, the item's discrimination vector is the direction (or composite of the directions) that is best measured by the item, and θ_{TT}, as discussed previously, is the weighted average of the item discriminations within the test. In this example, only Item 8 is solely measuring one trait (θ_1), all of the other items measure both θ_1 and θ_2 and best measure a composite ability somewhere between the two main traits.

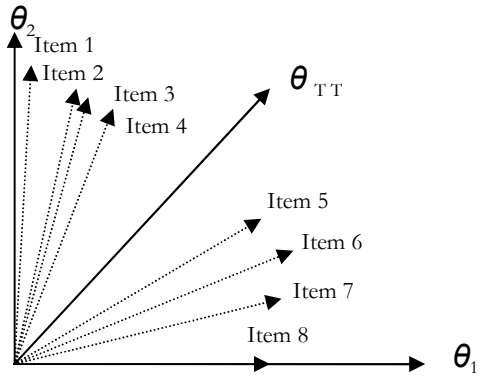


Figure 9. Item discrimination vector graph of a two-dimensional assessment.

The key result of this depiction from Zhang and Stout (1999a) is that the item pair conditional covariances based on the unidimensional θ_{TT} vector can be used to recover the multidimensional structure of the assessment as given by

$$CCov_{il} = \int_{-\infty}^{\infty} Cov(V_i, V_l | \Theta_{TT} = \theta_{TT}) f(\theta_{TT}) d\theta_{TT} \, . \quad (20)$$

With this equation, for a test with *d* dimensions, the vectors for *i* to *l* items are projected onto the (*d* - 1) dimensional hyperplane ($\Theta_{TT}^{\perp}$) that is orthogonal to the vector Θ_{TT}. If the angle between the pair of item vectors is less than 90° ($\pi/2$), then $CCov_{il}$ is positive, if the angle is more than 90°, the $CCov_{il}$ is negative, and if it is equal to 90°, it is equal to 0. The magnitude of the $CCov_{il}$ is related to the closeness of the two item vectors to each other, their closeness to the Θ_{TT} vector, and the length of their discrimination vectors. When there are only two dimensions, $\Theta_{TT}^{\perp}$ would be the vector through the origin orthogonal to Θ_{TT}. In this case, the item pairs whose vectors that lie on the same side of Θ_{TT} would have a projected angle of 0 which would create a positive conditional covariance. However, if the item pair has vectors that lie on opposite sides of Θ_{TT} , then their projected angle would be π (180°), creating a negative conditional covariance.

HCA/CCPROX. Using this theory in an iterative process, HCA/CCPROX identifies possible AT and PT subtests. It uses a step-wise method, starting with each item as a unique cluster then, at each step, the two closest clusters are joined together. The next to final stage contains two clusters with the final stage containing all items in one cluster. In our example of 8 items in Figure 8, in the first stage all eight items would each be in their own distinct cluster then, since Item 2 and Item 3 are items that are closest to each other (the degree of separation is the smallest), they would be joined in the second stage. The next to final stage would contain a cluster containing Items 1 through 4 and another containing Items 5 through 8, where the final stage would contain all eight items.

HCA/CCPROX uses a proximity measure $\rho_{CCOV}(V_i, V_l)$ to determine which two item clusters are closest to each other. This measure is an estimate of

$$-1(CCOV_{i,l}) + constant \tag{21}$$

for a pair of items *i* and *l*. The constant is added in the computation so that the proximity measure that results is non-negative. If any two items are on the same side of Θ_{TT}, they would have a positive conditional covariance, which after the sign reversal by multiplying by -1, would result in a small value of $\rho_{CCOV}(V_i, V_l)$. If, however, they were on opposite sides of Θ_{TT}, they would have a negative conditional covariance, which, after the sign reversal by multiplying by -1, would result in a large value of $\rho_{CCOV}(V_i, V_l)$. This creates a situation where the vectors of items that are near each other have small proximity values, where those that are far away from each other have large proximity values. To determine the distance between two clusters HCA/CCPROX, choosing single items from each of the two clusters, averages the distances between all of the item pairs. HCA/CCPROX returns a list of all item cluster pairings (all possible AT/PT pairings) but does not have a way to determine the optimal partitioning.

DETECT. The DETECT (Kim, 1994; Zhang & Stout, 1999b) procedure's goal is to determine the optimal partitioning of the test, assuming that it has approximate simple structure. Using the same underlying conditional covariance theory as described above (where all of the items within a cluster have a positive conditional covariance with each other, and a negative conditional covariance with any item from any other cluster), DETECT partitioning optimizes the DETECT index function

$$D(P, \Theta_{TT}) = \frac{2}{n(n-1)} \sum_{1 \le i \le l n} \int \delta_{i,l} CCOV_{i,l} \ . \quad (22)$$

When the items are in the same cluster in partition, *P*, $\delta_{i,l}$ is assigned the value of 1, and when they are in different clusters, $\delta_{i,l}$ is assigned the value of -1. In theory, the DETECT index could be computed for all possible pairings but, in practice, it only computes the index for the clusters provided by HCA/CCPROX. DETECT determines the pair of clusters that has both the most positive conditional covariances between the items in AT, while at the same time the most negative conditional covariances between the AT and PT items. For our example, this would create an AT containing Items 1 through 4 and a PT containing Items 5 through 8. DETECT does not, however, have a hypothesis test to determine whether the optimal clusters are dimensionally distinct (Froelich & Habing, 2008). This process is left to DIMTEST's procedure.

Study Structure

The study will be organized in two parts. The first part will be an investigation of ATFIND's ability to select a set of items that do not contain DIF, and the second part will be a comparison of the performance of three standard non-parametric DIF analyses. The performance of the DIF analyses with the matching subtests selected as the PT List by ATFIND will be compared to one generated to not contain DIF. The conditions used and the base data remain the same across both parts of the study.

CHAPTER 3

METHODOLOGY

The primary motivation for this study is the observed lack of available methodologies to efficiently select a matching subtest for DIF analysis when multiple groups and multiple potential causes for DIF are present. The purposes of this study are 1) to investigate the use of a current, easily accessible tool, ATFIND, to select a DIF free matching subtest and 2) explore SIBTEST's and MH's ability to identify DIF items among the suspect items, when the magnitude of DIF varies by group. The factors manipulated in the simulation study are selected such that they reflect multiple different, plausible, testing conditions (including data that contains no DIF items) and build upon existing research, especially that of Nandakumar (1994) and Froelich and Habing (2008).

Study Design

The factors manipulated in the study are: a) percent of DIF items, b) number of groups, c) percent of focal group simulees, d) ability distributions, and e) sample size. The study design, which is summarized in Table 2 resulted in 64 different conditions. Following Dinis (2008) and Guler and Penfield (2009), to allow for controlled magnitude of both uniform and non-uniform DIF items for the two focal groups, the unidimensional three parameter logistic model was used to create the data rather than a multidimensional model.

Number of Groups. Data were generated for a total of three groups. DIF analyses, by definition, always compare at least two groups, one reference and one focal. The cases for two groups acted as a baseline for the extension to three groups, one reference and two focal. This extension to three groups is to aid in the comparability of DIF analyses across

Table 2

Factors Manipulated for Data Generation

Factors	Levels	# of Levels
% of DIF Items	0%, 10% or 20% (1/2 uniform & 1/2 non-uniform)	3
Number of Groups	2 (Referent and Focal 1 or Focal 2) or 3	3
% of Focal Simulees	10% or 50%	2
Ability Distributions	*N*(0,1) or *N*(0,1), Mean =-1, SD=1, Skewness= .5, or *N*(0,1), Mean=-1.5, SD=1, Skewness = 1, or *N*(0,1), Mean=-1, SD=1, Skewness = .5, and Mean=-1.5, SD=1, Skewness = 1	4
Sample Size	750 or 2000	2
	Total Number of Conditions	64

multiple focal-reference pairings by selecting one "purified" subtest that is the same across those pairings.

Percent of DIF Items. Three percentages of DIF items were examined: 0 percent, 10 percent, and 20 percent of the 60 items modeled as containing DIF. The 10 and 20 percentage cases, which replicate Nandakumar's (1994) minimum and maximum percent of DIF items, are within the range studied by Furlow, Ross, and Gagne (2009; 10 and 25

percent) and are well within the 30 percent that has been found in application (Clauser, et al., 1993; Oshima & Miller, 1992; Hambleton & Rogers, 1989). The 0 percent cases were generated for each sample size and distribution for each focal group separately as well as for conditions when both focal groups were modelled within datasets. These were included to allow for the examination of Type I error of the proposed procedure for use with the three DIF analyses.

Percent of Focal Simulees. Two different percentages of focal simulees were modeled, 10 and 50 percent. With three groups, half of the focal simulees were from each of the two focal groups. These percentages follow Furlow, Ross, and Gagne (2009) study of DIF with disproportionate groups (e.g., Whites as compared to Asians or Hispanics) as well as approximately equal groups (e.g., genders).

Ability Distributions. There were two levels of ability distributions. For a baseline, the referent and both focal groups were all drawn from a normal $N(0,1)$ distribution. The second level maintained the ability distribution of the referent group at the normal $N(0,1)$, but drew the ability of the first focal group from a distribution with mean of -1.5, standard deviation (SD) of 1.0, and skew of 1.0., and the second focal group from a distribution with mean of -1.0, SD of 1.0, and skew of 0.5. The transformation for the focal group skewed distributions used Fleishman's (1972) method of transforming a $N(0,1)$ distribution to one with a specified skew and kurtosis explained within the Data Generation section below. These skewed distributions approximate those of a state high school mathematics assessment taken by students without an identified need as well as those identified with a disability and those identified as new to the English language, respectively.

Sample Size. Two sample sizes were used, 750 and 2000. The first 250 of the 750 simulees and the first 750 of the 2000 simulees were used in the analysis of ATFIND and the remaining 500 and 1250 were used in the DIF analyses. The total sample size and the number to use in the ATFIND analysis were the maximum and minimum used by Froelich and Habing (2008), who utilized the remaining simulees to run DIMTEST.

Data Generation

Model. Item responses using the conditions outlined above are generated using R (R Development Core Team, 2010) from a unidimensional 3-PL item response theory model. The items modeled to not contain DIF maintained the same parameters across the three groups (the reference and both focal groups); whereas the items modeled to contain DIF had at least one different parameter for each of the three groups. The difference in magnitude in the modification of the parameters for each focal group models a difference in effect associated with the cause for DIF. For cases containing both focal groups, these differences resulted in items generated to both DIF causes included within one set of data. The complete two-part study design (for subtest selection and DIF analysis) of 72 conditions is replicated 100 times (Froelich & Habing, 2008, Narayanan & Swaminathan, 1996).

Each 60-item dataset for each condition was generated using a group (*g*) specific version of the unidimensional 3-PL IRT equation given by

$$P(V_{is} = 1|\theta_s, a_{gi}, b_{gi}, c_{gi}) = c_{gi} + \frac{(1 - c_{gi})}{1 + e^{(-1.7a_{gi}(\theta_s - b_{gi}))}} \quad (23)$$

where a_{gi} is the item's discrimination for the group, b_{gi} is the item's difficulty parameter for the group, and c_{gi} is the item's pseudo-guessing parameter for the group. The simulee ability, θ_s, was taken from distributions, as discussed in Appendix E.

Item Parameters. The selected non-DIF item parameters remain fixed across all conditions. DIF item parameters are consistent across conditions for a particular group. The discrimination and difficulty item parameters for the 60 non-DIF items presented in Table 3 were randomly drawn from the parameters for the 80 items presented by Clauser, et al. (1993). As reported in Clauser, et al., these were originally estimated from the 1985 administration of the Graduate Management Admissions Test studied and published by Kingston, Leary and Wightman (1985). The pseudo-guessing parameter for all non-DIF items was set at .20.

Two different DIF models were incorporated within the data to simulate both uniform and non-uniform DIF. In each case, one-half of the DIF items were modeled as exhibiting uniform DIF favoring the focal group and one half of the DIF items were modeled as exhibiting non-uniform DIF. To determine appropriate parameter modifications for focal-group favoring uniform and non-uniform DIF items a study was undertaken using a real state assessment high school mathematics data where students were allowed to use various accommodations, such as calculators. Items in this data had previously been shown to exhibit DIF favoring the focal group that used accommodations (Scott, 2009). The investigation of focal-group favoring DIF item parameters is presented as Appendix B. For uniform DIF modification, the difference in difficulty parameters was randomly drawn from a uniform distribution from -0.20 to -0.50 for moderate DIF and from a uniform distribution from -0.50 to -0.80 for large DIF.

For non-uniform DIF modification the concern is to create differences in item characteristic curves that would ensure both that they cross at some point within the range of ability being studied and also that the resulting cross would advantage the focal group across more than an insignificant portion across that range. For the current study, that portion was defined as at least 20%. Therefore, the three parameter modifications, while being drawn separately from distributions, have those distributions influenced by the modification of the other parameters.

In an attempt to ensure that the ICCs crossed within the ability range of interest, lower differences in difficulty parameters were paired with higher ratios of discrimination parameters (greater than 1) and negative changes to the pseudo-guessing parameters (the Low *b*-difference modification). Similarly, higher differences in difficulty parameters were paired with lower ratios of discrimination parameters (less than 1) and positive changes to the pseudo-guessing parameters (the High *b*-difference modification). Each of the items modified for non-uniform DIF was assigned one of the two modification schemes. The only criteria for placement within one or the other modification subset, was that with average referent difficulty parameters for each subset would be as close to the same as possible. The resultant sets' average difficulty were equivalent at -0.06, however their standard deviations did vary (0.92 and 1.42).

For each item, two changes for each parameter were randomly drawn from distributions detailed below, then the changes that would resulted in parameters closest to that of the referent group (closest to 0 change for difficulty difference, 1 for ratio of discriminations, and 0 for pseudo-guessing) were identified as the "moderate" DIF modification and used for Focal Group 2 and the remaining changes were combined to be

identified as the "large" DIF used for Focal Group 1. After selection of the parameter changes, the resultant ICCs were examined and two (Items 20 and 60) failed to cross at any point within the ability range, -3.00 to 3.00. For these two items, the discrimination and pseudo-guessing parameter changes were redrawn until the curves were observed to cross. The item parameters for the DIF items for the two focal groups are presented in Table 4. The DIF items were purposefully dispersed throughout the simulated assessment to model an accommodation that would influence a student's response throughout the test.

The items modified to exhibit DIF were selected to approximate the difficulty and discrimination parameters of those that were randomly selected as non-DIF items. The sixty-item simulated non-DIF assessment difficulty and discrimination parameters had a mean and SD (in parentheses) of -0.10 (1.34) and 0.77 (0.28), respectively. For the items selected to be modified for uniform DIF, these values were -0.09 (0.94) and.90 (0.15), and for the items selected to be modified for nonuniform DIF they were -0.06 (1.08) and 0.77 (0.26), respectively.

It has been shown that dimensionality assessment and DIF analysis methods sometimes vary in performance when examinee ability distributions differ (Gierl, Gotzmann, & Boughton, 2004; Nandakumar & Yu, 1994; Seraphine, 2000; Walker, Azen, & Schmitt, 2006; Zwick, Thayer, & Mazzeo, 1997). Also, as exemplified by the state data discussed above, ability distributions for special populations such as SWDs and ELLs can be both disparate and skewed. Because of these factors, a comparison of ATFIND's performance for these distributions is incorporated. The person parameters for all three groups (the reference and both focal) are initially generated randomly from normal distributions with a mean of 0 and a variance of 1. To approximately model data found within real mathematics assessment

Table 3

Item Parameters for Non-DIF Items

#	a_{gi}	b_{gi}	c_{gi}	#	a_{gi}	b_{gi}	c_{gi}	#	a_{gi}	b_{gi}	c_{gi}
1	0.29	-2.95	0.20	21	1.04	2.11	0.20	41	0.50	0.80	0.20
2	0.75	-1.97	0.20	22	1.01	0.81	0.20	42	0.29	-1.00	0.20
3	0.36	-2.63	0.20	23	0.98	1.67	0.20	43	1.02	0.64	0.20
4	0.41	-2.93	0.20	24	0.65	1.68	0.20	44	1.16	1.11	0.20
5	0.56	-1.77	0.20	25	0.93	-0.23	0.20	45	0.48	2.12	0.20
6	0.73	-1.60	0.20	26	0.35	-1.12	0.20	46	0.65	1.19	0.20
7	0.94	-1.21	0.20	27	0.31	-1.37	0.20	47	0.79	-1.41	0.20
8	0.96	-2.70	0.20	28	0.39	-1.17	0.20	48	0.53	0.87	0.20
9*	0.64	-1.55	0.20	29*	1.05	0.10	0.20	49*	0.94	0.03	0.20
10**	0.75	-1.01	0.20	30**	0.51	-0.09	0.20	50**	1.01	0.91	0.20
11	0.82	0.61	0.20	31	0.55	1.26	0.20	51	1.11	0.35	0.20
12	0.86	-0.57	0.20	32	0.73	0.61	0.20	52	0.56	-1.41	0.20
13	0.42	-1.15	0.20	33	0.88	0.95	0.20	53	0.59	-1.29	0.20
14	0.74	0.60	0.20	34	1.40	1.64	0.20	54	1.01	0.22	0.20
15	0.44	-0.30	0.20	35	1.35	0.82	0.20	55	0.88	0.93	0.20
16	0.55	-1.06	0.20	36	0.92	1.13	0.20	56	1.32	0.57	0.20
17	0.82	1.02	0.20	37	0.73	1.18	0.20	57	1.09	1.11	0.20
18	0.52	-1.96	0.20	38	0.87	-0.75	0.20	58	0.83	1.54	0.20
19*	1.02	1.28	0.20	39*	0.81	-0.62	0.20	59*	0.94	0.25	0.20
20**	0.78	-0.05	0.20	40**	0.45	-1.49	0.20	60**	1.12	1.35	0.20

Note: * modified for uniform DIF. ** modified for non-uniform DIF.

Table 4

Item Parameters for DIF Items

	Focal Group 1			Focal Group 2		
#	a_{gi}	b_{gi}	c_{gi}	a_{gi}	b_{gi}	c_{gi}
9	0.64	-2.13	0.20	0.64	-1.82	0.20
10*	1.05	-1.28	0.12	0.91	-1.19	0.15
19	1.02	0.58	0.20	1.02	0.99	0.20
20**	0.54	-0.82	0.29	0.57	-0.60	0.24
29	1.05	-0.57	0.20	1.05	-0.14	0.20
30*	0.73	-0.58	0.10	0.58	-0.38	0.11
39	0.81	-1.42	0.20	0.81	-1.11	0.20
40**	0.32	-2.04	0.26	0.36	-2.02	0.23
49	0.94	-0.60	0.20	0.94	-0.22	0.20
50*	1.46	0.57	0.11	1.22	0.65	0.15
59	0.94	-0.31	0.20	0.94	-015	0.20
60**	0.74	0.82	0.30	0.77	0.85	0.21

Note: * assigned to the Low *b*-difference subset of non-uniform DIF items. ** assigned to the High *b*-difference subset of non-uniform DIF items.

data, a second set of distributions was used which retained the reference group's distribution at $N(0,1)$ while transforming the first focal group, which is intended to replicate students with special needs, to a distribution with mean -1.5, variance 1.0, and a skew of 1, and the second focal group, intended to replicate ELL students, to a distribution with mean -1,

variance 1.0, and a skew of .5. These focal distributions were transformed using Fleishman's (1978) process as described in Appendix E.

Analysis Methods

The analysis methods portion of the study is broken into two distinct parts, the first for the evaluation of ATFIND's performance in selecting a matching subtest and the second for an evaluation of the methods of identifying DIF using the selected subtest. For these purposes, MH D was computed and ATFIND, SIBTEST, and Crossing-SIBTEST, with their default options, were utilized. ATFIND selects between four and one-half of the total number of items that are most homogeneous and dimensionally, distinctly different from the rest of the items as the AT subtest (Froelich & Habing, 2008). A portion of the total sample (250 of 750 and 750 of 2000) is used to run ATFIND, while the remaining is used to run all three DIF analyses in a confirmatory mode. These quantities are both the maximum and minimum number Froelich and Habing (2008) used to run ATFIND as well as their total samples.

ATFIND Analysis. While it is expected that the AT subtest produced by ATFIND will contain the DIF items, since ATFIND (and DIMTEST) are single-group analyses, it remains unclear that this is the case when used with multiple groups.

Following Nandakumar (1994), this study uses the AT subtest as a list of suspect items and the PT subtest as the matching subtest for DIF analysis. However, as discussed above, her process of repeatedly eliminating items to purify the matching set might be less than ideal within the normal production constraints of typical state assessments systems which were the impetus for this study. These constraints, include high turn times (usually days to a few weeks), a high number of annual assessments (21 in one state), and many

regularly studied focus groups (typically, gender plus each ethnic minority), without even considering the workload demand for exploring DIF for the 22 different accommodations crossed with 19 different needs in our example. While Nandakumar's conditions are possible when an accommodation is allowed and used for the whole test by only one of two groups of students, the exploration of this extension of the use of the proposed methodology is reserved for future study.

DIF Analysis. The second part of the study will to be to examine the efficacy of the matching subtest procedure in standard DIF analysis. This part of the study will again be broken into two parts. The first is an examination of Type I error using the 0% DIF item cases and running SIBTEST, Crossing SIBTEST, and MH using the items selected for PT as the matching set. The second part of the analysis will be to run the three DIF analyses twice on each of the 100 datasets for each condition which contains DIF items.

For the DIF cases, the first run for each dataset will use the items not identified in the associated AT list as the matching set for the analysis. For the second run for each dataset, forty items consistently generated to be free of DIF will be included in the matching subtest where the remaining 20 items will be included in the suspect subtest. The twenty-item suspect list will include the 12 items modified for DIF for any condition, as well as 8 additional non-DIF items. The 8 non-DIF items were chosen to have discrimination and difficulty parameter distributions that were similar to both the whole test and the DIF modified items and will be included in the suspect list to allow for Type I error rate to be computed in all conditions for all DIF analyses. This second set of runs allows the new methodology's application within DIF analyses to be evaluated against the best possible scenario for each DIF analysis so that confounding of the applications can be reduced.

Outcome Variables

ATFIND. Both the PT List and the AT List for each dataset will be inspected for DIF items. The number and type of items selected for each of these subtests will be tabulated and various statistics will be presented. For the PT List, the mean number and standard deviation will be computed for each item type for each of the percentage DIF conditions for both the Small sample and Large sample. The percentage of each type of item in the PT List will also be computed for each condition generated. The matching subtest purity rate, will be determined by dividing the number of non-DIF items selected for the PT List by the total number of items selected for the PT List. .

AT List DIF item hit rates will be computed by sample size and degree of DIF modification for both percent DIF conditions and also by referent difficulty range ($b \leq -1.0$, $-1.0 < b \leq 1.0$, and $b > 1.0$) and for all conditions. Additionally, the overall AT List selection rate will be computed for each of the 12 DIF items.

DIF Analyses. The DIF analyses outcome measures will be computed for MH, SIBTEST, Crossing SIBTEST, and the combination of SIBTEST and Crossing SIBTEST results. This "Both SIBTESTs" is included since it is a reasonable expectation that a practitioner who used SIBTEST to check for uniform DIF will follow up that analysis with Crossing SIBTEST to check for non-uniform DIF. Both Type I error and Power rates will be computed for the three original analyses as well as the combined "Both SIBTESTs".

Since all conditions, including those with a pure matching subtest, should contain at least some non-DIF items inspected for DIF, Type I error rate will be computed for all conditions. It is expected that ATFIND will select at least a few non-DIF items for most of the AT Lists in each of the conditions and, as explained above, the suspect item list for the

pure matching subtest conditions will contain either 6 DIF and 14 non-DIF items or 12 DIF and 8 non-DIF items based on the percentage of DIF items modeled.

However, because of the differences in the default examination of suspect items between MH and the SIBTEST procedures, two different Type I error and Power rate computations will be produced for each matching subtest (Best and PT). The first, those which are standardly observed within the literature, especially when items are examined using all other items as a matching subtest, can be written as:

$$Type\ I\ error_{TOTAL} = \frac{\#\ of\ items\ falsely\ identified\ as\ containting\ DIF}{Total\ \#\ of\ non-DIF\ items\ in\ the\ dataset}$$

and

$$Power_{TOTAL} = \frac{\#\ of\ DIF\ items\ identified\ as\ containting\ DIF}{Total\ \#\ of\ DIF\ items\ in\ the\ dataset}$$

In the context of the current study, these rates are most appropriate in the examination of how the PT List matching subtest functions in conjunction with each of the DIF analyses, MH will inspect more of both DIF and non-DIF items than the SIBTEST procedure for DIF. Therefore, it is expected that MH will identify a higher number of both of these types of items as having DIF. This in turn will lead to both of these rates for both matching subtests being higher for MH than any of the SIBTEST procedures. The Total Power rates for the SIBTEST procedures are limited by the number of DIF items in the suspect list.

Because of the differences, following work by Zhou (2006) and Zhou, Gierl, and Tan (2006) in evaluating different aspects of SIBTEST procedures, a second set of Type I error and Power rates will be computed. These alternative rates will be based on the actual items analyzed for DIF by the different analyses. They can be written as:

$$Type\ I\ error_{ANALYZED} = \frac{\#\ of\ items\ falsely\ identified\ as\ containting\ DIF}{\#\ of\ nonDIF\ items\ analyzed\ for\ DIF}$$

and

$$Power_{ANALYZED} = \frac{\#\ of\ DIF\ items\ identified\ as\ containting\ DIF}{\#\ of\ DIF\ items\ analyzed\ for\ DIF}$$

In the context of the current study, these rates are most appropriate in the comparisons between the different DIF analyses as well as within each analysis between the use of a pure matching subtest (Best) and one that is contaminated by DIF items (PT). Since MH always analyzes all items, its Total Type I error is the same as its Analyzed Type I error and its Total power and its Analyzed power will both be the same. Therefore, only one of these rates will be computed for each of the matching subtests. For the SIBTEST procedures, however, the two sets of outcome variables are expected to be very different. With the reduction in number of items used in the denominator, for all of the SIBTEST procedures, both the Analyzed Type I error and Analyzed Power rates are expected to be much higher than the Total Type I error and Total Power rates.

Both the Total and Analyzed Type I error and Power rates will be computed for the three primary analyses (MH, SIBTEST, and Crossing SIBTEST) as well as the combined "Both SIBTESTs" for all conditions for both matching subtests. Additionally, the Total Power will be computed for all four analyses for each of the three referent difficulty ranges using the Best matching subtest. These will be computed within the 10 and 20 percent DIF conditions for all conditions as well as by sample size and overall.

CHAPTER 4

RESULTS

The results are provided in two main sections, first for those associated with the ATFIND analyses and then those for the three DIF analyses, MH, SIBTEST (SIB), and Crossing SIBTEST (X-SIB). The results, when both SIBTEST analyses (Both-SIB) are combined, are also included. Both main sections are subdivided into three subsections.

From this point forward, the categories of distributions for 0% DIF conditions will be noted as:

Normal focal distributions	Norm
Moderately Skewed focal distribution	MS
Large Skew focal distribution	LS
Both Skewed focal distributions	BS

For both the 10% and 20% DIF modified conditions, the following notation will be adopted:

Normal focal distribution with Moderate DIF modification	N-MDIF
Normal focal distribution with Large DIF modification	N-LDIF
Normal focal distribution with both DIF modifications	N-BDIF
Moderately Skewed focal distribution with Moderate DIF	MS-MDIF
Large Skewed focal distribution with Large DIF	LS-LDIF
Both Skewed focal distributions with Both DIFs	BS-BDIF

ATFIND Analysis

This section is divided into three subsections. The first presents the numbers of items (both DIF and non-DIF) selected for the PT List as a measure of purity of the

matching subtest to be used in the following DIF analyses. The second presents the distribution of the number of DIF items identified by the PT List in the 10 and 20 percent DIF conditions. The third presents the percentage of items ATFIND selected for the AT List based on difficulty range when items are divided by Low, Medium, and High referent item difficulty.

Matching subtest purity. This section describes the results of the investigation into the ability of ATFIND to identify a set of items that are DIF-free and to place them on the PT List. The intent of identifying this set of items is to use them as a matching set within DIF analysis.

Table 5 presents the number and percentage of items selected for the PT List by item type, aggregated by sample size and then the percentage of DIF items simulated. Regardless of the number of items simulated to exhibit DIF or the size of the sample, when aggregated by sample size within the three percent DIF conditions, on average approximately 36.3 (SD=4.3) items were selected for the PT List. This is about 60% of the 60 items included on each simulated test, leaving the remaining approximately 40% of the items selected for the AT List. This percentage varied slightly across the percentage of DIF items simulated and between sample sizes. However, in no case at the percent DIF/sample size aggregation level was the percent of items selected for the PT List lower than 59.7 or higher than 60.9.

There were only slight differences observed between the percentage of DIF and non-DIF items selected as well as between the type of DIF modification, uniform or non-uniform. Across the percent DIF/sample size aggregates, the percentage of non-DIF items selected for the PT List was slightly higher than that of DIF items. On average, across

sample size and both percent DIF conditions, 60.7% of non-DIF items were selected for the PT List where 58.9% of the DIF items were selected. Also, across these aggregations, on average, non-uniform DIF (NU-DIF) items (59.8%) were selected slightly more often than uniform DIF (U-DIF, 58.1%). The differences in these comparisons were extremely small in Small samples; they all increased, however, in Large samples. This was particularly true for the U-DIF to NU-DIF comparison which went from a maximum difference of .04% in the Small sample (10% DIF) to a difference of 3.5% in that same aggregate in the Large sample. While differences this small may be non-significant, the increase may indicate that they would get even larger as sample size increased beyond the Large sample size of 750 simulees that was used to run ATFIND.

Regardless of item type, for both percent DIF/sample size and condition level aggregations the percentage of items selected for the PT List was quite consistent and all were around 60%. The item type with the lowest PT List selection rate tended to be uniform DIF items (in 27 of the 48 DIF conditions). At the condition level, while the differences between the item type selection rate generally remained small (< 5% in 29 of the 48 DIF conditions), there were 19 instances where the difference in selection rate among the three items types (Non-DIF, U-DIF, and NU-DIF) was 5% or higher. Most of these (16 of the 19) were found with the Large sample, and in most instances (15), the item type with the lowest selection rate was U-DIF. In the comparisons with larger differences, the U-DIF items had an average selection rate of 56.3; whereas, the Non-DIF and NU-DIF items had selection rates of 61.5 and 60.6, respectively.

Table 5

Percent of Items Selected for PT List by Item Type.

Aggregation -Severity of DIF	0% DIF	10% DIF					20% DIF				
Item Type	**All**	**All**	**No DIF**	**U DIF**	**NU DIF**	**% Pure**	**All**	**No DIF**	**U DIF**	**NU DIF**	**% Pure**
Overall	**60.5**	**60.6**	**60.8**	**58.2**	**60.2**	**90.2**	**60.3**	**60.7**	**58.0**	**59.4**	**80.5**
***N*=250**	**60.2**	**60.3**	**60.4**	**59.2**	**59.6**	**90.1**	**59.7**	**59.9**	**58.6**	**58.9**	**80.3**
Normal Distribution (All Groups)											
10% Focal	59.6										
Moderate		59.6	59.2	63.0	63.3	89.4	60.3	60.1	62.8	59.5	79.7
Large		61.4	62.0	54.7	57.3	90.9	58.9	59.4	56.8	56.8	80.7
Both		59.8	59.7	60.7	59.7	89.9	60.4	60.4	58.2	62.2	80.1
50% Focal	59.9										
Moderate		60.4	60.7	59.7	56.7	90.4	60.1	59.9	61.3	60.3	79.8
Large		60.0	60.1	58.7	61.0	90.0	60.2	60.4	60.2	58.3	80.3
Both		60.5	60.6	61.7	58.0	90.1	59.9	60.1	58.0	60.3	80.3
Focal (Moderately Skewed Distribution - Moderate DIF)											
10% Focal	61.1	61.6	61.3	61.0	67.7	89.6	59.0	59.1	58.0	58.5	80.2
50% Focal	60.3	60.6	60.9	53.3	61.0	90.6	60.1	60.4	59.2	58.5	80.4
Focal (Large Skewed Distribution - Large DIF)											
10% Focal	61.0	60.4	60.7	60.0	56.3	90.4	59.0	59.3	57.2	58.0	80.5
50% Focal	59.7	60.2	60.1	60.0	62.0	89.9	59.2	59.9	55.3	57.7	80.9
Focal (Both Skewed Distributions - Both DIF)											
10% Focal	60.5	59.0	59.3	57.7	55.7	90.4	59.0	59.4	55.3	59.8	80.5
50% Focal	60.8	59.9	60.6	60.3	56.3	90.3	59.9	60.1	60.8	57.2	80.3

Aggregation -Severity of DIF	0% DIF	10% DIF					20% DIF				
Item Type	**All**	**All**	**No DIF**	**U DIF**	**NU DIF**	**% Pure**	**All**	**No DIF**	**U DIF**	**NU DIF**	**% Pure**
***N*=750**	**60.8**	**60.9**	**61.2**	**57.2**	**59.6**	**90.3**	**60.9**	**61.4**	**57.4**	**59.8**	**80.7**
Normal Distribution (All Groups)											
10% Focal	60.4										
Moderate		59.7	60.0	52.3	60.7	90.5	60.3	60.5	60.3	58.5	80.3
Large		60.0	59.8	61.7	60.3	89.8	60.9	61.9	54.7	59.3	81.3
Both		61.2	61.7	57.3	56.7	90.7	59.6	60.6	54.2	56.7	81.4
50% Focal	61.2										
Moderate		61.2	61.7	53.0	58.7	90.9	60.4	60.1	62.0	60.3	79.7
Large		61.0	61.9	51.0	56.0	91.2	62.8	63.4	56.8	63.8	80.8
Both		62.0	62.3	60.0	59.3	90.4	60.8	61.4	56.3	60.2	80.8
Focal (Moderately Skewed Distribution - Moderate DIF)											
10% Focal	58.8	60.6	60.9	56.0	59.7	90.5	61.1	61.7	57.3	59.8	80.8
50% Focal	61.4	61.5	61.6	58.7	61.7	90.2	62.1	62.8	62.2	56.2	80.9
Focal (Large Skewed Distribution - Large DIF)											
10% Focal	61.2	61.0	61.2	54.3	63.3	90.4	60.7	61.0	56.2	63.3	80.3
50% Focal	61.1	61.3	61.0	62.0	66.7	89.5	61.3	61.7	57.0	62.7	80.5
Focal (Both Skewed Distributions - Both DIF)											
10% Focal	60.8	60.5	60.7	53.3	63.7	90.3	60.5	60.6	58.7	61.2	80.2
50% Focal	61.9	61.4	61.1	66.7	62.0	89.5	60.2	61.5	53.5	56.0	81.8

Note: Data for all analyses consisted of combined Referent and Focal group generated responses. Referent group abilities were consistently drawn from a $N(0,1)$ distribution.

The percentage of items selected for the PT List was generally consistent regardless of the amount of the percent of simulee responses associated DIF items, the percent of DIF items simulated, the type of DIF simulated, or the severity of DIF modification imposed on the items. This generally consistent selection of items resulted in PT List purity rates that were very close to that of the percentages of non-DIF items. As can be seen in Table 5, the purity rate of the PT List in 10% DIF cases is, on average, 90.2%, where the purity rate in the 20% DIF cases, on average, is 80.5%. Purity rates for the various conditions fluctuate slightly around these averages.

In summary, a generally consistent 60-40 percent split for the PT and AT Lists was observed across sample sizes, percents of focal simulees included, focal distributions, percents of DIF items included, DIF item types, and severity of DIF modifications, as well as DIF item difficulties presented below. This surprising approximately 60% - 40% split led to an investigation to verify that the simulated data was created as specified prior to proceeding on to DIF analysis of the data. This investigation (presented in Appendix G) incorporated four analyses. Confirmatory and exploratory principal axis factor analysis was performed using Mplus (Muthen, & Muthen, 1998) and SPSS (SPSS, Inc., 1989), respectively. MH analysis (difMH within the difR package, Magis, Beland, & Raiche, 2013) was performed using R (The R Project for Statistical Computing, 2012). Finally, a parameter recovery check was performed using BILOG-MG 3 (Zimowski, Muraki, Mislevy, & Bock, 2003). The results of these four additional analyses confirmed that the simulee response data had indeed been generated as intended.

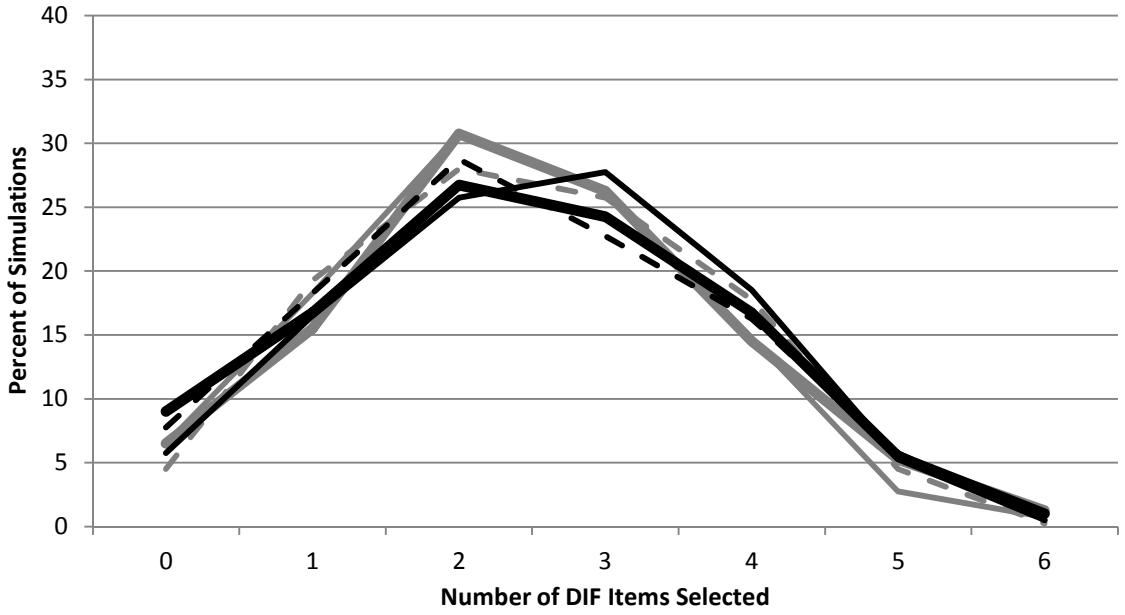
Items selected for AT List. This section is divided into two subsections. The first presents the frequencies and percentages of DIF items selected for the AT List by condition

aggregated across focal distributions. The second disaggregates these by focal distribution and referent item difficulty and compares these disaggregates to the selection for 0% DIF conditions.

Hit counts. Figure 10 presents the percentage of simulations for which the various number of DIF items were selected for 10% (a) and 20% (b) DIF cases. These values were aggregated across focal distributions. For each graph, the maxima appear between what are 33 to 50 percent of the DIF items. Rarely are none of the DIF items selected; and rarely are all, or even nearly all, of the DIF items selected. The overall number of items selected represents 40.8% (Mean = 2.4, SD=1.3) and 41.3% (Mean = 5.0, SD=2.0) of the DIF items in the 10% and 20% DIF conditions, respectively.

Referent item difficulty. Table 6 presents the percentage of all items selected for AT List by referent item difficulty range in 0%, and for the DIF modified items in the 10%, and 20%, DIF conditions. The items were divided into three difficulty ranges Low ($b \leq -1$), Medium ($-1.0 < b \leq 1.0$) and High ($b > 1.0$). For 10% DIF conditions, there was 1 DIF item with a referent item difficulty in the Low range, 3 items in the Medium range, and 2 items in the High range. In the 20% DIF conditions, there were 3, 7, and 2 DIF items in these ranges, respectively. Overall, regardless of sample size, simulee ability distribution, percentage of simulees in the focal group, or the percentage of DIF items in the dataset, items in the Medium difficulty range had a slightly higher AT List selection rate than either of the other two difficulty ranges. While these differences were generally very small, they tended to increase with sample size and with the number of DIF items included.

a)



b)

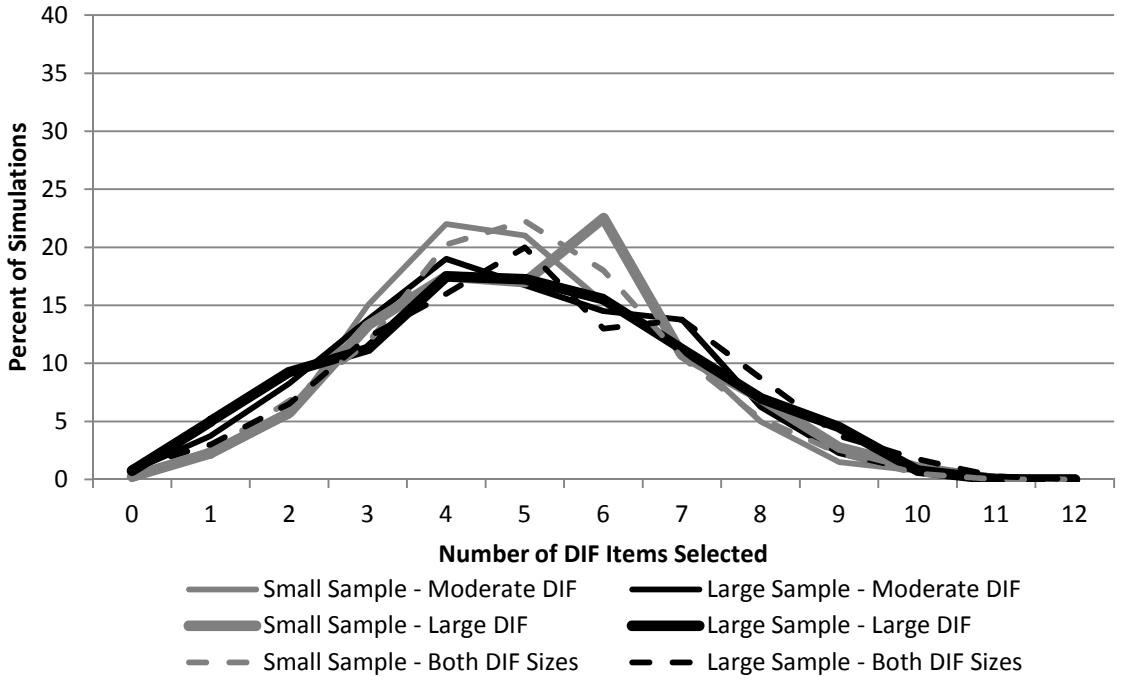


Figure 10. The percentage of simulations for which the various number of DIF items were selected for 10% (a) and 20% (b) DIF cases

Table 6

Percentage of items selected for AT List by referent item difficulty.

Aggregation-Severity of DIF	0% DIF			10% DIF			20% DIF		
	Low	Med	High	Low	Med	High	Low	Med	High
Overall	**37.6**	**41.3**	**38.9**	**38.4**	**41.7**	**40.7**	**38.4**	**42.9**	**40.2**
***N*=250**	**38.8**	**40.6**	**39.2**	**38.7**	**41.7**	**39.9**	**39.9**	**42.2**	**39.8**
Normal Distribution									
10% Focal	38.3	42.0	40.6						
Moderate DIF				35.0	37.7	36.5	33.7	40.7	40.0
Large DIF				38.0	47.0	42.5	42.3	44.0	41.5
Both DIF				42.0	42.0	35.5	39.0	39.7	41.5
50% Focal	38.2	41.4	40.5						
Moderate DIF				37.0	41.3	45.0	37.7	41.1	34.5
Large DIF				39.0	42.0	38.0	39.0	41.9	39.5
Both DIF				40.0	39.3	41.5	39.3	42.7	36.5
Focal (Moderately Skewed Distribution - Moderate DIF)									
10% Focal	34.8	41.4	40.8	37.0	39.3	29.5	41.3	42.4	40.0
50% Focal	39.5	41.4	37.2	35.0	45.3	43.0	38.0	41.6	44.5
Focal (Large Skewed Distribution - Large DIF)									
10% Focal	36.5	39.7	41.3	41.0	40.7	44.0	42.0	42.6	42.5
50% Focal	42.8	39.7	37.9	40.0	37.7	40.5	43.7	43.3	44.0
Focal (Both Skewed Distributions - Both DIF)									
10% Focal	39.3	40.5	38.1	40.0	47.7	38.5	41.0	44.7	36.5
50% Focal	41.1	38.9	37.3	40.0	40.7	44.0	42.0	41.9	36.5

Aggregation-Severity of DIF	0% DIF			10% DIF			20% DIF		
	Low	Med	High	Low	Med	High	Low	Med	High
***N*=750**	**36.4**	**42.0**	**38.5**	**38.1**	**41.7**	**41.5**	**36.8**	**43.5**	**40.6**
Normal Distribution									
10% Focal	35.7	43.5	38.9						
Moderate DIF				36.0	46.0	43.5	38.7	38.6	50.5
Large DIF				44.0	37.3	39.0	36.7	46.0	42.0
Both DIF				43.0	41.7	45.0	37.7	48.7	40.5
50% Focal	31.9	43.1	41.5						
Moderate DIF				38.0	42.7	49.5	33.3	40.3	42.0
Large DIF				38.0	47.3	49.5	29.3	43.9	40.5
Both DIF				44.0	39.7	39.5	37.3	44.6	38.5
Focal (Moderately Skewed Distribution - Moderate DIF)									
10% Focal	37.0	46.1	39.3	34.0	45.7	41.0	37.7	43.4	40.0
50% Focal	35.8	42.1	36.9	38.0	42.0	37.5	41.0	41.3	39.0
Focal (Large Skewed Distribution - Large DIF)									
10% Focal	36.5	40.8	38.7	35.0	42.7	42.0	33.0	43.7	39.0
50% Focal	41.1	40.0	34.3	30.0	38.0	35.0	36.3	41.1	42.5
Focal (Both Skewed Distributions - Both DIF)									
10% Focal	34.7	41.4	42.2	37.0	43.7	40.5	35.7	42.4	38.5
50% Focal	38.3	39.1	36.1	40.0	34.0	36.0	45.0	48.4	34.5

Note: Data for all analyses consisted of combined referent and focal group generated responses. Referent group abilities were consistently drawn from a $N(0,1)$ distribution. For 10% and 20% DIF conditions, results contain only for DIF items.

DIF Analysis.

The DIF analysis results section is presented in three subsections. The first presents the percent of time each analysis falsely identified items as containing DIF (Type I error rate). The second subsection contains the results of the analyses appropriately identifying all DIF items for both a DIF-free matching subtest and the PT List matching subtest (Power rates). The third describes the results of the analyses appropriately identifying DIF items (Power) by referent difficulty range using only the DIF-free matching subtest. In each of these subsections, the results for MH, SIB, X-SIB, and Both-SIB are presented. If an item was identified as containing DIF by either SIBTEST analysis, it was counted and presented as a percentage under Both-SIB.

The Total Type I error and Power rates are most appropriate in the examination of how the PT List matching subtest functions in conjunction with each of the DIF analyses. For this examination, the Total rates for the PT matching subtest are compared to those for the Best matching subtest. However, for an examination of how DIF contamination within the matching subtest influences each analyses' ability to identify DIF items, or for a comparison across the different DIF analyses, it is most appropriate to take into consideration the number of items available for it to identify. For these purposes, the Analyzed rates are used. For the examination of DIF contamination in the matching subtest, the Analyzed rates for the PT List matching subtest (which are expected to contain at least some DIF items) are compared to the rates for the Best matching subtest which were created to be DIF-free. For the comparison of the ability of each analysis to select DIF items within the three referent difficulty ranges, only the "Best" conditions were used. These Best

conditions ensure that all DIF items in each difficulty range were examined by each of the analyses.

