

ABSTRACT

Title of Dissertation: A PROPOSED INDEX TO DETECT RELATIVE ITEM PERFORMANCE WHEN THE FOCAL GROUP SAMPLE SIZE IS SMALL

Kari Kristine Hansen, Doctor of Philosophy, 2017

Dissertation directed by: Professor Laura Stapleton
Professor Hong Jiao
Measurement, Statistics and Evaluation Program
Department of Human Development & Quantitative Methodology

When developing educational assessments, ensuring that the test is fair to all groups of examinees is an essential part of the process. The primary statistical method for identifying potential bias in assessments is known as differential item functioning (DIF) analysis, where DIF refers to differences in performance on a specific test item between two groups assuming that the two groups have an overlap in their ability distribution. However, this requirement may be less likely to be feasible if the sample size for the focal group is small.

A new index, relative item performance, is proposed to address the issue of small focal group sample sizes without the requirement of an overlap in ability distribution. This index is calculated by obtaining the effect size of the difference in item difficulty estimates between the two groups. A simulation study was conducted to compare the proposed method with the Mantel-Haenszel test with score group widths and the Differential Item Pair Functioning in terms of Type I error rates and power. The following factors were manipulated: the sample size of the focal group, the mean of the ability distribution, the amount of DIF, the number of items on the assessment, and the number of items that have different item difficulties.

For all three methods, the main factors that affect the Type I error rates are the amount of item contamination, the size of the DIF, the ability mean for the focal group, and the item parameters. The sample size and the number of items were found not to have an effect on the Type I error rates for all methods. As the Type I error rate overall for the RI method is much lower than that of the MH1 and MH2 methods and not controlled across the simulation factors, power was only evaluated for the MH1 and MH2 methods. The median power of these methods were .203 and .181, respectively. It is recommended that the MH1 and MH2 methods be used only when the sample size is larger than 100 and in conjunction with expert and cognitive review of the items on the assessment.

A PROPOSED INDEX TO DETECT RELATIVE ITEM PERFORMANCE WHEN
THE FOCAL GROUP SAMPLE SIZE IS SMALL

By

Kari Kristine Hansen

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:

Professor Laura M. Stapleton, Chair
Professor Hong Jiao, Co-Chair
Professor Jeffrey Haring
Professor Ana Taboada Barber
Professor Ji Seung Yang

©Copyright by

Kari Kristine Hansen

2017

Acknowledgements

Completing this doctoral dissertation is the culmination of seven years of classes, discussions, and research. Through the ups and downs of this process, this would not have been possible without the amazing support and guidance of people from the first day of this journey.

First, I am grateful to my dissertation committee for all of your involvement in making this happen. A few of the members of this committee have shown support from the beginning as they have taught me in their classes such as Dr. Jeffrey Harring in Mathematical Foundations and Simulation Techniques and Dr. Hong Jiao in Modern Measurement Theory. Classes like these became the framework for this dissertation. I especially must thank my co-chairs for their advice and guidance in helping to shape this dissertation. I could always count on Dr. Laura Stapleton as she would frame her comments on the dissertation in the form of a question that would make me evaluate how I was explaining the research.

Second, teaching at Gallaudet University full-time while working on this dissertation required time management and the amazing support of my Department of Business colleagues. My colleagues would always encourage me to stay focused and be willing to take over classes when I needed time off. I especially thank Ms. Emilia Chukwuma (department chair), Dr. Khadijat Rashid (Dean of the School of Education, Business, and Human Services (SEBHS)), and Dr. Carol Erting (Provost) for their support.

Third, I could not have done this without my friends who became my cheerleaders, fans, and supporters. My time with this dissertation has been a long journey and I could not have done this without your understanding and patience. Without your cheers and words of encouragement, I would not have made it today!

Finally, my family has been my bedrock on which I stand throughout this process. My parents, my sister, my grandparents, aunts, uncles, and cousins have all sent words of encouragement, words of pride, and love even from 3,000 miles away! I love you all!

Table of Contents

Acknowledgements.....	ii
Table of Contents.....	iii
List of Tables.....	v
List of Figures.....	vi
Chapter 1 – Introduction.....	1
Chapter 2 – Literature Review.....	6
Empirical Studies with Small Focal Group Sample Sizes.....	6
Explanation of Differential Item Functioning.....	9
Simpson’s paradox.....	9
Classical test theory.....	10
Item response theory.....	12
Summary.....	15
Methods to Detect DIF within the CTT Framework.....	15
Mantel-Haenszel test.....	16
Standardization.....	21
Logistic regression.....	23
Performance of the CTT methods.....	24
Methods to Detect DIF within the IRT Framework.....	30
Wald test.....	30
Differential item pair functioning.....	32
Raju’s signed or unsigned areas.....	36
Sample size requirements.....	38
Rationale for the Proposed Method.....	40
The Proposed Method to Detect Relative Item Performance.....	41
Research Questions.....	44
Chapter 3 – Methods.....	45
Choosing DIF Detection Methods for Comparison.....	45
Simulation Design.....	46
Data generation.....	46
DIF detection methods.....	51
Analysis.....	52
Addressing the Research Questions.....	53
Chapter 4 – Results.....	55
Type I Error Rates Without Item Contamination.....	55
Mean of the focal ability distribution.....	58
Item parameters.....	59
Interaction between ability mean and item parameters.....	62
Summary.....	67