Type I error for DIF identification.

Total Type I error. In this section, all rates were computed using the number of non-DIF items in the analysis: 60 for 0% DIF, 54 for 10% DIF, and 48 for 20% DIF conditions. Figures 11 through 13 present the percentages of non-DIF modified items that were falsely identified as containing DIF by each of the analyses for the 60, 54, and 48 non-DIF modified items in the 0%, 10%, and 20% DIF conditions, respectively. The tabulated results are available in Appendix H, Tables H1a through H1c.

Comparison of analyses to a criterion of 5% Total Type I error rate. When Total Type I error rates were computed across all conditions, generally all four analyses had rates below or only slightly above the nominal .05 regardless of matching subtest. For the Best matching subtest, the overall Total Type I error rates were .059, .018, .017, and .030 for MH, SIB, X-SIB, and Both-SIB, respectively. With the PT List matching subtest, they were slightly higher (.072, .035, .030, .052, respectively). With Total Type I error rates, both SIB and X-SIB generally identified less than the nominal 5% of non-DIF items as containing DIF. The only exceptions were all found when using the PT matching subtest in skewed distribution conditions, with 50% focal simulees.

When the results from both of the individual SIBTESTs were combined into Both-SIB, it maintained this general trend in conditions using the Best matching subtest. Both-SIB obtained a Type I error rate higher than 5% in none of the 28 conditions with normal distributions and in only 5 of the 36 conditions with skewed distributions. In conditions using the PT matching subtest with normal distributions Both-SIB Total Type I error rate

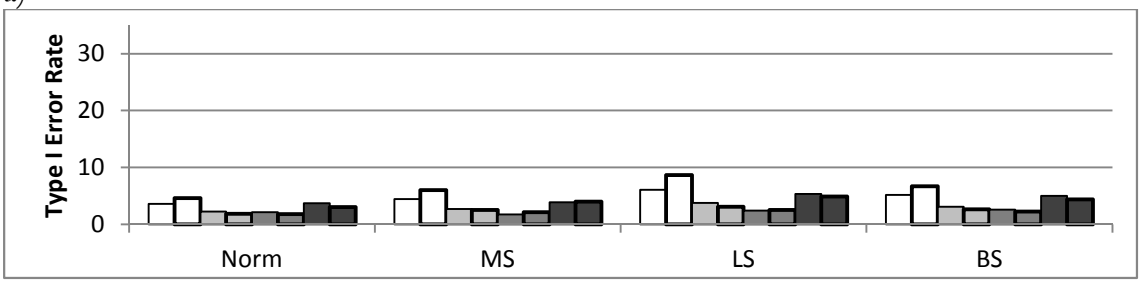
exceeded the 5% criterion in only 2 of the 28 conditions (20% DIF, Large sample, 50% focal simulees, N-LDIF and N-BDIF). It did, however, generally exceed the 5% criterion when the PT matching subtest was used with skewed distributions (in 31 out of the 36 conditions).

MH generally had a Total Type I error rate greater than the 5% criterion when data was modeled with skewed distributions regardless of matching subtest (in 32 and 34 of 36 conditions for Best and PT matching subtests, respectively). All six conditions where MH had a Total Type I error rate less than or equal to the 5% criterion in skewed conditions were with Small samples and 10% focal simulees. With data that was modeled using normal distributions, MH performed very differently. When the Best matching subtest was used with normal focal distributions, MH had Total Type I error rates less than or equal to the 5% criterion in all 28 conditions. When using the PT matching subtest it had a Total Type I error rate greater than the criterion in 13 of the 28 conditions.

Comparison between analyses. When comparing Total Type I error rates by matching subtest within conditions, regardless of sample size or percent of DIF items included in the data, MH consistently had a higher error rate than either SIB or X-SIB. MH also generally had a higher Total Type I error rate than Both-SIB (in 52 of 64 comparisons). The exceptions were all found in Small samples with 10% focal simulees. Eleven of these were found with PT matching subtest (spread approximately equally across the three percent DIF conditions) and one with Best matching subtest (0% DIF, Norm).

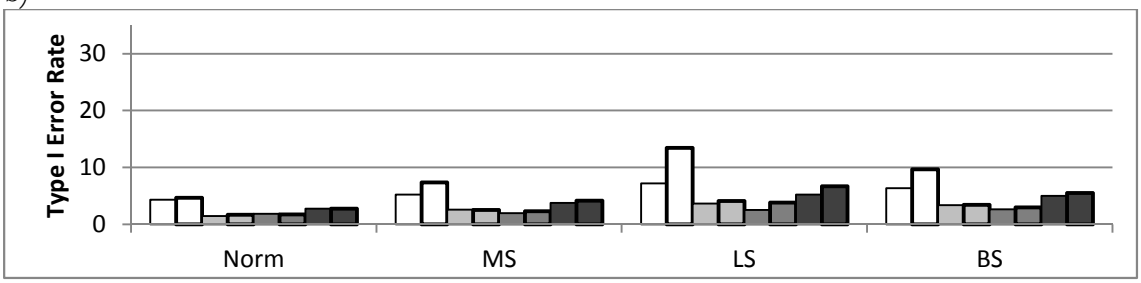
Comparison between Small and Large sample. When comparing the Total Type I error rates in Small versus Large samples, MH performed very differently than the SIB procedures. Generally, MH had a higher Total Type I error rate in Large than Small sample

a)



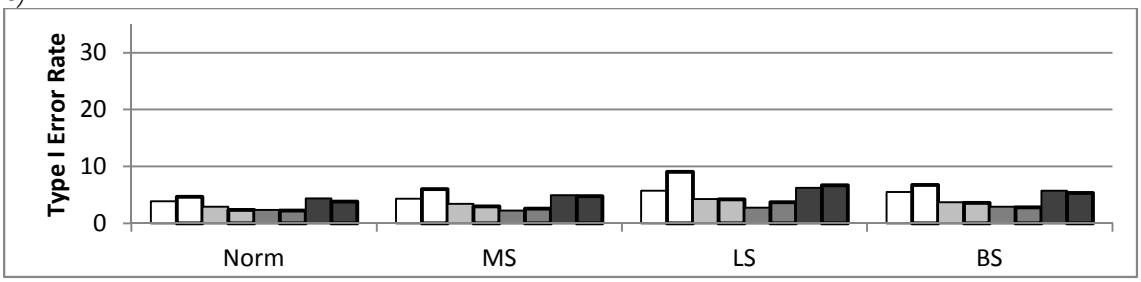
10% Focal simulees using Best subtest

b)



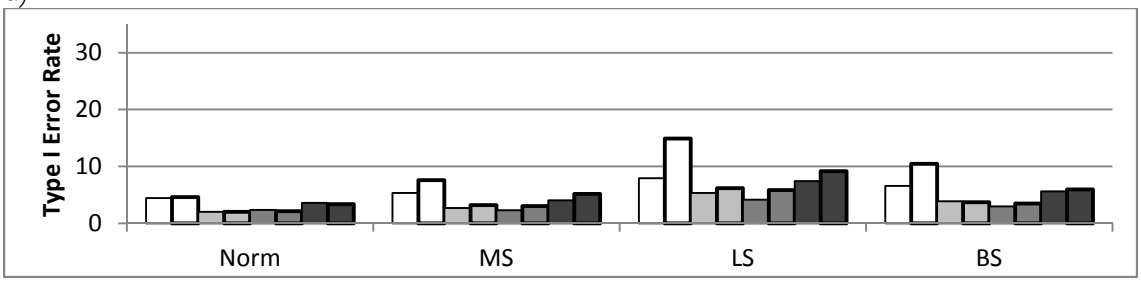
50% Focal simulees using Best subtest

c)



10% Focal simulees using PT subtest

d)

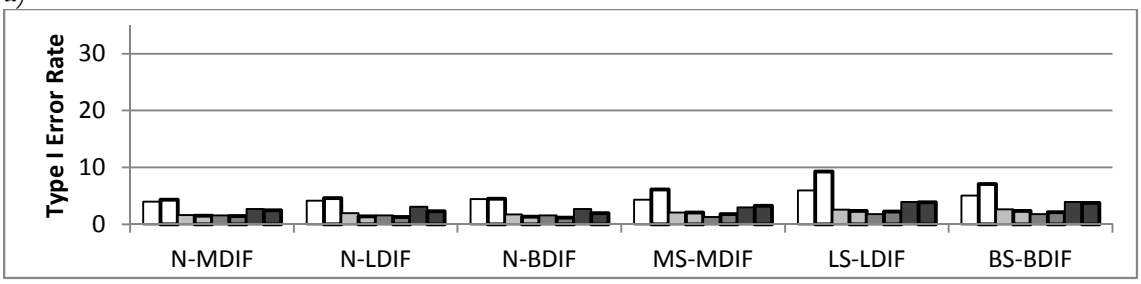


50% Focal simulees using Best subtest

Small Samples: Mantel-Haenszel SIBTEST Crossing SIBTEST Both SIBTESTs
Large Samples: Mantel-Haenszel SIBTEST Crossing SIBTEST Both SIBTESTs

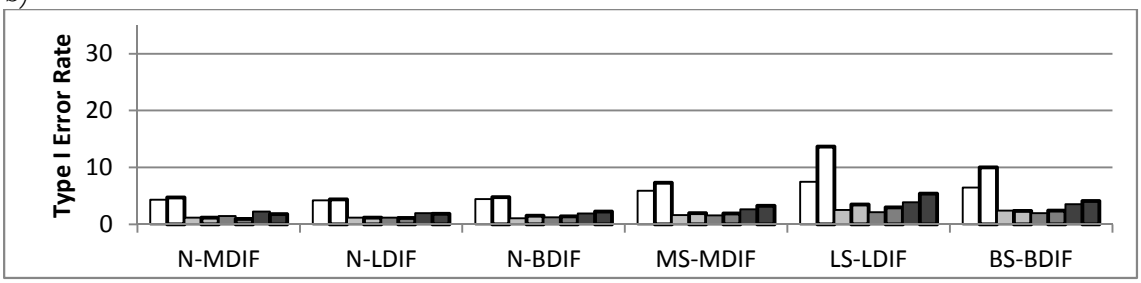
Figure 11. Total Type I error rates for 0% DIF conditions.

a)



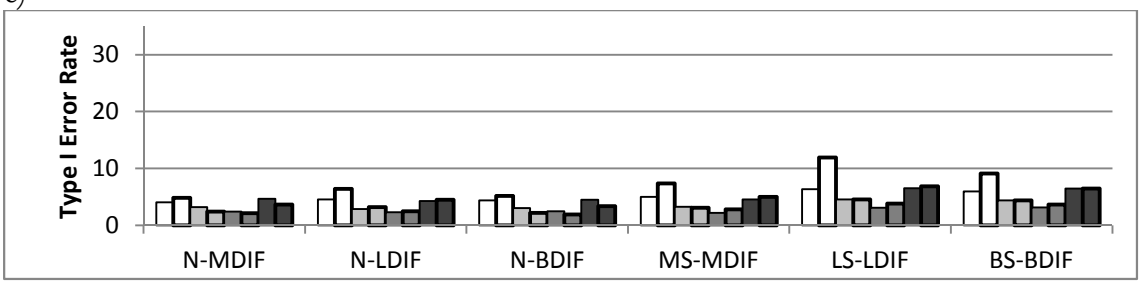
10% Focal simulees using Best subtest

b)



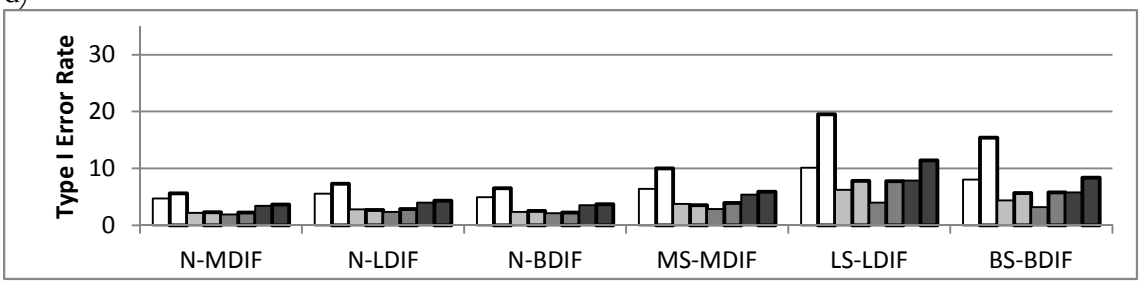
50% Focal simulees using Best subtest

c)



10% Focal simulees using PT subtest

d)

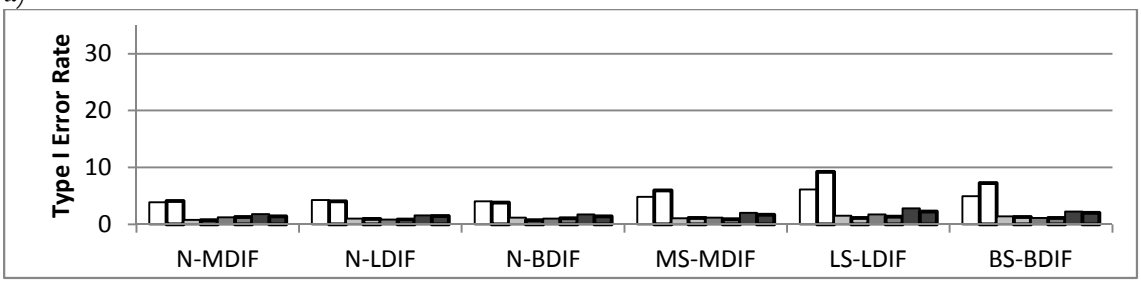


50% Focal simulees using Best subtest

Small Samples: Mantel-Haenszel SIBTEST Crossing SIBTEST Both SIBTESTs
Large Samples: Mantel-Haenszel SIBTEST Crossing SIBTEST Both SIBTESTs

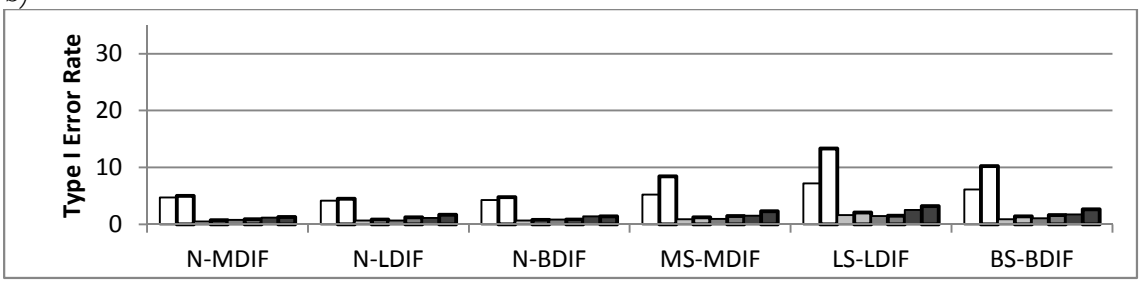
Figure 12. Total Type I error rates for 10% DIF conditions.

a)



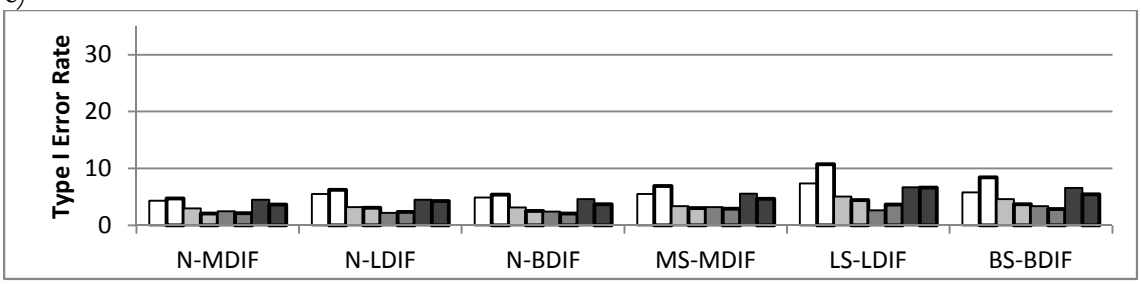
10% Focal simulees using Best subtest

b)



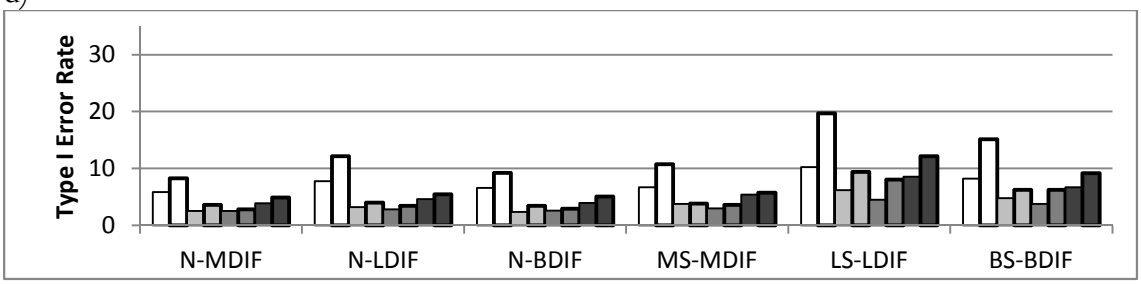
50% Focal simulees using Best subtest

c)



10% Focal simulees using PT subtest

d)



50% Focal simulees using Best subtest

Small Samples: Mantel-Haenszel SIBTEST Crossing SIBTEST Both SIBTESTs
Large Samples: Mantel-Haenszel SIBTEST Crossing SIBTEST Both SIBTESTs

Figure 13. Total Type I error rates for 20% DIF conditions.

conditions. There were only two conditions where this did not hold, both in 20% DIF conditions with 10% focal simulees using Best matching subtests.

SIB, X-SIB, and Both-SIB however, generally tended to have as high or higher Total Type I error rate with 10% focal simulees in Small sample conditions but with 50% focal simulees in Large sample conditions. For the Best matching subtest, across the 32 sample size condition level comparisons, there were very few conditions where this trend did not hold. These differences were quite small reaching a maximum of 1.5 percentage points with Both-SIB (6.7 for Large sample and 5.2 for Small). Most of the exceptions to the trend of Total Type I error rates in Large sample than in Small for SIBTEST were found in 10% DIF or skewed distribution conditions.

The one major difference between SIB and X-SIB, when comparing Total Type I error rates between Large and Small samples, was observed in 10% focal simulee, Skewed distribution conditions. In these conditions, for SIB there were no instances where Total Type I error rates in Large sample conditions was higher than in Small sample conditions. X-SIB however, had a higher Total Type I error rates in Large sample conditions than in Small in 11 out of the 18 comparisons.

Analyzed Type I error. In this section percentages were computed for MH using a denominator of 60, 54, and 48 for 0%, 10%, and 20% DIF conditions, respectively. The "Best" percentages for all three SIBTEST analyses (SIB, X-SIB, and Both-SIB) were computed using the number of non-DIF items in the static suspect list as the denominator (20, 14, and 8 for 0%, 10%, and 20% DIF conditions, respectively). The SIBTEST PT percentages were computed by dividing by the number of non-DIF items in the AT List. The Analyzed Type I error rates for each of the analyses in the 0%, 10%, and 20% DIF

conditions are presented in Figures 14, 15, and 16, respectively. The tabulated results are available in Tables H2a through H2c in Appendix H.

Comparison of analyses to a criterion of 5% Analyzed Type I error rate. With Analyzed Type I error rate, SIB and X-SIB, generally, and the combined Both-SIB consistently, had a higher Analyzed Type I error rate than the 5% criterion. This is in stark contrast to how the SIBTEST analyses performed with Total Type I error rate. While the MH error rate was close to the nominal .05, when Analyzed Type I error rates were computed across all conditions all SIBTEST analyses had elevated error rates, especially with the PT matching subtest. For the Best matching subtest, the overall Analyzed Type I error rates were .059, .0.70, .065, and .114 for MH, SIB, X-SIB, and Both-SIB, respectively. With the PT List matching subtest, they were even higher (.072, .090, .077, and .134, respectively).

In skewed distribution conditions, both SIB and X-SIB had Analyzed Type I error rates higher than the criterion in virtually every instance. The one exception was with X-SIB using the Best matching subtest in a 10% DIF condition (Small sample, 10% focal, MS-MDIF). With normal distribution conditions, using the Best matching subtest, SIB had an Analyzed Type I error rate higher than the criterion in half of the conditions (14 out of 28) where X-SIB had a rate higher than the criterion in 21 out of the 28 conditions. When the PT matching subtest was used in normal distribution conditions, SIB and X-SIB, each had Analyzed Type I error rates higher than the criterion in all but one condition. Both of these were found in Small sample with 50% focal simulees.

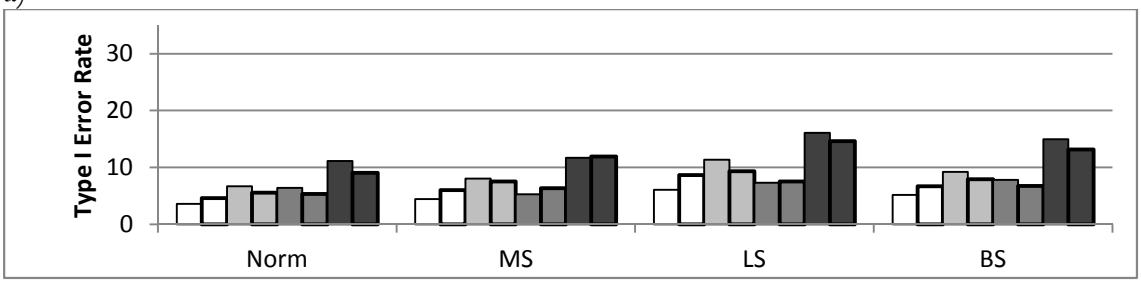
Comparison between analyses. With the Analyzed Type I error rates, SIB was more likely than either MH or X-SIB to have the highest Type I error rate with both the Best and PT matching subtests. With the Best and the PT matching subtests, SIB had the highest

Analyzed Type I error rate in 37 and 52, respectively. This compares to MH having the highest Total rate in 9 conditions with the Best matching subtest and 7 conditions with the PT matching subtest and X-SIB having the highest rate in 18 and 6 conditions using those subtests, respectively. Both-SIB consistently had the highest Analyzed Type I error rate compared to any of the other methods. The Analyzed Type I error rates for Both-SIB ranged from a low of 6.6% (20% DIF, Small sample, N-LDIF, with 50% focal simulees using the Best matching subtest) to a high of 31.8% (20% DIF, Large sample, LS-LDIF, with 50% focal simulees with the PT matching subtest).

Comparison between Best and PT matching subtests. Within conditions, analyses generally had lower Analyzed Type I error rates with the Best matching subtest than with the PT matching subtest across all analyses and all percent DIF conditions. This was especially true in the 10% and 20% DIF conditions, where in 95% of the comparisons the analyses had as low or lower Analyzed Type I error rates with the Best matching subtest as when the PT matching subtest was used. In datasets that contained DIF modified items, six of the seven conditions where the Best matching subtest was associated with a higher Type I error rate were observed in X-SIB analyses. Most of these six X-SIB exceptions were in conditions with 10% focal simulees. When DIF items were present in the dataset, MH consistently had the same or higher Type I error rate with the PT matching subtest than with the Best matching subtest.

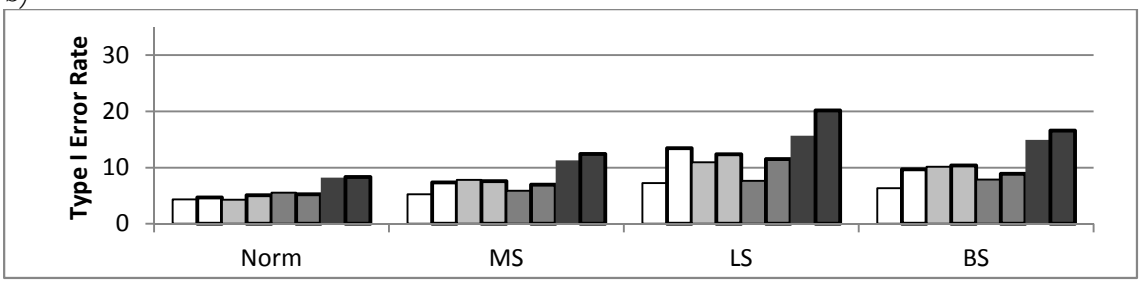
In 0% DIF conditions, both MH (with Total rate) and SIB (with Analyzed rate) generally had higher Type I error rates with the PT matching subtest than with the Best matching subtest regardless of sample size. While X-SIB followed this trend in 0% DIF conditions with Large sample, it did not do so in 0% DIF conditions with the Small sample.

a)



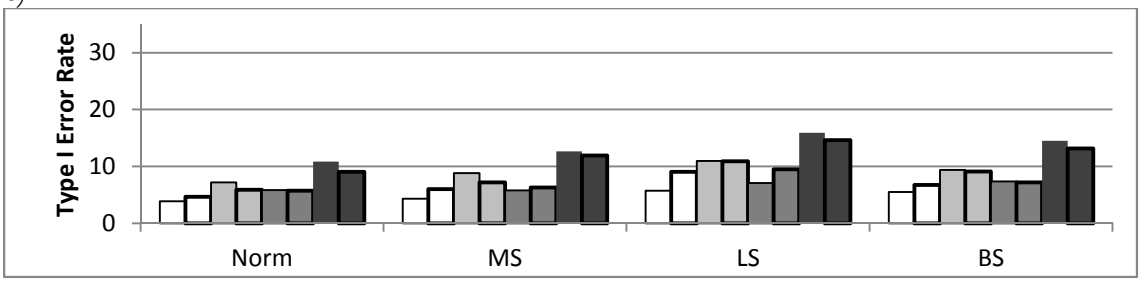
10% Focal simulees using Best subtest

b)



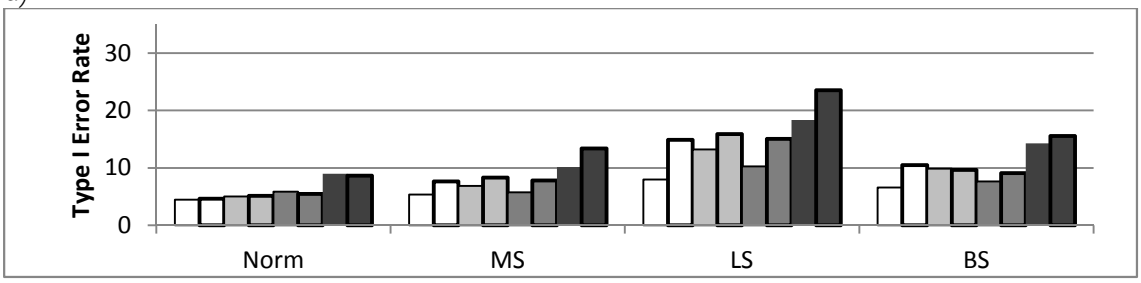
50% Focal simulees using Best subtest

c)



10% Focal simulees using PT subtest

d)

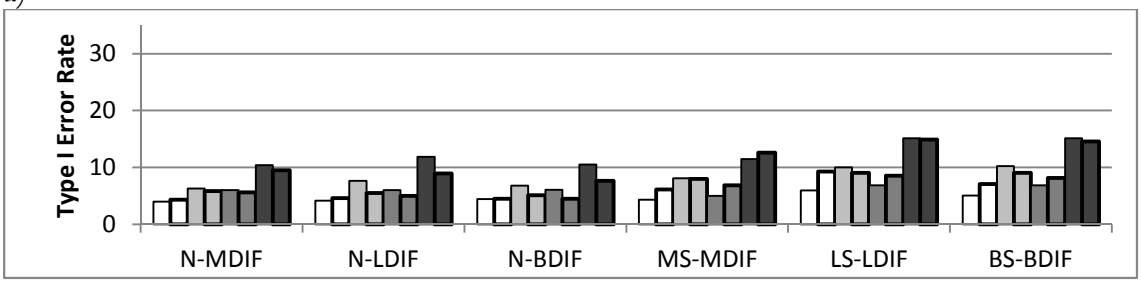


50% Focal simulees using Best subtest

Small Samples: Mantel-Haenszel SIBTEST Crossing SIBTEST Both SIBTESTs
Large Samples: Mantel-Haenszel SIBTEST Crossing SIBTEST Both SIBTESTs

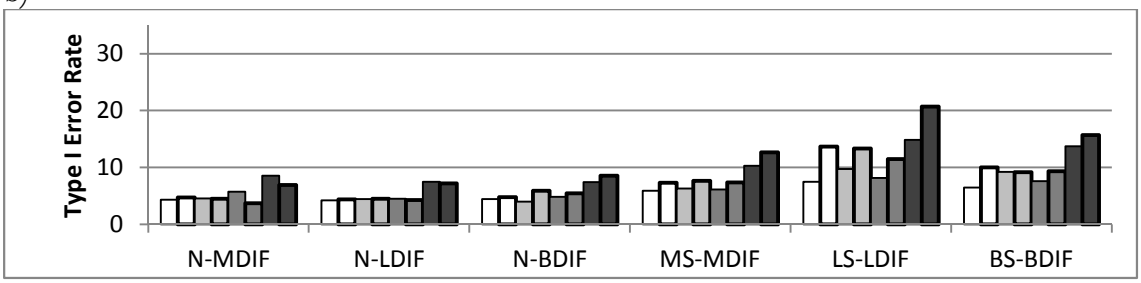
Figure 14. Analyzed Type I error rates for 0% DIF conditions.

a)



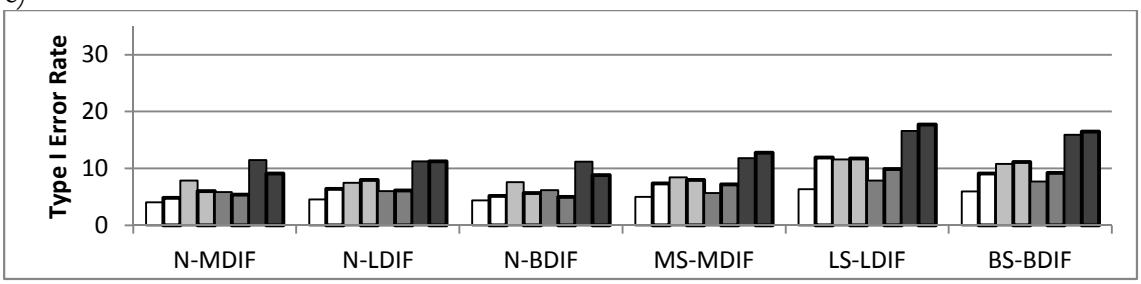
10% Focal simulees using Best subtest

b)



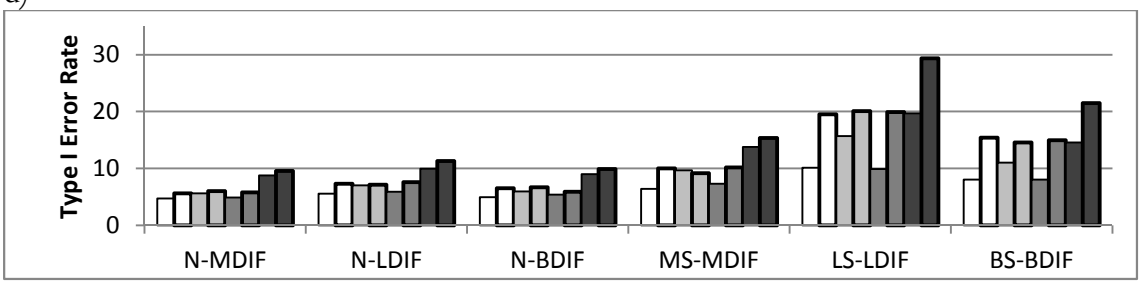
50% Focal simulees using Best subtest

c)



10% Focal simulees using PT subtest

d)

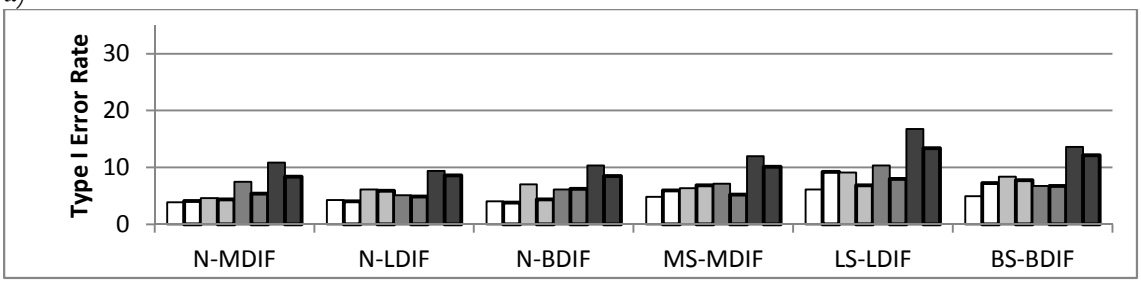


50% Focal simulees using Best subtest

Small Samples: Mantel-Haenszel SIBTEST Crossing SIBTEST Both SIBTESTs
Large Samples: Mantel-Haenszel SIBTEST Crossing SIBTEST Both SIBTESTs

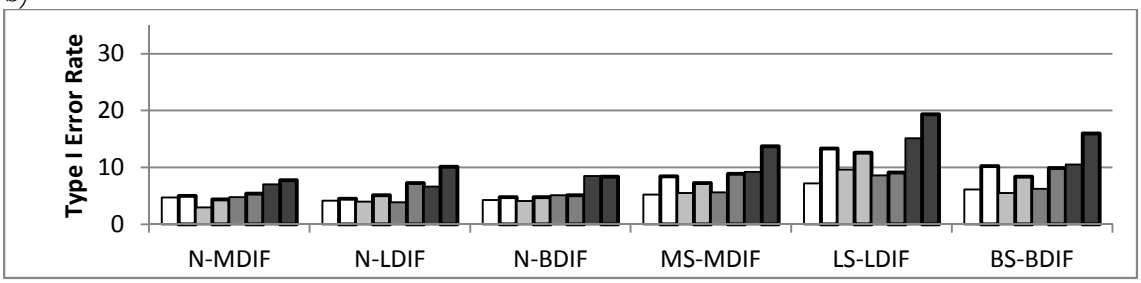
Figure 15. Analyzed Type I error rates for 10% DIF conditions.

a)



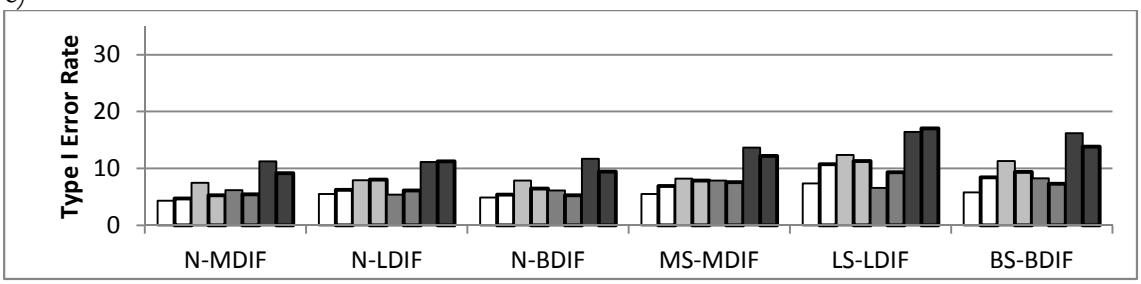
10% Focal simulees using Best subtest

b)



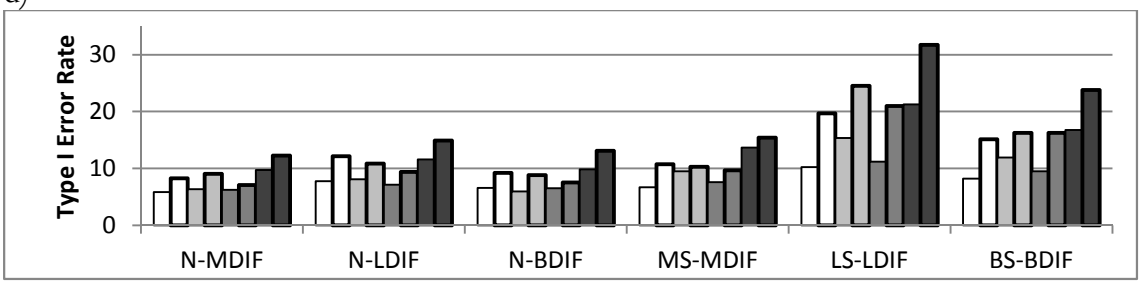
50% Focal simulees using Best subtest

c)



10% Focal simulees using PT subtest

d)



50% Focal simulees using Best subtest

Small Samples: Mantel-Haenszel SIBTEST Crossing SIBTEST Both SIBTESTs
Large Samples: Mantel-Haenszel SIBTEST Crossing SIBTEST Both SIBTESTs

Figure 16. Analyzed Type I error rates for 20% DIF conditions.

In the Small sample 0% DIF conditions, X-SIB generally had as high or higher Analyzed Type I error rate with the Best matching subtest (5 out of 8 times).

Power for DIF identification. This section contains the results of each of the analyses correctly identifying DIF items (power). As with Type I error rates, SIB and X-SIB results were combined and presented as Both-SIB. The section is divided into two main subsections. The first section presents power rates for all DIF items without regard to item type of difficulty range, and the second contains the power rates for DIF types within referent difficulty ranges.

Percent of DIF items correctly identified. Figures 17 and 18 present the Total and Analyzed power rates. The associated tabulated results are available in Table I1a and Table I1b in Appendix I. The Total power rates are presented as Best and PT matching subtest power rates. These were computed using the number of DIF items in the dataset as the denominator (6 and 12 for 10% and 20% DIF conditions, respectively). Since for MH, the default was to analyze all items for DIF only, these two power rates are presented. For all three SIBTEST analyses, in addition to the Total power rates for Best and PT matching subtests, the Analyzed power rates for the PT matching subtest are presented (labeled as DIF-AT within Figures 17 and 18).

Power rate across conditions using the Best matching subtest. When Total power is computed for the Best matching subtest across all conditions for each of the analyses, none of them performed very well. MH, SIB, and the combined Both-SIB did, however, perform very similarly, with Both-SIB performing slightly better than the other two analyses. X-SIB performed least well. Aggregating across all conditions, the overall Total power rates for MH, SIB, X-SIB, and Both-SIB were 46.8%, 45.9%, 35.8%, and 50.5%, respectively. When

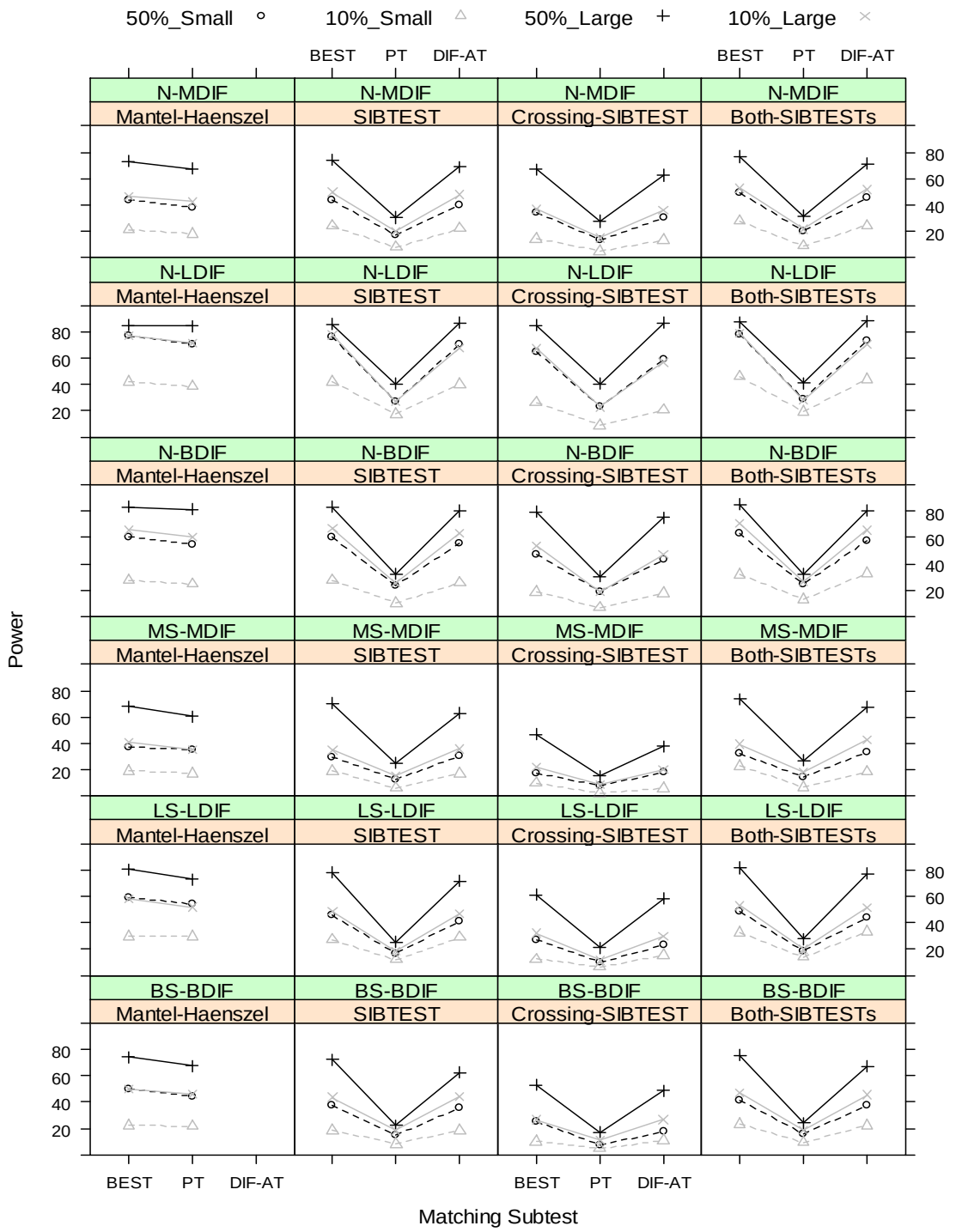


Figure 17. Power rates for DIF items within all difficulty ranges with 10% DIF. BEST = Best matching subtest; PT = PT List matching subtest; DIF-AT = Analyzed power rate with PT List matching subtest.

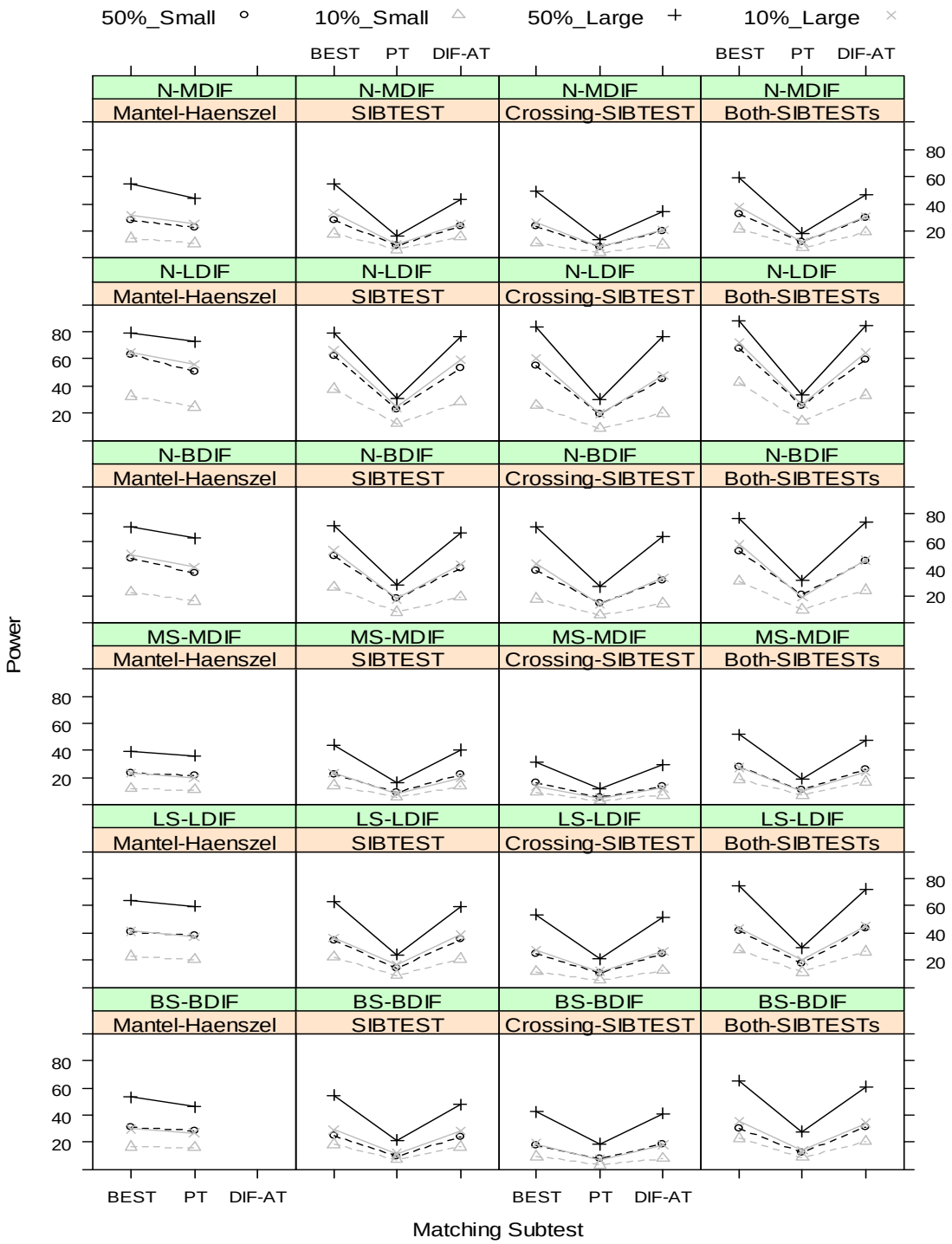


Figure 18. Power rates for DIF items within all difficulty ranges with 20% DIF. BEST = Best matching subtest; PT = PT List matching subtest; DIF-AT = Analyzed power rate with PT List matching subtest.

Total power was computed at the percent DIF level (across both sample sizes), all four analyses consistently performed better with 10% DIF than with 20% DIF conditions. When Total power was computed at the sample level within the percent DIF levels, all four analyses continued to show better performance in Large sample than in Small sample aggregations and in 10% DIF than in 20% DIF aggregations. The maximum Total power rates for the sample size within percent DIF aggregation were found in the 10% DIF/Large-sample aggregation for all four analyses (65.9%, 63.1%, 50.7% and 66.1% for MH, SIB, X-SIB, and Both-SIB, respectively). The minimums for all four analyses were found in the 20% DIF/Small-sample aggregation (31.2%, 33.5%, 25.9%, and 38.0%, respectively). The conditions where MH had higher Total power than Both-SIB with the Best matching subtest were generally in 10% DIF conditions with skewed distributions. Across all aggregations, X-SIB generally performed least well.

Total power rate comparisons between matching subtests. In this section, the results presented in the previous section for the Best matching subtest are compared to the Total power rates for the PT matching subtest. These are presented in the Best and PT columns in Figures 17 and 18 (and Appendix Tables I1a and I1b) for 10% and 20% DIF conditions, respectively. As described above, these Total power rates are particularly informative in the evaluation of how the ATFIND selected PT List matching subtest performed in conjunction with the DIF analyses.

Comparison of Best to the PT matching subtest within and between analyses. With Total power rates, generally for MH, and consistently across all conditions for all SIBTEST analyses, the use of the Best matching subtest produced higher power rates than the use of the PT matching subtest. The overall Total power rates with the PT matching subtest were 41.9%,

17.3%, 13.2%, and 19.3% for MH, SIB, X-SIB, and Both-SIB, respectively. Across all analyses and conditions regardless of matching subtest, Large samples yielded higher power rates than Small samples. When the PT List matching subtest was used, MH consistently had a much higher Total power rate than any of the SIBTEST analyses.

Analyzed power rates. Analyzed power rates, which account for the restricted number of items investigated by the SIBTEST procedures, are presented as DIF-AT in Figures 17 and 18 (and Appendix Tables I1a and I1b) for 10% and 20% DIF conditions respectively. Analyzed power rates are particularly informative in the evaluation of the impact DIF contamination within the matching subtest had on SIBTEST analysis performance. First the Analyzed power rates for the PT matching subtest will be compared to those for the Best matching subtest within the three SIBTEST analyses. The SIBTEST procedures' Analyzed power rates for the PT matching subtest will then be compared to MH power rates for the PT matching subtest.

Comparison of Best to the PT matching subtest within SIBTEST analyses. With Analyzed power rates, while all three SIBTEST procedures still generally had higher power rates with the Best than with the PT matching subtest, these rates within analyses were very similar. As was described above, SIBTEST procedures' overall power rates with the Best matching subtest were found to range from a low of 35.8% for X-SIB to a high of 50.5% for Both-SIB. Their overall Analyzed power rates for the PT matching subtest were consistently, slightly lower (32.1%, 41.1%, and 46.9% for X-SIB, SIB, and Both-SIB, respectively).The maximum difference between the two rates within a condition was 14.6 percentage points found for X-SIB in 20% DIF, Large sample, 50% focal simulee, N-MDIF. The maximum difference for the PT matching subtest having a higher Analyzed power rate than the Best

was a difference of 3.5 percentage points, found for Both-SIB in 10% DIF, Small sample, MS-MDIF with 50% focal simulees.

Comparison of PT matching subtests between analyses. When the Analyzed power rates for the PT matching subtest were used, similar to the finding with power rates with the Best matching subtest, Both-SIB and especially SIB had rates that were very similarly to MH's 41.9%. Both-SIB consistently performed slightly better than the other two analyses in 20% DIF conditions and generally better in 10% DIF conditions. Unlike with the Best matching subtest, with the PT matching subtest, there was no notable pattern, other than the percentage of DIF items in the dataset, to the conditions where MH had higher Total power than Both-SIB. The maximum difference between the Analyzed power rates for the PT matching subtest for either SIB or Both-SIB and those of MH was 15.0 percentage points. In this condition, Both-SIB had an Analyzed power rate of 61.0 and MH had a power rate of 46.0. X-SIB again performed least well, having power rates that were much lower in general than any other analysis. X-SIB performed most similarly to MH and SIB in Large sample conditions with 50% focal simulees, N-LDIF or N-BDIF.

Power rates for DIF items within referent difficulty ranges. This section contains the results of the analyses correctly identifying DIF items by referent difficulty range. The results presented in this section are only for the Best matching subtest; therefore, only the Total power rates are used. This restriction is for comparability as these were the only conditions that ensure that all DIF items in each difficulty range were examined by each of the analyses. The section is divided into three subsections, first for both U-DIF and NU-DIF modified items together and then for each of these two different modifications

separately. Within each subsection, results are presented for the analyses first across conditions and then within conditions.

Power rates for all DIF items. To compute power rates by referent difficulty, DIF modified items were divided into three ranges by the difficulty parameter assigned to the referent group. In 20% DIF cases, where all DIF modified items were used, three items (numbers 9, 10, and 40) had referent difficulty parameters less than -1.00, seven items (numbers 20, 29, 30, 39, 49, 50, and 59) had referent difficulty parameters greater than -1.00 but less than 1.00, and two items (numbers 19, and 60) had referent difficulty parameters greater than 1.00. For 10% DIF cases, items 9, 10, 29, 30, 49, and 50 were not modified to exhibit DIF. This modification design resulted in the referent difficulty power rates for Low, Medium, and High ranges for 20% DIF conditions being computed with a denominator of 3, 7, and 2, and for 10% DIF conditions with a denominator of 1, 3, and 2, respectively. Figures 19 and 20 present the power rates for the analyses within each of the referent difficulty ranges for 10 and 20 percent DIF conditions, respectively. The associated tabular results are available in Table I2a and Table I2b within Appendix I.

Total power rate across conditions. When power is computed across all conditions for each of the analyses, none of them performed well in all difficulty ranges. The overall power rates (aggregated across both sample size and percent DIF conditions) were consistently highest in the High difficulty range and lowest in the Low difficulty range for all four analyses. The overall power rates in the High difficulty range for MH, SIB, X-SIB, and Both-SIB were 60.8%, 54.9%, 42.8% and 58.6%, respectively. Those for the Medium difficulty range were slightly lower for each analysis (53.6%, 52.6%, 40.8%, and 57.3%, respectively).

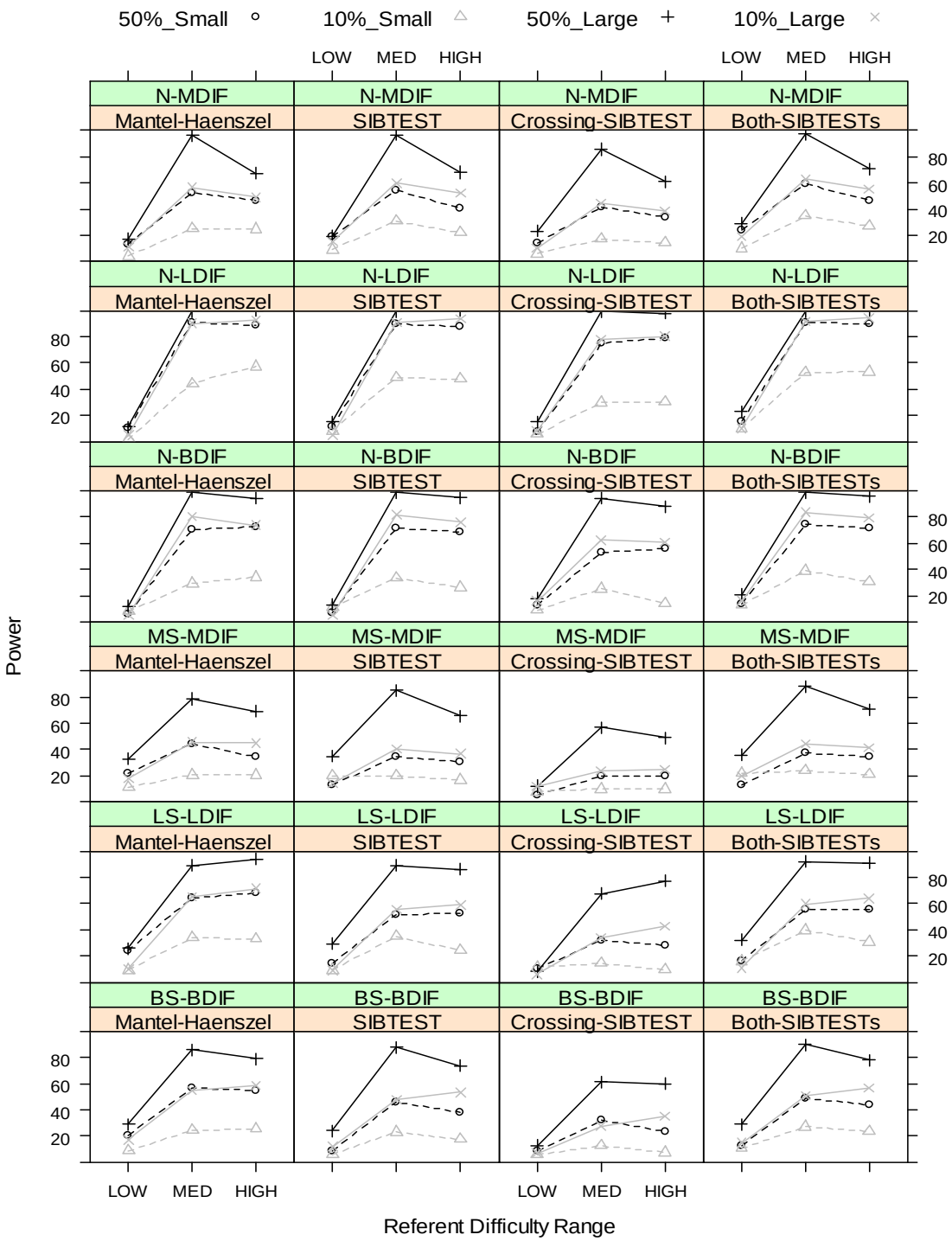


Figure 19. Power rates for 10% DIF conditions for all DIF items in the Low, Medium, and High referent difficulty ranges.

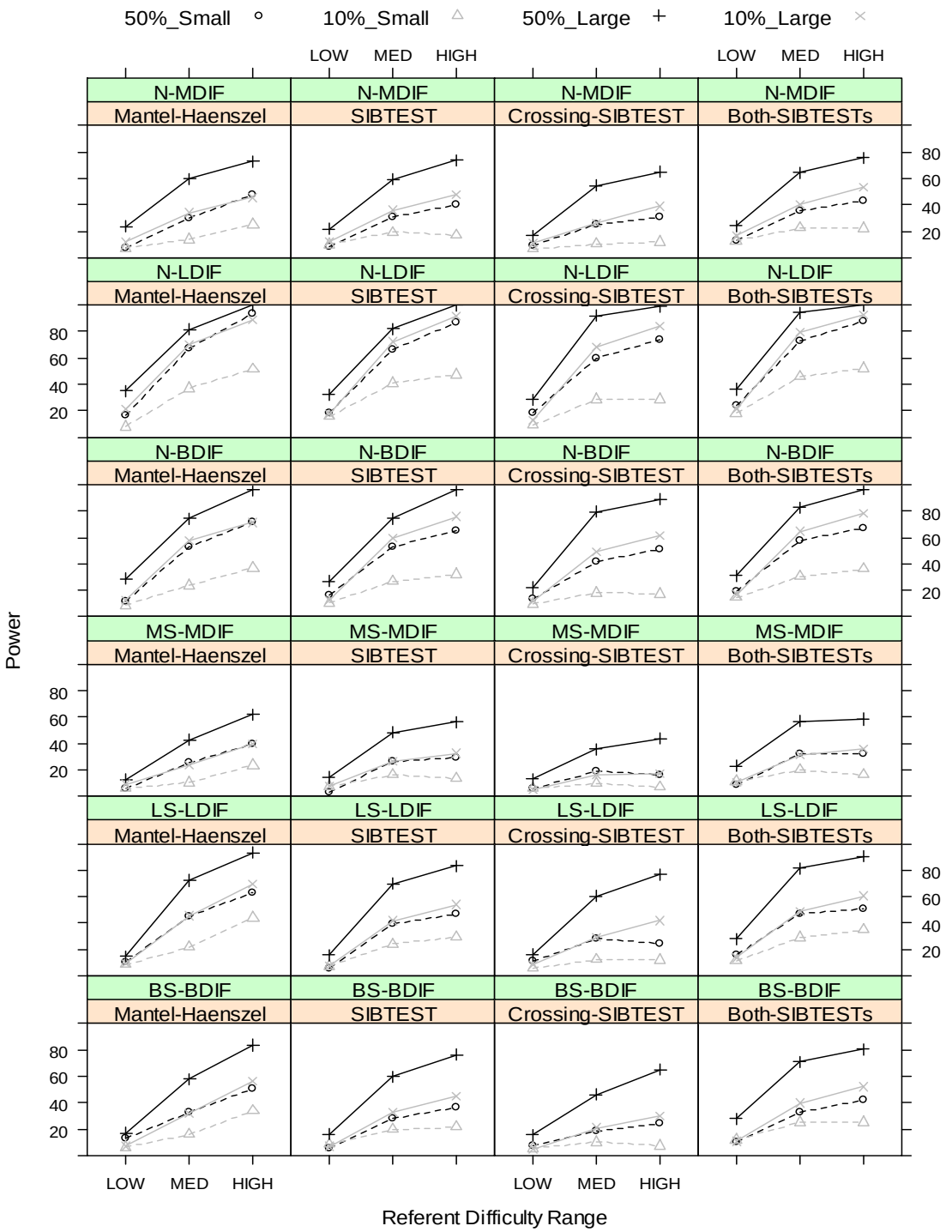


Figure 20. Power rates for 20% DIF conditions for all DIF items in the Low, Medium, and High referent difficulty ranges.

The power rates for the Low difficulty range were much lower than those in the Medium range (13.4%, 13.3%, 11.0%, and 18.1%, respectively).

Across all conditions for all analyses, increases in power were generally observed both as sample size increased and as the percent of focal simulees increased with 10% focal simulee/Large-sample results being very similar to those for 50% focal simulee/Small-sample. Additionally, all analyses tended to have higher power in normal distribution conditions than in the comparable skewed distribution and with the inclusion of more items with Large DIF modification.

Total power rates within conditions. Again, MH, SIB and Both-SIB performed very similarly, regardless of percent DIF, sample size, simulee distribution or severity of the DIF modification with X-SIB tending to have the lowest power rates. Strikingly, all four analyses performed much differently with the 6 DIF items in the 10% DIF conditions than with the 12 DIF items in the 20% DIF conditions. With 10% DIF, they all generally had the highest power rates in the Medium range, with somewhat lower rates in the High range and much lower rate in the Low range. With 20% DIF, while the power rates in the Low range remained low, those in the Medium were generally lower and those in the High range generally higher than those found with 10% DIF. This resulted in graphs for 20% DIF tending to be monotonically increasing as referent item difficulty range increased.

Power rates for Uniform DIF items. To compute Total power rates for U-DIF, the number of U-DIF modified items in each of the ranges was used as a denominator. In 20% DIF cases, where all U-DIF modified items were used, one item (number 9) had a referent difficulty parameter less than -1.00, four items (numbers 29, 39, 49, and 59) had referent difficulty parameters greater than -1.00 but less than 1.00, and one item (number 19)

had a referent difficulty parameters greater than 1.00. For 10% DIF cases, items 9, 29, and 49 were not modified to exhibit DIF. This modification design resulted in the referent difficulty power rates for Low, Medium, and High ranges for 20% DIF conditions being computed with a denominator of 1, 4, and 1. For 10% DIF conditions, where there were no items in the Low range, power rates for the Medium and High ranges were computed with a denominator of 2 and 1, respectively. Figures 21 and 22 present the U-DIF power rates for the analyses within each of the referent difficulty ranges for 10 and 20 percent DIF conditions, respectively. The associated tabulated results are available in Table I3a and Table I3b within Appendix I.

Total power rate across conditions. When Total power is computed across all conditions for each of the analyses, again none of them performed well in all difficulty ranges. The overall power rates (aggregated across both sample size and percent DIF conditions) for uniform DIF items were consistently highest in the Medium difficulty range and again lowest in the Low difficulty range for all four analyses. The overall power rates in the Medium range for MH, SIB, X-SIB, and Both-SIB were 55.6%, 56.8%, 41.8%, and 59.8%, respectively. Power rates for the one U-DIF item (number 19) in the High range were all over 10 percentage points lower than that found in the Medium range (41.2%, 36.9%, 27.8%, and 41.8%, respectively). U-DIF power rates for the one item (number 9) in the Low difficulty range in 20% DIF conditions were consistently higher than those when both U-DIF and NU-DIF items were included, however this range was still the lowest of the three difficulty-ranges (22.2%, 28.4%, 19.5%, and 27.5%, respectively for the four analyses).

Unlike the results found when power rates were computed for both uniform and non-uniform DIF items together, with power computed across all conditions for uniform

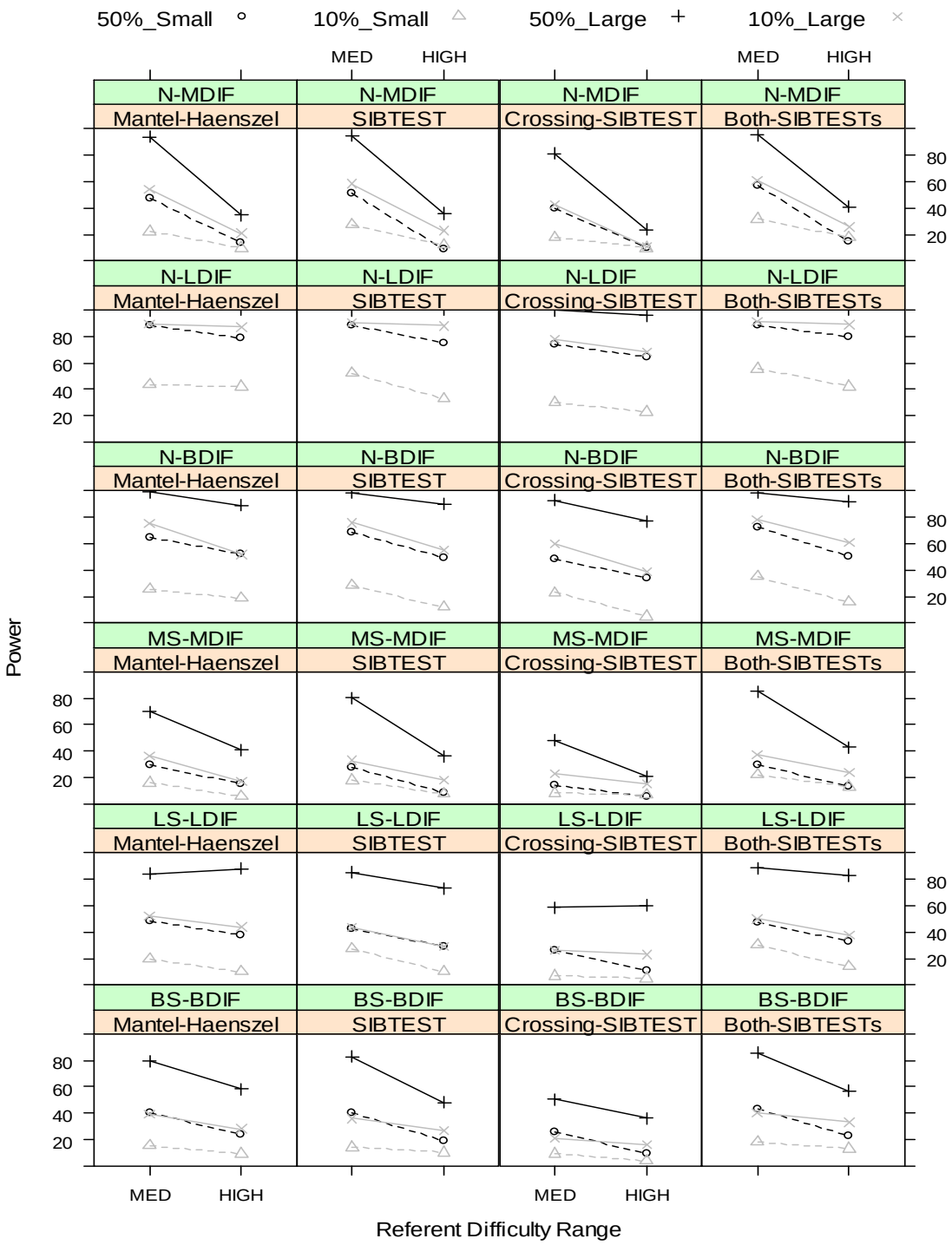


Figure 21. Power rates for 10% DIF conditions for Uniform DIF items in the Low, Medium, and High referent difficulty ranges.

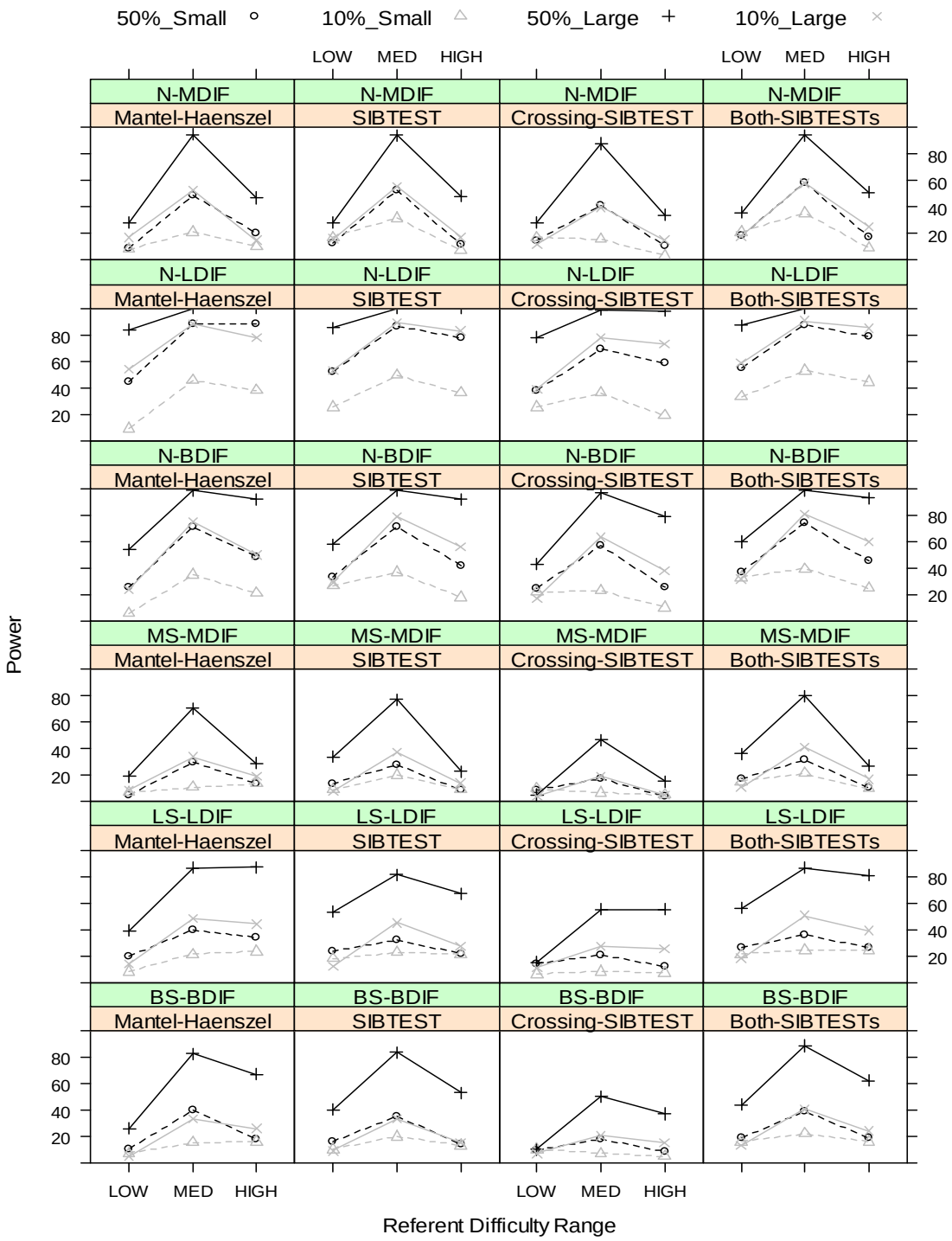


Figure 22. Power rates for 20% DIF conditions for Uniform DIF items in the Low, Medium, and High referent difficulty ranges.

DIF items alone, rates in 10% DIF conditions were very similar to those in the 20% DIF conditions. However, as was seen previously, power tended to increase as sample size or percentage of focal simulees increased, as well as with focal simulee ability drawn from a normal distribution rather than either of the skewed distributions. With U-DIF items, once again regardless of difficulty range, 50% focal simulee/Large-sample tended to have the highest power with 10% focal simulee/Small-sample having the lowest power. The other two combinations (50% focal with Small sample and 10% focal with Large sample) generally had power rates that were very similar.

Total power rates within conditions. When Total power rates for Best matching subtest were used to compare the results for the analyses aggregated across difficulty ranges, MH, SIB, and Both-SIB performed very similarly with X-SIB having slightly lower power rates. With Total power rates for U-DIF items, all four analyses performed very similarly within each of the difficulty ranges when focal group simulee ability was drawn from a normal distribution. When focal ability was drawn from a skewed distribution, however, the decrease in X-SIB's power was generally larger in each difficulty range than the other analyses. X-SIB's decreases were especially notable with 50% focal simulees in combination with Large sample. With normal distribution, all four analyses generally had extremely high power in the Medium range in Large sample with 50% focal simulees, regardless of the percentage of DIF items or of the severity of U-DIF modelled. In these conditions, they also tended to have relatively high power in the High difficulty range as long as at least some of the items were modelled with Large DIF (i.e., N-LDIF and N-BDIF).

Interestingly at the condition level, while all four analyses tended to have their highest power rate in the Medium difficulty range regardless of sample size, percentage of

DIF, percentage of focal simulees, or simulee distributions, there was an exception in one condition. Both MH and X-SIB had as high or higher Total power rates for the one item (number 19) in the High range than the Medium range in Large sample, LS-LDIF, with 50% focal simulees with both 10% DIF and 20% DIF. For this High range item, all four analyses had notably higher power with Large DIF than with Moderate DIF. When both DIF modifications were included within the data, a general increase in power over that with Moderate DIF was observed.

Power rates for Non-uniform DIF items. To compute Total power rates for NU-DIF, the number of NU-DIF modified items in each of the ranges was used as a denominator. In 20% DIF cases, where all NU-DIF modified items were used, two items (numbers 10 and 40) had referent difficulty parameters less than -1.00, three items (numbers 20, 30, and 50) had referent difficulty parameters greater than -1.00 but less than 1.00, and one item (number 60) had referent difficulty parameters greater than 1.00. For 10% DIF cases, items 10, 30, and 50 were not modified to exhibit DIF. This modification design resulted in the referent difficulty power rates for Low, Medium, and High ranges for 20% DIF conditions being computed with a denominator of 2, 3, and 1, respectively, and for 10% DIF conditions with a denominator of 1 for all ranges.

For NU-DIF modification, in addition to a change in difficulty from the referent item difficulty parameter, there was also a change in both discrimination and pseudo-guessing parameters. These changes were performed at two levels, demarcated as Low and High. In the Low NU-DIF modification (used with items 10, 30, and 50), the pseudo-guessing parameter was lower and the discrimination parameter was higher for the focal group than for the reference group. In the High NU-DIF modification (used with items 20,

40, and 60), the pseudo-guessing parameter was set higher and the discrimination parameter lower for the focal group. The 10% DIF conditions included only NU-DIF items that were simulated with the High modifications while the 20% DIF conditions included an equal percentage of both modifications. Figures 23 and 24 present the Total NU-DIF power rates for the analyses within each of the referent difficulty ranges for 10% and 20% DIF conditions, respectively. The associated tabulated results are available in Table I4a and Table I4b within Appendix I.

Total power rate across conditions. When Total power was computed across all conditions for each of the analyses, again, none of them performed well for all difficulty ranges. The overall power rates (aggregated across both sample size and percent DIF conditions) for NU-DIF items were consistently highest for the one item (number 60) in the High difficulty range and the lowest for the items in the Low difficulty range for all four analyses. The overall Total power rates for the one NU-DIF item in the High range for MH, SIB, X-SIB, and Both-SIB were 80.3%, 72.8%, 57.9%, and 75.3%, respectively. The Total NU-DIF power rates for the Medium difficulty range were much lower than those in the High range (54.3%, 51.7%, 43.2%, and 58.2%, respectively). The Total power rates for the NU-DIF items in the Low difficulty range were very low (15.5%, 16.5%, 13.9%, and 22.0%, respectively).

Unlike the results found for U-DIF, power rates in 10% DIF conditions were generally very different from those in the 20% DIF conditions in the Low and Medium ranges with the introduction of items modelled with the Low NU-DIF modification. However, as was seen previously with NU-DIF, power again tended to increase as sample size or percentage of focal simulees increased as well with focal simulee ability drawn from a

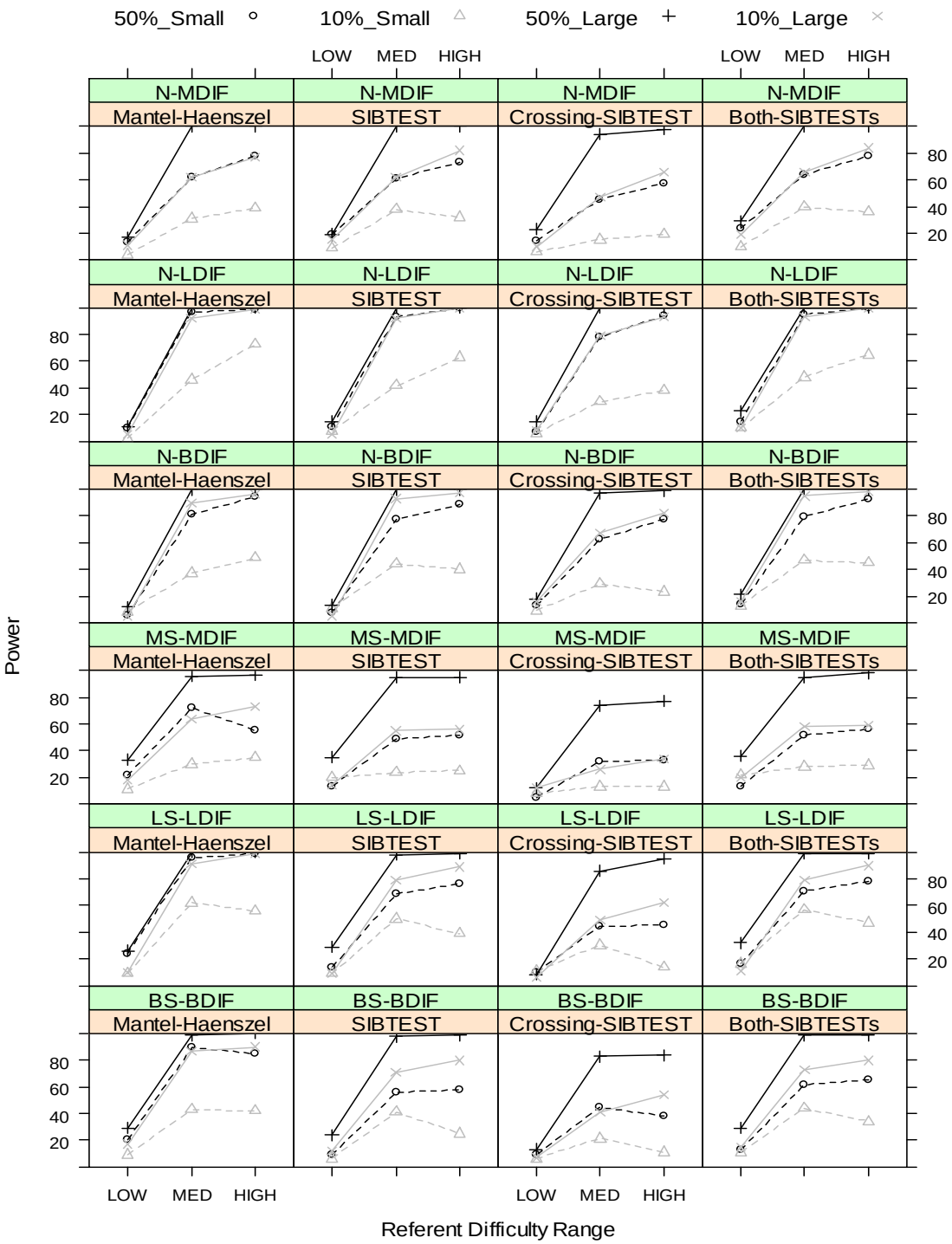


Figure 23. Power rates for 10% DIF conditions for Non-uniform DIF items in the Low, Medium, and High referent difficulty ranges.

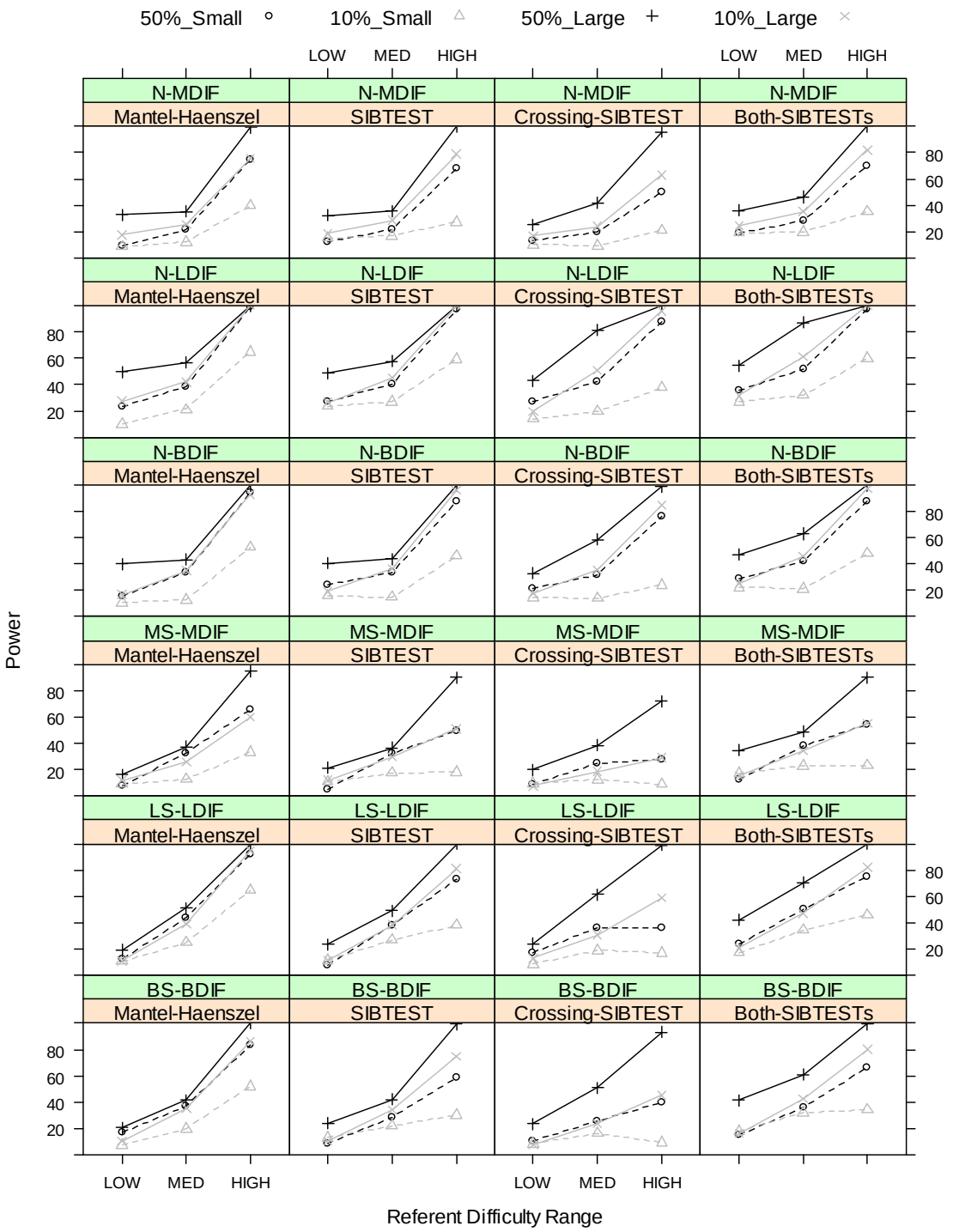


Figure 24. Power rates for 20% DIF conditions for Non-uniform DIF items in the Low, Medium, and High referent difficulty ranges.

normal distribution rather than either of the skewed distributions. With NU-DIF items, once again regardless of difficulty range, 50% focal simulee/Large-sample conditions tended to have the highest power with 10% focal simulee/Small-sample having the lowest power. The other two combinations (50% focal with Small sample and 10% focal with Large sample) again generally had power rates that were very similar.

Total power rates within conditions. With Total power rates for NU-DIF items, MH and SIB performed very similarly within each of the difficulty ranges regardless of focal ability distribution, percentage of DIF, or severity of DIF modification. X-SIB and Both-SIB also generally had power rates very similar to MH's and SIB's in 10% DIF conditions. With 10% DIF, especially with 50% focal simulees with Large sample, all four analyses generally had high power for the one item (number 20) in the Medium difficulty range in addition to the high power observed for the one item in the High difficulty range. In these conditions, X-SIB tended to have much lower power with skewed focal distributions than with normal distributions.

With the addition of the three items modelled with the Low NU-DIF modification in 20% DIF conditions, there were marked differences in power rates in those item's difficulty ranges (Low and Medium). With the addition of another item (number 10) in the Low difficulty range, the power rates for all four analyses generally increased with normal focal distribution but remained very low with skewed distributions. With the addition of two more Low NU-DIF modification items in the Medium difficulty range, regardless of severity of DIF modification, percentage of focal simulee, or sample size, all four analyses were observed to have lower power rates. In multiple instances, these decreases from the levels with 10% DIF were large. There were 19 instances where the decrease for an analysis was

greater than 50 percentage points with the maximum decrease found for MH in Large sample, 50% focal simulee, N-MDIF (64.7 percentage points). X-SIB tended to have the lowest decrease in power with the addition of the two items. Additionally, contrary to the general trend of X-SIB having the lowest power among the procedures, with either Large or Both DIF modifications, X-SIB tended to have the higher power than either MH or SIB in the Medium range with 20% DIF.

CHAPTER 5

DISCUSSION AND CONCLUSIONS

This research adds to the literature for four different statistical analysis programs. Importantly, it adds information about how each of these programs functions when DIF items are modelled to advantage a lower ability focal group over a higher ability reference group. An example where this might be the case is when students just learning English or with a disability are afforded test accommodations.

For ATFIND, the primary purpose was to examine its usefulness as a valid subtest selection tool. However, it also explores the influence of DIF items, item difficulty, and presence of multiple examinee populations with different ability distributions on the selection of the AT and PT Lists. For MH, SIB, and X-SIB, the primary purpose was to examine the ATFIND selected PT List's usefulness as a valid matching subtest. To compare the PT List's performance for the three DIF analyses, a static set of both DIF and non-DIF items were used in the suspect item-set and a set of DIF-free items were used in the matching subtest. Using this static suspect item-set also allowed for the comparison of the three analyses on the influence of DIF type, referent item difficulty, and presence of multiple examinee populations with different ability distributions on both their power and Type I error rates. Finally, since both SIB and X-SIB were included in the study, the influence of their joint use was examined.

ATFIND Analysis

Usefulness as a valid matching subtest selection tool. Regardless of simulee distribution, variability of simulee distributions, percentage of focal simulees, sample size, percent of items modelled to exhibit DIF, the type of DIF modeled, and referent item

difficulty, approximately 60% of items were selected by ATFIND for the PT List and 40% were selected for the AT List. While there were some slight variations in selection rates noted for these variables, ATFIND's general insensitivity to DIF modelled via unidimensional 3-PL item parameter change showed it to be unusable as a valid matching subtest selection tool for DIF analyses. It resulted in matching subtests that had approximately the same percentage DIF contamination as was in the original dataset (10% for 10% DIF conditions and 20% for the 20% DIF conditions).

ATFIND was developed to divide the items in a test into two sets that measure different composite abilities and uses two analysis software procedures HCA/CCPROX (Roussos, Stout, & Marden, 1998) and DETECT (Kim, 1994; Zhang & Stout, 1999b) to accomplish this (Froelich, & Habing, 2008). Therefore, it could be that it is susceptible to the same issues as found with its base procedures. In their comparative study of test dimensionality assessment procedures, van Abswoude, van der Ark, and Sijtsma (2004) found that the performance of both DETECT and HCA/CCPROX suffered as the correlation between latent abilities increased past 0.60.

To examine how strongly correlated the factors created by simulee score on DIF modified items were to their score on non-DIF items, a small preliminary study was conducted. For this study, the first 20 datasets within the 20% DIF, Large sample, 50% focal simulees in the N-LDIF condition were analyzed using SPSS (1989-2010). For each simulee, two subscores were computed; one for all DIF modified items and the other for all non-DIF items. Pearson correlations were then computed for the subscores. The factors in all twenty of the datasets had correlations at or above .70 (Mean=0.725, SD=0.017). These high

correlations indicate that this might be one cause of ATFIND's inability to select the DIF items for the AT List.

Since this is a simulation study, these results are also dependent on the DIF modeling method used and, as is often quoted, "Essentially, all models are wrong, but some are useful" (Box & Draper, 1987, p. 424). It could be that, while modeling DIF via unidimensional 3-PL item parameter non-invariance between groups is often useful, a different model may provide different results. Previous researchers had indicated that DIF might be (e.g., Lord, 1980; Ironson, Homan, Willis, & Signer, 1984) or always is (e.g., Shealy and Stout, 1993a; Ackerman, 1992) caused by the impact of a secondary dimension upon the examinee's response. As Ackerman explains, within this view, DIF is caused by differences between groups in the distributions on this secondary dimension. Multiple researchers have used the multidimensional item response theory put forward by Reckase and McKinley (Reckase, 1985; Reckase, 1996; Reckase & McKinley, 1991) to model DIF (e.g., Ackerman, & Evans, 1994; Gierl, Gotzmann, & Boughton, 2004; Mazor, Hambleton, & Clauser, 1998; Nandakumar, 1993; Stout, Li, Nandakumar, & Bolt, 1997). An example where multidimensionality might especially be the case is items affected by the various testing accommodations (e.g., calculators on a mathematics test, read-aloud on a reading test) since they might in fact be an additional dimension afforded to only a portion of the examinee population. There is some indication that DIF might not always cause the lack of essential unidimensionality within a dataset (Wu, 2009). It would be interesting to repeat this study using both uniform and non-uniform DIF modelled with multidimensional item response theory.

In 2009, Wu explicitly studied whether DIF would always cause a lack of essential unidimensionality as determined by DIMTEST. In addition to modeling both uniform and non-uniform DIF unidimensionally as was done in this study, she also simulated both types of DIF multidimensionally. Additionally, she varied not only the proportion of DIF items in the dataset (up to 60%) but also the proportion of uniform DIF items in relation to non-uniform DIF items. For unidimensionally modelled DIF with moderate DIF (an area between ICCs of .4), regardless of the percentage of DIF items included in the dataset or the proportion of uniform DIF items, DIMTEST's power remained extremely low (around 5%). When DIF was modelled unidimensionally with large DIF (an area between ICCs of .8), DIMTEST's power increased as the percentage of DIF items in the dataset, the proportion of uniform DIF items, the number of items in the test, the sample size, and the percentage of focal simulees increased. When DIF was modelled multidimensionally, in addition to the factors found to cause an increase in DIMTEST's power with large DIF in unidimensional data, increases were also observed as the angle between the traits increased or the correlation between the traits decreased regardless of severity of DIF induced. While these results are specifically for DIMTEST, both the integral nature of ATFIND within the DIMTEST procedure, and the fact that they are both based on the same conditional covariance-based theory (Froelich & Habing, 2008) might lead to the conjecture that at least some the same factors that Wu found to impact DIMTEST's power might also impact ATFIND's. Teasing the results for the two procedures apart would perhaps lead to more useful applications for ATFIND than were found in the current study.

ATFIND's selection process was observed to be relatively robust to multiple previously unexplored variables. The percentage of DIF items, sample size, simulee ability

distribution, percentage of focal simulees, and severity of DIF had little impact on the selection of items for the PT List. Two variables (DIF type and referent item difficulty), however, were shown to have very slight influence on the items selected for the PT List. There was a slight propensity for ATFIND to exclude U-DIF items and items in the Medium referent difficulty range from the PT List and place them instead in the AT List.

To examine how the results obtained here might compare to those with the DIF modifications using a similar process to that proposed by Nandakumar (1994), a brief auxiliary study was performed to examine the results of selecting a matching subtest based on factor analysis. The first version of DIMTEST used factor analysis to identify its assessment subtests and partitioning subtest. While the original analysis was performed using Principal Components, Principal Axis factor (PAF) analysis was used based on research showing that it is more sensitive to additional factors when more than one factor is expected (Crawford, et al., 2010). For this study, the first 20 datasets within the 20% DIF, Large sample, 50% focal simulees in the N-LDIF condition used in the correlation study above were again analyzed using SPSS (1989-2010). After PAF extraction using correlation, VARIMAX rotation was performed and any item that weighted at least .300 on the second factor was identified as a suspect item and all other items were identified as a matching subtest item.

While DIF purity rates for the matching subtest were generally higher than the 80% DIF purity for the whole test, and individual runs achieved purity rates above 90%, the average rate was only slightly higher (83.4%) than that of the whole test. While this study is only preliminary, it would seem to indicate that even with optimal conditions (Large sample size, Large DIF, equal percent of both subgroups, and normal distribution) focal-

advantaging DIF modification is extremely difficult for different dimensionality analyses to identify.

Influences on the selection of subtests. While ATFIND generally selected approximately the same proportion of both DIF items types as non-DIF items, across conditions, uniform DIF items were selected for the AT List (and excluded from the PT List) at a slightly higher rate than both non-uniform DIF and non-DIF items. Additionally, ATFIND's uniform DIF AT List selection rate generally increased as sample size increased where its non-uniform DIF and non-DIF AT List selection rate remained relatively stable. This slight propensity for selecting uniform DIF items for the AT List, however, might be a consideration with a much larger sample in the analysis, a short test, or with a higher percentage of uniform DIF items in the dataset. It would be appropriate to explore influence of these additional variables on ATFIND's selection of uniform DIF items.

Skewness, percentage in focal group, and degree of DIF. With the exception of both uniform and non-uniform DIF items in the Large sample, 10% DIF, 50% focal group, Large DIF condition, ATFIND's subtest selection rate was virtually the same for normally distributed simulees as it was for the other three distribution conditions. In this one condition, ATFIND's AT List selection rate was over 10 percentage points higher in the normal distribution condition than in the Large Skewed condition. In all other conditions, including the rest with that most severe distribution, the ability distribution of the simulees seemed to have little or no effect on ATFIND's selection rate.

Similarly, the percentage of focal simulees seemed to have little impact on ATFIND's selection rates. Only two conditions exhibited a selection rate for 10% focal simulees greater than 10 percentage points different from that of the 50% focal simulee

condition. Both of these were for uniform DIF with the Large sample but were in opposite directions.

The degree of DIF induced had even less impact on ATFIND's selection rates. The only difference greater than 10 percentage points occurred in Small sample, 10% DIF, Moderately Skewed distribution with 10% focal simulees. In this condition, ATFIND selected a much lower percentage of non-uniform DIF items with Moderate DIF for the AT List than either those with a Large Skewed distribution and Large DIF or Both Skewed distributions and Both DIF modifications.

Taken together, these results present a picture of a software procedure that is relatively robust to some of the most influential unidimensional conditions that might occur in real data.

Item difficulty. Regardless of item type, percentage of DIF items, simulee distribution, percentage of focal simulees in the dataset, or sample size, ATFIND showed a slight tendency to select items for the AT List in the Medium difficulty range ($-1 < b \leq 1$) more often than those in either the Low or the High difficulty ranges. Since this tended to increase as sample size increased, perhaps the difference would continue to increase as the sample size increased past the 750 used for Large sample in the current study's ATFIND analyses.

An examination of ATFIND's selection rates within difficulty ranges in 0% DIF conditions provides a hint to the reason behind it choosing items in the Medium range most frequently. In these conditions, when the simulee ability was drawn from either the normal or the Moderately skewed distributions, or when 10% focal simulees were included with either of the other two distributions, ATFIND consistently selected more items in the

Medium range. It is in these conditions that most of the simulees will have abilities that have been drawn from the associated Medium ability range. However, when 50% focal simulees were used in association with the Large skew and group mean (-1.5), ATFIND selected the most items in the Low difficulty range. Combined, these trends seem to indicate that ATFIND's selection of items is influenced at least to a degree by the ability distributions of the examinees.