Type I Error Rates with Item Contamination	68
Item contamination and the size of the DIF.....	73
Interaction of amount of item contamination with other simulation factors.....	76
Interaction of the size of the DIF with other factors.....	79
Interaction between the ability mean and the item parameters.....	83
Summary.....	88
Analysis of Effect Size for the RI Method.....	89
Power	92
Sample size.....	97
Mean of the focal ability distribution.....	99
Item contamination and size of the DIF.....	100
Item parameters.....	102
Interaction of amount of item contamination with other simulation factors.....	105
Interaction of the size of the DIF with other factors.....	108
Interaction between the ability mean and the item parameters.....	111
Summary.....	116
Answering the Research Questions	116
Chapter 5 – Discussion	118
Summary of Results	118
Post Analysis.....	119
Alignment with Literature Review	122
Implications.....	123
Limitations	124
Future Research	126
Conclusion	127
Appendix A.....	128
Appendix B.....	157
References.....	165

List of Tables

Table 1	Example of Simpson’s Paradox	10
Table 2	2×2 Contingency Table for Score Interval k	18
Table 3	Example Effect Sizes for 10 Items	43
Table 4	Difficulty and Discrimination Parameters	49
Table 5	Summary of Simulation Factors	51
Table 6	Summary Statistics of Type I Error Rates Across Cells	56
Table 7	Logistic Regression Odds Ratios for Items with No Item Contamination	57
Table 8	Summary Statistics of Type I Error Rates Across Cells	68
Table 9	Logistic Regression Odds Ratios for 20 Items with Item Contamination	70
Table 10	Logistic Regression Odds Ratios for 40 Items with Item Contamination	71
Table 11	Odds Ratios for Interaction between Item Contamination and Item Parameters	79
Table 12	Effect of Simulation Factors on Type I Error Rates	89
Table 13	Descriptive Statistics of Effect Size for Items with No DIF	90
Table 14	Descriptive Statistics of Effect Size for Items with No DIF and No Item Contamination	90
Table 15	Descriptive Statistics of Effect Size for Items with No DIF and Item Contamination	91
Table 16	Descriptive Statistics of Effect Size for Items with DIF	91
Table 17	Summary Statistics for Power Across All Cells	93
Table 18	Logistic Regression Odds Ratios for Power	95
Table 19	Power for MH1 and MH2 Methods Across Sample Size	99
Table 20	Median Power Across Discrimination Parameters	104
Table 21	Odds Ratios for Interaction Between Ability Mean and Item Parameters	108
Table 22	Type I Error Rates Across Methods for Data Generated from the 1PL and 2PL IRT Models	120
Table 23	Overlap for MH1 Score Group Widths	121
Table 24	Overlap for MH2 Score Group Widths	122
Table A1	Item Parameters for 20 Items (Rounded to Four Places After the Decimal) .	128
Table A2	Item Parameters for 40 Items (Rounded to Four Places After the Decimal) .	129
Table A3	Item Parameters for 80 Items (Rounded to Four Places After the Decimal) .	131
Table A4	Median Type I Error Rates for All Conditions without Item Contamination	135
Table A5	Median Type I Error Rates with Item Contamination	136
Table A6	Median Power Across All Conditions	147

List of Figures

Figure 1. Two example item characteristic curves.	13
Figure 2. Heat map for differential item pair functioning.	35
Figure 3. Median Type I error rates across focal ability distribution.	59
Figure 4. Median Type I error rates across difficulty parameters.....	60
Figure 5. Median Type I error rates across discrimination parameters.	60
Figure 6. Median Type I error rates across item parameters.	61
Figure 7. Median Type I error rates across ability mean and item difficulty.	63
Figure 8. Median Type I error rates across ability mean and item discrimination.	65
Figure 9. Median Type I error rates across ability mean and item parameters.....	66
Figure 10. Median Type I error rates across item contamination.	73
Figure 11. Median Type I error rates across the size of the DIF.	74
Figure 12. Median Type I error rates across the size of the DIF and amount of item contamination.....	76
Figure 13. Median Type I error rates across the ability mean and amount of item contamination.....	78
Figure 14. Median Type I error rates across the item parameters and amount of item contamination.....	79
Figure 15. Median Type I error rates across the size of the DIF and the ability mean.....	81
Figure 16. Median Type I error rates across the size of the DIF and the item difficulty parameter.....	82
Figure 17. Median Type I error rates across the size of the DIF and the item discrimination parameter.	82
Figure 18. Median Type I error rates across ability mean and item difficulty.	85
Figure 19. Median Type I error rates across ability mean and item discrimination.	86
Figure 20. Median Type I error rates across ability mean and item parameters.....	87
Figure 21. Comparison of Empirical and Standard Normal Distributions.	92
Figure 22. Median power rates across focal ability distribution.....	99
Figure 23. Median power across item contamination.....	100
Figure 24. Median power rates across item contamination and size of DIF.....	101
Figure 25. Median power across difficulty parameters.	102
Figure 26. Median power across item parameters.	104
Figure 27. Median power rates across the ability mean and amount of item contamination.....	106
Figure 28. Median power rates across the item parameters and amount of item contamination.....	107
Figure 29. Median power rates across the size of the DIF and the ability mean.	109
Figure 30. Median power rates across the size of the DIF and the item difficulty parameter.....	110
Figure 31. Median power across the size of the DIF and the item discrimination parameter.....	110
Figure 32. Median power rates across ability mean and item difficulty.....	112
Figure 33. Median power rates across ability mean and item discrimination.	114
Figure 34. Median power across ability mean and item parameters.....	115