DIF Analysis

Influence of using the PT List as a matching subtest. This section will be restricted to the Total Type I error and Total power rates, computed with all of the non-DIF items or with all of the DIF items used as a denominator, respectively. These rates are most appropriate in the examination of how the PT List matching subtest functions in conjunction with each of the DIF analyses.

Total Type I error. When the PT List was used as the matching subtest, and the AT List was used as the suspect item list, both SIB and X-SIB generally had Type I error rates no greater than the nominal 5%. The only exceptions to this were with a skewed focal ability distribution with 50% focal simulees. When the results of these two analyses were combined, as a researcher might do in practice, the combined Both-SIB generally maintained error rates of less than 5% with simulees with normal distributions, but generally exceeded the nominal rate with focal simulees with skewed distributions. The SIBTEST analyses tended to have a higher Type I error rate with 10% focal simulees when the Small sample was used and with 50% focal simulees with the Large sample. The impact of sample size on SIB and X-SIB was very different however. SIB consistently had lower Type I error rates with the Large sample

size where X-SIB had a higher Type I error rate with the Large sample in approximately half of the conditions.

When the PT List was used as the matching subtest, and all items were inspected for DIF as is MH's default, it exhibited Type I error rates greater than the 5% nominal rate in approximately half of the 28 conditions with normal distributions and almost all of the skewed distribution conditions. Across distributions, it tended to have lower Type I error rates with the Small sample in combination with 10% focal simulees than with Large sample especially in combination with 50% focal simulees.

These findings generally support studies (e.g., Bolt & Gierl, 2006; Finch & French, 2007; Jaing & Stout, 1998; Li, 1995; Li & Stout, 1996; Narayanan & Swaminathan, 1996; Roussos & Stout, 1996b) that have shown that MH's, SIB's, and X-SIB's Type I error rates tend to increase as sample size increases and/or as the difference in the groups' mean ability increases. While the Type I error rates for MH were very similar to those found in the current study, those for SIB and X-SIB were not. Previously reported Type I error rates for SIB's and X-SIB's were either similar to or slightly higher than MH's; however, their Total Type I error rates in the current study were much lower than MH's regardless of matching subtest.

While MH consistently had a higher Total Type I error rate than SIB, X-SIB, and Both-SIB, its overall error rate (.07) with the PT List matching subtest was only slightly higher than the nominal .05 and in many conditions, especially with normal distributions, was well under .05. Additionally, MH generally maintain an error rate of less than 10% except in conditions with a Large sample combined with Large skewed distributions and 50% focal simulees. Based only on the Total Type I error rate computation, one might

conclude that SIBTEST procedures are superior to MH, but Type I error is only part of the story.

Some may argue that statistical analyses should only be used in situations where their Type I error rates are well-maintained (Newton & Rudestam, 1999). With DIF analyses, however, the statistical identification of an item as containing DIF is only the first step in the determination that the item is biased, and should be modified or discarded. The second step entails the item being reviewed by content experts who may determine that there in fact is nothing wrong with the item (Camilli & Shepard, 1994). Therefore, the threat of over-identifying items as having DIF is not as egregious as large Type I error rates might be within other types of analyses. Because of the second level of inspection generally embodied within DIF studies, power for all analyses for all conditions will be discussed regardless of the error rates for those conditions.

Total Power. When the PT List was used as the matching subtest, and power was computed using the total number of DIF items in the dataset, none of the analyses was able to identify the focal advantaging DIF items consistently. However, as expected, MH far outperformed any of the SIBTEST procedures. For MH, Total power rates with the PT List matching subtest averaged 41.9% (SD 19.9) where none of the SIBTEST procedures, including Both-SIB, had an average Total power of over 20%. With the PT matching subtest, while the power rates for MH were generally similar to but slightly lower than those found in the literature (e.g., Bolt & Gierl, 2006; Clauser, et al., 1993; Gierl, Jodoin, Ackerman, 2000; Li & Stout, 1996; and Narayanan & Swaminathan, 1996), the Total power rates for the SIBTEST analyses were much lower than those previously reported. As an example, one previous study, which compared the performance of SIBTEST to those of

MH and logistic regression, characterized SIBTEST as "clearly the most powerful DIF detection method" (Gierl, Jodoin, Ackerman, p. 16). This was definitely not the case for the three SIBTEST analyses here as indicated by Total power rate when the PT matching subtest was used.

While Total power rates within conditions with the use of the PT List as the matching subtest consistently improved as sample size increased, the constant low power for the SIBTEST procedures and the general low power for the MH provides strong evidence that this list should not be used as a matching subtest in practice, regardless of the generally low Type I error rates obtained.

Impact of DIF contamination in the matching subtest. The Type I error and power rates discussed in this section were computed using the number of non-DIF or DIF items, respectively, that each analysis actually examined for DIF. In the context of this study, using the rates that take into consideration the number of items an analysis is given to examine are most appropriate both in the examination of how DIF contamination in the matching subtest influences their performance and in the comparison across the different analyses. As explained above, for MH, only Total power and Total Type I error rate were computed for each of the subtests. However, for the three SIBTEST procedures, the denominator for the rates with the Best matching subtest is the number of that type of item in the 20 item static suspect list, and the denominator for the rates with the PT matching subtest is the number of that type of item selected for the AT List.

Analyzed Type I error. Unlike when the Total Type I error rate was examined, with the Analyzed Type I error rate both SIB and X-SIB generally had Type I error rates slightly higher than the 5% nominal rate regardless of matching subtest. These rates were much

more consistent with MH's Total Type I error rates, where Both-SIB not only generally had the highest Analyzed Type I error rate, but consistently had a rate higher than 5% with the PT matching subtest. While the Analyzed Type I error rates for all four analyses tended to increase when the PT matching subtest was used as compared to the Best subtest, regardless of the presence of DIF items in the matching subtest, the increase was amplified as the percentage DIF items in the dataset increased. The highest Analyzed Type I error rates for all four analyzes were observed with the PT matching subtest in 20% DIF conditions with 50% focal simulees. The amplification of Analyzed Type I error rate with the increase in the percentage of DIF items was not observed when the Best matching subtest was used. These results provide evidence that DIF contamination in the matching subtest causes items to be identified as containing DIF, when in fact they do not, and support the findings of Lopez (2012) that Type I error rates generally increase for X-SIB with the presence of differential test functioning within the dataset, especially as the sample size increases. The current findings are an indication that using either all items in the assessment or all other items in the assessment as the matching subtest, as is generally used for MH and SIBTEST procedures, respectively, might best be reconsidered, since they would most likely contain DIF items if there were any DIF items in the dataset. Perhaps, however, there are other methodologies available (or being developed) that might have better success with identifying suspect items for DIF analysis. Based on the findings within this study, these methodological options should be explored.

With normal distributions, the Analyzed Type I error rates found in this study for MH, SIB, and X-SIB are very similar to those found in the literature (e.g., Clauser, et al., 1993; Finch & French, 2007; Li, 1995; Li & Stout, 1996; and Narayanan & Swaminathan,

1996). Also similar to the literature, Analyzed Type I error rates for all of the analyses tended to increase as the difference between abilities increased. In the literature, however, the maximum ability departure was generally 1.0 SD, where in this study (based on distributions found within real data), this was increased to a maximum of 1.5 SDs. It was in the conditions with the largest differences in abilities that the highest Analyzed Type I error rates were observed. In this study, not only was the difference between mean group ability increased but, because it was observed in the real data for groups that are expected to take advantage of testing accommodations, the ability distributions of the groups was also skewed (.5 and 1.0 for Moderate and Large, respectively). The current study found that as the combination of difference between abilities and skewness in ability increased, Type I error increased for all of the analyses regardless of matching subtest but was generally amplified with the presence of DIF items in the PT matching subtest. This would seem to be an indication that the search for a reliable statistical purifying process for use with these analyses remains a worthy goal.

The ability distributions of the referent group and both focal groups modelled within the current study's data are based on those found in a real state content assessment for regular education, English language learners, and students with disabilities, respectively. Because these students continue to be groups of interest, especially as new accommodations are proposed to facilitate their ability to demonstrate what they know and can do, it is important to understand how those distributions impact the identification of DIF items by each of the analyses. For the study of DIF caused by the use of accommodations, perhaps a more appropriate reference group would be a separate sample from the same subgroup that did not access the accommodation. Then at least the ability distributions would be similar. A

simulation study with data modelled for both focal and referent groups with the same moderate and/or large positive skewed ability distributions with one group having DIF items similar to those in this study might be able to inform a better selection of referent group for accommodation DIF studies. It could be that since the ability distributions of both groups in this future study would similar, that the Type I error rates found would be more similar to those found with normal ability distributions in this study.

Analyzed power. The large Analyzed Type I error rates observed for all four procedures, especially with mismatch of abilities between groups, might be a contraindication for the examination of power for the procedures. However, as explained above, since DIF identification is only the first step in the DIF study process, Type I error is much less of a threat than Type II error which is informed by power rates. With Analyzed power rates, all of the procedures had rates for the PT matching subtest that were very similar to, but generally lower than, those with the Best matching subtest.

Regardless of matching subtest, MH, SIB, and Both-SIB performed very similarly, with Both-SIB generally obtaining the highest power followed by MH, then SIB. Again X-SIB generally had the lowest power. MH generally achieved higher power rates with skewed simulee distributions and with 10% focal simulees, where SIB generally had the highest rates in normal distribution conditions and with 50% focal simulees. While none of the procedures consistently achieved even an acceptable level of power (e.g., > 70%), across all conditions even with the Best matching subtest, they were more successful in some conditions than in others. Unlike Total power rates, with Analyzed power rates, not only were MH's rates similar to those in the literature but SIB's and X-SIB's were also (e.g., Bolt & Stout, 2006; Clauser, et al., 1993; Finch & French, 2007; Li, 1995; Lopez, 2012; Narayanan

& Swaminathan, 1996; and Wu, 2009). While as might be expected, there were differences in power rates between the results found in the current study which focused on focal-advantaging DIF and those in previous studies which focused on referent-advantaging DIF, the factors that caused those differences tended to be the same.

In the current study, regardless of matching subtest, all four analyses generally had higher power with Large sample than with Small, with normal distributions for both groups than with skewed focal distributions, with Large DIF modification than Moderate, with 10% DIF items than with 20% DIF, and with 50% focal simulees than 10%. Increases were most evident in skewed ability conditions or if only Moderate DIF was modelled with focal ability drawn from a normal distribution. In previous studies, power was found to increase as sample size increased and decrease as the difference between focal and referent ability increased (Bolt & Stout, 2006; Finch & French, 2007; Gierl, Jodoin, & Ackerman, 2000; Li, 1995; Lopez, 2012; and Narayanan & Swaminathan, 1996). Also, they found that increases in the magnitude of DIF induced within sample sizes improved power (Gierl, Jodoin, & Ackerman, 2000; Li, 1995; Lopez, 2012; Narayanan & Swaminathan, 1996; and Wu, 2009), and some found that the percentage of focal to referent simulees seemed to impact power as well. However, because both sample size and percentage of focal simulees in previous studies varied, it is very difficult to tease apart the impact each has on the power rates achieved. Narayana and Swaminathan (1994) did find that both SIB and MH had higher power as the percentage of focal simulees increased as compared to a static referent group size. The current study, however, systematically varied the percentage of focal simulees within a sample size and showed that using an equal percentage of each group increases power over and above increases in sample size.

For the impact of the percentage of DIF items in the dataset, the comparison between the results found in the current study and that of previous studies is somewhat muddier. Some previous studies found that there was little or no difference between power rates for MH and/or X-SIB as the percentage of DIF items increased in the matching subtest (Finch & French, 2007; and Narayanan & Swaminathan, 1996). These findings are inconsistent with the current study that found that not only does the presence of any DIF items in the matching subtest negatively influence power, but also that the number of DIF items in the matching subtest had a negative influence. Other studies however, found that the impact of percentage of DIF contamination was additionally influenced by test length (Clauser, et al., 1993) or by the severity of DIF modification (Wu, 2009). While test length was not a variable of interest in the current study, the inclusion of items with Large DIF in the dataset was found to partially mitigate the effect of DIF contamination in the matching subtest. This was evidenced by the comparison of decrease in power for conditions with Large DIF to those of either Moderate or Both DIF. When the groups were of equal size, regardless of distribution, all four analyses consistently had less of a loss of power with Large DIF than with either of the other two modifications when the number of DIF items increased. While they all also generally had an associated increase in Type I error in Large DIF conditions, those increases tended to be quite small.

While power increased for both matching subtests with larger sample, higher percentage of focal simulees, normal distribution for both groups, and more severe DIF induced, power with the PT matching subtest remained lower than that with the Best matching subtest. The highest power rates for all four analyses were generally found with 10% DIF, Large sample, 50% focal simulee, and N-LDIF with the Best matching subtest.

Interestingly, this was the one condition where the combined Both-SIB performed slightly better with the 20% DIF than with 10% DIF indicating that having more DIF items to identify allowed the two individual SIBTEST analyses more opportunity to select different items for this condition. In this condition, where both the ability distributions and the percentages of simulees are the same in both groups, it is completely arbitrary as to which group might be considered the focal and referent. Therefore, it is this condition in the current study, which focuses on focal-advantaging DIF that is most comparable to the referent-advantaging DIF studies within the literature.

The generally higher power observed for all four analyses with the Best matching subtest than with the PT subtest is another indication that the presence of DIF items in the matching subtest negatively affects the performance of each of these analyses. This supports previous findings of differential test functioning's (or contamination's) negative influence on power sometimes observed for these analyses (Clauser, et al.,1993; Lopez, 2012; and Wu, 2009). The Type I error and power rates taken together seem to indicate that none of the four procedures is generally very successful at identifying focal-advantaging DIF items of the sort that might be found with the use of accommodations. Overall, their performance was better with a DIF free matching subtest than with one that was contaminated with DIF items. However, it could be that they are most successful in some referent difficulty ranges than in others in combination with one of the types of DIF modifications (e.g. uniform DIF) than the others (e.g., High non-uniform DIF or Low non-uniform DIF). These trends will be discussed in the following section.

Examination of power rates by DIF item characteristics. This section is divided in two by DIF type, with a discussion of the impact of item characteristics on analyses'

power for uniform DIF items first, followed by that for non-uniform DIF items. Since the focus here is on the item characteristics, and because power generally decreases to varying degrees with smaller sample, lower percentage of focal simulees, skewed distributions, and with DIF contamination in the matching subtest, these conditions have been excluded and only the "best case scenario" conditions will be discussed. The section addresses only the power rates in Large sample conditions with 50% focal simulees with the normal ability distribution.

Uniform DIF items. All four analyses had the lowest power with the Low difficulty item. While this item also had a low discrimination (0.64), the analyses were highly successful with the two medium difficulty items (numbers 39 and 59) with low to medium discriminations (0.81 and 0.94), respectively. All the analyses had very high power for the rest of the uniform DIF items, regardless of difficulty or item discrimination, as long as at least some of the focal simulees were associated with Large DIF. These results seem contrary to those of Narayanan and Swaminathan (1994) who found that higher power was achieved by both SIB and MH when items were modelled with low or medium difficulty in association with medium or high discrimination, and relatively lower power (e.g., < .70) for high difficulty with either low or medium discrimination. The current study found no evidence that the uniform DIF item's discrimination played a part in determining the analyses' ability to identify it. The current study did find, however, that both item difficulty and severity of DIF induced were key factors in the analyses' ability to identify uniform DIF items.

Non-uniform DIF items. With non-uniform DIF items, in addition to referent difficulty and the severity of DIF modification influencing analysis power, the type of non-

uniform DIF modification seemed to be a factor. Both of the items (number 20 and 60) that were modelled with the High non-uniform DIF modification, had high power for all analyses regardless of severity of DIF induced. For these items, the guessing parameter was set higher and the discrimination parameter was set lower than that of the referent parameters. These items were in the Medium and High difficulty ranges, respectively, where the one item with High non-uniform DIF modification in the Low difficulty range was observed to have very low power. It could be that the DIF modification interacts with item difficulty to make it harder for analyses to identify the item as containing DIF. Additional evidence that this might be the case is provided by the power rates for the Low non-uniform DIF modification. All of the analyses generally had low power for these three items, especially when some or all of the simulee DIF was modelled as Moderate. Interestingly, of the three, the item (number 10) with the highest power was the only one in the Low difficulty range. However, its rates were very similar to the Low difficulty uniform DIF item (number 9), so it could be that this item is functioning as a uniform DIF item rather than a non-uniform DIF item for the analyses. Since DIF modification was found to influence power in addition to referent difficulty, and since the DIF parameters used in this study were informed by real data, the low power rates for these items are particularly concerning when considering the fairness of accommodations.

One might expect that based on X-SIB's purpose and prior studies (e.g., Narayanan & Swaminathan, 1996) that X-SIB would have higher power rates than either MH or SIB when examining non-uniform DIF items. While this was generally not the case with the High non-uniform DIF modification, there was a marked increase for X-SIB over that of MH and SIB for the Low non-uniform DIF modification items in the Medium range. The

results for these items support the finding by Narayanan & Swaminathan that X-SIB had both much higher power than MH when non-uniform DIF items were modelled by only changing the discrimination parameter. While their study found that X-SIB was most successful with the High discrimination items (with Low or Medium difficulty) and least successful with Low discrimination items (with Medium or High difficulty), there was no relationship between the level of item discrimination and power found in the current study.

To explore the reasons ATFIND performed so differently than expected and to check the results found for the DIF analyses, several additional conditions were run. These conditions increased both sample size (to 20,000) and severity of DIF (difference in difficulty parameter of 1.00 between groups), separately and together. All 12 DIF items for these conditions were modeled for uniform DIF, which is different from the original study where in every condition, one-half of the DIF items was modeled for uniform DIF, and the other half were modeled for non-uniform DIF. The increases in sample size and DIF severity were compared to the sample size that used as Large sample for ATFIND (750, with an additional 750 for DIF analyses), and severity of DIF (.80 difference in difficulty) that was the maximum of the range selected in this original study. Since the DIF analyses did better with normal distribution and 50% focal simulees, the 20% DIF, Large sample, 50% focal, N-LDIF condition was used as a comparison between the additional conditions and the original study. The documentation of these conditions and the results are provided in Appendix J.

In these additional conditions, ATFIND was found to select DIF items very differently for the two sample sizes. When a total sample size of 20,000 (10,000 for ATFIND) was used with the Larger DIF modification (difference of 1.00 between focal and

referent item difficulty), ATFIND almost consistently selected either all of the DIF items or none of the DIF items for the AT List. As either sample size or severity of DIF decreased, ATFIND was found to increasingly split the DIF items between the AT List and the PT List. When the lower sample size of 1,500 (750 for ATFIND) was used with the Smaller DIF modification (difference of .80 between focal and referent item difficulty), ATFIND selected DIF items in a distribution that was very similar to that found in the comparison original study condition.

To explore possible reasons for ATFIND's all or nothing selection of the DIF items in the Large sample, Large DIF additional condition, a series of multiple regression analyses were performed. These evaluated how well the discrimination and difficulty item parameters of the 48 non-DIF items predicted their placement either with the DIF items on the AT List or on the AT List when all of the DIF items were placed on the PT List. The linear combination of the two parameters was significantly related to their placement on the AT List both with and without the DIF items. However, only the non-DIF discrimination item parameter was significantly correlated (-.52, $p < .01$) with placement on the AT List with the DIF items, and only the non-DIF difficulty item parameter was significantly correlated (.61, $p < .01$) with placement on the AT List when the DIF items were all placed on the PT List.

Because of the apparent quadratic nature of the non-DIF item difficulty parameter's relationship to selection for the AT List without the DIF items, a curve estimation regression procedure was also performed. This also was statistically significant and accounted for approximately 67% of the variance of the non-DIF items placement on the AT List when all of the DIF items were placed on the PT List. The fact that ATFIND selected this group of medium difficulty non-DIF items for the AT List as a similar, but

slightly lower, rate (41%) as it did all of the DIF items (57%) would seem to indicate that ATFIND was having difficulty identifying which of the two groups of items was most "relatively dimensionally homogeneous" (Froelich & Habing, 2008, p. 144).

The DIF analyses' power and Type I error results for the additional conditions also generally confirm those of the original study. With the additional conditions, MH, SIB, and X-SIB all had consistently high power (over 90%) with the Best matching subtest for the uniform DIF items, as well as for that type of DIF item in the original condition. They all also tended to have slightly lower Analyzed power when DIF items were included in the matching subtest, especially with the smaller sample size (750 for the additional conditions and 1,250 for the original condition). The amplified Type I error rates with DIF contamination found in the original study were magnified with DIF contamination in conjunction with the Large sample size in the additional conditions. All three analyses had Analyzed Type I error rates of over 45% in both Large sample conditions with SIB and X-SIB having Analyzed Type I error rates of over 90% when Large sample was combined with the Largest DIF modification. These results confirm the original study's finding that the PT List is unusable as a matching subtest, regardless of any additional content examination of the DIF items identified by the analyses, prior to determination of item bias.

Limitations

While the current research explored many conditions found in real assessment data, it did not include all conditions for all examinees. In this study, the ability distributions were selected based on only three groups of high school students, those with no special need (regular education), those with an identified disability, and those who were registered in an English language learning program. For examination of performance, as well as

accountability and research purposes, each of these broad groups can, and often are, divided into multiple subgroups based on student characteristics such as ethnicity, type of disability, and home language. The current study did not address any of these additional differences.

Also, the results indicate that additional conditions would be worth exploring in future studies. For ATFIND, it would be interesting to explore the influence both the correlation and the degree of angle between abilities have on its selection of its subtests. Using a multidimensional model of both uniform and non-uniform DIF in addition to the unidimensional used here might help lead to more useful applications for ATFIND. While the current study was limited to a "Large" sample size of 750, results indicate that both uniform DIF and Medium referent difficulty range items are selected more frequently for the AT List as sample size increases. It would be prudent to examine larger sample sizes for this analysis especially as new state and national level large sample tests are being developed. In this study, the guessing parameter was set for ATFIND at .20 since that was the referent group guessing parameter. However, based on Socha and DeMars (2013a) finding, this might have influenced ATFIND's Type I error rate. It would be interesting to see how much setting the ATFIND guessing parameter at .00 and at .30 influences item selection when DIF items are included in the dataset.

It seems, based on the current study and the small axillary study, that at least some dimensionality analyses have difficulty identifying DIF items modified to advantage the lower ability focal group. It could be that an analysis other than those used here would have better success, especially if a larger sample size is used. Also, the inclusion of an equal number of non-uniform DIF items and uniform DIF items in each dataset might have influenced ATFIND's power rates. Based on its slight propensity to selected uniform DIF

items over both non-DIF and non-uniform DIF items for the AT List, especially as sample size increased, it would be interesting to see how changing the ratio of these items influences ATFIND's power.

With DIF analyses, there were indications that contamination in the matching subtest influenced both power and Type I error rate but the current study only explored a limited number of conditions where impact against the focal group was combined with focal-advantaging DIF items. With more and more accommodations being introduced in testing situations, especially with the new, innovative item types in computer-based testing, a further exploration of the relationship between impact and focal-advantaging DIF would seem prudent. Associated with this, is the lack within the current study of referent groups with the same skewed distributions as the focal groups. An example of these groups in real data these might be two groups of students with disabilities, one who had access to the accommodation under investigation and the other who did not. It might be that, if the two groups had the same positively skewed ability distribution, the DIF analyses would have been more successful in finding the DIF items in the Low referent difficulty range.

The results of this study indicate that research on non-uniform DIF needs to be expanded. A prior study had shown that the item discrimination parameter was a major factor in the analyses ability to identify the item as containing DIF, where in this study no evidence of this as a factor was found. The factor that did influence power was the type of non-uniform DIF modification induced, with the High non-uniform DIF modification seeming to be much easier for MH and SIB to identify than the Low non-uniform DIF modification. However, only six items were used to explore these differences. It would seem

that, since accommodations might generally cause non-uniform DIF rather than uniform DIF (Scott, 2009), these differences should be further explored.

Lastly, the current study did not explore test length or the amplification of DIF across item bundles. It could be that the results here would vary as test length varied, especially in combination with Moderate DIF. Also, it is possible that when similar items are bundled together that focal-advantaging DIF would be more evident.

Conclusions

The focus of this research was to examine how ATFIND, MH, SIB, and X-SIB function when DIF items in the dataset are modelled to advantage a lower ability focal group over a higher ability reference group. While the primary purpose for ATFIND was to examine its usefulness as a valid subtest selection tool, it also explored the influence of DIF items, item difficulty, and presence of multiple examinee populations with different ability distributions on the selection of the AT and PT Lists.

ATFIND was found to be a less than effective matching subtest selection tool with DIF items that are modelled unidimensionally. Regardless of sample size, percentage of DIF items in the dataset, type of DIF modelled or the referent difficulty of the DIF items, simulee distribution, variability of simulee distribution within the dataset, or percentage of focal simulees, ATFIND selected approximately 60% of the items for the PT List and 40% for the AT List. The resultant PT List matching subtests contained approximately the same percentage of DIF items as were in the original dataset. This led to matching subtests that were shorter, but had similar contamination (and purity) rates as the original set of items. Interestingly, ATFIND's lack of usefulness as a matching subtest selection tool for DIF studies also indicated that it tended to be robust to some of the most influential

unidimensional conditions that might occur in real data. When examining the characteristics of the items ATFIND selected for the two subtests, only two characteristics seemed to influence selection. Both uniform DIF items and items in the Medium referent difficulty range were selected slightly more often for the AT List than the PT List. These trends were seen to increase as sample size increased and might become a consideration when ATFIND sample exceeds the 750 used as Large sample in the current study.

For MH, SIB, X-SIB, and Both-SIB, the primary purpose was to examine the ATFIND selected PT List's usefulness as a valid matching subtest. This was evaluated using the Total Type I error rates and the Total power rates for each analysis and subtest combination. The PT List was less than successful as a valid matching subtest, especially for the SIBTEST analyses that only examined the suspect items in the AT List for DIF. MH's default of examining all items, regardless of their subtest, was found to aid in its much higher power.

The use of both PT List DIF-contaminated matching subtest and a static set DIF-free matching subtest that included both DIF and non-DIF items in the suspect list allowed for the examination DIF contamination on power and Type I error. Overall, all four analyses had higher power and lower Type I error with the DIF-free matching subtest than with the use of the DIF-contaminated PT matching subtest. While generally these differences were relatively small, they tended to be amplified when more DIF items were present in the dataset. The amplification was especially found when DIF was modelled only with Moderate severity or when focal distributions were skewed.

The use of the static DIF-free matching subtest with both DIF and non-DIF items in the suspect list also allowed for the comparison between the analyses' Type I error and

power rates not only by examinee ability distributions but also by the influence DIF type, DIF modification, and referent item difficulty. Overall, for all four analyses, as focal simulee ability varied from that of the referent group, Type I error rates increased. However, the results for power were not as clear. These were complicated by the combination of the severity of DIF modification being associated with a degree of difference in mean ability in the skewed conditions, as well as differences in DIF modification, especially for non-uniform DIF.

Finally, since both SIB and X-SIB were included in the study, the influence of their joint use was examined. Generally, the combined Both-SIB had both the highest power and the highest Type I error rate of any of the four procedures regardless of all other factors. Since some of the Type I error rates found were extremely high (up to 20.7% with the Best matching subtest and even higher with the contaminated PT matching subtest), practitioners should use extreme caution when using the results of the two analyses in combination. While the results from MH were not combined with X-SIB within this study, since MH and SIB have been found to identify similar items (Fidalgo, Ferreres, Muniz, 2004), one might conjecture that similar Type I error and power rates would be found combining the results from MH with X-SIB.

END NOTES

1. Ackerman (1992, p. 69) defines formally defines impact as "a between-group difference in test performance caused by a between-group difference on a valid skill (e.g., the differences between the proportion correct for two groups of interest on a valid item)." Where in his case-wise analysis of his regression equation of the difference between the expected value of the conditional distributions for a specific ability level, he concluded that a difference in group means on the intended construct would signal U-DIF rather than impact. He explains that this surprising result as a function of the amount and direction of the correlation between θ, the valid skill, and η, the nuisance skill, stating that "by examining the expected value of an η 'conditional slice'… [that] Because of the positive correlation between θ and η, the focal group [the group with the lower ability] will have the higher expected value" (p. 81). Using this reasoning then, if there is no correlation between the two traits as in our example, no potential for DIF would be expected. The table was modified to reflect the Ackerman's definition rather than his case-wise conclusion.

REFERENCES

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

Ackerman, T. A. (1994). Using multidimensional Item Response Theory to understand what items and tests are measuring. *Applied Measurement in Education*, *7*, 255-278.

Ackerman, T. A., & Evans, J. A. (1994). The influence of conditioning scores in performing DIF analyses. *Applied Psychological Measurement, 18*(4), 329-342.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.) *Differential Item Functioning* (pp. 3-23). Hillside, NJ: Lawrence Erlbaum.

Arizona Department of Education. (2011a). *Investigation of standard accommodation usage on the Spring 2010 administration of Arizona's Instrument to Measure Standards: Science results.* Retrieved June 24, 2012, from http://www.azed.gov/standards-development-assessment/files/2011/12/accommodationreport_011011.pdf.

Arizona Department of Education. (2011b). *Arizona: Arizona's Instrument to Measure Standards 2011 technical report.* Retrieved June 24, 2012, from http://www.azed.gov/standards-development-assessment/files/2011/12/aims_tech_report_2011_final.pdf.

Arizona Department of Education. (2011c). *Testing accommodations: Guidelines for school year 2011-2012*. Retrieved May 16, 2012, from http://www.azed.gov/standards-development-assessment/files/2011/08/testingaccommodations2011-12.pdf.

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement*, *11*, 111-141.

Blais, J., & Laurier, M. (1995). The dimensionality of a placement test from several analytical perspectives. *Language Testing 12*(1), 72-98.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, *46*, 443-449.

Bock, R. D., Muraki, E. & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275-285.

Bolt, D. M., & Gierl, M. J. (2006). Testing features of graphical DIF: Application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement*, *43*(4), 313-333.

Bolt, D. & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika, 23*(1), 67-95.

Bolt, S. E., & Thurlow, M. L. (2007). Item-level effects of the read-aloud accommodation for students with reading disabilities. *Assessment for Effective Intervention*, *33*, 15-28.

Box, G. E. P., & Draper, N. R. (1987). Empirical Modeling Building and Response Surfaces. John Wiley & Sons, New York, NY.

Byrne, B. M. (2006). *Structural Equation Modeling with EQS: Basic Concepts, Applications, and Programming, 2nd ed.* Mahwah, NJ: Lawrence Erlbaum.

Camara, W. (2009). Validity evidence in accommodations for English language learners and students with disabilities. *Journal of Applied Testing Technology*, *10*(2) Retrieved June 24, 2012 from http://www.testpublisher.org/assets/documents/Special%20Issue%20article%207.pdf.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.) *Educational Measurement, 4th ed.* Westport, CT: Praeger.

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

Clauser, B., & Mazor, K. (1998). Using statistical procedures to identify differential 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 the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, *6*, 269-279.

Cohen, A. S., Gregg, N., Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research & Practice*, *20*, 225-233.

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice*, *10*(3), 37-45.

Crawford, A. V., Green, S. B., Levy, R., Lo, W. J., Scott, L., Svetina, D., & Thompson, M. S. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Measurement, 70*(6), 885-901.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-18). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Dainis, A. M. (2008). Methods of identifying differential item and test functioning: An investigation of Type I error rates and power. (Unpublished doctoral dissertation). James Madison University, Harrisonburg, VA.

de Ayala, R. J. (2009). *The theory and practice of Item Response Theory*. New York: The Guilford Press.

Deng, H., & Ansley, T. N. (2000, April). Detecting compensatory and noncompensatory multidimensionality using DIMTEST. Paper presented at the annual meeting of the National Council on Measurement in Educational, New Orleans, LA.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 137-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Douglas, J. A., Roussos, L. A., Stout, W. Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, *33*, 465-484.

Elliot, S. N., & Marquart, A. M. (2004). Extended time as a testing accommodation: Its effects and perceived consequences. *Exceptional Children*, *70*, 349-367.

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Fidalgo, A. M., Ferreres, D., & Muniz, J. (2004). Liberal and conservative differential item functioning detection using Mantel-Haenszel and SIBTEST: Implications for Type I and Type II error rates. *The Journal of Experimental Education*, *73*(1), 23-39.

Fidalgo, A. M., Hashimoto, K., Bartram, D., & Muniz, J. (2007). Empirical Bayes versus standard Mantel-Haenszel statistics for detecting differential item functioning under small sample conditions. *The Journal of Experimental Education*, *75*(4), 293-314.

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.

Finch, W. H. & French, B. F. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement*, *47*, 299-317.

Franciosi, R. (2008). *Arizona's School Accountability System 2007 Technical Manual Volume II: Adequate Yearly Progress*. Phoenix, AZ: Arizona Department of Education. Retrieved July 22, 2012 from http://www.azed.gov/research-evaluation/files/2011/09/2007-nclb-technical-manual.pdf.

Froelich, A. G. (2000). *Assessing unidimensionality of test items and some asymptotics of parametric item response theory.* (Unpublished doctoral dissertation). University of Illinois, Urbana, IL.

Froelich, A. G, & Habing, B. (2008). Conditional covariance-based subtest selection for DIMTEST. *Applied Psychological Measurement*, *32*, 138-155.

Furlow, C. F., Ross, T. R., & Gagné, P. (2009). The impact of multidimensionality on the detection of differential bundle functioning using Simultaneous Item Bias Test. *Applied Psychological Measurement*, *33*, 441-464.

Gierl, M. J., Bisanz, J., Bisanz, G. L., Boughton, K. A., & Khaliq, S. N. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*(2), 26-36.

Gierl, M. J., Gotzmann, A., Boughton, K. A. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education*, *17*, 241-264.

Gierl, M. J., Jodoin, M. G., & Ackerman, T. A. (April, 2000). Performance of Mantel-Haenszel, Simultaneous Item Bias Test, and Logistic Regression when the proportion of DIF items is large. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Glass, G. V., & Hopkins, K. D. (1996). Statistical Methods in Education and Psychology. 3rd ed. Allyn and Bacon, Boston, MA.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent test score models. *British Journal of Mathematical and Statistical Psychology*, *33*, 234-246.

Gorin, J. S. & Embretson, S. E. (2008). Item Response Theory and Rasch models. In D. Mc Kay (Ed.) *Handbook of research methods in abnormal and clinical psychology* (271-291). Thousand Oaks, CA: Sage.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, *44*, S78-S94.

Güler, N., & Penfield, R. D. (2009). A comparison of logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement, 46* (3), 314-329.

Headrick, T. C, & Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the Fleishman Power method. *Psychometrika*, *64*, 25-35.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care*, *44*, S182-S188.

Hambleton, R., & Rogers, H. J. (1989). Detecting potentially biased items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education*, *2*, 313-334.

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Identification of potentially biased test items. In *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hanson, B. A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics*, *23*, 244-253.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, *35*, 57-63.

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

Holland, P. W., & Thayer, D. T. (1988). Influence of the criterion variable on the identification of differentially functioning test items using the Mantel-Haenszel statistic. *Applied Psychological Measurement*, *15*, 353-359.

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*, 249-260.

Ilich, M. O. (2013). *Differential item functioning (DIF) among Spanish-speaking English language learners (ELLs) in state science tests.* (Unpublished doctoral dissertation). University of Washington, Seattle, WA.

Ironson, G., Homan, S., Willis, R., & Signer, B. (1984). The validity of item bias techniquests with math word problems. *Applied Psychological Measurement, 8*(4), 391-396.

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.

Johnstone, C. J., Altman, J., Thurlow, M. L., & Thompson, S. J. (2006). *A summary of research on the effects of test accommodations: 2002 through 2004 (Technical Report 45).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved June 23, 2012 from http://www.cehd.umn.edu/NCEO/OnlinePubs/Tech45/Technical45.pdf.

Jones, R. N. (2006b). Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: Detecting differential item functioning using MIMIC modeling. *Medical Care*, *44*, S124-S133.

Jöreskog, K. G. (1979). Basic ideas of factor and component analysis. In K. G. Jöreskog and D. Sorbom (Eds.) *Advances in Factor Analysis and Structural Equation Models*. Cambridge, MA: Abt Books.

Kanjee, A. (2007). Using logistic regression to detect bias when multiple groups are tested. *South African Journal of Psychology*, *37*(1), 47-61.

Kim, H. R. (1994). *New techniques for the dimensionality assessment of standardized test data.* (Unpublished doctoral dissertation). University of Illinois, Urbana, IL.

Kingston, N., Leary, L., & Wightman, L. (1985). *An exploratory study of the applicability of item response theory methods to the Graduate Management Admissions Test* (GMAC Occasional Papers). Princeton, NJ: Graduate Management Admissions Council.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking.* New York, NY: Springer.

Li, H-H. (1995). *New nonparametric statistical procedures for analyzing bias/DIF and dimensionality in item response data.* (Unpublished doctoral dissertation). University of Illinois, Urbana, IL.

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

Linn, R. (1993). The use of differential item functioning statistic: A discussion of current practice and future implications. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 349-364). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lopez, G. E. (2012). *Detection and classification of DIF types using parametric and nonparametric methods: A comparison of the IRT-Likelihood Ratio test, Crossing-SIBTEST, and Logistic Regression procedures.* (Unpublished doctoral dissertation). University of South Florida, Tampa, FL.

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

Magis, D., Beland, S., & Raiche, G. (2013). difR: A collection of methods to detect dichotomous differential item functioning (DIF) in psychometrics [Computer software]. Flanders, Belgium: Katholieke Universiteit Leuven.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment*, *13*, 160-179.

Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22*(4), 357-367.

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1993). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement*, *32*, 131-144.

McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of general theory. *Multivariate Behavioral Research 14*, 21-38.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology 34*, 100-117.

McDonald, R. P. (1982). Linear versus nonlinear models in Item Response Theory. *Applied Psychological Measurement* 6(4), 379-396.

McDonald, R. P. (1996). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 258-270). New York: Springer.

McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis 17*(3), 305-322.

Meredith, W. & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*, S69-S77.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33-46). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Middleton, K., & Cahalan Laitusis, C. (2007). *Examining test items for differential distractor functioning among students with learning disabilities* (Report RR-07-43). Princeton, NJ: ETS. Retrieved June 23, 2012, from http://www.ets.org/research/policy_research_reports/rr-07-43.

Millsap, R. E. (2006). Comments on methods for the investigation of measurement bias in the Mini-Mental State Examination. *Medical Care*, *44*, S171-S175.

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York, NY: Routledge Academic.

Millsap, R. E. & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297-334.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, *13*, 57-75.

Mokken, R. J. (1996). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern Item Response Theory* (pp. 251-268). New York: Springer.

Mulaik, S. A. & Millsap, R. E. (2000). Doing the four-step right. *Structural Equation Modeling*, *7*, 36-73.

Muthén, B. O. (2002). Beyond SEM. General latent variable modeling. *Behaviormetrika*, *29*, 81-117.

Muthén, B. & Muthén, L. (1998-2010). *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement 28*(2), 99-117.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30*(4), 293-311.

Nandakumar, R. (1994, April). *Development of a valid subtest for assessment of DIF/Bias*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, *18*, 41-68.

Nandakumar, R., & Yu, F. (1994, April). *Testing the robustness of DIMTEST on nonnormal ability distribution*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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

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

Newton, R. R., & Rudestam, K. E. (1999). *Your statistical consultant: Answers to your data questions*. Thousand Oaks, CA: Sage Publications.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 U.S.C. § 1425 (2002). Retrieved June 23, 2012, from http://www.ed.gov/policy/elsec/leg/esea02/begining.html#sec1.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.255-279). Hillsdale, NJ: Erlbaum.

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

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measure of differential functioning of items and tests. *Journal of Educational Measurement*, *34*(3), 253-272.

Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement*, *43*(1), 1-17.

R Project for Statistical Computing. (2012). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 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. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*(4), 401-412.

Reckase, M. D. (1996). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York: Springer.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, *25*, 193-203.

Reckase, M. D., Carlson, J. E., Ackerman, T. A., & Spray, J. A. (1986, June). *The interpretation of unidimensional IRT parameters when estimated from multidimensional data.* Paper presented at the Psychometric Society annual meeting, Toronto.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*(4), 361-373.

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

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, *49*, 425-435.

Roussos, L. A., Schnipke, D. L, & Pashley, P. J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics*, *24*(3), 293-322.

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

Roussos, L., & Stout, W. (1996b). Simulation studied of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement, 33*(2), 215-230.

Roussos, L., Stout, W., & Marden, J. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement*, *35*, 1-30.

Scott, L. (2009). *The relationship between non-standard accommodations and item content in item functioning within a state high school mathematics assessment.* (Unpublished master's thesis). Arizona State University, Tempe, AZ.

Scheuneman, J. D., Camara, W. J., Cascallar, A. S., Wendler, C., & Lawrence, I. (2002). Calculator access, use, and type in relation to performance on the SAT 1: Reasoning Test in mathematics. *Applied Measurement in Education*, *15*, 95-112.

Schmitt, A. P., Doran, N. J., Crone, C. R., & Maneckshana, B. T. (1990, April). *Differential speededness and item omit patterns on the SAT.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Seong, T-J., Suh, Y., Lee, Y-S., & Cohen, A. S. (2004, March). *Examining Type I error and power for detection of differential item and testlet functioning.* Paper presented at the annual conference of the National Council on Measurement in Education, San Diego, CA.

Seraphine, A. E. (2000). The performance of DIMTEST when latent trait and item difficulty distributions differ. *Applied Psychological Measurement*, *24*, 82-94.

Shealy, R., & Stout, W. (1993a). An item response theory model for test bias. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 349-364). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shealy, R., & Stout, W. (1993b). 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*, 159-194.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*(3), 249-275.

Socha, A., & DeMars, C. E. (2013a). A note on specifying the guessing parameter in ATFIND and DIMTEST. *Applied Psychological Measurement, 37*(1), 87-92.

SPSS, Inc. (1989-2010). IBM SPSS Statistics: Statistical software package (Version 19.0) [Computer software]. Somers, NY: IBM Corporation.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292-1306.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293-325.

Stout, W. F., Douglas, J., Habing, B., Kim, H. R., Roussos, L. A., & Zhang, J. (1996). Conditional covariance based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*, 331-354.

Stout, W., Li, H-H., Nandakumar, R., & Bolt, D. (1997). MULTISIB: A procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement*, *21*(3), 195-213.

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

Takane, Y. & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393-408.

Teresi, J. A. (2006a). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical Care*, *44*, S152-S170.

Teresi, J. A. (2006b). Overview of quantitative measurement methods: Equivalence, invariance, and differential item functioning in health applications. *Medical Care*, *44*, S39-S49.

Thissen, D. & Mooney, J. (1989). Loglinear item response models, with applications to data from social surveys. *Sociological Methodology*, *19*, 299-330.

Thompson, M. S. & Green, S. B. (2006). Evaluating between-group differences in latent variable means. In G. R. Hancock & R. O. Mueller (Eds.) Structural Equation Modeling: A Second Course. Greenwich, CT: Information Age Publishing.

Thurlow, M. L., & Bolt, S. (2001). *Empirical support for accommodations most often allowed in state policy (Synthesis Report No. 41).* Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved June 23, 2012 from http://education.umn.edu/NCEO/OnlinePubs/Synthesis41.html.

Tian, F. (1999). *Detecting DIF in Polytomous Item Responses.* (Unpublished doctoral dissertation). University of Ottawa, Ottawa, Canada.

Toland, M. D. (2008). *Determining the accuracy of item parameter standard error of estimates in BILOG-MG 3*. (Unpublished doctoral dissertation). University of Nebraska, Lincoln, NE.

United States Department of Education (2000). *Elementary and Secondary Education Summary Guidance on the Inclusion Requirement for Title I Final Assessments*. Retrieved June 23, 2012, from http://www.ed.gov/policy/elsec/guid/inclusion.html.

van Abswoude, A. A. H., van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement*, *28*(1), 3-24.

van der Linden, W. J., & Hambleton, R. K. (1996). Handbook of Modern Item Response Theory. New York, NY: Springer-Verlag.

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.) *Differential Item Functioning* (pp. 3-23). Hillside, NJ: Lawrence Erlbaum.

Walker, C. M. (2001). A review of DIFPACK: Dimensionality-based DIF analysis package. *International Journal of Testing*, *1*, 305-317.

Walker, C. M., Azen, R., & Schmitt, T. (2006). Statistical versus substantive dimensionality: The effect of distributional differences on dimensionality assessment using DIMTEST. *Educational and Psychological Measurement*, *66*, 721-738.

Whitmore, M. L., & Schumacker, R. E. (1999). A comparison of logistic regression and analysis of variance differential item functioning detection methods. *Educational and Psychological Measurement*, *59*(6), 910-927.

Williams, V. S. L. (1997). The "unbiased" anchor: Bridging the gap between DIF and item bias. *Applied Measurement in Education*, *10*, 253-267.

Woods, C. M. (2011). Ramsay-curve differential item functioning. Applied Psychological Measurement, 35, 536-556.

Wu, N. (2009). *Does DIF signal a lack of essential unidimensionality?* (Unpublished doctoral dissertation). Purdue University, West Lafayette, Indiana.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187-213.

Yen, W. M. & Fitzpatrick, A. R. (2006). Item Response Theory. In R. L. Brennan (Ed.) *Educational Measurement, 4th ed.* (pp. 111 – 153) Westport, CT: Praeger.

Zhang, J. & Stout, W. F. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika 64*(2),129-152.

Zhang, J., & Stout, W. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 212-249.

Zhou, J. (2006). *Evaluating the performance of SIBTEST and MULTISIB for a multidimensional test.* (Unpublished master's thesis). University of Alberta, Alberta, Canada.

Zhou, J., Gierl, M. J., & Tan, X. (April, 2006). Evaluating the performance of SIBTEST and MULTISIB using Different Matching Criteria. Paper presented at the annual conference of the National Council on Measurement in Education, San Francisco, CA. Retrieved May 14, 2014, from http://www2.education.ualberta.ca/educ/psych/crame/files/NCME06_JZ.pdf.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multi-group IRT analysis and test maintenance for binary items [Computer software]. Chicago: Scientific Software International.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education*, *10*, 321-344.

APPENDIX A

IRB APPROVAL LETTER

ASU Knowledge Enterprise Development

Office of Research Integrity and Assurance

To:	Roy Levy EDB
From:	Mark Roosa, Chair Soc Beh IRB
Date:	03/01/2013
Committee Action:	**Expedited Approval**
Approval Date:	03/01/2013
Review Type:	Expedited F5 F7
IRB Protocol #:	1302008883
Study Title:	An Investigation of Item Parameters of Focal Advantaging DIF Items for Simulating such Items to Explore an Analytic Selection of a Valid Subtest for DIF Analysis when DIF has Multiple Potential Causes Among Multiple Groups
Expiration Date:	02/28/2014

The above-referenced protocol was approved following expedited review by the Institutional Review Board.

It is the Principal Investigator's responsibility to obtain review and continued approval before the expiration date. You may not continue any research activity beyond the expiration date without approval by the Institutional Review Board.

Adverse Reactions: If any untoward incidents or severe reactions should develop as a result of this study, you are required to notify the Soc Beh IRB immediately. If necessary a member of the IRB will be assigned to look into the matter. If the problem is serious, approval may be withdrawn pending IRB review.

Amendments: If you wish to change any aspect of this study, such as the procedures, the consent forms, or the investigators, please communicate your requested changes to the Soc Beh IRB. The new procedure is not to be initiated until the IRB approval has been given.

Please retain a copy of this letter with your approved protocol.

APPENDIX B

INVESTIGATION OF REAL-DATA FOCAL-ADVANTAGING DIF

The investigation to inform the modeling of differential item functioning (DIF) items for simulation research contained four distinct phases. These phases were, in order, a review of the literature, DIF analysis of real data to identify items that favored the focal group, the equating of each groups' item parameters to that of the reference group, and finally, an examination of the changes between the reference group's item parameters and that of the focal groups' for those focal group favoring DIF items. Each of these phases will be described, in order, within this document followed by a presentation of the resultant parameters to be used to simulate focal advantaging DIF items.

Review of the Literature

A brief review of the literature was performed to investigate the methods that have been used by researchers to model DIF for simulation research. The type of DIF analysis was specifically excluded from the criteria for selecting studies resulting in those that included not only SIBTEST and MH DIF analyses but also, Raju's area, logistic regression, analysis of variance, confirmatory factor analysis, and empirical Bayes MH analyses. Particular attention was paid to the type of DIF (U-DIF or NU-DIF) that was modeled and to the type of modeling (unidimensional or multidimensional) that was performed. While DIF can be viewed as a secondary dimension on which two of more groups have different distributions, because it has been previously determined that the data for the current research would be modeled unidimensionally, those studies modeling DIF using multidimensional item response theory were noted but are not included within this discussion.

Modifications used to model DIF items. Of the ten studies selected that simulated unidimensional data to research DIF analyses, six modeled both U-DIF and NU-

DIF, one modeled only NU-DIF, and three modeled only U-DIF. For U-DIF, researchers generally chose to add or subtract a static value from the reference difficulty parameter (e.g., add .5, Fidalgo, Ferreres, & Muniz, 2004; add 1.0, Seong, Suh, Lee, & Cohen, 2004; subtract .93, Raju, 1990; and add .48 or .64, Swaminathan, & Rogers, 1990). Sometimes the researchers determined the amount to add or subtract from the difficulty parameter based on the area between the item characteristic curves (ICCs) for the focal and referent groups (Kanjee, 2007; Raju, 1990; and Whitmore, & Schumacker, 1999) with areas of .4 referring to small DIF, .6 as moderate DIF, and .8 as large DIF. Some researchers chose not only to specify a static difference between item difficulties but also chose specific standard item difficulties for the reference group. Bolt & Gierl (2006) used this process when they chose respective referent and focal item difficulties as displayed in Table B1. Of the studies reviewed only one research team, Fidalgo, Hashimoto, Bartram, and Muniz (2007), chose the amount they added to the referent item difficulty from a distribution of values, using a $N(0, 0.18)$ distribution to compare an empirical Bayes to the standard MH statistic.

Table B1

Difficulty Parameters for DIF Items Simulated by Bolt and Gierl, 2006

Referent Difficulty	1st Focal Difficulty	2nd Focal Difficulty	3rd Focal Difficulty
-1.00	-1.25	-0.75	-0.50
0.00	-0.25	0.25	0.50
1.00	0.75	1.00	1.50

For NU-DIF, where a modification to the discrimination parameter is added to that of the difficulty parameter, researchers again generally chose either to add a value from the

reference group's discrimination parameter or select specific parameter combinations. Stark, Chernyshenko, and Drasgow (2006) added -.25 or -.40, Swaminathan, and Rogers (1990) added 1.76 or 2.36, and Whitmore, and Schumacker (1999) added 0.76 to the referent discrimination parameter. Where Oshima, Raju, and Nanda (2006) crossed an addition of -0.50 to the discrimination parameter with the addition of 0.00, 0.50 and 1.0 to the difficulty parameter and Bolt and Gierl (2006) matched the addition of -0.25 to the discrimination parameter with the addition of -0.50, 0.00, and 0.25 to the difficulty parameter. Only one researcher specified the use of a ratio modification of the referent discrimination parameter to obtain that of the focal (1.13 multiplied to the discrimination parameter, Raju, 1990). The researchers that used the area between ICCs to inform their choice of changes to the difficulty parameter also used it to inform the changes to the discrimination parameter. Only one study (Kanjee, 2007) mentioned using a pseudo-guessing parameter which was set at 0.20 for all items for both referent and focal groups.

Relationships between MH Δ and IRT. Researchers have investigated and derived formulas that relate the MH delta statistic (Δ) for U-DIF to one or more IRT parameters. A summary of these relationships follows with each IRT model discussed separately.

Rasch and 1PL IRT model. Holland and Thayer (1988) derived a relationship conditioned on total score for a Rasch one parameter logistic (1PL) IRT model as follows:

$$\hat{\Delta} = e^{b_{iF} - b_{iR}} \quad (1)$$

where b_{iF} and b_{iR} are the studied item's difficulty parameters for the focal and referent group, respectively. Roussos, Schnipke, and Pashley (1999) reframe this equation as:

$$\hat{\Delta}_{1PL} = 4(b_R - b_F) \quad (2)$$

and found that $\hat{\Delta}_{1PL}$ follows this formula and that an absolute difference in b values of 0.375 or more would result in a Δ corresponding to a "C" MH Δ flag (a Δ statistic of at least 1.50) and a "B" flag (a Δ statistic of between 1.00 and 1.50) would correspond to an absolute difference of at least 0.25 and less than 0.375.

2PL IRT model. Roussos et al. (1999) derived and examined a similar formula for a two parameter logistic (2PL) IRT model which incorporates the discrimination parameter (a, assumed to be the same for both groups for U-DIF) as follows:

$$\hat{\Delta}_{2PL} = 4a(b_R - b_F). \quad (3)$$

With this equation the amount of difference in difficulty parameters required to flag DIF at a specific level with the Δ statistic would be proportionally reduced if the discrimination parameter is greater than 1 and proportionally increased if it is less than 1.

3PL IRT model. Roussos at al. (1999) found a very different effect when they studied the relationship between the 3PL model and MH Δ. They numerically evaluated the integrals within a theoretical formula for the MH odds ratio, written in terms of examinee ability, and then computed Δ by multiplying the natural log of that value by -2.35. When they incorporated the pseudo-guessing parameter (at 0.20 for both groups), they found that as the

average of the difficulty parameters of the two groups increased, the absolute value of Δ decreased from the base level associated with that difference between the difficulty parameters. Increases in the discrimination parameter were related to a decrease in the absolute value of Δ, which was the *opposite* (authors' italics) of that found when the pseudo-guessing parameter was set at 0. They also noted effects for both ability distribution variation and the ratio of reference to focal group size as the difficulty parameters increased but noted that there was no evidence to indicate that a larger ratio for the groups affect the value of Δ.

Analysis of Real Focal Advantaging DIF Items.

Scott (2009) identified a set of items that advantaged disabled students who used non-standard accommodations such as calculators or manipulatives, over either regular education or disabled students who did not access non-standard accommodations. Because the current study is intended to model focal advantaging DIF items not only for disabled students (SWD) but also for English language Learners (ELL) who accessed non-standard accommodations, which had not been previously studied, a DIF analysis was conducted with new random samples taken from the same High School AIMS Mathematics 2006 dataset used by Scott (2009). Two new 200 and 500 case random samples were drawn for a reference group, obtained from students, not designated as either SWD or ELL, who took the examination without the use of a non-standard accommodation and two focal groups (SWD or ELL exclusively) who took the assessment with a non-standard accommodation. For each comparison, referent versus SWD and referent versus ELL, DIMTEST (Nandakumar & Stout, 1993; Stout, 1987) was used on the 200 case samples combined to identify items in the AT and PT lists. SIBTEST and Crossing-SIBTEST (Jiang & Stout, 1998; Shealy & Stout, 1993) analyses were then performed using the 500 case samples, first

with the items on the PT list as the matching subtest and then the items on the AT list as the matching subtest. The additional use of the AT list items was to ensure that all items that exhibit DIF favoring a focal group were identified for the following parameter investigation.

The DIF analyses resulted in 22 of the 84 items on the test being identified as advantaging one of the two focal groups uniformly (β_{uni}) or across at least 20% of the analysis region as given within Crossing-SIBTEST's output as the percent of the magnitude favoring a group above (and/or below) the computed crossing point. Table B2 presents the SIBTEST and Crossing-SIBTEST results for these 22 focal-favoring DIF items. Included in the table for each item is the crossing point (where the difference in probability of correct response for the two groups is equal to zero), the magnitude of B_{cro} both above and below the crossing point, the percentage and group favored by that magnitude of B_{cro}, followed by the B_{cro} and B_{uni} statistics along with their associated *p*-values. Nine items favored the focal group over between 29% (item # 66) to 84% (items # 18 and 34) of the analysis region. Thirteen items favored the focal group across the whole analysis region (e.g., items # 3 and 32, 100% either below or above the crossing point, and item # 13 favored the focal group 59% below the crossing point and 41% above the crossing point).

While item #13's results might be surprising when taken from the ICC view, both SIBTEST and Crossing-SIBTEST analyses are based solely observed probabilities rather than estimates of an ICC. With Crossing-SIBTEST, the probability of correct response for each for each level of ability for each comparison group is computed using a regression correction. Starting at the lowest ability level, these probabilities are then compared and the first instance where they are equal (their difference is equal to zero) is determined to be the crossing point for that item. The SIBTEST statistic is then computed piecewise below and

Table B2

Crossing and Uniform DIF Analysis Results for Focal Advantaging AIMS Items

Item #	Cross Point	B_{cro} below	Favor	% Favor	B_{cro} above	Favor	% Favor	B_{cro}	B_{cro} *p*-value	β_{uni}	β_{uni} *p*-value
3	38	.120	F	100	0		0	.120*	.045	-.119*	.037
6	38	.150	F	100	0		0	.150*	.012	-.153*	.033
11	33	.135	F	98	.002	F	2	.133*	.006	-.146*	.042
13	30	.077	F	59	.053	F	41	.023	.737	-.129*	.041
18	15	.021	R	16	.107	F	84	-.127*	.038	-.092*	.050
19	23	.059	R	44	.074	F	56	-.133*	.015	-.033	.052
22	22	.036	R	27	.097	F	73	-.133*	.048	-.058	.053
28	20	.024	F	25	.072	F	75	-.048	.166	-.107*	.036
32	0	0		0	.063	F	100	-.063	.218	-.082*	.037
33	0	0		0	.119	F	100	-.119*	.014	-.109*	.039
34	28	.049	F	84	.009	R	16	.058	.068	-.061*	.030
44	17	.010	F	8	.126	F	92	-.116	.101	-.141*	.044
46	8	0		0	.129	F	100	-.129*	.006	-.116*	.038
51	12	.004	F	4	.090	F	96	-.087	.156	-.096*	.035
55	19	.052	R	32	.112	F	68	-.163**	.000	-.045*	.048
62	13	.016	R	22	.054	F	78	-.070*	.003	-.026*	.045
65	38	.128	F	100	0		0	.128*	.045	-.135	.051
66	23	.072	R	71	.030	F	29	-.102*	.031	.093*	.042
71	27	.047	R	52	.043	F	48	-.090*	.026	.042*	.039
79	11	.006	F	6	.092	F	94	-.085	.136	-.096*	.041
82	14	.005	F	4	.110	F	96	-.105	.058	-.110*	.043
83	21	.090	R	60	.060	F	40	-.150*	.007	.050	.055

Note: *Indicates significance at a .05 level where ** indicates a significance level of less than .001. A negative β_{uni} indicates that examinees in the focal group are favored. A negative β_{cro} indicates the examinees in the focal group above the crossing point are favored.

above this point using the difference between expected probabilities of the focal and referent groups below this point and the difference between the expected probabilities of referent and focal groups above the point. There is nothing in the computation that requires that different groups are favored above and below the crossing point. In the case of item #13, where the crossing point was approximately in the middle of the range, the magnitude of B_{cro} below the crossing point was 0.077 and above the crossing point was 0.053, both favoring the focal group. This resulted in a composite B_{cro} of 0.023 (p-value = 0.737) after adding the negative value of 0.053 since it favored the focal group above the crossing point. While this result is non-significant, the B_{uni} statistic for this item which used the difference between focal and referent expected probabilities across the whole range was -0.129 (p-value = 0.041) which was significant. Item #11, which favored the focal group both above and below the crossing point, was found to have significant DIF by both SIBTEST and Crossing-SIBTEST.

Of the 22 focal advantaging DIF items, 8 exhibited U-DIF alone, 7 exhibited NU-DIF alone, and the remaining 7 were identified by both SIBTEST and Crossing-SIBTEST. Item number 66, which was found significant by both analyses, is interesting in that while 29% of the magnitude of B_{cro} indicates that the item favors the focal group, the B_{uni} statistic indicates that it exhibits U-DIF favoring the reference group.

Estimation and Equating of Item Parameters.

To obtain comparable estimates of the difficulty, discrimination, and pseudo-guessing item parameters for focal-favoring DIF items, the three 500 case samples used to identify the DIF items were separately analyzed using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) and then the resultant estimates were equated using methodology

described by Cook and Eignor (1991). To equate the parameters estimates of the focal groups to that of the referent group a non-equivalent group anchor test (NEAT) design was used. The anchor set, containing 30 non-DIF items, were selected by eliminating any of the 84 items on the test that exhibited DIF for any group in either the original Scott (2009) or the current analysis. These 30 items comprised 35.7% of the total test. This number and percentage of anchor items meets Kolen and Brenan's (2004) recommendation that at least 20% of the overall test be included as anchors for a test of at least 40 items in length and at least 30 anchor items for very long tests.

An examination of the anchor items' content, revealed a distribution similar to that of the full test, with a maximum difference in percentage of 3.6 within the content area of algebra. The percentage of items aligned to each of the five mathematics content strands for the full test and the anchor test are presented in Figure 1. This content area contains the most items both on the full test (25 items, 29.8%) and on the anchor set (10 items, 33.3%). The distribution of content for the anchor set approximately meets the criteria put forth or

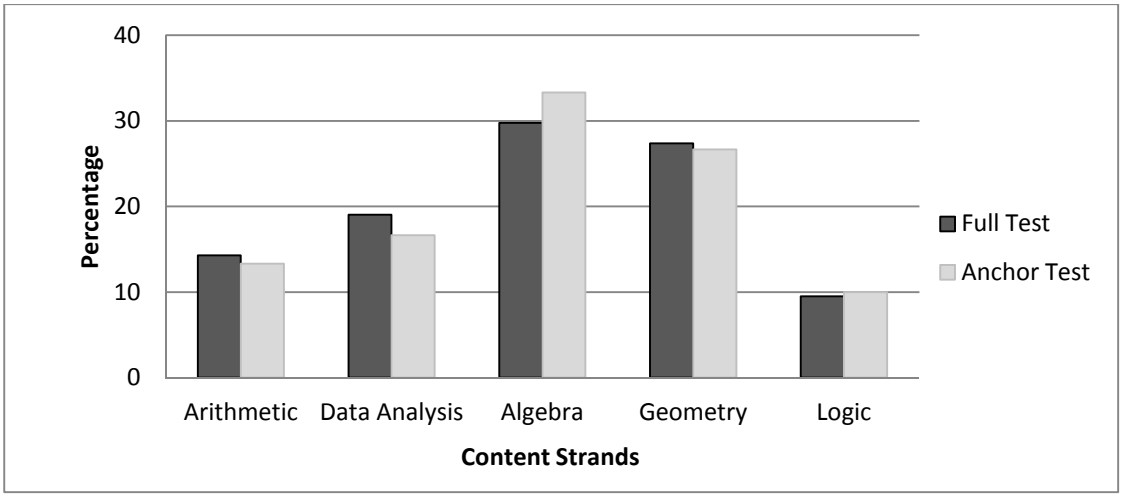


Figure B1. Percentage of items by content strand in full and anchor tests.

explored by multiple authors (e.g., Cook, & Eignor, 1991; Hambleton, Swaminathan, & Rogers, 1991; Kolen, & Brennan, 2004; and Sinharay, & Holland, 2007) that it proportionally match the operational test's content specifications.

Following Cook and Eignor (1991), the mean and SD of the difficulty parameters of the anchor set was computed for each group. These were then used to compute a slope and intercept for the conversion of the focal groups' parameters to that of the referent group. Cook and Eignor give the relationship to create the slope and intercept as:

$$\frac{b_1 - Mean_1}{Standard\ Deviation_1} = \frac{b_2 - Mean_2}{Standard\ Deviation_2}. \quad (4)$$

Reformatted, this equation becomes:

$$b_1 = \frac{Standard\ Deviation_1}{Standard\ Deviation_2}\ b_2 + \frac{-\ Mean_2}{Standard\ Deviation_2} +\ Mean_1. \quad (5)$$

Within the reformatted equation the transformative slope (noted as A below) is the ratio of the two SDs where the intercept (noted as B below) is the difference between of the mean of Group 1 and the mean of Group 2 divided by its SD. Each group's mean, SD, transformation slope and intercept are displayed in Table B3. The resultant transformation constants were used with all of the items for the group to convert both the difficulty and discrimination parameters to a scale comparable with that of the referent group. The following transformation equations:

$$\hat{b}_i^* = \mathrm{A}\hat{b}_i + \mathrm{B} \tag{6}$$

and

$$\hat{a}_i^* = \frac{\hat{a}_i}{\mathrm{A}} \tag{7}$$

where $\hat{b}_i$ is the focal group's difficulty parameter for the item as determined by individual group analysis, $\hat{b}_i^*$ is the focal group's equated difficulty parameter, $\hat{a}_i$ is the focal group's discrimination parameter for the item as determined by individual group analysis, and $\hat{a}_i^*$ is the focal group's equated discrimination parameter were used. Since the pseudo-guessing parameter is read from the probability axis rather than the ability axis, transformation of this parameter is not necessary (Cook, & Eignor, 1991). Item characteristic curves for the 8 NU-DIF items, created using the equated parameters are presented in Appendix C.

Table B3

Means and Standard Deviations of Each Group's Anchor Set

Group	Mean Difficulty	Difficulty SD	Transformation Slope	Transformation Intercept
Referent	-0.5602	1.0642		
SWD	1.2592	1.0506	0.9872	-1.759
ELL	1.0557	1.0345	0.9721	-1.581

Note: Displayed are each group's mean and standard deviation (SD) of difficulty parameters for the 30 item anchor set plus the transformation slope and intercept that were used to equate each focal group's parameters to that of the referent group.

Examination of Changes to Parameters.

The item parameters of the referent group were compared to that of the focal group favored (either SWD or ELL) for each of the 22 focal-advantaging items. For each item, the difference between each group's difficulty and pseudo-guessing parameters was taken and the ratio of the focal group's to the referent group's discrimination parameter was found. The comparisons between the difficulty, pseudo-guessing, and discrimination parameters for the groups will be discussed in that order separately below.

Difficulty parameter. Figure 2 displays the distribution of the difference between the difficulty parameters for the two groups as compared with the difficulty parameter value for the referent group by DIF type. The referent group's difficulty parameters, on the whole test, range from -4.24 to 1.67 with a mean of -0.47 and SD of .96. While no pattern to the comparison of difficulty difference to referent difficulty was discerned, the DIF items were observed over a slightly smaller range and at a higher average referent difficulty (minimum = -2.14, maximum= 1.67, mean=0.03, and SD= 0.84). The 8 NU-DIF items were found almost exclusively in the upper region of the ability scale with a minimum referent difficulty of -0.14. The mean difficulty for the NU-DIF items was 0.70, with a standard deviation of 0.61 and again a maximum of 1.67. The range, mean, and standard deviation of differences between the difficulty parameters were approximately same for both U-DIF and NU- DIF items, with a minimum of -0.79 and -0.78, a maximum of -0.18 and -0.04, a mean (and SD) of -0.49 (0.20) and -0.45 (0.28), respectively.

While the difference in the referent difficulty location between the U-DIF and NU-DIF focal-favoring DIF items is interesting, the main goal of the current investigation is to

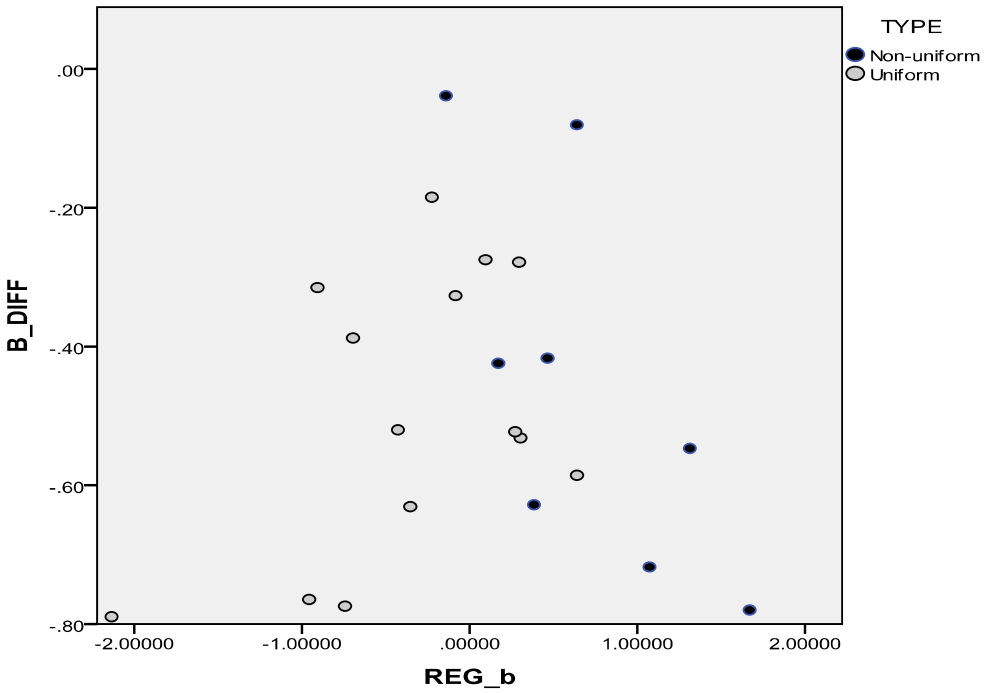


Figure B2. The difference in difficulty parameters versus the difficulty parameter for the referent group by DIF type.

determine the amount of change that should be induced to model both types of DIF items in a subset of items that is similar to that of a whole test. It is evident, based on the distribution of differences between the group difficulty parameters, that the difficulty modification for both types of DIF items would be very similar, with NU-DIF having a slightly lower minimum.

Pseudo-guessing parameter. Figures 3 and 4 present the comparison of the difference between pseudo-guessing parameters versus the referent difficulty parameters and the difference between difficulty parameters by DIF type, respectively. When the difference in pseudo-guessing parameters (*c*) was examined for all 22 focal-favoring DIF items, no discernible change to the range or distribution of differences was evident either across the

referent difficulty parameters range or the difference in difficulty parameters range. The differences in *c*-parameters, with the exception of that for one item (# 3), ranged from -0.09 to 0.08. The mean of the 21 differences (excluding that of item #3 that had a difference of 0.23) was essentially zero (-0.0012). An examination of the differences for the NU-DIF focal-favoring items showed results very similar to those for all 22 focal-favoring DIF items.

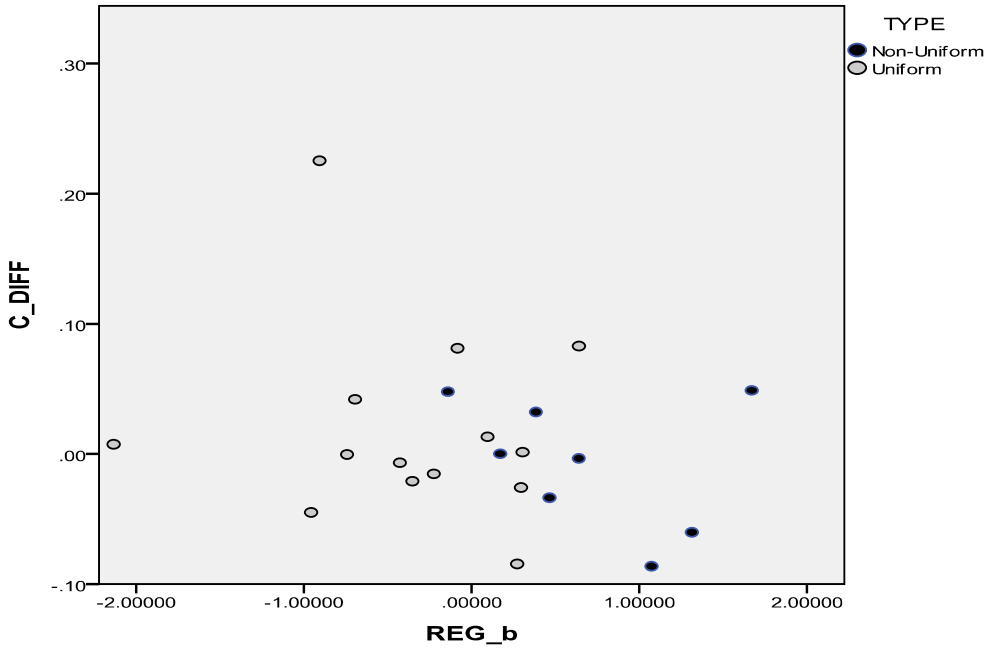


Figure B3. The difference in pseudo-guessing parameters versus the difficulty parameter for the referent group by DIF type.

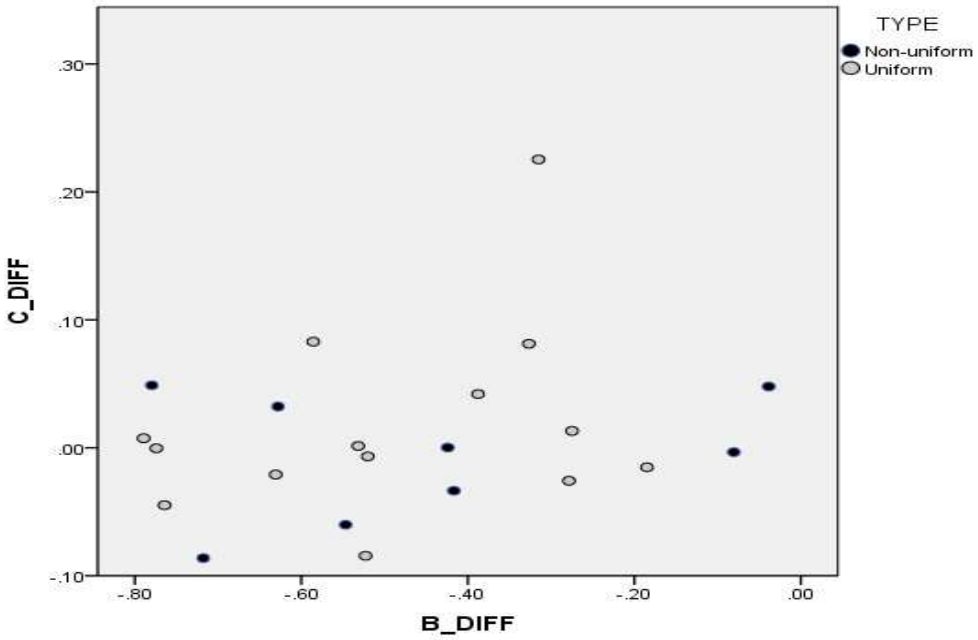


Figure B4. The difference in pseudo-guessing parameters versus the difference in difficulty parameters for all 22 focal-favoring DIF items.

Discrimination parameter. The ratios of the focal to the referent discrimination parameters were examined both for all 22 focal advantaging DIF items and for the 8 that were identified as exhibiting NU-DIF. These ratios as compared to the referent difficulty parameters and the difference between the difficulty parameters by DIF type are displayed in Figures 5 and 6, respectively. The ratios for the NU-DIF items had a larger range of values and, generally, had larger values than those for the U-DIF items. The one exception to this trend was the ratio of 0.37 for item # 66 which was also the item that was found to favor the focal group for 29% of the magnitude of β_{cro} while being flagged as having U-DIF favoring the referent group. There were three items that had difficulty differences of less than 0.25, which is the lower bound identified by Roussos et al. (1999) of the difference in difficulties required to obtain a Mantle-Haenszel Δ "B" flag in a 1PL model. These items had

discrimination ratios that ranged from 1.24 to 1.44, which indicates that the item discrimination parameter for the focal group was much larger than that for the reference group.

Similar to the comparison that was performed for the differences between difficulty parameters to the difficulty parameters for the referent group; the ratios of focal to referent discrimination parameters were compared with the discrimination parameters for the referent group as presented in Figure 7. Most of both types of DIF items discrimination ratios are grouped in a seemingly random distribution between the referent discrimination values of 1.17 and 2.33. While the referent discrimination parameters on the 84 item test ranged from 0.59 to a maximum of 3.12, all of the focal advantaging DIF items identified had referent discrimination parameters above 1.00, with a minimum of 1.17 and maximum of 3.12. The minimum referent discrimination parameter for NU-DIF items was even higher at 1.36.

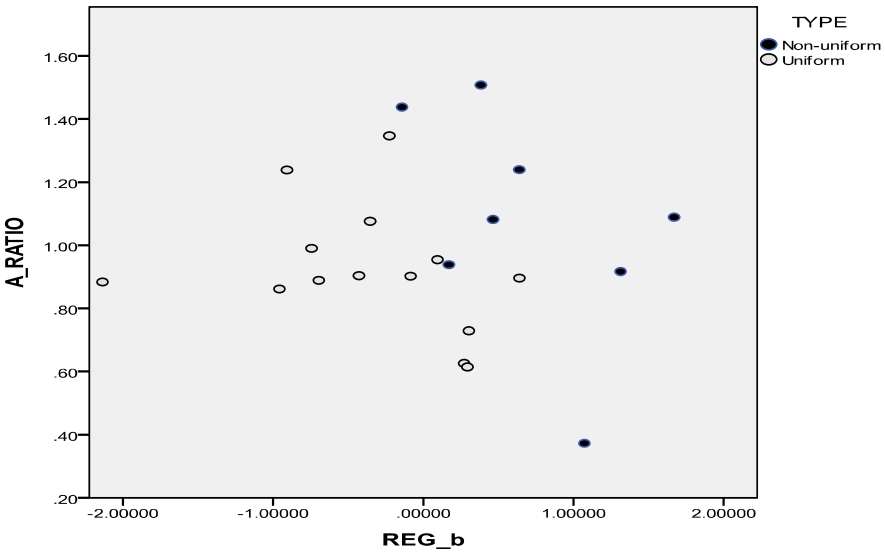


Figure B5. The ratio of the focal to the referent discrimination parameters versus the difficulty parameter for the referent group by DIF type

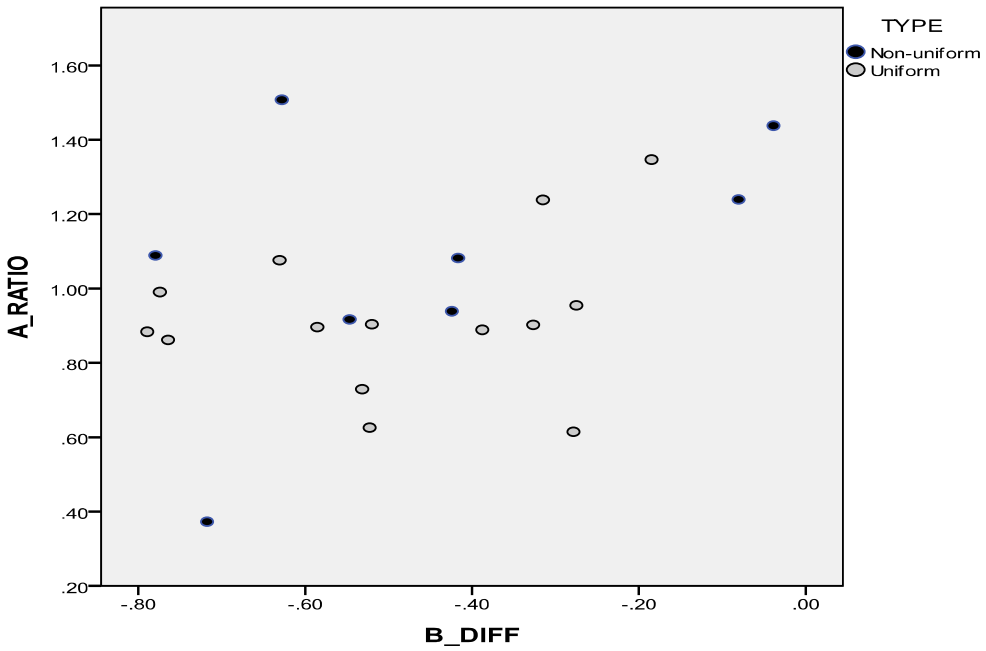


Figure B6. The ratio of the focal to the referent discrimination parameters versus the difference in difficulty parameters by DIF type.

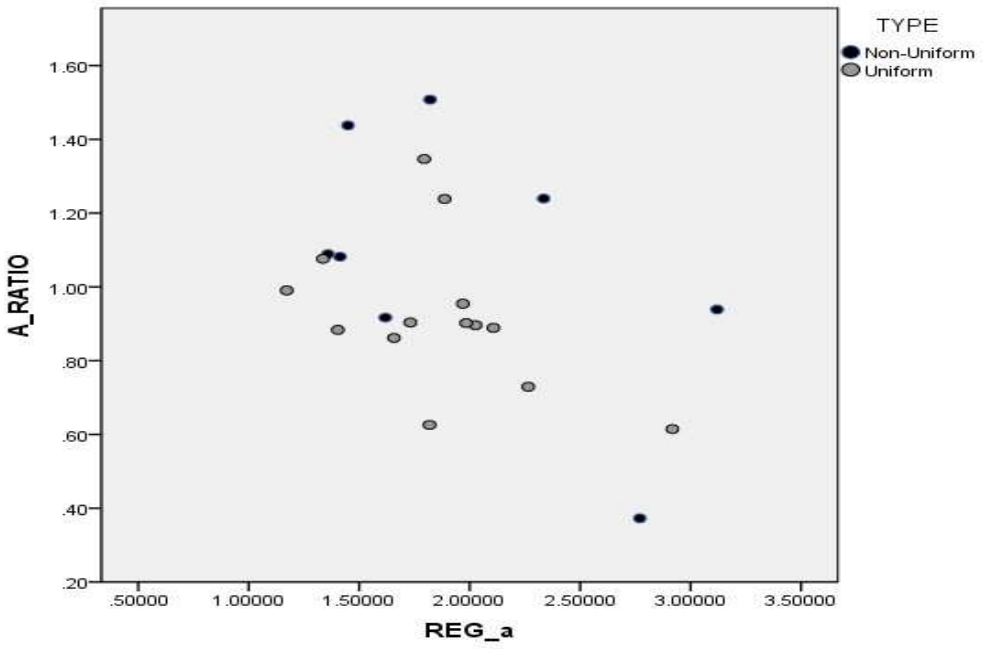


Figure B7. The ratio of the focal to the referent discrimination parameters versus the discrimination parameter for the referent group by DIF type.

Interaction of discrimination and pseudo-guessing parameters. To model NU-DIF with a 3PL IRT model, parameters need to be considered not only separately but also in combination. The ratios of the focal to referent discrimination parameters were examined in reference with the difference between the focal and the referent pseudo-guessing parameters as displayed in Figure 8. For NU-DIF items, the graph reveals that there are very different patterns of this interaction below and above the ratio of 1 for the discrimination parameters, where they are the same for both groups. Below this level, for all cases of NU-DIF items, the referent pseudo- guessing parameter (*c*) is less than or equal to zero (e.g., the discrimination ratio of 0.37 has a referent *c*-parameter of 0.373 and a focal *c*-parameter of 0.287 resulting in a *c*-parameter difference of -0.086). However, where the focal discrimination

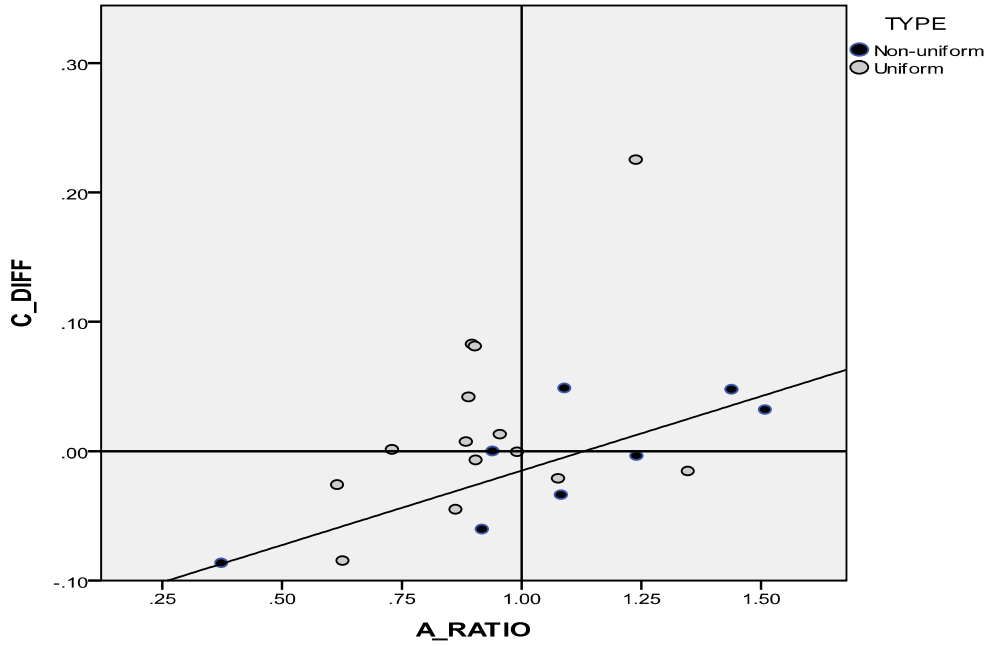


Figure B8. The difference in the pseudo-guessing parameters versus the ratio of the focal to the referent discrimination parameters by DIF type.

parameter is larger than that of the referent group (a ratio greater than 1), there seems to be a higher likelihood that the focal *c*-parameter will be larger than that of the associated referent parameter.

A linear regression analysis of the NU-DIF item discrimination parameter ratios to the associated differences of *c*-parameters was significant indicating that the two variables are linearly related. The scatterplot for the two variables in Figure 8, indicates that the variables are related such that as the ratio of NU-DIF item discrimination parameters for the focal to referent groups increases, the difference in the focal and referent pseudo-guessing parameters increases. The regression equation for predicting the *c*-parameter based on the ratio of discrimination parameters is

$$c\ parameter = .115\ discrimination\ ratio - .130$$

The 95% confidence interval for the slope, 0.033 to 0.197, does not contain the value of zero, and therefore the discrimination ratio is significantly related to the difference in *c*-parameters. The correlation between the discrimination ratio and *c*-parameter difference variables was 0.81 with approximately 66% of the variance of the *c*-parameter difference accounted for by the linear relationship with the ratio of discrimination parameters.

Parameter Modification Distributions.

This section presents the distributions from which the values used for modifying the referent item parameters to achieve focal advantaging DIF item parameters were drawn. These distributions were based mainly on the work of Roussos et al. (1999) as discussed above and on the parameters changes observed in the focal-advantaging AIMS DIF items.

The modification for difficulty parameters for U-DIF items will be addressed first with the discussion of the modifications for the difficulty, discrimination, and pseudo-guessing parameters for NU-DIF items following. In both cases, each item was modified from that for the referent group to exhibit two levels of DIF, one intended to be moderate and the other large.

Uniform DIF. For U-DIF modification, the difference in difficulty parameters was randomly drawn from a uniform distribution from -0.20 to -0.50 for moderate DIF and for large DIF from a uniform distribution from -0.50 to -0.80.

Non-Uniform DIF. The concern when modeling NU-DIF is to create differences in item characteristic curves (ICCs) that would ensure both that they cross at some point within the range of ability being studied and also that the resulting cross would advantage the focal group across more than an insignificant portion across that range. For the current study, that portion was defined as at least 20%. Therefore, the three parameter modifications, while being drawn separately from distributions, have those distributions influenced by the modification of the other parameters.

In an attempt to ensure that the ICCs crossed within the ability range of interest, lower differences in difficulty parameters were paired with higher ratios of discrimination parameters (greater than 1) and negative changes to the pseudo-guessing parameters (the Low *b*-difference modification). Similarly, higher differences in difficulty parameters were paired with lower ratios of discrimination parameters (less than 1) and positive changes to the pseudo-guessing parameters (the High *b*-difference modification). Each of the six items modified for NU-DIF was assigned one of the two modification schemes. The only criteria for placement within one or the other modification subset was that with average referent

difficulty parameters for each subset be as close to the same as possible. The resultant sets' average difficulty were equivalent at -0.06, however their standard deviations did vary (0.92 and 1.42). For each item, two changes for each parameter were randomly drawn from distributions detailed below, then the changes that would resulted in parameters closest to that of the referent group (closest to 0 change for difficulty difference, 1 for ratio of discriminations, and 0 for pseudo-guessing) were identified as the "moderate" DIF modification and used for Focal Group 2 and the remaining changes were combined to be identified as the "large" DIF used for Focal Group 1. After selection of the parameter changes, the resultant ICCs were examined and two (Items 20 and 60) failed to cross at any point within the ability range, -3.00 to 3.00. For these two items, the discrimination and pseudo-guessing parameter changes were redrawn until the curves were observed to cross. The change drawn to modify each item parameter for DIF for each group is presented in Table B4, the resultant item parameters for the referent and both focal groups for all of the DIF items are presented in Table B5, and the ICCs for the six NU-DIF items are presented in Appendix D.

Difference in difficulty parameters. The Low *b*-difference change of difficulty parameter was randomly drawn from a uniform distribution from -0.05 to -0.50 and the High *b*-difference change of difficulty parameter was randomly drawn from a uniform distribution from -0.50 to -0.80. This decreases lower bound of the distribution of the change in difficulty parameters from that of the uniform modification but keeps all of the other bounds consistent with that modification. It is also consistent with the differences of difficulty parameters found for non-uniform focal advantaging AIMS DIF items.

Table B4

Changes Drawn to Modify Item Parameters for DIF Items

	Focal 1			Focal 2		
#	a_{gi}	b_{gi}	c_{gi}	a_{gi}	b_{gi}	c_{gi}
9		-0.58			-0.27	
10*	1.40	-0.27	-0.08	1.21	-0.18	-0.05
19		-0.70			-0.29	
20**	0.69	-0.77	0.09	0.73	-0.55	0.04
29		-0.67			-0.24	
30*	1.43	-0.49	-0.10	1.14	-0.29	-0.09
39		-0.80			-0.49	
40**	0.72	-0.55	0.06	0.81	-0.53	0.03
49		-0.63			-0.25	
50*	1.45	-0.34	-0.09	1.21	-0.26	-0.05
59		-0.56			-0.40	
60**	0.66	-0.53	0.10	0.69	-0.50	0.01

Note: *assigned to the Low *b*-difference subset of NU-DIF items. ** assigned to the High *b*-difference subset of NU-DIF items.

Table B5

Item Parameters for DIF Items

	Referent			Focal 1			Focal 2		
#	a_{gi}	b_{gi}	c_{gi}	a_{gi}	b_{gi}	c_{gi}	a_{gi}	b_{gi}	c_{gi}
9	0.64	-1.55	.20	0.64	-2.13	.20	0.64	-1.82	.20
10*	0.75	-1.01	.20	1.05	-1.28	.12	0.91	-1.19	.15
19	1.02	1.28	.20	1.02	0.58	.20	1.02	0.99	.20
20**	0.78	-0.05	.20	0.54	-0.82	.29	0.57	-0.60	.24
29	1.05	0.10	.20	1.05	-0.57	.20	1.05	-0.14	.20
30*	0.51	-0.09	.20	0.73	-0.58	.10	0.58	-0.38	.11
39	0.81	-0.62	.20	0.81	-1.42	.20	0.81	-1.11	.20
40**	0.45	-1.49	.20	0.32	-2.04	.26	0.36	-2.02	.23
49	0.94	0.03	.20	0.94	-0.60	.20	0.94	-0.22	.20
50*	1.01	0.91	.20	1.46	0.57	.11	1.22	0.65	.15
59	0.94	0.25	.20	0.94	-0.31	.20	0.94	-0.15	.20
60**	1.12	1.35	.20	0.74	0.82	.30	0.77	0.85	.21

Note: *assigned to the Low *b*-difference subset of NU-DIF items. ** assigned to the High *b*-difference subset of NU-DIF items.

Ratio of discrimination parameters. As stated above, items in the Low *b*-difference subset were paired with ratios that were higher than 1. These were randomly drawn from a uniform distribution from 1.13 (the ratio used by Raju, 1990) to 1.50 (approximately the maximum observed in the AIMS data). The uniform distribution from

which the ratios were randomly drawn for the items in the High *b*-difference subset was the reciprocal of that used for the Low *b*-difference subset, 0.66 (1/1.50) to 0.88 (1/1.13).

Differences in pseudo-guessing parameters. For each item two random draws from a uniform distribution from 0.01 to 0.10 were used to modify the 0.20 value assigned for the referent pseudo-guessing item parameters. To achieve a negative value of this change for items in the Low b-difference subset the value drawn was multiplied by -1 before adding to the referent *c*-parameter.

APPENDIX C

ITEM CURVES FOR REAL-DATA NON-UNIFORM DIF ITEMS

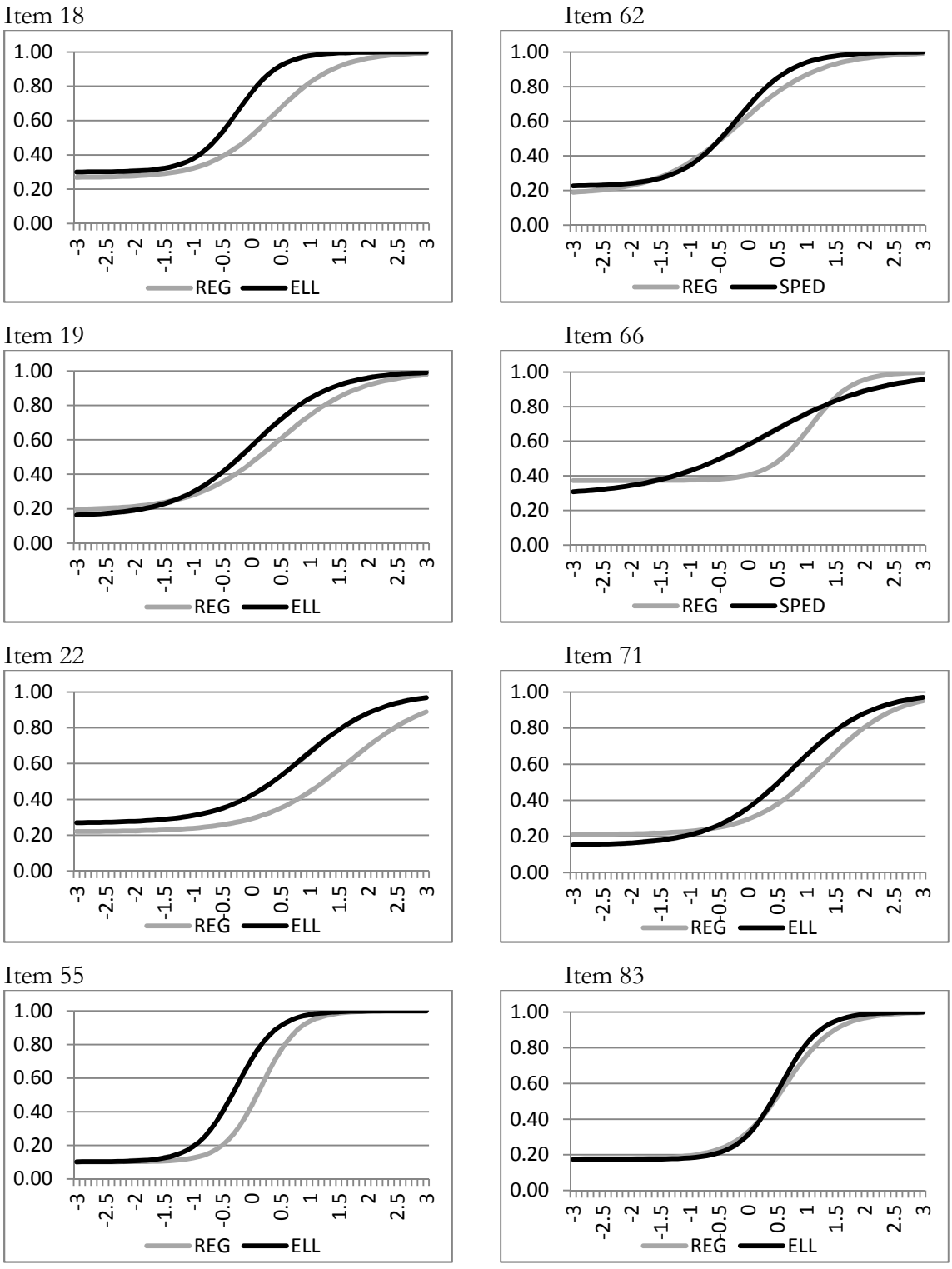


Figure C1. Item characteristic curves for the 8 AIMS-Mathematics non-uniform Focal-advantaging DIF items.

APPENDIX D

ITEM CURVES FOR SIMULATED NON-UNIFORM DIF ITEMS

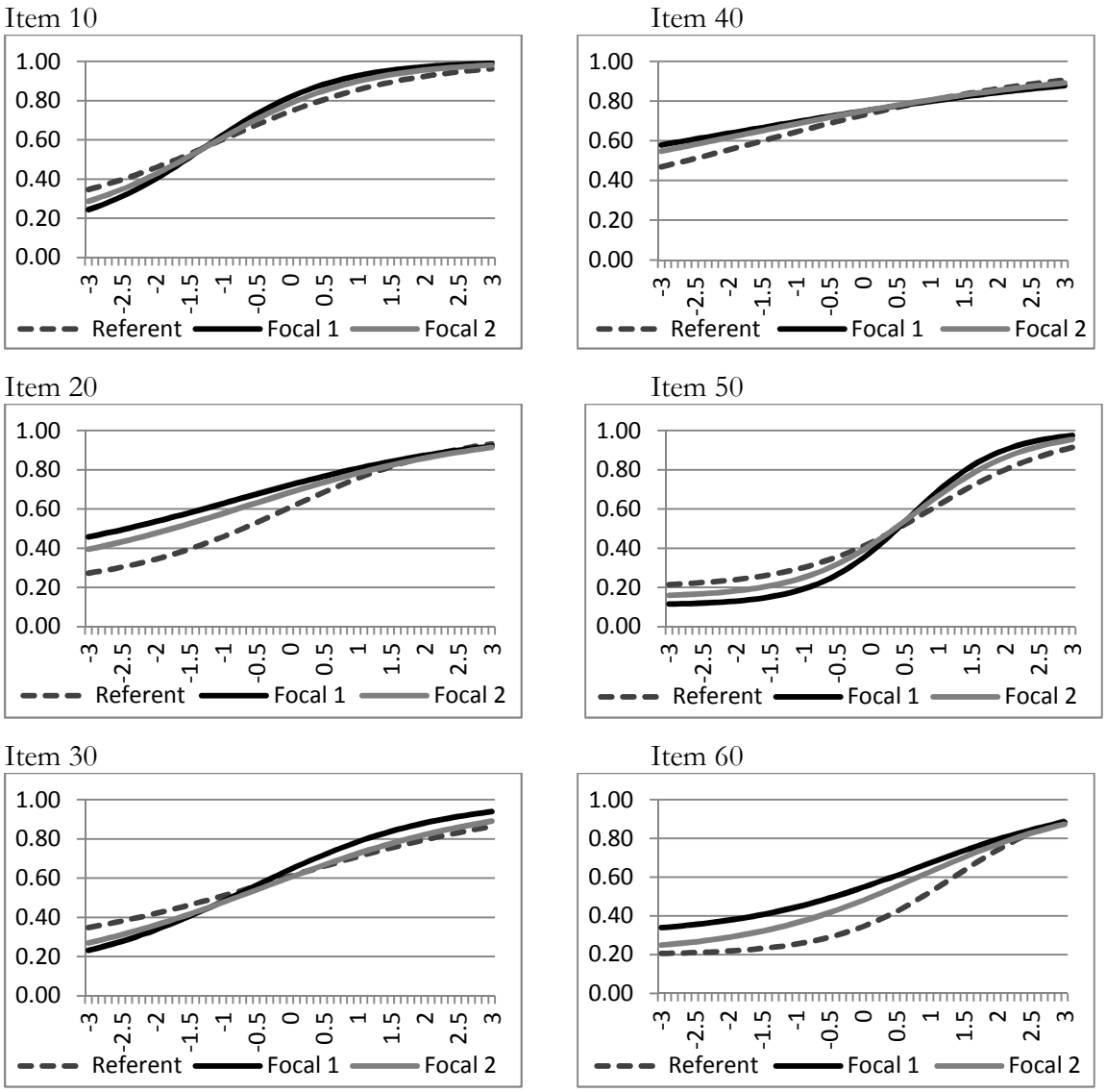


Figure D1. Item characteristic curves for the 6 simulated non-uniform Focal-advantaging DIF items.

APPENDIX E

TRANSFORMATION FOR NON-NORMAL DATA

Fleishman (1978) provided an equation and table of parameters that allowed for the transformation of the *N*(0,1) distribution to contain various degrees of skewness and kurtosis. His transformation equation, as simplified by Headrick and Sawilowsky (1999) was

$$T = a + bZ + (-a)Z^2 + dZ^3, \quad (24)$$

where *T* is the univariate case and Z ~ *N*(0, 1). Fleishman's parameters were computed through the use of a system of equations derived from formulas for the first four moments, arbitrarily setting the first and second moment at 0 and 1, respectively for convenience, and then solving for the third and fourth moments (γ_1, γ_2). These parameters, which are not available for all values of skewness and kurtosis, are then substituted into his transformation equation to produce *T* with a mean of 0, a variance of 1 and the designated skewness and

Table E1

Transformation Parameters for Focal Distributions

Skew	Kurtosis	a	b	d
Focal Group 1				
1	.5	-.2585	1.1147	-.0660
Focal Group 2				
.5	0	-.0926	1.0399	-.0165

Note: Following Headrick and Sawilowsky (1999), while Fleishman (1978) presented these parameters to 14 decimal places, they were rounded to four decimal places for use in this study.

kurtosis. The Fleishman parameters used within this study are displayed in Table E1. The kurtosis of .50 for skew of 1 was chosen since this was the least value of kurtosis available for the specified skew.

APPENDIX F

EXAMPLES OF COMPUTER CODE USED

R Code for generating dichotomous skewed or Normal distribution IRT 3-PL response data.

```
#####################################################
#########################
##
## Code for generating dichotomous IRT data.
##
#####################################################
#########################


#SPECIFY THE SAMPLE SIZE AND TEST LENGTH.
#SPECIFY TNE PERCENT OF FOCAL SIMULEES AND DISTRIBUTION
#SPECIFY THE NUMBER OF SIMULEES FOR AT
#SPECIFY THE MULTIDIMENSIONAL PARAMETERS FOR REFERENCE, AND
        TWO FOCAL GROUPS
#SPECIFY THE NUMBER OF REPLICATIONS.
#SPECIFY ALL SIMULATION PARAMETERS.


Sample = 2000 #2000 or 750
Per_Focal = 50 #50 or 10
Per_DIF = 20 #20
DIST = "Skew" #Norm or Skew
Num_AT = 750 #750 or 250
Num_DIF = Sample-Num_AT # = 1250 or 500
Num_Items = 60
nreps = 100 #number of replications.
Num_Groups = 3

# LOAD NECESSARY PACKAGES.

 library(MASS)
 library(mvtnorm)

for (Group in 1: Num_Groups) #A=1, B=2, or AB=3
{ #to create "Group Loop"

# To zero out previous numbers for AT & DIF for each group
Num_AT_A = 0
Num_AT_B = 0
Num_AT_R = 0
Num_DIF_A = 0
Num_DIF_B = 0
Num_DIF_R = 0
```

```
# To set new AT & DIF numbers for each group
Num_AT_R = Num_AT*(1-Per_Focal/100)
Num_DIF_R = Num_DIF*(1-Per_Focal/100)

if (Group==1)
   {Num_AT_A = Num_AT - Num_AT_R
         Num_AT_B = 0
         Num_DIF_A = Num_DIF - Num_DIF_R
    Num_DIF_B = 0 }

if (Group==2)
   {Num_AT_A = 0
    Num_AT_B = Num_AT - Num_AT_R
         Num_DIF_A = 0
         Num_DIF_B = Num_DIF - Num_DIF_R }

if (Group==3)
   {Num_AT_B = round((Num_AT - Num_AT_R)/2,0)
    Num_AT_A = (Num_AT-Num_AT_R)-Num_AT_B
         Num_DIF_B = round((Num_DIF - Num_DIF_R)/2,0)
         Num_DIF_A = (Num_DIF - Num_DIF_R) - Num_DIF_B }

#DIRECTORIES. You must specify where you want things to go in gen.dir.

condition = paste( "DATA", Group, Sample, Per_Focal, Per_DIF, DIST, sep="_")
gen.dir = paste("C:\\Users\\lmscot2\\Desktop\\Data\\", condition, "\\", sep="")
         #Desktop.

#create subdirectories.
dir.create(gen.dir) #create the general directory.
dir.create(paste(gen.dir, "items\\", sep="")) #store item parameter information.
dir.create(paste(gen.dir, "at_thetas\\", sep="")) #store theta values.
dir.create(paste(gen.dir, "at_data\\", sep="")) #store generated data files.
dir.create(paste(gen.dir, "DIF_thetas\\", sep="")) #store theta values.
dir.create(paste(gen.dir, "DIF_data\\", sep="")) #store generated data files.

item_fold=paste(gen.dir, "items\\", sep="") #NAME of item folder.
at_data_fold=paste(gen.dir, "at_data\\", sep="") #NAME of data folder.
at_theta_fold=paste(gen.dir, "at_thetas\\", sep="") #NAME of theta folder.
DIF_data_fold=paste(gen.dir, "DIF_data\\", sep="") #NAME of data folder.
DIF_theta_fold=paste(gen.dir, "DIF_thetas\\", sep="") #NAME of theta folder.

setwd(gen.dir)

reference_at_data = matrix(NA,Num_AT_R,Num_Items) #SET UP AT DATA MATRIX.
focus1_at_data = matrix(NA,Num_AT_A,Num_Items) #SET UP AT DATA MATRIX.
```

```
focus2_at_data = matrix(NA,Num_AT_B,Num_Items) #SET UP AT DATA MATRIX.
reference_DIF_data = matrix(NA,Num_DIF_R,Num_Items) #SET UP DIF DATA
        MATRIX.
focus1_DIF_data = matrix(NA,Num_DIF_A,Num_Items) #SET UP DIF DATA
        MATRIX.
focus2_DIF_data = matrix(NA,Num_DIF_B,Num_Items) #SET UP DIF DATA
        MATRIX.

# SET ITEM PARAMETERS FOR REFERENCE, FOCAL1 AND FOCAL2 GROUPS

# Non-DIF difficulty parameters.

  for (Num_Items in 1:Num_Items) bp = as.matrix(c(
              -2.95, -1.97, -2.63, -2.93, -1.77, -1.60, -1.21, -2.70,-1.55, -1.01,
               .61, -.57, -1.15, .60, -.30, -1.06, 1.02, -1.96,1.28, -.05,
              2.11, .81, 1.67, 1.68, -.23, -1.12, -1.37, -1.17,.10, -.09,
              1.26, .61, .95, 1.64, .82, 1.13, 1.18, -.75,-.62, -1.49,
               .80, -1.00, .64, 1.11, 2.12, 1.19, -1.41, .87,.03, .91,
               .35, -1.41, -1.29, .22, .93, .57, 1.11, 1.54,.25, 1.35), ncol=1, nrow=Num_Items)


# Non-DIF discrimination parameters.

  for (Num_Items in 1:Num_Items) ap = as.matrix(c
        (.29,.75,.36,.41,.56,.73,.94,.96,.64,.75,
        .82,.86,.42,.74,.44,.55,.82,.52,1.02,.78,
        1.04,1.01,.98,.65,.93,.35,.31,.39,1.05,.51,
        .55,.73,.88,1.40,1.35,.92,.73,.87,.81,.45,
        .50,.29,1.02,1.16,.48,.65,.79,.53,.94,1.01,
        1.11,.56,.59,1.01,.88,1.32,1.09,.83,.94,1.12), ncol=1, nrow=Num_Items)

# Non_DIF pseudo-guessing parameter.

  for (Num_Items in 1:Num_Items) cp = as.matrix(runif(Num_Items, min=.20, max=.20))

#WRITE OUT GENERATED PARAMETERS FOR REFERENCE, FOCUS 1, &
        FOCUS 2

item.id = matrix(seq(from=1, to=Num_Items, by=1), ncol=1, nrow=Num_Items)
items = matrix(cbind(item.id, ap, bp, cp), ncol=4, nrow=Num_Items)
items_reference <- items
items_focus1 <- items
items_focus2 <- items

for(z in 1:Num_Items)  # items_focus1
{
```

```
if(z==9)
{
        items_focus1[z,3] <- -2.13 # bp
}
if(z==10)
{
        items_focus1[z,2] <- 1.05  # ap
        items_focus1[z,3] <- -1.28 # bp
        items_focus1[z,4] <- .12   # cp
}
if(z==19)
{
        items_focus1[z,3] <- .58   # bp
}
if(z==20)
{
        items_focus1[z,2] <- .54   # ap
        items_focus1[z,3] <- -.82  # bp
        items_focus1[z,4] <- .29   # cp
}
if(z==29)
{
        items_focus1[z,3] <- -.57  # bp
}
if(z==30)
{
        items_focus1[z,2] <- .73   # ap
        items_focus1[z,3] <- -.58  # bp
        items_focus1[z,4] <- .10   # cp
}
if(z==39)
{
        items_focus1[z,3] <- -1.42 # bp
}
if(z==40)
{
        items_focus1[z,2] <- .32   # ap
        items_focus1[z,3] <- -2.04 # bp
        items_focus1[z,4] <- .26   # cp
}
if(z==49)
{
        items_focus1[z,3] <- -.60 # bp
}
if(z==50)
{
```

```
		items_focus1[z,2] <- 1.46 # ap
		items_focus1[z,3] <- .57  # bp
		items_focus1[z,4] <- .11  # cp
	}
	if(z==59)
	{
		items_focus1[z,3] <- -.31 # bp
	}
	if(z==60)
	{
		items_focus1[z,2] <- .74  # ap
		items_focus1[z,3] <- 0.82 # bp
		items_focus1[z,4] <- .30  # cp
	}
}

for(z in 1:Num_Items)  # items_focus2
{
	if(z==9)
	{
		items_focus2[z,3] <- -1.82  # bp
	}
	if(z==10)
	{
		items_focus2[z,2] <- .91  # ap
		items_focus2[z,3] <- -1.19  # bp
		items_focus2[z,4] <- .15   # cp
	}
	if(z==19)
	{
		items_focus2[z,3] <- .99  # bp
	}
	if(z==20)
	{
		items_focus2[z,2] <- .57  # ap
		items_focus2[z,3] <- -.60  # bp
		items_focus2[z,4] <- .24   # cp
	}
	if(z==29)
	{
		items_focus2[z,3] <- -.14  # bp
	}
	if(z==30)
	{
		items_focus2[z,2] <- 0.58  # ap
		items_focus2[z,3] <- -.38  # bp
```

```
		items_focus2[z,4] <- .11   # cp
	}
	if(z==39)
	{
		items_focus2[z,3] <- -1.11  # bp
	}
	if(z==40)
	{
		items_focus2[z,2] <- .36  # ap
		items_focus2[z,3] <- -2.02  # bp
		items_focus2[z,4] <- .23   # cp
	}
	if(z==49)
	{
		items_focus2[z,3] <- -.22  # bp
	}
	if(z==50)
	{
		items_focus2[z,2] <- 1.22  # ap
		items_focus2[z,3] <- .65  # bp
		items_focus2[z,4] <- .15   # cp
	}
	if(z==59)
	{
		items_focus2[z,3] <- -.15  # bp
	}
	if(z==60)
	{
		items_focus2[z,2] <- .77  # ap
		items_focus2[z,3] <- .85  # bp
		items_focus2[z,4] <- .21   # cp
	}
}

write.table(items_reference, paste(item_fold, "items_r", ".dat", sep=""),
col.names=c("item", "ap", "bp", "cp"),
row.names=F,
sep="\t",
quote=F)

write.table(items_focus1, paste(item_fold, "items_f1", ".dat", sep=""),
col.names=c("item", "ap", "bp", "cp"),
row.names=F,
sep="\t",
quote=F)
```

```
write.table(items_focus2, paste(item_fold, "items_f2", ".dat", sep=""),
col.names=c("item", "ap", "bp", "cp"),
row.names=F,
sep="\t",
quote=F)

for(which.rep in 1:nreps){

if (DIST == "Norm") {

at_theta_r = as.matrix(rnorm(Num_AT_R, mean=0, sd=1), ncol=1, nrow=Num_AT_R)

if (Num_AT_A > 0)
at_theta_f1 = as.matrix(rnorm(Num_AT_A, mean=0, sd=1), ncol=1, nrow=Num_AT_A)

if (Num_AT_B > 0)
at_theta_f2 = as.matrix(rnorm(Num_AT_B, mean=0, sd=1), ncol=1, nrow=Num_AT_B)

DIF_theta_r = as.matrix(rnorm(Num_DIF_R, mean=0, sd=1), ncol=1,
        nrow=Num_DIF_R)

if (Num_DIF_A > 0)
DIF_theta_f1 = as.matrix(rnorm(Num_DIF_A, mean=0, sd=1), ncol=1,
        nrow=Num_DIF_A)

if (Num_DIF_B > 0)
DIF_theta_f2 = as.matrix(rnorm(Num_DIF_B, mean=0, sd=1), ncol=1,
        nrow=Num_DIF_B)
}

if (DIST == "Skew") {

at_theta_r = as.matrix(rnorm(Num_AT_R, mean=0, sd=1), ncol=1, nrow=Num_AT_R)

if (Num_AT_A > 0){
at_theta_f1 = as.matrix(rnorm(Num_AT_A, mean=0, sd=1), ncol=1, nrow=Num_AT_A)
for(y in 1:Num_AT_A)  # Skew theta
{
        at_theta_f1[y,] <- ( (-1.5) + (-.2585)+ 1.1147*(at_theta_f1[y,])+
        .2585*((at_theta_f1[y,])^2) + (-.0660)*((at_theta_f1[y,])^3))
} }

if (Num_AT_B > 0){
at_theta_f2 = as.matrix(rnorm(Num_AT_B, mean=0, sd=1), ncol=1, nrow=Num_AT_B)
for(y in 1:Num_AT_B)  # Skew theta
{
```

```
        at_theta_f2[y,] <- ( (-1) +(-.0926)+ 1.0399*(at_theta_f2[y,])+
        .0926*((at_theta_f2[y,])^2) + (-.0165)*((at_theta_f2[y,])^3))
} }

DIF_theta_r = as.matrix(rnorm(Num_DIF_R, mean=0, sd=1), ncol=1,
        nrow=Num_DIF_R)

if (Num_DIF_A > 0){
DIF_theta_f1 = as.matrix(rnorm(Num_DIF_A, mean=0, sd=1), ncol=1,
        nrow=Num_DIF_A)
for(y in 1:Num_DIF_A)  # Skew theta
{
        DIF_theta_f1[y,] <- ( (-1.5)+(-.2585)+ 1.1147*(DIF_theta_f1[y,])+
        .2585*((DIF_theta_f1[y,])^2) + (-.0660)*((DIF_theta_f1[y,])^3))
} }

if (Num_DIF_B > 0){
DIF_theta_f2 = as.matrix(rnorm(Num_DIF_B, mean=0, sd=1), ncol=1,
        nrow=Num_DIF_B)
for(y in 1:Num_DIF_B)  # Skew theta
{
        DIF_theta_f2[y,] <- ( (-1)+ (-.0926)+ 1.0399*(DIF_theta_f2[y,])+
        .0926*((DIF_theta_f2[y,])^2) + (-.0165)*((DIF_theta_f2[y,])^3))
} } }

#write out separate theta files for the at files.
  write.table(at_theta_r,
    paste(at_theta_fold, "at_thetas_r_", which.rep, ".dat", sep=""),
    row.names=T,
    col.names=c( "              theta"),
    quote=F,sep="\t")

if (Num_AT_A > 0)
write.table(at_theta_f1,
    paste(at_theta_fold, "at_thetas_f1_", which.rep, ".dat", sep=""),
    row.names=T,
    col.names=c("              theta"),
    quote=F,sep="\t")

if (Num_AT_B > 0)
write.table(at_theta_f2,
    paste(at_theta_fold, "at_thetas_f2_", which.rep, ".dat", sep=""),
    row.names=T,
    col.names=c("              theta"),
    quote=F,sep="\t")
```

```
#write out separate theta files for the DIF files.
write.table(DIF_theta_r,
   paste(DIF_theta_fold, "DIF_thetas_r_", which.rep, ".dat", sep=""),
   row.names=T,
   col.names=c( "              theta"),
   quote=F,sep="\t")

if (Num_DIF_A > 0)
write.table(DIF_theta_f1,
   paste(DIF_theta_fold, "DIF_thetas_f1_", which.rep, ".dat", sep=""),
   row.names=T,
   col.names=c("              theta"),
   quote=F,sep="\t")

if (Num_DIF_B > 0)
write.table(DIF_theta_f2,
   paste(DIF_theta_fold, "DIF_thetas_f2_", which.rep, ".dat", sep=""),
   row.names=T,
   col.names=c("              theta"),
   quote=F,sep="\t")

#GENERATE THE DATA.
for(a in 1:Num_AT_R){
      for(Num_Items in 1:Num_Items){
            r_numerator = (1 - items_reference[Num_Items, 4])
            r_denominator = (1 + exp(-1.7*items_reference[Num_Items, 2]*(at_theta_r[a]-
         items_reference[Num_Items, 3])))
            r_p_at=(items_reference[Num_Items,4] + (r_numerator/r_denominator))
      r_uni_at = runif(1)
      if (r_p_at > r_uni_at) reference_at_data[a,Num_Items]=1
      if (r_p_at < r_uni_at) reference_at_data[a,Num_Items]=0
            } #closes IRT AT item loop.
   } #closes IRT AT person loop.

         if (Num_AT_A > 0) {
         for(a in 1:Num_AT_A){
      for(Num_Items in 1:Num_Items){
            f1_numerator = (1 - items_focus1[Num_Items, 4])
            f1_denominator = (1 + exp(-1.7*items_focus1[Num_Items, 2]*(at_theta_f1[a]-
         items_focus1[Num_Items, 3])))
            f1_p_at=(items_focus1[Num_Items,4] + (f1_numerator/f1_denominator))
      f1_uni_at = runif(1)
      if (f1_p_at > f1_uni_at) focus1_at_data[a,Num_Items]=1
      if (f1_p_at < f1_uni_at) focus1_at_data[a,Num_Items]=0
      } #closes IRT AT item loop.
   } #closes IRT AT person loop.
```

```
      } #closes if.

      if (Num_AT_B > 0) {
      for(a in 1:Num_AT_B){
  for(Num_Items in 1:Num_Items){
        f2_numerator = (1 - items_focus2[Num_Items, 4])
        f2_denominator = (1 + exp(-1.7*items_focus2[Num_Items, 2]*(at_theta_f2[a]-
      items_focus2[Num_Items, 3])))
        f2_p_at=(items_focus2[Num_Items,4] + (f2_numerator/f2_denominator))
  f2_uni_at = runif(1)
  if (f2_p_at > f2_uni_at) focus2_at_data[a,Num_Items]=1
  if (f2_p_at < f2_uni_at) focus2_at_data[a,Num_Items]=0
  } #closes IRT AT item loop.
} #closes IRT AT person loop.
      } #closes if.

      for(a in 1:Num_DIF_R){
  for(Num_Items in 1:Num_Items){
        r_DIF_numerator = (1 - items_reference[Num_Items, 4])
        r_DIF_denominator = (1 + exp(-1.7*items_reference[Num_Items,
      2]*(DIF_theta_r[a]-items_reference[Num_Items, 3])))
        r_p_DIF=(items_reference[Num_Items,4] + (r_numerator/r_denominator))
  r_uni_DIF = runif(1)
  if (r_p_DIF > r_uni_DIF) reference_DIF_data[a,Num_Items]=1
  if (r_p_DIF < r_uni_DIF) reference_DIF_data[a,Num_Items]=0
        } #closes IRT AT item loop.
} #closes IRT AT person loop.

      if (Num_DIF_A > 0) {
      for(a in 1:Num_DIF_A){
        for(Num_Items in 1:Num_Items){
        f1_DIF_numerator = (1 - items_focus1[Num_Items, 4])
        f1_DIF_denominator = (1 + exp(-1.7*items_focus1[Num_Items,
      2]*(DIF_theta_f1[a]-items_focus1[Num_Items, 3])))
        f1_p_DIF=(items_focus1[Num_Items,4] + (f1_numerator/f1_denominator))
  f1_uni_DIF = runif(1)
  if (f1_p_DIF > f1_uni_DIF) focus1_DIF_data[a,Num_Items]=1
  if (f1_p_DIF < f1_uni_DIF) focus1_DIF_data[a,Num_Items]=0
        } #closes IRT AT item loop.
} #closes IRT AT person loop.
      } #closes if.

      if (Num_DIF_B > 0) {
      for(a in 1:Num_DIF_B){
        for(Num_Items in 1:Num_Items){
        f2_DIF_numerator = (1 - items_focus2[Num_Items, 4])
```

```
            f2_DIF_denominator = (1 + exp(-1.7*items_focus2[Num_Items,
          2]*(DIF_theta_f2[a]-items_focus2[Num_Items, 3])))
            f2_p_DIF=(items_focus2[Num_Items,4] + (f2_numerator/f2_denominator))
      f2_uni_DIF = runif(1)
      if (f2_p_DIF > f2_uni_DIF) focus2_DIF_data[a,Num_Items]=1
      if (f2_p_DIF < f2_uni_DIF) focus2_DIF_data[a,Num_Items]=0
            } #closes IRT AT item loop.
   } #closes IRT AT person loop.
         } #closes if.

         if (Num_AT_B == 0)
at_data = as.matrix(rbind(reference_at_data,focus1_at_data), ncol=Num_Items,
         nrow=(Num_AT_A+Num_AT_R))

         if (Num_AT_A == 0)
at_data = as.matrix(rbind(reference_at_data,focus2_at_data), ncol=Num_Items,
         nrow=(Num_AT_B+Num_AT_R))

         if (Num_AT_A > 0)
         {
         if (Num_AT_B > 0)
at_data = as.matrix(rbind(reference_at_data,focus1_at_data, focus2_at_data),
         ncol=Num_Items, nrow=(Num_AT_A+Num_AT_B+Num_AT_R))     }

#write out data files for AT
write.table(at_data,
  paste(at_data_fold,"at_data_", which.rep, ".dat", sep=""),
  row.names=F,
  col.names=F,
  quote=F,   sep="")

if (Num_AT_B == 0)
f_DIF_data <- focus1_DIF_data

if (Num_AT_A == 0)
f_DIF_data <- focus2_DIF_data

if (Num_AT_A > 0)
  {
if (Num_AT_B > 0)
f_DIF_data = as.matrix(rbind(focus1_DIF_data, focus2_DIF_data), ncol=Num_Items,
         nrow=(Num_DIF_A+Num_DIF_B))    }

#write out data files for the DIF
write.table(reference_DIF_data,
  paste(DIF_data_fold,"DIF_data_r_", which.rep, ".dat", sep=""),
```

```
  row.names=F,
  col.names=F,
  quote=F,   sep="")

write.table(f_DIF_data,
  paste(DIF_data_fold,"DIF_data_f_", which.rep, ".dat", sep=""),
  row.names=F,
  col.names=F,
  quote=F,   sep="")

} #close nreps loop.
} #close group loop.
```

Batch Code for running ATFIND

```
@ECHO ON

set NUMBEROFFILESTOPROCESS=100
set /A PROCESSINGFILENUMBER=0
set /A sample=2000
set /A ratio=50
set /A difitems=0
set /A groups=3
set size=750
set guess=.2
set items=60

set OUTPUTDIR =
        C:\Users\lmscot2\Desktop\WData\DATA_%groups%_%sample%_%ratio%_%d
        ifitems%_Norm\atlist_output

MD %OUTPUTDIR%

:begin

set /A PROCESSINGFILENUMBER+=1

ECHO Processing ATFIND loop %PROCESSINGFILENUMBER% of
        %NUMBEROFFILESTOPROCESS%

DEL atlist.in

DEL atfind.in
ECHO
        C:\Users\lmscot2\Desktop\WData\DATA_%groups%_%sample%_%ratio%_%d
        ifitems%_Norm\at_data\at_data_%PROCESSINGFILENUMBER%.dat >>
        .\atfind.in
ECHO %size% >>.\atfind.in
ECHO %items%  >>.\atfind.in
ECHO %guess% >>.\atfind.in

atfind_v.1.3.exe

TYPE .\atlist.in >> %outputdir%\atlist%PROCESSINGFILENUMBER%.in

IF %PROCESSINGFILENUMBER% LSS %NUMBEROFFILESTOPROCESS% GOTO
        begin
:EndOfProgram
ECHO Finished.
```

Code for running Mantel-Haenszel with a designated matching subtest using difR package

```
#########################################################
##
## Code for running Mantel Haenszel with a designated matching subtest using difR
## package.
##
#########################################################

#SPECIFY THE SAMPLE SIZE AND TEST LENGTH.
#SPECIFY TNE PERCENT OF FOCAL SIMULEES AND DISTRIBUTION
#SPECIFY THE NUMBER OF SIMULEES FOR AT
#SPECIFY THE NUMBER OF REPLICATIONS.
#SPECIFY ALL SIMULATION PARAMETERS.

Sample = 750 #2000 or 750
Per_Focal = 50 #50 or 10
Per_DIF = 20 #0, 10, or 20
DIST = "Skew" #Norm or Skew
Num_AT = 250 #750 or 250
Num_DIF = Sample-Num_AT # = 1250 or 500
Num_Items = 60
nreps = 100 #number of replications.
Num_Groups = 3
Group = 3

# LOAD NECESSARY PACKAGES.

 library(difR)
 source ("C:\\Users\\lmscot2\\Desktop\\difMH_ls.r")

#DIRECTORIES. You must specify where you want things to go in gen.dir.

condition = paste( "DATA", Group, Sample, Per_Focal, Per_DIF, DIST, sep="_")
gen.dir = paste("C:\\Users\\lmscot2\\Desktop\\Data_WORKING\\", condition,
        "\\DIF_combined\\", sep="") #Desktop.

#create subdirectory.

dir.create(paste("C:\\Users\\lmscot2\\Desktop\\Data_WORKING\\", condition,
        "\\MH_RESULTS\\", sep="")) #store results from MH.

MH_RESULTS_fold=paste("C:\\Users\\lmscot2\\Desktop\\Data_WORKING\\",
        condition, "\\MH_RESULTS\\", sep="") #NAME of MH_RESULTS folder.
DIF_combined_fold=paste("C:\\Users\\lmscot2\\Desktop\\Data_WORKING\\",
        condition, "\\DIF_combined\\", sep="") #NAME of DIF_Combined folder.
```

```
setwd(gen.dir)

anchor_1

anchor_default <-
        c(1,2,3,4,5,6,8,11,12,13,14,15,16,17,21,22,23,24,26,27,28,31,32,33,34,35,36,38,41,42,
          43,44,45,47,48,51,52,53,54,55,56,57)

for (iteration in 1:nreps) {

input_file = paste("DIF_data_combined_",iteration,".dat",sep="")
input_file_route = paste(DIF_combined_fold, input_file, sep="")

MH_output_AT=paste( "difMH_ATout_", iteration, sep="")
MH_output_Default= paste("difMH_DEFout_", iteration, sep="")

read_file <- as.matrix(read.fortran(input_file_route,"62I1"))

response = matrix(NA,Num_DIF,Num_Items) #SET UP AT DATA MATRIX.
Focal = matrix(NA,Num_DIF,1) #SET UP AT DATA MATRIX.

for (Person in 1:Num_DIF) {
        for (Item in 1:60){
                response[Person, Item] <- read_file [Person,Item]
        } # coses loop over i
} # closes loop over person

for (Person in 1:Num_DIF) {
        Focal[Person] <- read_file [Person,62]
} # closes loop over person

# Reference ="Ref", F1 & F2 = Focal
use_names<-c("Ref","Focal")

if (iteration==1)
  {anchor_list <- anchor_1}
if (iteration==2)
  {anchor_list <- anchor_2}
if (iteration==3)
  {anchor_list <- anchor_3}
if (iteration==4)
  {anchor_list <- anchor_4}
if (iteration==5)
  {anchor_list <- anchor_5}
if (iteration==6)
  {anchor_list <- anchor_6}
```

```
if (iteration==7)
   {anchor_list <- anchor_7}
if (iteration==8)
   {anchor_list <- anchor_8}
if (iteration==9)
   {anchor_list <- anchor_9}
if (iteration==10)
   {anchor_list <- anchor_10}
if (iteration==11)
   {anchor_list <- anchor_11}
if (iteration==12)
   {anchor_list <- anchor_12}
if (iteration==13)
   {anchor_list <- anchor_13}
if (iteration==14)
   {anchor_list <- anchor_14}
if (iteration==15)
   {anchor_list <- anchor_15}
if (iteration==16)
   {anchor_list <- anchor_16}
if (iteration==17)
   {anchor_list <- anchor_17}
if (iteration==18)
   {anchor_list <- anchor_18}
if (iteration==19)
   {anchor_list <- anchor_19}
if (iteration==20)
   {anchor_list <- anchor_20}
if (iteration==21)
   {anchor_list <- anchor_21}
if (iteration==22)
   {anchor_list <- anchor_22}
if (iteration==23)
   {anchor_list <- anchor_23}
if (iteration==24)
   {anchor_list <- anchor_24}
if (iteration==25)
   {anchor_list <- anchor_25}
if (iteration==26)
   {anchor_list <- anchor_26}
if (iteration==27)
   {anchor_list <- anchor_27}
if (iteration==28)
   {anchor_list <- anchor_28}
if (iteration==29)
   {anchor_list <- anchor_29}
```

```
if (iteration==30)
   {anchor_list <- anchor_30}
if (iteration==31)
   {anchor_list <- anchor_31}
if (iteration==32)
   {anchor_list <- anchor_32}
if (iteration==33)
   {anchor_list <- anchor_33}
if (iteration==34)
   {anchor_list <- anchor_34}
if (iteration==35)
   {anchor_list <- anchor_35}
if (iteration==36)
   {anchor_list <- anchor_36}
if (iteration==37)
   {anchor_list <- anchor_37}
if (iteration==38)
   {anchor_list <- anchor_38}
if (iteration==39)
   {anchor_list <- anchor_39}
if (iteration==40)
   {anchor_list <- anchor_40}
if (iteration==41)
   {anchor_list <- anchor_41}
if (iteration==42)
   {anchor_list <- anchor_42}
if (iteration==43)
   {anchor_list <- anchor_43}
if (iteration==44)
   {anchor_list <- anchor_44}
if (iteration==45)
   {anchor_list <- anchor_45}
if (iteration==46)
   {anchor_list <- anchor_46}
if (iteration==47)
   {anchor_list <- anchor_47}
if (iteration==48)
   {anchor_list <- anchor_48}
if (iteration==49)
   {anchor_list <- anchor_49}
if (iteration==50)
   {anchor_list <- anchor_50}
if (iteration==51)
   {anchor_list <- anchor_51}
if (iteration==52)
   {anchor_list <- anchor_52}
```

```
if (iteration==53)
   {anchor_list <- anchor_53}
if (iteration==54)
   {anchor_list <- anchor_54}
if (iteration==55)
   {anchor_list <- anchor_55}
if (iteration==56)
   {anchor_list <- anchor_56}
if (iteration==57)
   {anchor_list <- anchor_57}
if (iteration==58)
   {anchor_list <- anchor_58}
if (iteration==59)
   {anchor_list <- anchor_59}
if (iteration==60)
   {anchor_list <- anchor_60}
if (iteration==61)
   {anchor_list <- anchor_61}
if (iteration==62)
   {anchor_list <- anchor_62}
if (iteration==63)
   {anchor_list <- anchor_63}
if (iteration==64)
   {anchor_list <- anchor_64}
if (iteration==65)
   {anchor_list <- anchor_65}
if (iteration==66)
   {anchor_list <- anchor_66}
if (iteration==67)
   {anchor_list <- anchor_67}
if (iteration==68)
   {anchor_list <- anchor_68}
if (iteration==69)
   {anchor_list <- anchor_69}
if (iteration==70)
   {anchor_list <- anchor_70}
if (iteration==71)
   {anchor_list <- anchor_71}
if (iteration==72)
   {anchor_list <- anchor_72}
if (iteration==73)
   {anchor_list <- anchor_73}
if (iteration==74)
   {anchor_list <- anchor_74}
if (iteration==75)
   {anchor_list <- anchor_75}
```

```
if (iteration==76)
   {anchor_list <- anchor_76}
if (iteration==77)
   {anchor_list <- anchor_77}
if (iteration==78)
   {anchor_list <- anchor_78}
if (iteration==79)
   {anchor_list <- anchor_79}
if (iteration==80)
   {anchor_list <- anchor_80}
if (iteration==81)
   {anchor_list <- anchor_81}
if (iteration==82)
   {anchor_list <- anchor_82}
if (iteration==83)
   {anchor_list <- anchor_83}
if (iteration==48)
   {anchor_list <- anchor_84}
if (iteration==85)
   {anchor_list <- anchor_85}
if (iteration==86)
   {anchor_list <- anchor_86}
if (iteration==87)
   {anchor_list <- anchor_87}
if (iteration==88)
   {anchor_list <- anchor_88}
if (iteration==89)
   {anchor_list <- anchor_89}
if (iteration==90)
   {anchor_list <- anchor_90}
if (iteration==91)
   {anchor_list <- anchor_91}
if (iteration==92)
   {anchor_list <- anchor_92}
if (iteration==93)
   {anchor_list <- anchor_93}
if (iteration==94)
   {anchor_list <- anchor_94}
if (iteration==95)
   {anchor_list <- anchor_95}
if (iteration==96)
   {anchor_list <- anchor_96}
if (iteration==97)
   {anchor_list <- anchor_97}
if (iteration==98)
   {anchor_list <- anchor_98}
```

```
if (iteration==99)
   {anchor_list <- anchor_99}
if (iteration==100)
   {anchor_list <- anchor_100}

anchor_list_DEFAULT <- anchor_default

difMH_ls (response[,1:60], group=Focal[,1], focal.name=1, anchor_list, alpha=0.05, exact =
          TRUE, purify = FALSE, save.output = TRUE,
      output=c(MH_output_AT,MH_RESULTS_fold))

difMH_ls (response[,1:60], group=Focal[,1], focal.name=1, anchor_list_DEFAULT,
          alpha=0.05, exact = TRUE, purify = FALSE, save.output = TRUE,
      output=c(MH_output_Default,MH_RESULTS_fold))
} # Closes loop over iteration
```

Code for modifying difMH that is called within difR to allow for specifying particular items as the matching subtest.

```
#########################################################
#
# difMH_ls
# This is a modification of difMH within difR (Magis, Beland, & Raiche, 2013)
# to allow for specifying only a particular anchor set.
#
#########################################################

difMH_ls <- function (Data, group, focal.name,anchor_list,MHstat = "MHChisq", correct =
TRUE,
   exact = FALSE, alpha = 0.05, purify = FALSE, nrIter = 10,
   save.output = FALSE, output = c("out", "default"))
{
   internalMH <- function() {
      if (length(group) == 1) {
         if (is.numeric(group) == TRUE) {
            gr <- Data[, group]
            DATA <- Data[, (1:ncol(Data)) != group]
            colnames(DATA) <- colnames(Data)[(1:ncol(Data)) !=
              group]
         }
         else {
            gr <- Data[, colnames(Data) == group]
            DATA <- Data[, colnames(Data) != group]
            colnames(DATA) <- colnames(Data)[colnames(Data) !=
              group]
         }
      }
      else {
         gr <- group
         DATA <- Data
      }
      Group <- rep(0, nrow(DATA))
      Group[gr == focal.name] <- 1
      Q <- switch(MHstat, MHChisq = qchisq(1 - alpha, 1), logOR = qnorm(1 -
         alpha/2))
      if (is.null(Q))
         stop("'MHstat' argument not valid", call. = FALSE)
      if (exact) {
         if (!purify) {
            PROV <- mantelHaenszel(DATA, Group, correct = correct,  anchor =
anchor_list,
               exact = exact)
```

```
        STATS <- PROV$resMH
        if (min(PROV$Pval) >= alpha)
          DIFitems <- "No DIF item detected"
        else DIFitems <- (1:ncol(DATA))[PROV$Pval < alpha]
        RES <- list(MH = STATS, Pval = PROV$Pval, alpha = alpha,
          DIFitems = DIFitems, correct = correct, exact = exact,
          purification = purify, names = colnames(DATA),
          save.output = save.output, output = output)
      }
      else {}
    }
    else {
      if (!purify) {
        PROV <- mantelHaenszel(DATA, Group, correct = correct,
          exact = exact)
        if (MHstat == "MHChisq")
          STATS <- PROV$resMH
        else STATS <- log(PROV$resAlpha)/sqrt(PROV$varLambda)
        if (max(abs(STATS), na.rm = TRUE) <= Q)
          DIFitems <- "No DIF item detected"
        else DIFitems <- (1:ncol(DATA))[is.na(STATS) ==
          FALSE & abs(STATS) > Q]
        RES <- list(MH = STATS, alphaMH = PROV$resAlpha,
          varLambda = PROV$varLambda, MHstat = MHstat,
          alpha = alpha, thr = Q, DIFitems = DIFitems,
          correct = correct, exact = exact, purification = purify,
          names = colnames(DATA), save.output = save.output,
          output = output)
      }
      else {}
    }
    class(RES) <- "MH"
    return(RES)
}
resToReturn <- internalMH()
if (save.output == TRUE) {
    if (output[2] == "default")
      wd <- paste(getwd(), "/", sep = "")
    else wd <- output[2]
    fileName <- paste(wd, output[1], ".txt", sep = "")
    capture.output(resToReturn, file = fileName)
}
return(resToReturn)}
```

SIBTEST '.in" code for running SIBTEST with the pure matching subtest.

```
60
"C:\Data\DATA_2_2000_50_20_Norm\DIF_data\DIF_data_r_1.dat"
"C:\Data\DATA_2_2000_50_20_Norm\DIF_data\DIF_data_f_1.dat"
1
"C:\Data\DATA_2_2000_50_20_Norm\sib_Def\sibDef_out_1"
2
20
0
0

1
7
'e'
40
1   2   3   4   5 6   8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
9
'e'
40
1   2   3   4   5 6   8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
10
'e'
40
1   2   3   4   5 6   8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2
```

1
18
'e'
40
1 2 3 4 5 6 8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
19
'e'
40
1 2 3 4 5 6 8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
20
'e'
40
1 2 3 4 5 6 8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
25
'e'
40
1 2 3 4 5 6 8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
28
'e'
40
1 2 3 4 5 6 8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
29
'e'
40
1 2 3 4 5 6 8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
30
'e'
40
1 2 3 4 5 6 8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
37
'e'
40
1 2 3 4 5 6 8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
39
'e'
40
1 2 3 4 5 6 8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
40
'e'
40
1 2 3 4 5 6 8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
46
'e'
40
1 2 3 4 5 6 8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
49
'e'
40
1 2 3 4 5 6 8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
50
'e'
40
1 2 3 4 5 6 8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
54
'e'
40
1 2 3 4 5 6 8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
58
'e'
40
1 2 3 4 5 6 8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
59
'e'
40
1 2 3 4 5 6 8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

1
60
'e'
40
1 2 3 4 5 6 8 11 12 13
14 15 16 17 21 22 23 24 26 27
31 32 33 34 35 36 38 41 42 43
44 45 47 48 51 52 53 55 56 57
0.2

Batch code for running SIBTEST and Crossing SIBTEST

```
@ECHO ON

set NUMBEROFFILESTOPROCESS=100
set /A PROCESSINGFILENUMBER=0

:begin

set /A PROCESSINGFILENUMBER+=1

ECHO Processing SIBTEST loop %PROCESSINGFILENUMBER% of
        %NUMBEROFFILESTOPROCESS%

DEL sib.in
copy .\sib_AT\sibAT%PROCESSINGFILENUMBER%.in sib.in

nsibtfl2.exe

DEL sib.in
copy .\sib_Def\sibDef%PROCESSINGFILENUMBER%.in sib.in

nsibtfl2.exe

DEL sib.in
copy .\csib_AT\csibAT%PROCESSINGFILENUMBER%.in sib.in

csib2.exe

DEL sib.in
copy .\csib_Def\csibDef%PROCESSINGFILENUMBER%.in sib.in

csib2.exe

IF %PROCESSINGFILENUMBER% LSS %NUMBEROFFILESTOPROCESS% GOTO
        begin

:EndOfProgram

ECHO Finished
```

APPENDIX G

STUDY TO VERIFY GENERATED RESPONSE DATA

The surprising result of the overall observed 40 - 60 percent split for AT-PT items regardless of condition, DIF status, or region of difficulty for the item lead to an examination of what other analyses might find within the data and a desire to confirm that the data had been generated to the desired specifications. A random iteration (run number 10) of the condition with the largest AT sample (750), largest difference in the modification of DIF items (referent versus Focal Group 1), largest number of DIF items (20%), equal proportion of referent and focal group simulees, and both groups of simulee abilities drawn from normal $N(0,1)$ distributions was chosen for four additional analyses. The iteration was selected without regard to ATFIND results. For the selected iteration, ATFIND placed 5 DIF items (2 modified for U-DIF and 3 modified for NU-DIF, 41.7%) in the ATLIST and the remaining 7 DIF items (58.3%) in the PTLIST. Four additional analyses were performed on this dataset: 1) a confirmatory factor analysis using Mplus (Muthen, & Muthen, 1998), 2) an exploratory principal axis factor analysis using SPSS (SPSS, Inc., 1989), 3) a MH analysis (difMH within the difR package, Magis, Beland, & Raiche, 2013) using R (The R Project for Statistical Computing, 2012), and 4) a parameter recovery check using BILOG-MG 3 (Zimowski, Muraki, Mislevy, & Bock, 2003). The results of these additional analyses are discussed in order below.

The confirmatory factor analysis undertaken with MPLUS (Muthen & Muthen, 1998) revealed a unidimensional structure as would be expected given that the data were generated using a unidimensional 3-parameter IRT model. Both the one factor and two factor models (second factor made up of the items modified to exhibit DIF) exhibited good fit (RMSEA = 0.009, CFI = 0.985, and TFI= 0.984 for both, with a weighted root mean square residual of 0.932 and 0.931, respectively). Within the two factor analysis, the first

factor (made up of non-DIF modified items) had a 0.966 correlation to the second factor, which indicates that the two factors function essentially as one. The covariance matrix however, revealed that the second factor, made up of items modified for DIF, had a larger variance (0.178) than that of the first factor (0.047) indicating that while the factors might be highly correlated; the items with which they are comprised function very differently. An analysis of a three factor structure with the first factor again non-DIF modified items, a second factor of U-DIF modified items, and a third factor of NU-DIF modified items produced a latent variable covariance matrix that was not positive definitive due to estimated correlations of the third factor with both of the other two factors of over the maximum at 1.008 and 1.047, respectively.

When the exploratory principal axis factor analysis using SPSS (SPSS, Inc., 1989) was performed the DIF modified items loaded most heavily on three different rotated factors. This analysis was performed using Varimax rotation with Kaiser normalization. The first factor, consisted of 24 of the 60 items and contained all of the U-DIF modified items and three of the NU-DIF modified items (initial Eigenvalue = 6.823) accounted for 11.37% of the variance observed. The second factor which consisted of 18 of the 60 items (initial Eigenvalue = 1.623) and contained two of the remaining three NU-DIF items, accounted for 2.70% of the remaining variance. Where the third factor which contained the remaining 18 items but only the 1 remaining NU-DIF modified item had a slightly lower initial Eigenvalue (1.475) and accounted for a slightly lower amount (2.58) of the remaining variance. Since 37.5% of the first factor is comprised of DIF modified items (all located within the top 17 rotated factor loadings) it might not seem unreasonable to name this factor as the DIF factor with the other two factors collapsed to one "non-DIF factor" since the

addition of the third factor does not increase the amount of variance accounted for greatly. Interestingly, this 24/36 split is exactly the 40-60 percent split generally found between the AT and PT lists within the ATFIND analyses.

To examine if the items had actually been modified for DIF within the response generation, a MH analysis (difMH within the difR package, Magis, Beland, & Raiche, 2013) coded in R (The R Project for Statistical Computing, 2012) was performed using the available standard purification process. MH identified 14 items as exhibiting U-DIF, 5 of which had not been modified as well as all 6 items that had been modified to exhibit U-DIF and 3 (Item Numbers 10, 20, and 60) of the 6 that had been modified to exhibit NU-DIF. Two of the NU-DIF modified items that the analysis did not identify (Item Numbers 30 and 50) had both been modified using a small difference between the referent and focal difficulty parameters (less than 0.50) and an increase from the referent discrimination parameters (by a factor of over 1.40). Given that MH is primarily a U-DIF detection analysis; its insensitivity to these smaller differences in difficulty parameter might be expected (Donoghue, Holland, & Thayer, 1993). The last NU-DIF modified item that MH failed to detect (Item Number 40) had been coded to have a slightly higher difference in difficulty parameter (-0.55) and had a discrimination parameter that had been reduced by a factor of 0.72 from that of the referent parameters. The changes to this item's parameters resulted in very low discrimination and difficulty parameters (a = 0.324, b= -2.04, respectively) from which to generate focal simulee responses. Some researchers have found that this combination of low discrimination and low difficulty, especially paired with a relatively low difference between the difficulty parameters of the referent and focal groups, impacts MH's ability to identify DIF items (Donoghue, Holland, & Thayer, 1993; Rogers, & Swaminathan, 1993).

As an additional check of the data generation process that produced the data that was analyzed by ATFIND, an estimation, and following equating, of parameters was performed using BILOG-MG 3 (Zimowski, Muraki, Mislevy, & Bock, 2003). To achieve estimates for all three item parameters (difficulty, discrimination, and pseudo-guessing) for each of the groups (referent and focal), the ATFIND simulee responses for iteration number 10 used for the other analyses was separated by group and then analyzed separately. The parameter estimates for the two groups were then equated using a non-equivalent group anchor test (NEAT) design (Cook, & Eignor, 1991) with the 48 non-DIF modified comprising 80% of the test as the anchor. The difference between the expected difference in parameters between referent and focal groups (based on item response generation code) and the observed difference in parameters between the two groups as they related to the distribution of simulee abilities was explored.

For the difficulty parameter, the magnitude of the difference between the observed and expected difficulty parameter difference was generally low (less than 0.15) in the area of the simulee ability distribution where the most simulees might be expected in a random draw between -1.01 to 0.91. The magnitude increased markedly, however, outside this region. There was one exception to the general trend, NU-DIF modified Item Number 30 (simulee ability, referent-generated difficulty of -0.09) which was expected to have a focal to referent parameter difference of -0.49 was observed to have a difference of -0.85 resulting in a magnitude difference of 0.36.

The analysis of discrimination parameter changes led to similar, though slightly more discrepant findings. Here again, the general trend was for the magnitude of difference between the expected and observed changes in the parameter to be less than 0.15 in the

central region of the ability scale and increase the referent item difficulty increased or decreased (where fewer and fewer simulees might be expected to be selected in a random draw from a normal distribution). With this parameter, there were three items in the middle of the scale that showed abnormally high (greater than 0.20 magnitude) difference between expected and observed difference. While two of the items (both NU-DIF modified, Item Number 20 and again Item Number 30) had a difference in the moderate range, 0.23 and 0.25, respectively, one item (U-DIF modified Item Number 49, with a referent difficulty location of 0.03) had an observed difference for the discrimination parameter of 0.46 where none was expected.

The magnitude of difference between the observed and expected differences between the referent and focal group for the pseudo-guessing parameter were all less than 0.10 with only a very slight increase noticeable for NU-DIF modified items as referent item difficulty/simulee ability increased. For U-DIF modified items, the magnitude of difference was consistently less than 0.05 which is very close to the expect 0 change for these items.

There are indications that a mismatch between item location and ability distribution increases the error of estimation in item parameter recovery (Hulin, Lissak, & Drasgow, 1982; Baker, 1987) and in particular that BILOG-MG 3 analysis' performance is dependent on both the distribution of examinees and the location of the items (Mislevy, & Stocking, 1989; Toland, 2008). Therefore, the simulee-distribution, dependent, results observed in the estimation of the difficulty and discrimination parameter changes may be a function not of the data generation but of the estimation procedure. The results of both of the factor analyses and of the MH DIF analysis seem to support this alternative hypotheses, that most

if not all of the items intended to be modified to exhibit DIF for this random iteration of this selected condition actually were.

APPENDIX H

TABULAR RESULTS FOR TYPE I ERROR RATES

Table H1a

Total Type I Error Rate for the 60 Non-DIF Modified Items in 0% DIF.

Aggregation	Mantel-Haenszel		SIBTEST		Crossing SIBTEST		Both SIBTESTS	
	Best	PT	Best	PT	Best	PT	Best	PT
Overall	**6.5**	**6.7**	**2.8**	**3.5**	**2.3**	**3.0**	**4.4**	**5.4**
***N*=500**	**5.3**	**5.5**	**2.9**	**3.5**	**2.2**	**2.8**	**4.3**	**5.2**
Normal Distribution (All Groups)								
10% Focal	3.6	3.9	2.2	2.9	2.1	2.4	3.7	4.4
50% Focal	4.3	4.5	1.4	2.0	1.8	2.4	2.7	3.6
Focal - Moderately Skewed Distribution								
10% Focal	4.4	4.3	2.7	3.4	1.8	2.3	3.9	4.9
50% Focal	5.2	5.3	2.6	2.7	2.0	2.3	3.8	4.0
Focal - Large Skewed Distribution								
10% Focal	6.1	5.7	3.8	4.3	2.4	2.8	5.4	6.2
50% Focal	7.2	8.0	3.7	5.3	2.6	4.1	5.2	7.4
Focal - Both Skewed Distributions								
10% Focal	5.2	5.5	3.1	3.7	2.6	2.9	5.0	5.7
50% Focal	6.4	6.6	3.4	3.9	2.6	3.0	5.0	5.6

Aggregation	Mantel-Haenszel		SIBTEST		Crossing SIBTEST		Both SIBTESTS	
	Best	PT	Best	PT	Best	PT	Best	PT
N=1250	**7.6**	**8.0**	**2.7**	**3.5**	**2.4**	**3.2**	**4.4**	**5.5**
Normal Distribution (All Groups)								
10% Focal	4.6	4.7	1.9	2.3	1.8	2.3	3.0	3.8
50% Focal	4.7	4.6	1.7	2.0	1.8	2.1	2.8	3.4
Focal - Moderately Skewed Distribution								
10% Focal	6.0	6.0	2.5	3.0	2.1	2.6	4.0	4.8
50% Focal	7.4	7.6	2.5	3.2	2.3	3.0	4.1	5.2
Focal - Large Skewed Distribution								
10% Focal	8.7	9.0	3.1	4.2	2.5	3.7	4.9	6.7
50% Focal	13.4	14.9	4.1	6.2	3.8	5.9	6.7	9.2
Focal - Both Skewed Distributions								
10% Focal	6.7	6.8	2.6	3.6	2.3	2.8	4.4	5.3
50% Focal	9.7	10.5	3.5	3.7	3.0	3.5	5.5	5.9

Note: Regardless of the Focal group distribution, Referent group abilities were consistently drawn from a $N(0,1)$ distribution. All percentages calculated by dividing by 60 (the number of items in the test).

Table H1b

Total Type I Error Rate for the 54 Non-DIF Modified Items in 10% DIF.

Aggregation	Mantel-Haenszel		SIBTEST		Crossing SIBTEST		Both SIBTESTS	
	Best	PT	Best	PT	Best	PT	Best	PT
Overall	**5.6**	**7.1**	**1.7**	**3.4**	**1.6**	**2.9**	**2.8**	**5.0**
***N*=500**	**4.5**	**5.0**	**1.5**	**3.1**	**1.4**	**2.4**	**2.5**	**4.5**
Normal Distribution (All Groups)								
10% Focal								
Moderate DIF	4.0	4.0	1.6	3.2	1.6	2.4	2.7	4.7
Large DIF	4.2	4.5	2.0	2.9	1.6	2.3	3.1	4.3
Both DIF	4.4	4.4	1.8	3.1	1.6	2.5	2.7	4.5
50% Focal								
Moderate DIF	4.3	4.7	1.2	2.2	1.5	1.9	2.2	3.4
Large DIF	4.2	5.5	1.1	2.8	1.2	2.4	1.9	4.0
Both DIF	4.5	4.9	1.0	2.4	1.3	2.1	1.9	3.6
Focal - Moderately Skewed Distribution								
10% Focal	4.3	5.0	2.1	3.3	1.3	2.2	3.0	4.6
50% Focal	5.9	6.4	1.6	3.8	1.6	2.9	2.7	5.4
Focal - Large Skewed Distribution								
10% Focal	6.0	6.3	2.6	4.6	1.8	3.1	3.9	6.5
50% Focal	7.5	10.1	2.5	6.3	2.1	4.0	3.9	7.9
Focal - Both Skewed Distributions								
10% Focal	5.1	5.9	2.6	4.4	1.8	3.1	3.9	6.5

Aggregation	Mantel-Haenszel		SIBTEST		Crossing SIBTEST		Both SIBTESTS	
	Best	PT	Best	PT	Best	PT	Best	PT
50% Focal	6.4	8.0	2.4	4.4	2.0	3.2	3.6	5.8
N=1250	**6.7**	**9.1**	**1.9**	**3.7**	**1.7**	**3.5**	**3.0**	**5.6**
Normal Distribution (All Groups)								
10% Focal								
Moderate DIF	4.3	4.9	1.5	2.4	1.5	2.1	2.5	3.6
Large DIF	4.6	6.4	1.4	3.2	1.3	2.5	2.3	4.5
Both DIF	4.5	5.2	1.3	2.2	1.2	1.9	2.0	3.4
50% Focal								
Moderate DIF	4.7	5.6	1.2	2.3	1.0	2.2	1.8	3.6
Large DIF	4.4	7.3	1.2	2.7	1.1	2.9	1.9	4.3
Both DIF	4.8	6.5	1.5	2.5	1.4	2.2	2.2	3.7
Focal - Moderately Skewed Distribution								
10% Focal	6.1	7.4	2.1	3.1	1.8	2.8	3.3	5.0
50% Focal	7.3	10.0	2.0	3.5	1.9	3.9	3.3	5.9
Focal - Large Skewed Distribution								
10% Focal	9.3	11.9	2.4	4.6	2.2	3.8	3.9	6.9
50% Focal	13.7	19.5	3.5	7.8	3.0	7.8	5.4	11.4
Focal - Both Skewed Distributions								
10% Focal	7.1	9.1	2.4	4.4	2.1	3.6	3.8	6.5
50% Focal	10.0	15.4	2.4	5.7	2.4	5.8	4.1	8.4

Note: Regardless of the Focal group distribution, Referent group abilities were

consistently drawn from a $N(0,1)$ distribution. All percentages calculated by dividing by 54, the number of items not modified to exhibit DIF.

Table H1c

Total Type I Error Rate for the 48 Non-DIF Modified Items in 20% DIF.

Aggregation	Mantel-Haenszel		SIBTEST		Crossing SIBTEST		Both SIBTESTS	
	Best	PT	Best	PT	Best	PT	Best	PT
Overall	**5.6**	**7.9**	**1.0**	**3.6**	**1.1**	**3.2**	**1.7**	**5.3**
***N*=500**	**4.4**	**6.0**	**0.9**	**3.1**	**1.0**	**2.8**	**1.6**	**4.8**
Normal Distribution (All Groups)								
10% Focal								
Moderate DIF	3.9	4.4	0.8	3.0	1.3	2.5	1.8	4.5
Large DIF	4.3	5.5	1.0	3.2	0.9	2.2	1.6	4.5
Both DIF	4.1	4.9	1.2	3.1	1.0	2.4	1.7	4.6
50% Focal								
Moderate DIF	4.7	5.9	0.5	2.5	0.8	2.5	1.2	3.9
Large DIF	4.2	7.8	0.7	3.2	0.6	2.8	1.1	4.6
Both DIF	4.3	6.6	0.7	2.4	0.9	2.6	1.4	3.9
Focal - Moderately Skewed Distribution								
10% Focal	4.8	5.5	1.1	3.4	1.2	3.2	2.0	5.6
50% Focal	5.2	6.7	0.9	3.8	0.9	3.0	1.5	5.4
Focal - Large Skewed Distribution								
10% Focal	6.1	7.4	1.5	5.0	1.7	2.7	2.8	6.7
50% Focal	7.2	10.2	1.6	6.2	1.4	4.5	2.5	8.5
Focal - Both Skewed Distributions								
10% Focal	4.9	5.8	1.4	4.6	1.1	3.4	2.3	6.6

Aggregation	Mantel-Haenszel		SIBTEST		Crossing SIBTEST		Both SIBTESTS	
	Best	PT	Best	PT	Best	PT	Best	PT
50% Focal	6.1	8.2	0.9	4.8	1.0	3.8	1.8	6.7
N=1250	**6.7**	**9.8**	**1.1**	**4.1**	**1.2**	**3.6**	**1.9**	**5.9**
Normal Distribution (All Groups)								
10% Focal								
Moderate DIF	4.1	4.7	0.7	2.1	1.3	2.1	1.4	3.6
Large DIF	4.0	6.3	1.0	3.1	0.8	2.3	1.4	4.3
Both DIF	3.8	5.4	0.7	2.5	1.0	2.1	1.4	3.7
50% Focal								
Moderate DIF	5.0	8.3	0.7	3.6	0.9	2.8	1.3	4.9
Large DIF	4.5	12.1	0.9	4.0	1.2	3.4	1.7	5.5
Both DIF	4.8	9.2	0.8	3.4	0.9	2.9	1.4	5.1
Focal - Moderately Skewed Distribution								
10% Focal	6.0	6.9	1.1	3.0	0.9	2.9	1.7	4.7
50% Focal	8.4	10.8	1.2	3.8	1.5	3.6	2.3	5.7
Focal - Large Skewed Distribution								
10% Focal	9.2	10.8	1.1	4.4	1.3	3.6	2.2	6.6
50% Focal	13.4	19.7	2.1	9.4	1.5	8.0	3.2	12.2
Focal - Both Skewed Distributions								
10% Focal	7.3	8.5	1.3	3.7	1.1	2.9	2.0	5.4
50% Focal	10.2	15.1	1.4	6.3	1.6	6.3	2.7	9.2

Note: Regardless of the Focal group distribution, Referent group abilities were

consistently drawn from a $N(0,1)$ distribution. All percentages calculated by dividing by 48, the number of items not modified to exhibit DIF.

Table H2a

Analyzed Type I Error Rate for the Non-DIF Items in 0% DIF Conditions.

Aggregation	Mantel-Haenszel		SIBTEST		Crossing SIBTEST		Both SIBTESTS	
	Best	PT	Best	PT	Best	PT	Best	PT
Overall	**6.5**	**6.7**	**8.4**	**9.0**	**7.0**	**7.6**	**13.1**	**13.7**
***N*=500**	**5.3**	**5.5**	**8.6**	**8.9**	**6.7**	**7.0**	**13.0**	**13.2**
Normal Distribution (All Groups)								
10% Focal	3.6	3.9	6.7	7.2	6.4	5.8	11.2	10.9
50% Focal	4.3	4.5	4.3	5.0	5.5	5.9	8.2	9.0
Focal - Moderately Skewed Distribution								
10% Focal	4.4	4.3	8.1	8.8	5.3	5.8	11.7	12.7
50% Focal	5.2	5.3	7.8	6.8	5.9	5.8	11.3	10.2
Focal - Large Skewed Distribution								
10% Focal	6.1	5.7	11.4	10.9	7.3	7.1	16.1	15.9
50% Focal	7.2	8.0	11.0	13.2	7.7	10.2	15.7	18.3
Focal - Both Skewed Distributions								
10% Focal	5.2	5.5	9.3	9.4	7.8	7.4	15.0	14.5
50% Focal	6.4	6.6	10.2	9.9	7.9	7.6	15.0	14.3

Aggregation	Mantel-Haenszel		SIBTEST		Crossing SIBTEST		Both SIBTESTS	
	Best	PT	Best	PT	Best	PT	Best	PT
N=1250	**7.6**	**8.0**	**8.2**	**9.0**	**7.3**	**8.3**	**13.3**	**14.1**
Normal Distribution (All Groups)								
10% Focal	4.6	4.7	5.6	5.9	5.4	5.7	9.1	9.6
50% Focal	4.7	4.6	5.1	5.1	5.3	5.5	8.3	8.6
Focal - Moderately Skewed Distribution								
10% Focal	6.0	6.0	7.6	7.2	6.4	6.3	12.0	11.5
50% Focal	7.4	7.6	7.6	8.3	7.0	7.8	12.4	13.4
Focal - Large Skewed Distribution								
10% Focal	8.7	9.0	9.4	10.9	7.6	9.5	14.6	17.3
50% Focal	13.4	14.9	12.4	15.9	11.5	15.1	20.2	23.5
Focal - Both Skewed Distributions								
10% Focal	6.7	6.8	7.9	9.1	6.8	7.2	13.2	13.6
50% Focal	9.7	10.5	10.4	9.7	8.9	9.1	16.6	15.6

Note: Regardless of the Focal group distribution, Referent group abilities were consistently drawn from a $N(0,1)$ distribution. Percentages were calculated by dividing by the number of items analyzed for DIF (60 for Mantel-Haenszel, 20 for all SIBTEST Best analyses, and the number of items in the AT List for SIBTEST PT analyses).

Table H2b

Analyzed Type I Error Rate for the Non-DIF Items in 10% DIF Conditions.

Aggregation	Mantel-Haenszel		SIBTEST		Crossing SIBTEST		Both SIBTESTS	
	Best	PT	Best	PT	Best	PT	Best	PT
Overall	**5.6**	**7.1**	**6.6**	**8.6**	**6.1**	**7.5**	**10.6**	**12.8**
***N*=500**	**4.5**	**5.0**	**6.0**	**7.8**	**5.4**	**6.0**	**9.6**	**11.2**
Normal Distribution (All Groups)								
10% Focal								
Moderate DIF	4.0	4.0	6.3	7.8	6.0	5.9	10.4	11.5
Large DIF	4.2	4.5	7.6	7.5	6.0	6.0	11.9	11.3
Both DIF	4.4	4.4	6.8	7.6	6.1	6.2	10.5	11.2
50% Focal								
Moderate DIF	4.3	4.7	4.6	5.6	5.7	4.9	8.6	8.8
Large DIF	4.2	5.5	4.4	7.0	4.5	5.9	7.5	10.0
Both DIF	4.5	4.9	4.0	6.0	4.9	5.4	7.4	9.0
Focal - Moderately Skewed Distribution								
10% Focal	4.3	5.0	8.1	8.4	5.0	5.7	11.5	11.8
50% Focal	5.9	6.4	6.3	9.7	6.1	7.3	10.3	13.8
Focal - Large Skewed Distribution								
10% Focal	6.0	6.3	10.0	11.6	6.9	7.9	15.1	16.6
50% Focal	7.5	10.1	9.7	15.7	8.1	9.9	14.9	19.7
Focal - Both Skewed Distributions								
10% Focal	5.1	5.9	10.2	10.8	6.9	7.7	15.1	15.9

Aggregation	Mantel-Haenszel		SIBTEST		Crossing SIBTEST		Both SIBTESTS	
	Best	PT	Best	PT	Best	PT	Best	PT
50% Focal	6.4	8.0	9.2	11.0	7.6	8.1	13.7	14.6
N=1250	**6.7**	**9.1**	**7.3**	**9.5**	**6.7**	**8.9**	**11.7**	**14.4**
Normal Distribution (All Groups)								
10% Focal								
Moderate DIF	4.3	4.9	5.9	6.0	5.6	5.4	9.5	9.1
Large DIF	4.6	6.4	5.5	8.0	5.0	6.1	8.9	11.2
Both DIF	4.5	5.2	5.1	5.7	4.5	5.0	7.6	8.8
50% Focal								
Moderate DIF	4.7	5.6	4.5	6.0	3.7	5.8	6.9	9.5
Large DIF	4.4	7.3	4.5	7.1	4.3	7.6	7.2	11.3
Both DIF	4.8	6.5	5.9	6.7	5.4	5.9	8.6	9.9
Focal - Moderately Skewed Distribution								
10% Focal	6.1	7.4	8.0	8.0	6.9	7.2	12.6	12.7
50% Focal	7.3	10.0	7.6	9.2	7.4	10.2	12.6	15.3
Focal - Large Skewed Distribution								
10% Focal	9.3	11.9	9.1	11.7	8.6	9.9	14.9	17.7
50% Focal	13.7	19.5	13.4	20.1	11.5	19.9	20.7	29.3
Focal - Both Skewed Distributions								
10% Focal	7.1	9.1	9.1	11.1	8.1	9.2	14.6	16.5
50% Focal	10.0	15.4	9.1	14.6	9.4	14.9	15.7	21.5

Note: Regardless of the Focal group distribution, Referent group abilities were

consistently drawn from a $N(0,1)$ distribution. Percentages were calculated by dividing by the number of non-DIF modified items analyzed for DIF (54 for Mantel-Haenszel, 14 for all SIBTEST Best analyses, and the number of non-DIF items in the AT List for SIBTEST PT analyses).

Table H2c

Analyzed Type I Error Rate for the Non-DIF Items in 20% DIF Conditions.

Aggregation	Mantel-Haenszel		SIBTEST		Crossing SIBTEST		Both SIBTESTS	
	Best	PT	Best	PT	Best	PT	Best	PT
Overall	**5.6**	**7.9**	**5.9**	**9.3**	**6.4**	**8.1**	**10.5**	**13.6**
***N*=500**	**4.4**	**6.0**	**5.3**	**7.9**	**5.9**	**6.9**	**9.6**	**11.9**
Normal Distribution (All Groups)								
10% Focal								
Moderate DIF	3.9	4.4	4.6	7.5	7.5	6.2	10.9	11.2
Large DIF	4.3	5.5	6.1	7.9	5.1	5.4	9.4	11.1
Both DIF	4.1	4.9	7.0	7.9	6.1	6.1	10.4	11.7
50% Focal								
Moderate DIF	4.7	5.9	3.0	6.3	4.8	6.2	7.0	9.7
Large DIF	4.2	7.8	4.0	8.1	3.9	7.2	6.6	11.6
Both DIF	4.3	6.6	4.1	6.0	5.1	6.5	8.5	9.8
Focal - Moderately Skewed Distribution								
10% Focal	4.8	5.5	6.4	8.2	7.1	7.9	12.0	13.7
50% Focal	5.2	6.7	5.5	9.5	5.6	7.6	9.3	13.7
Focal - Large Skewed Distribution								
10% Focal	6.1	7.4	9.1	12.4	10.4	6.6	16.8	16.4
50% Focal	7.2	10.2	9.6	15.4	8.6	11.2	15.1	21.2
Focal - Both Skewed Distributions								
10% Focal	4.9	5.8	8.4	11.3	6.8	8.3	13.6	16.2

Aggregation	Mantel-Haenszel		SIBTEST		Crossing SIBTEST		Both SIBTESTS	
	Best	PT	Best	PT	Best	PT	Best	PT
50% Focal	6.1	8.2	5.5	11.9	6.3	9.5	10.5	16.8
N=1250	**6.7**	**9.8**	**6.6**	**10.7**	**6.8**	**9.3**	**11.4**	**15.4**
Normal Distribution (All Groups)								
10% Focal								
Moderate DIF	4.1	4.7	4.4	5.3	5.4	5.4	8.4	9.2
Large DIF	4.0	6.3	5.9	8.0	4.9	6.1	8.6	11.3
Both DIF	3.8	5.4	4.4	6.4	6.3	5.3	8.5	9.5
50% Focal								
Moderate DIF	5.0	8.3	4.4	9.0	5.4	7.1	7.8	12.3
Large DIF	4.5	12.1	5.1	10.9	7.3	9.4	10.1	14.9
Both DIF	4.8	9.2	4.8	8.9	5.1	7.6	8.4	13.1
Focal - Moderately Skewed Distribution								
10% Focal	6.0	6.9	6.9	7.9	5.3	7.6	10.1	12.2
50% Focal	8.4	10.8	7.3	10.3	8.9	9.7	13.8	15.4
Focal - Large Skewed Distribution								
10% Focal	9.2	10.8	6.9	11.3	8.0	9.3	13.4	17.0
50% Focal	13.4	19.7	12.6	24.5	9.1	21.0	19.4	31.8
Focal - Both Skewed Distributions								
10% Focal	7.3	8.5	7.8	9.4	6.8	7.3	12.1	13.8
50% Focal	10.2	15.1	8.4	16.2	9.9	16.2	16.0	23.8

Note: Regardless of the Focal group distribution, Referent group abilities were

consistently drawn from a $N(0,1)$ distribution. Percentages were calculated by dividing by the number of non-DIF modified items analyzed for DIF (48 for Mantel-Haenszel, 8 for all SIBTEST Best analyses, and the number of non-DIF items in the AT List for SIBTEST PT analyses).

APPENDIX I

TABULAR RESULTS FOR POWER RATES

Table I1a

Percent of DIF Items Correctly Identified in 10% DIF Conditions

Aggregation/ Severity of DIF	Mantel-Haenszel		SIBTEST			Crossing SIBTEST			Both SIBTESTs		
	Best	PT	Best	PT	DIF-AT	Best	PT	DIF-AT	Best	PT	DIF-AT
Overall	**53.8**	**49.5**	**51.5**	**19.8**	**48.4**	**39.1**	**14.7**	**35.9**	**55.0**	**21.2**	**51.9**
***N*=500**	**41.6**	**38.1**	**40.0**	**15.5**	**38.1**	**27.5**	**10.2**	**24.9**	**43.8**	**17.1**	**41.8**
Normal Distribution (All Groups)											
10% Focal											
Moderate	21.5	18.2	24.3	8.2	22.2	14.3	4.8	13.1	28.0	9.0	24.4
Large	41.7	38.7	41.7	17.5	39.8	26.2	9.0	20.5	46.0	19.2	43.6
Both	27.3	24.7	27.3	10.2	25.5	18.7	7.0	17.6	31.8	13.0	32.6
50% Focal											
Moderate	46.5	42.5	49.8	20.0	47.8	36.7	15.0	35.9	52.8	21.7	51.8
Large	77.0	71.5	77.7	27.2	67.6	67.2	22.8	56.8	79.0	28.3	70.5
Both	65.7	60.2	67.0	25.3	63.1	54.0	19.0	47.3	71.0	26.3	65.6
Focal - Moderately Skewed Distribution											
10% Focal	19.0	16.8	18.8	6.0	16.8	9.7	2.0	5.6	22.7	6.5	18.2
50% Focal	40.8	35.2	34.8	15.3	35.8	22.2	8.5	19.8	39.3	18.3	42.8
Focal - Large Skewed Distribution											
10% Focal	29.5	29.7	27.2	12.2	29.1	12.3	6.5	15.5	32.5	13.8	33.1
50% Focal	58.0	51.8	48.8	18.2	46.6	32.0	11.5	29.5	53.0	20.0	51.3
Focal - Both Skewed Distributions											
10% Focal	22.3	22.0	18.3	8.0	18.5	10.0	4.7	10.8	23.2	9.7	22.3

Aggregation/ Severity of DIF	Mantel-Haenszel		SIBTEST			Crossing SIBTEST			Both SIBTESTs		
	Best	PT	Best	PT	DIF-AT	Best	PT	DIF-AT	Best	PT	DIF-AT
50% Focal	50.0	46.3	43.7	18.3	44.0	26.7	11.2	26.8	46.8	19.0	45.6
***N*=1250**	**65.9**	**60.9**	**63.1**	**24.1**	**58.7**	**50.7**	**19.3**	**46.8**	**66.1**	**25.3**	**61.9**
Normal Distribution (All Groups)											
10% Focal											
Moderate	43.7	38.5	44.0	17.3	39.8	34.2	13.2	30.3	49.2	19.7	45.2
Large	76.8	70.8	75.8	27.3	70.1	65.0	23.0	59.0	77.8	28.5	73.1
Both	60.5	54.3	60.2	23.8	55.4	47.5	18.8	43.8	63.3	24.8	57.8
50% Focal											
Moderate	73.3	67.5	74.2	30.5	69.1	67.0	27.7	62.6	76.8	31.5	71.3
Large	85.0	84.2	85.7	40.3	86.7	85.0	40.3	86.7	87.0	41.0	88.2
Both	83.2	81.2	83.3	32.3	80.2	79.5	30.3	75.2	85.0	32.5	80.6
Focal - Moderately Skewed Distribution											
10% Focal	37.3	34.8	29.8	12.7	30.0	17.5	7.8	18.6	32.3	14.2	33.6
50% Focal	67.8	60.3	70.3	25.0	62.8	46.7	15.0	37.7	73.7	26.8	67.4
Focal - Large Skewed Distribution											
10% Focal	58.8	54.0	45.7	16.8	40.9	27.2	9.5	23.1	48.8	18.0	43.7
50% Focal	80.3	72.8	78.2	25.5	71.5	61.0	20.8	58.4	81.7	27.5	77.1
Focal - Both Skewed Distributions											
10% Focal	50.0	44.5	37.2	14.7	35.3	25.3	7.3	17.7	41.5	15.7	37.8
50% Focal	74.2	67.5	72.3	22.3	62.6	52.8	17.3	48.6	75.8	24.0	67.3

Note: Regardless of the Focal group distribution, Referent group abilities were consistently

drawn from a $N(0,1)$ distribution. Percentages of Best and PT calculated by dividing by 6, where the divisors of DIF-AT percentages were the number of DIF items in the AT List.

Table I1b

Percent of DIF Items Correctly Identified in 20% DIF Conditions

Aggregation/ Severity of DIF	Mantel-Haenszel		SIBTEST			Crossing SIBTEST			Both SIBTESTs		
	Best	PT	Best	PT	DIF-AT	Best	PT	DIF-AT	Best	PT	DIF-AT
Overall	**39.8**	**34.4**	**40.2**	**14.8**	**35.9**	**32.4**	**11.7**	**28.3**	**46.1**	**17.3**	**41.9**
***N*=500**	**30.2**	**25.4**	**31.5**	**11.3**	**27.4**	**22.8**	**7.9**	**19.2**	**36.5**	**13.3**	**32.1**
Normal Distribution (All Groups)											
10% Focal											
Moderate	14.3	10.7	17.9	6.2	15.9	11.2	3.9	10.1	21.7	7.5	19.3
Large	32.6	24.6	37.8	12.3	28.4	25.9	8.6	19.9	42.8	14.3	33.2
Both	22.3	15.8	25.9	7.6	19.0	17.6	5.6	14.0	30.4	9.5	23.8
50% Focal											
Moderate	31.8	25.3	33.3	9.9	25.3	25.8	8.2	20.9	38.0	12.0	30.6
Large	65.0	55.8	66.3	24.3	59.5	60.3	19.6	48.1	72.0	26.3	64.6
Both	50.3	41.1	53.1	17.4	42.7	43.3	13.4	32.9	57.6	18.8	46.1
Focal - Moderately Skewed Distribution											
10% Focal	12.0	11.1	14.3	5.7	13.6	9.2	2.9	7.0	18.8	7.1	17.0
50% Focal	23.3	19.9	23.6	8.3	20.0	13.8	5.1	12.3	27.6	9.9	24.1
Focal - Large Skewed Distribution											
10% Focal	22.5	20.5	22.3	8.8	20.8	11.2	5.3	12.4	27.3	11.1	26.1
50% Focal	41.2	36.9	36.3	16.8	38.5	27.1	11.6	26.6	43.6	19.8	45.4
Focal - Both Skewed Distributions											
10% Focal	16.8	16.3	18.2	7.1	16.7	9.3	3.4	8.1	23.0	8.8	20.6

Aggregation/ Severity of DIF	Mantel-Haenszel		SIBTEST			Crossing SIBTEST			Both SIBTESTs		
	Best	PT	Best	PT	DIF-AT	Best	PT	DIF-AT	Best	PT	DIF-AT
50% Focal	29.9	27.0	29.0	11.8	28.7	19.0	7.4	18.1	35.5	14.0	34.1
***N*=1250**	**49.5**	**43.3**	**48.9**	**18.3**	**44.3**	**42.1**	**15.5**	**37.5**	**55.7**	**21.4**	**51.7**
Normal Distribution (All Groups)											
10% Focal											
Moderate	27.7	22.8	27.6	9.5	23.4	23.3	8.3	20.3	32.9	11.9	29.4
Large	62.7	50.8	62.4	22.8	52.9	54.9	19.4	45.2	67.6	25.6	59.5
Both	47.4	36.6	48.8	17.9	40.2	38.2	13.8	31.0	52.9	20.4	45.8
50% Focal											
Moderate	54.8	43.9	54.7	16.8	43.3	49.2	13.4	34.5	59.3	18.3	47.0
Large	79.5	72.9	79.6	30.5	76.9	83.8	30.4	76.7	88.1	33.5	84.5
Both	70.5	62.3	71.0	27.7	66.3	70.3	26.4	63.3	77.2	30.8	73.9
Focal - Moderately Skewed Distribution											
10% Focal	23.3	21.8	22.3	9.2	22.1	16.0	5.5	13.3	27.5	10.7	25.8
50% Focal	39.4	36.3	44.2	16.5	40.4	31.8	12.1	29.6	51.7	19.3	47.3
Focal - Large Skewed Distribution											
10% Focal	40.3	38.2	34.2	14.3	35.4	24.7	9.9	24.6	41.7	17.4	43.3
50% Focal	64.3	59.5	63.0	23.8	59.3	53.4	20.7	51.5	74.3	28.9	72.0
Focal - Both Skewed Distributions											
10% Focal	31.1	28.4	24.9	9.5	23.7	17.4	7.6	18.9	30.3	12.5	31.2
50% Focal	53.3	46.0	54.8	21.7	47.9	42.4	18.6	41.1	65.5	27.6	61.0

Note: Regardless of the Focal group distribution, Referent group abilities were

consistently drawn from a $N(0,1)$ distribution. Percentages of Best and PT calculated by dividing by 12, where the divisors of DIF-AT percentages were the number of DIF items in the AT List.

Table I2a

Power Rates for All DIF Items by Difficulty for Best Subtest 10% DIF.

Aggregation/ Severity of DIF	Mantel-Haenszel			SIBTEST			Crossing SIBTEST			Both SIBTESTS		
	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High
Overall	**13.8**	**62.5**	**60.7**	**13.8**	**61.4**	**55.6**	**10.5**	**45.8**	**43.4**	**18.3**	**64.3**	**59.4**
***N*=500**	**9.1**	**47.6**	**48.9**	**10.3**	**47.3**	**43.9**	**8.8**	**31.6**	**30.6**	**14.4**	**50.8**	**48.1**
Normal Distribution (All Groups)												
10% Focal												
Moderate	4.0	25.3	24.5	9.0	31.0	22.0	6.0	17.0	14.5	10.0	34.7	27.0
Large	2.0	44.3	57.5	8.0	48.7	48.0	6.0	30.0	30.5	10.0	53.0	53.5
Both	8.0	29.3	34.0	11.0	33.7	26.0	9.0	25.0	14.0	13.0	39.0	30.5
50% Focal												
Moderate	11.0	56.7	49.0	15.0	59.7	52.5	10.0	44.3	38.5	19.0	62.7	55.0
Large	5.0	90.3	93.0	5.0	91.0	94.0	8.0	78.0	80.5	10.0	91.7	94.5
Both	5.0	80.3	74.0	5.0	81.7	76.0	16.0	62.3	60.5	16.0	83.7	79.5
Focal - Moderately Skewed Distribution												
10% Focal	11.0	20.7	20.5	20.0	20.0	16.5	8.0	10.0	10.0	22.0	24.0	21.0
50% Focal	18.0	45.7	45.0	14.0	40.3	37.0	12.0	24.0	24.5	20.0	44.3	41.5
Focal - Large Skewed Distribution												
10% Focal	9.0	34.0	33.0	9.0	35.0	24.5	11.0	14.7	9.5	16.0	39.3	30.5
50% Focal	10.0	65.0	71.5	9.0	55.3	59.0	6.0	33.7	42.5	11.0	59.7	64.0
Focal - Both Skewed Distributions												
10% Focal	9.0	24.7	25.5	6.0	23.0	17.5	6.0	13.0	7.5	11.0	27.0	23.5

Aggregation/ Severity of DIF	Mantel-Haenszel			SIBTEST			Crossing SIBTEST			Both SIBTESTS		
	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High
50% Focal	17.0	55.0	59.0	12.0	47.7	53.5	7.0	27.7	35.0	15.0	51.0	56.5
***N*=1250**	**18.6**	**77.3**	**72.5**	**17.3**	**75.5**	**67.3**	**12.3**	**60.0**	**56.1**	**22.1**	**77.7**	**70.7**
Normal Distribution (All Groups)												
10% Focal												
Moderate	13.0	52.3	46.0	19.0	54.3	41.0	14.0	41.3	33.5	24.0	59.3	46.5
Large	10.0	91.0	89.0	11.0	89.7	87.5	7.0	75.0	79.0	15.0	90.7	90.0
Both	6.0	70.3	73.0	7.0	72.0	69.0	13.0	53.3	56.0	14.0	74.3	71.5
50% Focal												
Moderate	17.0	96.0	67.5	19.0	96.7	68.0	23.0	85.7	61.0	29.0	97.0	70.5
Large	11.0	99.7	100.0	15.0	99.7	100.0	15.0	99.7	98.0	23.0	99.7	100.0
Both	12.0	99.3	94.5	13.0	99.0	95.0	18.0	94.3	88.0	21.0	99.0	96.0
Focal - Moderately Skewed Distribution												
10% Focal	22.0	44.0	35.0	13.0	35.0	30.5	5.0	20.3	19.5	13.0	37.3	34.5
50% Focal	33.0	78.7	69.0	35.0	85.3	65.5	12.0	56.7	49.0	36.0	88.3	70.5
Focal - Large Skewed Distribution												
10% Focal	24.0	64.0	68.5	14.0	51.7	52.5	10.0	32.3	28.0	16.0	55.3	55.5
50% Focal	26.0	89.3	94.0	29.0	89.3	86.0	8.0	67.7	77.5	32.0	92.0	91.0
Focal - Both Skewed Distributions												
10% Focal	20.0	57.0	54.5	9.0	45.7	38.5	9.0	32.0	23.5	13.0	49.3	44.0
50% Focal	29.0	86.0	79.0	24.0	87.7	73.5	13.0	61.3	60.0	29.0	90.0	78.0

Note: Regardless of the Focal group distribution, Referent group abilities were consistently

drawn from a $N(0,1)$ distribution. Percentages were computed by dividing by the total number of items modified to exhibit DIF within the Referent difficulty range (Low = 1, Medium = 3, High = 2).

Table I2b

Power Rates for All DIF Items by Difficulty for Best Subtest 20% DIF.

Aggregation/ Severity of DIF	Mantel-Haenszel			SIBTEST			Crossing SIBTEST			Both SIBTESTS		
	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High
Overall	**13.0**	**42.8**	**60.8**	**12.8**	**43.9**	**54.1**	**11.5**	**35.8**	**42.3**	**17.9**	**50.3**	**57.7**
***N*=500**	**9.6**	**32.2**	**48.8**	**10.4**	**34.7**	**42.3**	**8.1**	**24.9**	**29.8**	**14.1**	**39.8**	**46.7**
Normal Distribution (All Groups)												
10% Focal												
Moderate	7.0	13.9	25.0	10.0	19.1	17.0	6.7	10.4	12.0	12.7	22.4	22.0
Large	8.0	37.0	51.5	15.7	40.9	47.5	9.3	28.7	28.5	18.0	46.1	52.0
Both	8.7	23.7	37.0	10.7	26.9	32.0	9.7	18.0	17.0	14.7	30.7	36.5
50% Focal												
Moderate	11.7	34.0	45.5	12.3	35.9	48.0	11.3	26.4	39.0	16.3	40.3	53.5
Large	21.0	70.4	89.0	17.7	72.3	91.5	13.0	68.1	84.5	21.0	79.6	92.5
Both	12.7	57.6	71.5	12.7	59.7	76.0	11.7	49.3	61.5	16.7	64.6	78.5
Focal - Moderately Skewed Distribution												
10% Focal	6.3	10.3	23.5	7.3	16.3	13.5	6.3	9.6	7.0	11.3	20.4	16.5
50% Focal	9.0	23.9	39.5	8.0	26.6	32.5	5.0	16.0	17.0	10.3	31.0	36.0
Focal - Large Skewed Distribution												
10% Focal	8.7	21.9	44.0	7.3	24.0	29.5	5.7	12.4	12.0	11.7	28.9	35.0
50% Focal	9.7	45.6	69.5	7.7	41.7	54.0	8.7	29.1	42.0	13.7	49.0	60.5
Focal - Both Skewed Distributions												
10% Focal	6.0	16.1	34.0	8.7	19.9	21.5	5.3	10.1	7.0	11.7	25.0	25.0

Aggregation/ Severity of DIF	Mantel-Haenszel			SIBTEST			Crossing SIBTEST			Both SIBTESTS		
	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High
50% Focal	7.0	31.7	56.0	6.7	32.7	45.0	5.0	20.9	30.0	10.7	39.6	52.0
***N*=1250**	**16.4**	**53.5**	**72.8**	**15.3**	**53.2**	**66.0**	**14.8**	**46.7**	**54.8**	**21.7**	**60.8**	**68.8**
Normal Distribution (All Groups)												
10% Focal												
Moderate	7.7	30.0	47.5	8.0	30.6	40.0	9.0	25.1	30.5	12.7	36.0	43.5
Large	16.7	67.7	93.5	18.3	66.7	87.5	18.0	60.0	73.5	23.7	72.7	88.0
Both	11.7	52.7	71.5	16.3	53.3	65.0	14.0	41.3	51.0	19.3	58.0	67.0
50% Focal												
Moderate	23.7	59.7	73.0	21.7	59.3	74.0	16.7	54.6	65.0	24.0	64.6	75.5
Large	35.3	81.6	100.0	32.3	81.9	100.0	28.7	91.9	99.0	36.3	94.4	100.0
Both	29.0	74.3	96.0	27.0	74.4	96.0	22.0	79.6	89.0	31.3	82.7	96.5
Focal - Moderately Skewed Distribution												
10% Focal	6.0	25.6	40.0	3.3	26.3	29.5	6.0	19.0	16.0	8.7	31.7	32.5
50% Focal	12.7	42.6	62.0	14.3	48.6	56.5	13.3	35.6	43.5	23.0	56.9	58.5
Focal - Large Skewed Distribution												
10% Focal	10.0	44.9	63.0	5.3	39.6	47.0	11.3	28.6	24.0	15.7	46.6	50.5
50% Focal	14.7	72.4	93.5	15.7	69.9	83.5	16.0	60.6	77.0	28.0	81.6	90.5
Focal - Both Skewed Distributions												
10% Focal	13.0	32.7	50.5	5.3	27.7	36.5	7.0	18.7	24.0	10.0	32.7	42.5
50% Focal	17.0	57.9	83.5	16.0	59.7	76.0	15.7	46.0	65.0	27.7	71.1	80.5

Note: Regardless of the Focal group distribution, Referent group abilities were consistently

drawn from a $N(0,1)$ distribution. Percentages were computed by dividing by the total number of items modified to exhibit DIF within the Referent difficulty range (Low = 3, Medium = 7, High = 2).

Table I3a

Power Rates for Uniform DIF Items by Difficulty in Best Subtest 10% DIF.

Aggregation/ Severity of DIF	Mantel-Haenszel			SIBTEST			Crossing SIBTEST			Both SIBTESTS		
	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High
Overall	-	**55.6**	**40.8**	-	**56.9**	**37.4**	-	**41.9**	**28.1**	-	**60.1**	**42.4**
***N*=500**	-	**40.8**	**28.8**	-	**42.1**	**27.1**	-	**28.8**	**18.8**	-	**45.9**	**32.3**
Normal Distribution (All Groups)												
10% Focal												
Moderate	-	22.5	10.0	-	27.5	12.0	-	18.0	10.0	-	32.0	18.0
Large	-	43.5	42.0	-	52.0	33.0	-	30.0	23.0	-	55.5	42.0
Both	-	25.5	19.0	-	28.5	12.0	-	23.0	5.0	-	35.0	16.0
50% Focal												
Moderate	-	54.0	21.0	-	58.5	23.0	-	43.0	11.0	-	61.0	26.0
Large	-	89.5	87.0	-	90.5	88.0	-	77.5	68.0	-	91.0	89.0
Both	-	75.5	52.0	-	76.0	55.0	-	60.0	39.0	-	78.0	61.0
Focal - Moderately Skewed Distribution												
10% Focal	-	16.0	6.0	-	18.0	8.0	-	8.5	7.0	-	22.0	13.0
50% Focal	-	36.5	17.0	-	33.0	18.0	-	23.0	15.0	-	37.5	24.0
Focal - Large Skewed Distribution												
10% Focal	-	20.0	10.0	-	27.5	10.0	-	7.0	5.0	-	30.5	14.0
50% Focal	-	52.0	44.0	-	43.5	29.0	-	26.0	23.0	-	50.0	38.0
Focal - Both Skewed Distributions												
10% Focal	-	15.5	9.0	-	14.0	10.0	-	9.0	4.0	-	18.5	13.0

Aggregation/ Severity of DIF	Mantel-Haenszel			SIBTEST			Crossing SIBTEST			Both SIBTESTS		
	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High
50% Focal	-	39.0	28.0	-	36.0	27.0	-	21.0	16.0	-	40.0	33.0
***N*=1250**	-	**70.4**	**52.8**	-	**71.7**	**47.8**	-	**54.9**	**37.3**	-	**74.2**	**52.5**
Normal Distribution (All Groups)												
10% Focal												
Moderate	-	47.5	14.0	-	51.0	9.0	-	39.5	10.0	-	57.0	15.0
Large	-	88.0	79.0	-	88.0	75.0	-	73.5	64.0	-	88.5	80.0
Both	-	65.0	52.0	-	69.0	49.0	-	48.5	34.0	-	72.0	50.0
50% Focal												
Moderate DIF	-	94.0	35.0	-	95.0	36.0	-	81.5	24.0	-	95.5	41.0
Large DIF	-	99.5	100.0	-	99.5	100.0	-	99.5	96.0	-	99.5	100.0
Both DIF	-	99.0	89.0	-	98.5	90.0	-	93.0	77.0	-	98.5	92.0
Focal - Moderately Skewed Distribution												
10% Focal	-	30.0	15.0	-	28.0	9.0	-	14.5	6.0	-	30.0	13.0
50% Focal	-	70.0	41.0	-	80.5	36.0	-	48.0	21.0	-	85.0	43.0
Focal - Large Skewed Distribution												
10% Focal	-	48.0	38.0	-	43.0	29.0	-	26.5	11.0	-	47.5	33.0
50% Focal	-	84.0	88.0	-	85.0	73.0	-	58.5	60.0	-	88.5	83.0
Focal - Both Skewed Distributions												
10% Focal	-	40.5	24.0	-	40.5	19.0	-	25.5	9.0	-	43.0	23.0
50% Focal	-	79.5	58.0	-	82.5	48.0	-	50.5	36.0	-	85.5	57.0

Note: Regardless of the Focal group distribution, Referent group abilities were consistently

drawn from a N(0,1) distribution. Percentages were computed by dividing by the number of items modified to exhibit U-DIF within the Referent difficulty range (Low = 0, Medium = 2, High = 1).

Table I3b

Power Rates for Uniform DIF Items by Difficulty in Best Subtest 20% DIF.

Aggregation/ Severity of DIF	Mantel-Haenszel			SIBTEST			Crossing SIBTEST			Both SIBTESTS		
	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High
Overall	**22.2**	**55.5**	**41.6**	**28.4**	**56.7**	**36.5**	**19.5**	**41.8**	**27.4**	**32.5**	**59.6**	**41.3**
***N*=500**	**14.0**	**40.0**	**29.5**	**19.4**	**43.3**	**26.3**	**15.0**	**28.7**	**18.3**	**24.2**	**46.4**	**31.5**
Normal Distribution (All Groups)												
10% Focal												
Moderate	8.0	21.0	10.0	17.0	31.5	7.0	17.0	16.0	3.0	21.0	35.5	9.0
Large	9.0	45.5	38.0	25.0	49.5	36.0	25.0	36.0	19.0	33.0	53.0	44.0
Both	6.0	35.0	21.0	27.0	37.0	18.0	22.0	23.0	10.0	33.0	39.5	25.0
50% Focal												
Moderate	17.0	52.5	15.0	16.0	55.5	17.0	12.0	40.0	15.0	18.0	58.0	25.0
Large	54.0	88.5	78.0	53.0	89.5	83.0	39.0	78.0	73.0	59.0	90.5	85.0
Both	24.0	75.5	50.0	29.0	79.0	56.0	17.0	63.5	38.0	32.0	81.0	60.0
Focal - Moderately Skewed Distribution												
10% Focal	7.0	11.0	14.0	9.0	20.0	9.0	10.0	6.5	5.0	15.0	21.5	10.0
50% Focal	9.0	34.0	19.0	8.0	37.5	14.0	4.0	19.0	5.0	11.0	41.0	17.0
Focal - Large Skewed Distribution												
10% Focal	8.0	21.0	23.0	18.0	22.5	21.0	6.0	8.0	7.0	21.0	24.0	24.0
50% Focal	14.0	48.0	44.0	12.0	45.0	27.0	11.0	27.0	25.0	18.0	50.5	39.0
Focal - Both Skewed Distributions												
10% Focal	7.0	15.5	16.0	10.0	19.5	13.0	10.0	7.0	5.0	16.0	22.0	16.0

Aggregation/ Severity of DIF	Mantel-Haenszel			SIBTEST			Crossing SIBTEST			Both SIBTESTS		
	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High
50% Focal	5.0	33.0	26.0	9.0	33.0	15.0	7.0	20.5	15.0	13.0	40.5	24.0
***N*=1250**	**30.3**	**71.0**	**53.8**	**37.4**	**70.1**	**46.6**	**24.1**	**54.8**	**36.5**	**40.9**	**72.8**	**51.0**
Normal Distribution (All Groups)												
10% Focal												
Moderate	9.0	49.0	20.0	13.0	52.5	12.0	15.0	41.0	11.0	18.0	58.5	17.0
Large	44.0	88.5	88.0	52.0	86.5	78.0	38.0	69.0	59.0	55.0	87.5	79.0
Both	26.0	71.0	49.0	33.0	71.5	42.0	25.0	57.0	26.0	37.0	74.0	46.0
50% Focal												
Moderate	28.0	95.0	47.0	28.0	94.5	48.0	28.0	88.0	34.0	36.0	94.5	51.0
Large	84.0	100.0	100.0	85.0	100.0	100.0	78.0	99.0	98.0	87.0	100.0	100.0
Both	54.0	99.0	92.0	58.0	99.0	92.0	43.0	97.0	79.0	60.0	99.0	93.0
Focal - Moderately Skewed Distribution												
10% Focal	5.0	30.0	14.0	14.0	28.0	9.0	9.0	17.5	4.0	17.0	31.5	11.0
50% Focal	19.0	70.5	29.0	34.0	77.0	23.0	5.0	46.5	15.0	36.0	80.0	27.0
Focal - Large Skewed Distribution												
10% Focal	20.0	40.0	34.0	23.0	31.5	21.0	14.0	20.5	12.0	26.0	35.5	26.0
50% Focal	39.0	86.5	87.0	53.0	81.5	67.0	15.0	55.0	55.0	56.0	86.5	81.0
Focal - Both Skewed Distributions												
10% Focal	10.0	39.5	18.0	16.0	35.5	14.0	9.0	17.5	8.0	19.0	38.5	19.0
50% Focal	26.0	82.5	67.0	40.0	83.5	53.0	10.0	50.0	37.0	44.0	88.5	62.0

Note: Regardless of the Focal group distribution, Referent group abilities were consistently

drawn from a N(0,1) distribution. Percentages were computed by dividing by the number of items modified to exhibit U-DIF within the Referent difficulty range (Low = 1, Medium = 4, High = 1).

Table I4a

Power Rates for Non-uniform DIF Items by Difficulty in Best Subtest 10% DIF.

Aggregation/ Severity of DIF	Mantel-Haenszel			SIBTEST			Crossing SIBTEST			Both SIBTESTS		
	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High
Overall	**13.8**	**76.2**	**80.6**	**13.8**	**70.3**	**73.7**	**10.5**	**53.7**	**58.6**	**18.3**	**72.7**	**76.4**
***N*=500**	**9.1**	**61.3**	**69.0**	**10.3**	**57.6**	**60.7**	**8.8**	**37.3**	**42.4**	**14.4**	**60.7**	**63.9**
Normal Distribution (All Groups)												
10% Focal												
Moderate	4.0	31.0	39.0	9.0	38.0	32.0	6.0	15.0	19.0	10.0	40.0	36.0
Large	2.0	46.0	73.0	8.0	42.0	63.0	6.0	30.0	38.0	10.0	48.0	65.0
Both	8.0	37.0	49.0	11.0	44.0	40.0	9.0	29.0	23.0	13.0	47.0	45.0
50% Focal												
Moderate	11.0	62.0	77.0	15.0	62.0	82.0	10.0	47.0	66.0	19.0	66.0	84.0
Large	5.0	92.0	99.0	5.0	92.0	100.0	8.0	79.0	93.0	10.0	93.0	100.0
Both	5.0	90.0	96.0	5.0	93.0	97.0	16.0	67.0	82.0	16.0	95.0	98.0
Focal - Moderately Skewed Distribution												
10% Focal	11.0	30.0	35.0	20.0	24.0	25.0	8.0	13.0	13.0	22.0	28.0	29.0
50% Focal	18.0	64.0	73.0	14.0	55.0	56.0	12.0	26.0	34.0	20.0	58.0	59.0
Focal - Large Skewed Distribution												
10% Focal	9.0	62.0	56.0	9.0	50.0	39.0	11.0	30.0	14.0	16.0	57.0	47.0
50% Focal	10.0	91.0	99.0	9.0	79.0	89.0	6.0	49.0	62.0	11.0	79.0	90.0
Focal - Both Skewed Distributions												
10% Focal	9.0	43.0	42.0	6.0	41.0	25.0	6.0	21.0	11.0	11.0	44.0	34.0

Aggregation/ Severity of DIF	Mantel-Haenszel			SIBTEST			Crossing SIBTEST			Both SIBTESTS		
	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High
50% Focal	17.0	87.0	90.0	12.0	71.0	80.0	7.0	41.0	54.0	15.0	73.0	80.0
***N*=1250**	**18.6**	**91.1**	**92.3**	**17.3**	**83.1**	**86.8**	**12.3**	**70.1**	**74.8**	**22.1**	**84.7**	**88.8**
Normal Distribution (All Groups)												
10% Focal												
Moderate	13.0	62.0	78.0	19.0	61.0	73.0	14.0	45.0	57.0	24.0	64.0	78.0
Large	10.0	97.0	99.0	11.0	93.0	100.0	7.0	78.0	94.0	15.0	95.0	100.0
Both	6.0	81.0	94.0	7.0	78.0	89.0	13.0	63.0	78.0	14.0	79.0	93.0
50% Focal												
Moderate	17.0	100.0	100.0	19.0	100.0	100.0	23.0	94.0	98.0	29.0	100.0	100.0
Large	11.0	100.0	100.0	15.0	100.0	100.0	15.0	100.0	100.0	23.0	100.0	100.0
Both	12.0	100.0	100.0	13.0	100.0	100.0	18.0	97.0	99.0	21.0	100.0	100.0
Focal - Moderately Skewed Distribution												
10% Focal	22.0	72.0	55.0	13.0	49.0	52.0	5.0	32.0	33.0	13.0	52.0	56.0
50% Focal	33.0	96.0	97.0	35.0	95.0	95.0	12.0	74.0	77.0	36.0	95.0	98.0
Focal - Large Skewed Distribution												
10% Focal	24.0	96.0	99.0	14.0	69.0	76.0	10.0	44.0	45.0	16.0	71.0	78.0
50% Focal	26.0	100.0	100.0	29.0	98.0	99.0	8.0	86.0	95.0	32.0	99.0	99.0
Focal - Both Skewed Distributions												
10% Focal	20.0	90.0	85.0	9.0	56.0	58.0	9.0	45.0	38.0	13.0	62.0	65.0
50% Focal	29.0	99.0	100.0	24.0	98.0	99.0	13.0	83.0	84.0	29.0	99.0	99.0

Note: Regardless of the Focal group distribution, Referent group abilities were consistently

drawn from a N(0,1) distribution. Percentages were computed by dividing by the number of items modified to exhibit NU-DIF within the Referent difficulty range (Low = 1, Medium = 1, High = 1).

Table I4b

Power Rates for Non-uniform DIF Items by Difficulty in Best Subtest 20% DIF.

Aggregation/ Severity of DIF	Mantel-Haenszel			SIBTEST			Crossing SIBTEST			Both SIBTESTS		
	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High
Overall	**17.2**	**32.4**	**80.0**	**19.3**	**33.1**	**71.8**	**17.2**	**32.7**	**57.2**	**26.8**	**43.7**	**74.2**
***N*=500**	**12.5**	**25.4**	**68.2**	**15.6**	**28.0**	**58.3**	**12.2**	**22.8**	**41.3**	**21.1**	**35.6**	**61.8**
Normal Distribution (All Groups)												
10% Focal												
Moderate	8.5	12.0	40.0	15.0	16.7	27.0	10.0	9.3	21.0	19.0	19.3	35.0
Large	10.0	21.0	65.0	23.5	27.0	59.0	14.0	20.0	38.0	27.0	32.3	60.0
Both	10.5	12.3	53.0	16.0	14.7	46.0	14.5	14.0	24.0	22.0	21.0	48.0
50% Focal												
Moderate	17.5	25.7	76.0	18.5	28.7	79.0	17.0	24.0	63.0	24.5	35.3	82.0
Large	27.5	42.0	100.0	26.5	45.3	100.0	19.5	50.7	96.0	31.5	61.3	100.0
Both	17.0	35.0	93.0	19.0	36.3	96.0	17.5	35.3	85.0	25.0	45.7	97.0
Focal - Moderately Skewed Distribution												
10% Focal	9.0	12.7	33.0	11.0	17.3	18.0	9.5	12.3	9.0	17.0	22.7	23.0
50% Focal	12.0	25.7	60.0	12.0	29.7	51.0	7.5	18.7	29.0	15.5	34.3	55.0
Focal - Large Skewed Distribution												
10% Focal	10.5	24.7	65.0	11.0	26.7	38.0	8.5	18.7	17.0	17.5	34.3	46.0
50% Focal	11.0	38.7	95.0	11.5	38.3	81.0	13.0	30.7	59.0	20.5	47.0	82.0
Focal - Both Skewed Distributions												
10% Focal	7.0	19.7	52.0	13.0	21.7	30.0	8.0	16.3	9.0	17.5	31.7	34.0

Aggregation/ Severity of DIF	Mantel-Haenszel			SIBTEST			Crossing SIBTEST			Both SIBTESTS		
	Low	Med	High	Low	Med	High	Low	Med	High	Low	Med	High
50% Focal	10.0	35.3	86.0	10.0	34.0	75.0	7.5	23.3	45.0	16.0	42.3	80.0
***N*=1250**	**21.9**	**39.4**	**91.9**	**23.0**	**38.2**	**85.3**	**22.2**	**42.6**	**73.1**	**32.5**	**51.8**	**86.6**
Normal Distribution (All Groups)												
10% Focal												
Moderate	9.0	21.3	75.0	12.0	21.3	68.0	13.5	19.7	50.0	19.0	28.0	70.0
Large	23.5	38.7	99.0	27.5	40.3	97.0	27.0	42.7	88.0	35.5	51.7	97.0
Both	15.5	33.3	94.0	24.5	33.7	88.0	21.0	31.3	76.0	29.0	41.7	88.0
50% Focal												
Moderate	33.0	35.3	99.0	32.5	35.7	100.0	25.0	41.7	96.0	36.0	46.3	100.0
Large	49.5	57.0	100.0	48.5	57.7	100.0	43.0	81.7	100.0	54.5	87.0	100.0
Both	40.0	43.3	100.0	40.5	43.7	100.0	33.0	58.7	99.0	47.0	63.0	100.0
Focal - Moderately Skewed Distribution												
10% Focal	8.0	32.7	66.0	5.0	32.0	50.0	9.0	24.7	28.0	13.0	38.3	54.0
50% Focal	16.0	37.3	95.0	21.5	36.3	90.0	20.0	38.3	72.0	34.5	48.7	90.0
Focal - Large Skewed Distribution												
10% Focal	12.0	44.0	92.0	8.0	38.3	73.0	17.0	35.7	36.0	23.5	50.0	75.0
50% Focal	19.0	51.0	100.0	23.5	49.3	100.0	24.0	61.3	99.0	42.0	70.3	100.0
Focal - Both Skewed Distributions												
10% Focal	16.5	37.3	83.0	8.0	28.0	59.0	10.5	25.3	40.0	15.0	36.3	66.0
50% Focal	20.5	41.3	100.0	24.0	41.7	99.0	23.5	50.7	93.0	41.5	60.7	99.0

Note: Regardless of the Focal group distribution, Referent group abilities were consistently

drawn from a N(0,1) distribution. Percentages were computed by dividing by the number of items modified to exhibit NU-DIF within the Referent difficulty range (Low = 2, Medium = 3, High = 1).

APPENDIX J

EXAMINATION OF ADDITIONAL CONDITIONS

To explore the reasons ATFIND performed so differently than expected, several additional conditions were run. These additional conditions focused on changes in sample size but also explored severity of DIF modification. Sample size was chosen as a variable of interest both because of the slight increase found in the Medium difficulty items chosen for the AT List with an increase in sample size and because the DIF analyses tended to have higher power as sample size increased. Severity of DIF modification was included in these conditions because it was found to influence the DIF analyses ability to identify items, especially when the matching subtest had DIF contamination.

Since the DIF analyses tended to have their best performance in Large sample, 20% DIF items, Large DIF conditions with 50% Focal simulees, and ability distributions for both the referent and focal groups drawn from the normal distribution, this condition was used as a basis for comparisons to the additional conditions. One of the differences between this original condition and the additional ones is that in all original conditions both uniform and non-uniform DIF items were modelled (with 20% DIF, 6 of each type), where in the additional conditions, all DIF items were modified to exhibit uniform DIF (only a change in the difficulty parameter for the focal group as compared to that of the referent). Both the non-DIF and the referent item parameters were held constant with those used in the original conditions.

The process for DIF modification selection also changed between the original conditions and the additional ones. With the original conditions, focal item parameters for each DIF item were randomly drawn from a uniform distribution and then applied consistently across the various conditions. In the additional conditions, only two values (1.00 and 0.80) were chosen to subtract from the referent item difficulty. These were held constant

for all items within a condition, but varied across conditions. The value of 1.00 was chosen to align with that used by Gierl, Jodoin, and Ackerman (2000), where the value of .80 was approximately halfway between their value and that used by Swaminathan and Rogers (.64) in their 1990 study. Both of these studies referred to the values they chose as a Large DIF modification. The values subtracted from the referent difficulty parameter for Large DIF in the original conditions were drawn from a uniform distribution of .50 to .80; however, only two values over .70 were chosen. The variables that were held constant across all of the additional conditions were 50% Focal simulees, *N*(0,1) ability distribution for both groups, 50% of the total sample used for the ATFIND analysis, 20% DIF (12 out of 60 items), and 100% of the DIF items modified with uniform DIF. Table J1 presents a summary of the original and additional conditions. The sample size of 1,500 was chosen to match the "Large" sample size for ATFIND, *N*=750, in the original condition and then was doubled so that an equal number of simulees was available for DIF analyses.

Table J1.

Summary of the original and additional conditions.

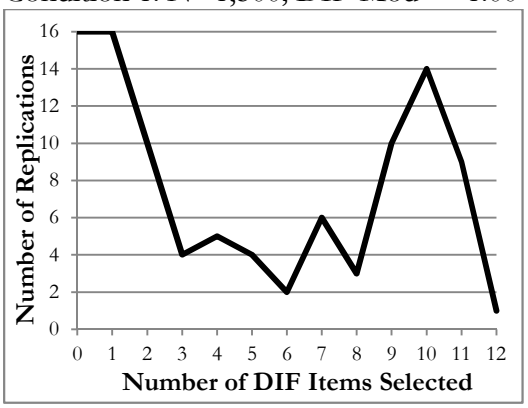
Condition	Total *N*	% ATFIND	% Focal	% DIF	% Uniform DIF	DIF Modification
Original	2000	37.5	50	20	50	Varied
1	1500	50	50	20	100	-1.00
2	20000	50	50	20	100	-1.00
3	1500	50	50	20	100	-0.80
4	20000	50	50	20	100	-0.80

Note: The total sample size of 1500 was chosen to match the large sample size of 750 used for ATFIND in the original condition.

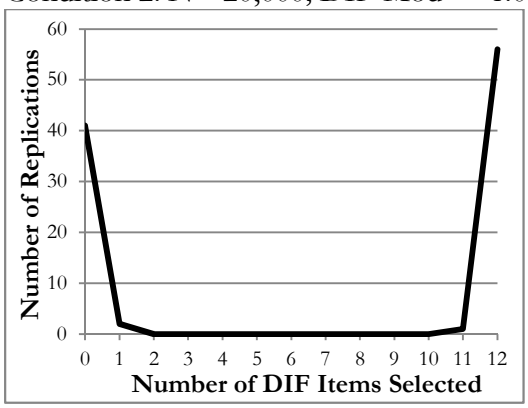
Figure J1 presents the hit frequencies for the number of DIF items selected for the AT List in the original condition and all four additional conditions. As can be seen in these graphs, ATFIND selects DIF items very differently for the two sample sizes. With the Large sample, especially with a DIF modification of 1.00 (Condition 2), ATFIND tended to select either all of the DIF items for the AT List, or none at all. Where with the Small sample and a DIF modification of 1.00 (Condition 1), ATFIND selected between 3 and 10 DIF items many more times. The slight decrease in severity of DIF modification to -0.80 was seen to impact the Small sample size (Condition 3) much more than that of the Large (Condition 4), with Condition 3 approaching the distribution of DIF hits seen with the Original condition. It seems that the sample size used in the Original condition simply did not provide ATFIND enough power to identify the DIF items, especially when DIF was modeled within the ranges found in real data.

An examination of item parameters for the non-DIF items chosen by ATFIND for the AT List when the DIF items were selected for that list, and when DIF items were all selected for the PT List provides a possible answer to ATFIND's behavior in Condition 2. To evaluate how well the discrimination and difficulty item parameters for the 48 non-DIF items predicted their placement 1) with the DIF items in the AT List or 2) in the AT List when all of the DIF items were placed in the PT List, a series of multiple regression analyses were performed. In Condition 2, there were 30 replications where only the 12 DIF items were selected for the AT List, 27 replications where all (or almost all, 11 in one condition) of the DIF items plus a few non-DIF items (between 1 and 11) were selected, and 41 replications where only non-DIF items (between 22 and 30) were selected for the AT List.

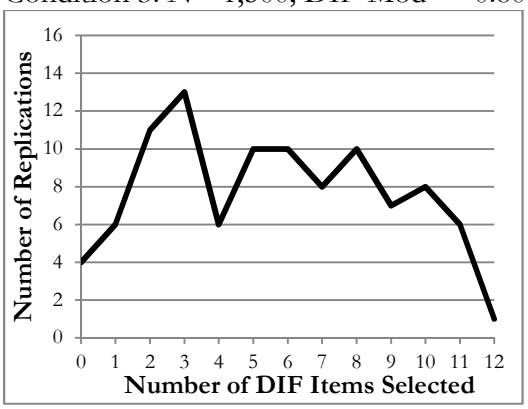
Condition 1: *N*=1,500; DIF-Mod = -1.00



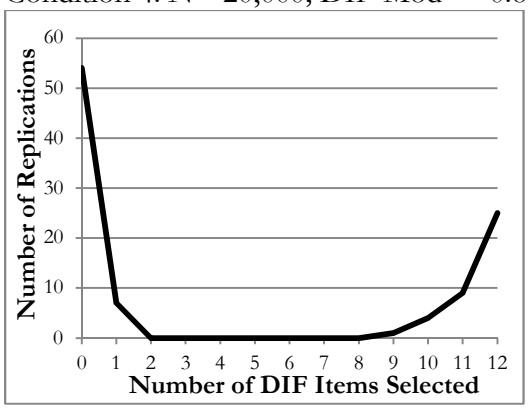
Condition 2: *N*= 20,000; DIF-Mod = -1.00



Condition 3: *N*= 1,500; DIF-Mod = -0.80



Condition 4: *N*= 20,000; DIF-Mod = -0.80



Original: *N*= 2,000; DIF-Mod = Various

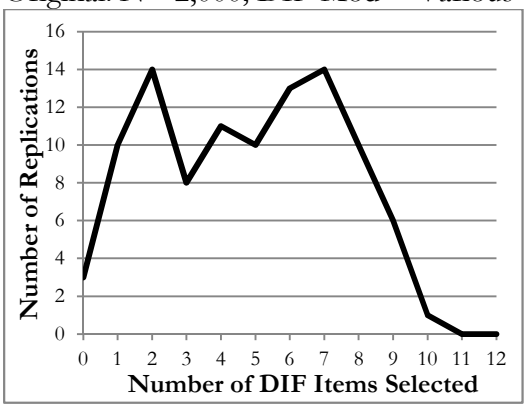


Figure J1. Hit frequencies for the number of DIF items selected for the AT List. These frequencies are for all DIF items. For Conditions 1 through 4, DIF items were only modelled for uniform DIF where for the Original Condition; only 50% of the DIF items had a uniform DIF modification.

The linear combination of the discrimination and difficulty parameters of the non-DIF items was significantly related to the item's placement with the DIF items on the AT List, $F(2,45) = 10.66$, $p < .01$, with the sample multiple correlation (.57) indicating that the linear combination of the parameters accounts for approximately 29% (adjusted) for the variance in placement with the DIF items on the AT List. Both of the item parameters were negatively correlated with placement on the AT List with the DIF items; however, only the discrimination parameter was significantly related (-.52, $p < .01$). Figure J2 presents a graph of the percentage of times non-DIF items were placed on the AT List along with the DIF items by discrimination parameter. As can be seen in this graph, items with lower discrimination parameters were much more likely than items with high discrimination parameters to be placed on the AT List along with the 12 DIF items. Since ATFIND places

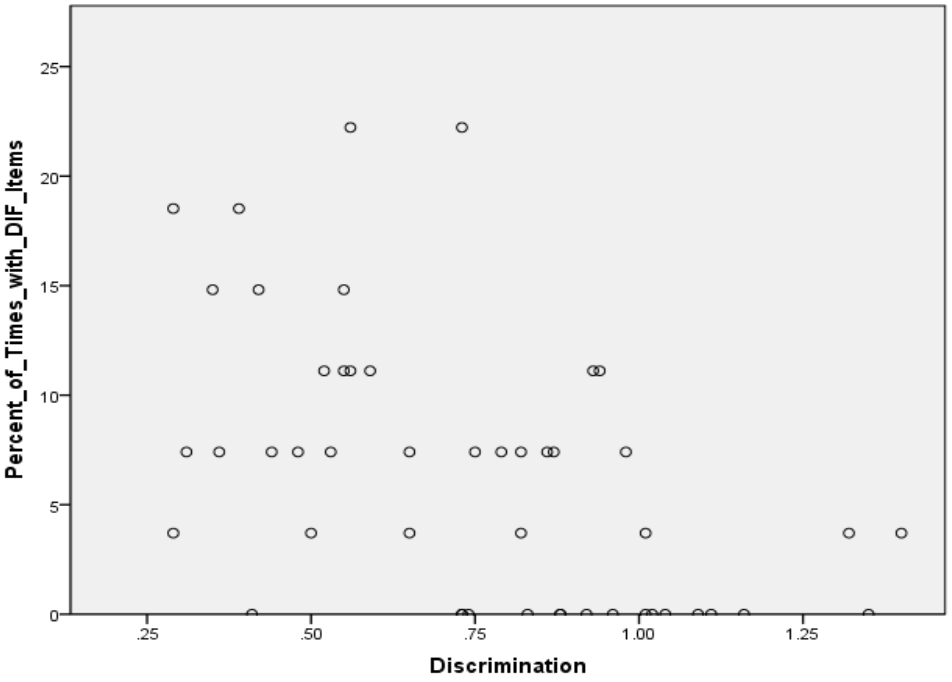


Figure J2. Percentage of times non-DIF items were placed on the AT List along with the 12 DIF items in Condition 2 by item discrimination parameter.

the items on the AT List that are dimensionally homogenously distinct from those on the PT List, it would seem, based on these results, that ATFIND is sometimes finding the DIF items to be of a similar dimension to non-DIF items that have low discrimination. Similarly, the linear combination of the two parameters was significantly related to their placement on the AT List when the DIF items were placed on the PT List, $F(2,45) = 19.38$, $p < .01$. The sample multiple correlation (.68) for this analysis indicates that the two item parameters for the non-DIF items accounts for approximately 44% (adjusted) of the variance in placement on the AT List when the DIF items were placed on the PT List. As expected, given the results of the first analysis, both of the item parameters were positively correlated to placement on the AT List without the DIF items. Here again, only one of the correlations was statistically significant; however, for this analysis the parameter was item difficulty (.61, $p < .01$). Figure J3 presents a graph of the percentage of times non-DIF items were placed on the AT List without the DIF items by difficulty parameter. As can be seen in this graph, items with difficulty parameters in the middle of the range (-.75 to 1.19) were much more likely than items with either high or low difficulty parameters to be placed on the AT List when the 12 DIF items were placed on the PT List.

Because of the apparent quadratic nature of the non-DIF difficulty parameter selection for the AT List, a curve estimation regression procedure was performed. This analysis was also statistically significant $F(2,45) = 49.43$, $p < .01$ and accounted for approximately 67% (adjusted) of the variance of non-DIF items placement on the AT List without the DIF items. The resultant regression equation was

$$Placement\ on\ AT\ List = 5.67\ Difficutly - 8.35\ Difficulty^2 + 65.01.$$

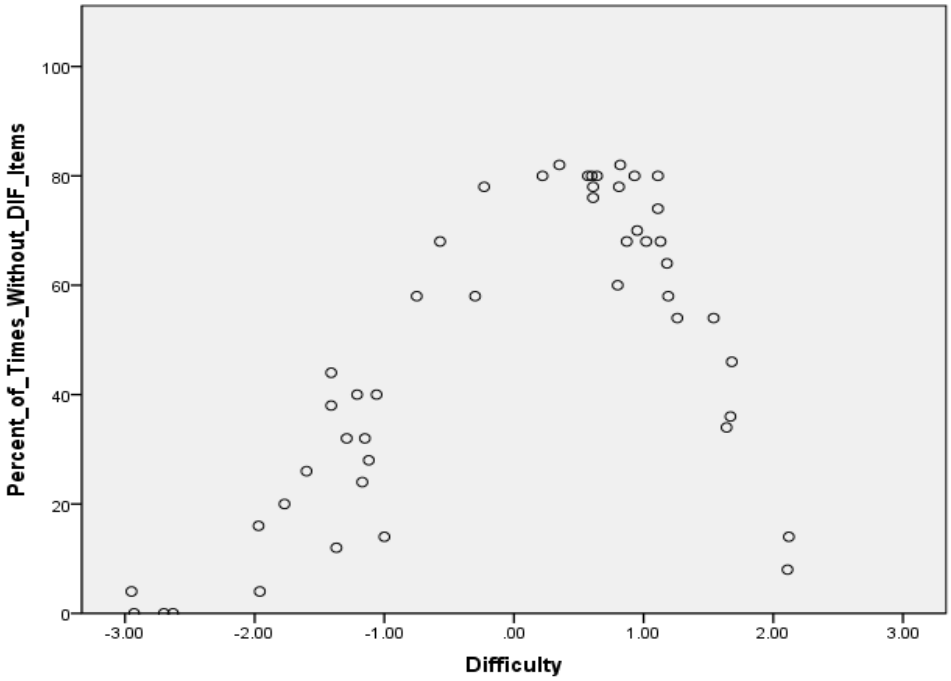


Figure J3. Percentage of times non-DIF items were placed on the AT List along without the 12 DIF items in Condition 2 by item difficulty parameter.

ATFIND selected this group of medium difficulty non-DIF items for the AT List at approximately the same, but lower, rate (41%) as it did the DIF items (57%). This indicates that ATFIND was having difficulty identifying which of the two groups of items was most "relatively dimensionally homogeneous" (Froelich & Habing, 2008, p. 144), even when both the sample size (10,000 used for ATFIND) and the DIF modification (-1.00 from difficulty) were quite large. Given the results found here for Condition 2, which used both a very large sample and a very large DIF modification, perhaps it should be less of a surprise that it had similar issues when a much smaller sample and lower DIF modification was used as in the Original condition.

Not only were ATFIND results examined for the additional conditions, but MH's, SIB's, and X-SIB's results were also examined and compared to the results found for all DIF items in the Original condition. Figure J4 presents the Total Power rates for the uniform DIF items in the Best matching subtest and the PT matching subtest for the three DIF analyses, as well as the Analyzed power rates for SIB and X-SIB for the PT matching subtest (noted DIF-AT). Also included in Figure J4 are the results of the Original condition where one-half of the DIF items were modeled for uniform DIF and the rest were modified to exhibit non-uniform DIF. As can be seen with these graphs, with the Best matching subtest, all three analyses have very high power (consistently100% with sample size of 20,000) and power of over 90% with a sample size of 750. This is comparable with the power rates in the Original condition for uniform DIF items using the Best matching subtest which were 97.3, 97.5, and 95.7 percent for MH, SIB, and X-SIB, respectively. It was only with the inclusion of non-uniform DIF items in this Original condition that the power for these three analyses dropped below 90%, 79.5, 79.6, and 83.8 percent, respectively.

When the PT matching subtest was used, MH was able to maintain its high power (over 90%) as long as the sample size was 20,000 or the DIF modification was 1.00. It dropped slightly below 90% to 88.6% when both sample size and DIF modification was reduced. As might be expected, due to SIBTEST's default of examining only the suspect items in the AT List, all of the SIB and X-SIB Total power rates were much lower with the PT matching subtest. Both of these analyses reached over 50% power only with a sample size of 20,000 and the 1.00 DIF modification. Their Analyzed power rates however, similar to MH's, were generally over 90%, with their lowest found with a sample size of 1500 and the .80 DIF modification (91.2 and 80.3 percent for SIB and X-SIB, respectively). While the

rates displayed within these graphs are generally very positive, power is only part of the story.

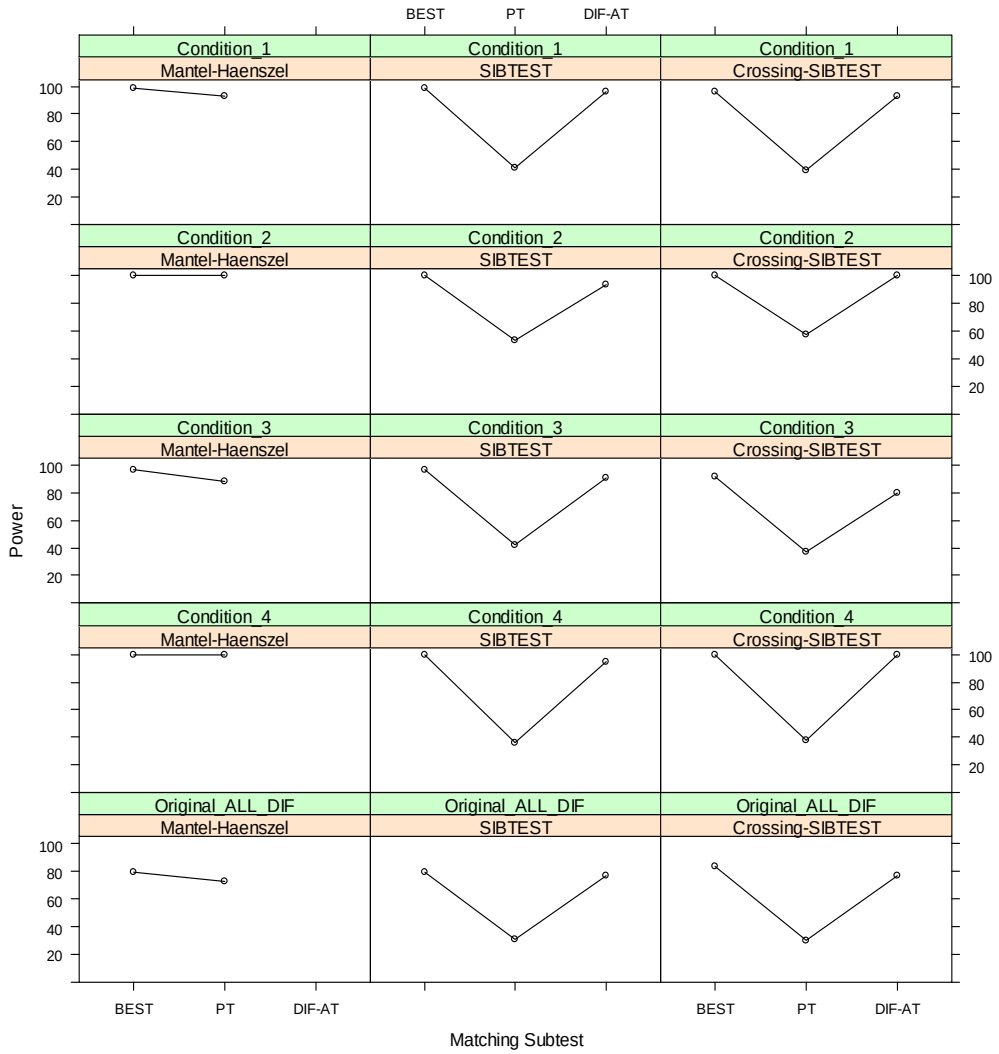


Figure J4. Power rates for DIF items within all difficulty ranges. BEST = Best matching subtest; PT= PT List matching subtest; DIF-AT= Analyzed power rate with PT List matching subtest. DIF items within Conditions 1 through 4 were consistently modified for uniform DIF.

Figure J5 presents Analyzed Type I error rates for the Best matching subtest and the PT matching subtest for the three DIF analyses. As can be seen within these graphs, while Type I error was generally low (rarely higher than 5%) with the Best matching subtest, it was sometimes extremely high (over 90% in two instances) with the PT matching subtest. With the PT matching subtest, Type I error rates were seen to increase with the higher DIF

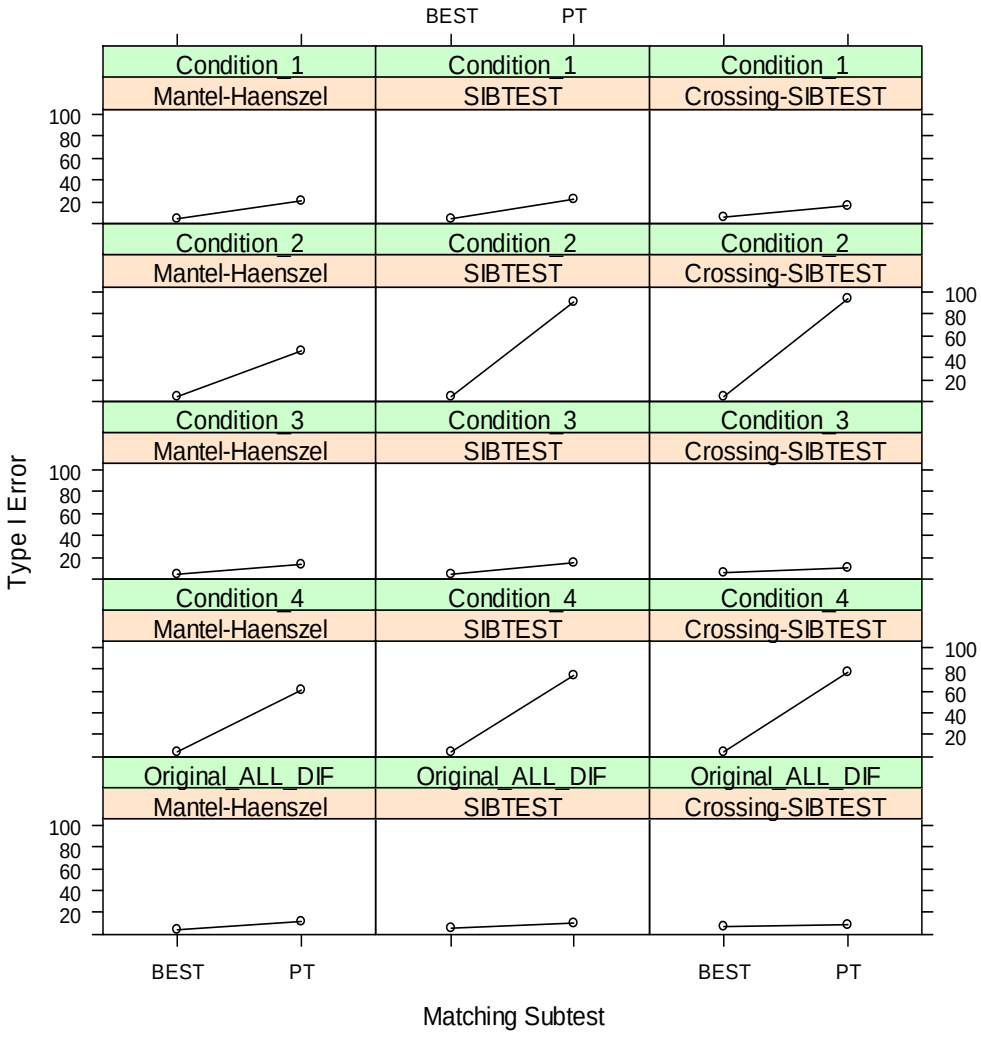


Figure J5. Type I error rates for non-DIF items within all difficulty ranges. BEST = Best matching subtest; PT= PT List matching subtest.

modification (from an average of 13.0% to an average of 20.1%), but were found to greatly increase with the largest sample size (from an average of 16.6% with the Small sample size to an average of 73.7% with the Large). Error rates this large confirm the results of the main part of this dissertation, which found that the PT List is unusable as a matching subtest. Here, error rates are so large as to strongly advise against its use regardless of the additional examination of content of DIF identified items that are usually carried out in bias studies.

Discussion and Opportunities for Future Investigations

This section provides a discussion of the results found within the additional conditions and highlights the findings and what remains to be researched, as well as some suggestions as to how those research projects might be tied to this study. Consistent with both the original study and this additional one, the ATFIND analysis will be discussed first, followed by the DIF analyses.

ATFIND. Within the results presented here, there were four findings. These were related to sample size, DIF modification, non-DIF item difficulty parameters, and non-DIF item discrimination parameters. These finding will be discussed in turn.

Sample size has a large influence on ATFIND's ability to identify DIF items as a homogenously dimensionally distinct group of items from the rest of the items. For example, when 10,000 simulees were used, it was generally able to select the DIF items together for either the AT or the PT List. It selected all of the DIF items together for one of these two lists in 97 of the 100 replications with DIF modification of -1.00, and in 79 of the 100 replications with DIF modification of -.80. However, ATFIND was not generally able to group the DIF items together with a sample size of 750. With this sample size, ATFIND only selected all of the DIF items together for one of the two lists in 17 of the 100

replications, and in 5 of the 100 replications with DIF modification of -.80. Since only two sample sizes (750 and 10,000) were used for this study, we don't know at what level of sample size ATFIND starts having difficulty grouping the DIF items together. One suggestion for a future study would be to use the large uniform DIF modification (1.00) and vary sample size to find out at what sample size level it stops selecting most of the DIF items together for one of the two lists.

Severity of uniform DIF induced in the items has some influence on ATFIND's ability to identify DIF items as a homogenously dimensionally distinct group of items from the rest of the items over and above sample size. An example of this was observed when 10,000 simulees were used for ATFIND. ATFIND was able to select the DIF items together for either the AT or the PT List at a higher rate with the -1.00 DIF modification than with the -.80 DIF modification (in 97 and in 79 of the 100 replications, respectively). Even with the smaller sample size of 750, there was an observed decrease in all of the DIF items selected together for one of the two lists (in 17 and in 5 of the 100 replications for the -1.00 and -.80 DIF modifications, respectively). The one replication in the smaller sample, lower DIF modification where all DIF items were selected for the AT List was still higher than found in the original study. Originally, with 20% DIF items in the 2400 replications, at no time did ATFIND select all 12 DIF items for the AT List and, it selected 11 DIF items in only two of the replications. Therefore, even with these additional conditions, we still don't know at what level of DIF severity ATFIND starts selecting no DIF items for the AT List. A follow-up study could use the large sample size (10,000) and vary uniform DIF severity to find the level of severity at which ATFIND stops selecting all twelve DIF items together for the AT List.

Among the non-DIF items ATFIND selects for the AT List, when selecting all of the DIF items for the PT List, it tends to select items in the Medium difficulty range most often. For example, when 10,000 simulees were used for ATFIND in conjunction with the larger DIF modification (-1.00), all 23 of the non-DIF items selected for the AT List in at least 29 of the 41 replications (approximately 60%) when all of the DIF items were selected for the PT List had difficulty parameters in the middle of the range (-.75 to 1.19). Additionally, in this condition, none of the 25 non-DIF items selected for the AT List in less than 28 of the replications when all of the DIF items were selected for the PT List had difficulty parameters within this range. Since ATFIND's tendency to select items in the Medium difficulty range was also found as a trend in the original study, especially when no DIF items were included in the dataset, we don't know whether this behavior is caused by the severity of DIF induced in 20% of the items, or if it would be a significant factor also when no DIF items were included in a large sample size dataset. One way this could be explored would be to use the large sample size (10,000) and vary both the percentage of DIF items and uniform DIF severity to find at what combination, ATFIND's selection of medium difficulty non-DIF items for the AT List, when selecting all DIF items for the PT List, becomes non-significant.

Among the non-DIF items that ATFIND selects for the AT List, when also selecting all of the DIF items for the AT List, the items with a low discrimination tend to be selected most often. This was evidenced by the significant negative relationship between the placement of non-DIF items with low discriminations on the AT List and all of the DIF items on that list, when 10,000 simulees were used in conjunction with the larger DIF modification (-1.00). Since the behavior of non-DIF items selected with the DIF items was

not explored in the original study or in any other of the additional conditions, we don't know whether this selection is generalizable to other conditions such as lower sample sizes or lower severity of DIF. To investigate the generalizability of this relationship, the selection of non-DIF items for the AT List in the rest of the additional conditions could be examined, as well varying both sample size and uniform DIF severity in more additional conditions. The goal would be to find the combination at which low-discrimination non-DIF items selection for the AT List along with the DIF items becomes non-significant.

DIF analyses. Both sample size and severity of DIF modification have an influence on MH's, SIB's, and X-SIB's ability to identify uniform DIF items. The evidence for this statement was provided when 10,000 simulees were used for the DIF analyses. With this large sample size, all three procedures were able to identify all DIF items in all replications (100% power) for both DIF modifications, as long as there were no DIF items in the matching subtest. However, when the sample size was dropped to 750, a slight drop in power for them all was observed. Also, when the severity of DIF was decreased, regardless of sample size, power for all three analyses decreased. Since both DIF modifications used in the additional conditions were considered to be in the large range by previous researchers, we don't know at what level of DIF modification, power rates would become unacceptable. The use of additional Condition 4 (.80 DIF modification and 10,000 simulees) as a base for more conditions that vary both sample size and severity of uniform DIF modification, may lead to a sample size for each of the DIF modifications where power becomes unacceptable.

Type I error rates for all of three of the DIF analyses are inflated with inclusion of items with large DIF modification that contaminate the matching subtest, especially as sample size increases. With no DIF items contaminating the matching subtest, as with the

use of the Best matching subtest, all of three DIF analyses consistently had Analyzed Type I error rates below the 5% nominal rate with the large sample size. Also, only X-SIB had a slightly elevated rate (approximately 6%) with a sample size of 750. However, when the PT matching subtest which contained items modified to exhibit large DIF was used with a sample size of 750 and the lower DIF modification (-.80), all of the analyses had Analyzed Type I error rates of over 10%. When larger DIF modification was used, these error rates increased to approximately 20% for the same sample size. The Analyzed Type I error rates increased even further when the sample size used for DIF analysis was increased to 10,000, reaching over 90% for both SIB and X-SIB with the larger DIF modification. We don't know the amount of DIF contamination (percentage of DIF items and the severity of DIF induced in those items) that the analyses can handle in their matching subtest at various sample sizes and still maintain an acceptable level of Type I error. Since Type I error increased with the larger sample size, the use of the Best matching subtest with Condition 4 (sample size of 10,000 and DIF modification of .80) might serve as a basis for additional conditions. These conditions might change the matching subtests by adding various percentages of DIF items at various DIF modifications. The goal would be to identify both the severity of DIF and the percentage of DIF items that the analyses can handle, while maintaining an acceptable Type I error rate.

Because all of the additional conditions were modeled only for uniform DIF, these suggestions are specifically for that type of DIF item. However, since previous studies have indicated that the use of accommodations could cause non-uniform DIF, all of these studies probably should also be done not only for uniform DIF items, but also for only non-uniform DIF items, as well as for combinations of uniform and non-uniform DIF items.