Chapter 1 – Introduction

When developing educational assessments, ensuring that the test is fair to all groups of examinees is an essential part of the process. Test fairness is defined in the Standards for Educational and Psychological Testing (American Educational Research Association, 2014) as the absence of bias, equitable treatment of all test takers in the testing process, and equity in opportunity to learn the material in an achievement test. Regarding the absence of bias, the development of test items should ensure that examinees from different backgrounds, but with similar ability levels, have similar probabilities of obtaining the correct response. To evaluate these test items, a comparison is made between two groups: a focal group that is comprised of examinees with specific backgrounds that the researcher is interested in, and a reference group which is comprised of examinees without those specific backgrounds and used as a benchmark. Another definition of the focal and reference group would be where the focal group is a subgroup that is suspected to be at risk of being disadvantaged by the test, while a reference group is a group that the test is expected to favor, and often serves as a basis for comparison (Jiao & Chen, 2014). An example would be a comparison of students with disabilities (focal) with students without disabilities (reference) such as evaluating ACT scores between deaf and hard-of-hearing students with their hearing peers. When Gallaudet University chose to use the ACT score as a criterion for admission, there were two reasons why the ACT exam may not be a valid measure for deaf and hard-of-hearing students (Gallaudet, 2007). First, the primary language used by deaf students is American Sign Language (ASL), which has its own grammar and syntax and is not a

simple translation of the English language (Valli, 2000). Second, there is a difference in learning styles as the deaf student learns visually and the hearing student learns through an auditory method (Marschark & Hauser, 2008). As a result, deaf students and hearing students may differ in their probabilities of obtaining a correct response because deaf students learn visually and the ACT test is administered through written English.

However, the sample size for the deaf/hard-of-hearing students is much smaller than that of their hearing peers, which could lead to invalid inferences about the fairness of the assessment.

Traditional statistical methods used to detect the potential bias in educational assessments require a large sample in both the focal and the reference groups in order to detect differences between groups and to reduce the risk of identifying test items as potentially biased when they are not (Zieky, 1993). These methods rely on an overlap between the two groups in terms of their knowledge or ability, usually indicated by their total test score. If the sample size for the focal group is small, meeting the requirement of an overlap of the ability distribution between the two groups may be less likely to be feasible (Dorans & Holland, 1992). Various revisions to current methods of potential bias detection have been proposed to address this issue of small focal group sizes such as using iterative procedures until a stable set of items that perform differently are identified.

There have been several empirical studies where the focal group ranged from 50 to 200 examinees (Bennett, Rock, & Kaplan, 1985; Maller, 1997; Martin, 2005; Steinberg, Cline, Ling, Cook, & Tognatta, 2014). Different methods were used to analyze the difference in item performance between the two groups where some items were

identified as being easier for one group than the other group. With small sample sizes, there is a concern that the results found may not be valid as there is a potential risk of identifying an item as performing differently when the item does not actually perform differently for the two groups. Another concern with small sample sizes is that there is the possible confounding of differences in ability with differences in item performance. When items are identified as performing differently, this could be due to the differences in the ability distributions between the two groups instead of actual differential item performance. Thus, there is the question of what is the appropriate method to use when the focal group sample size is small. The current methods, such as the Mantel-Haenszel test, standardization, logistic regression, and the Wald test, all have the requirement that there be an overlap in the abilities between the reference and the focal groups, as well as a minimum sample size of 100 if using the simplest method (the Mantel-Haenszel test) (Zieky, 1993). It would be advantageous to identify a method such that the overlap in the abilities is not a necessary requirement with the interpretation being that an item is performing differently from that of other items without regards to which group the item is easier for.

There are two separate approaches to detecting item performance: Classical Test Theory (CTT) and Item Response Theory (IRT). CTT uses only the observed data to measure how difficult and how discriminating the items are, while IRT assumes that the observed data come from a specified statistical distribution and uses that distribution to estimate the difficulty and discrimination of the items. As there are no distributional assumptions under CTT, the sample size requirement is smaller than that required for IRT.

An example of a method that uses the CTT approach is known as the Mantel-Haenszel (MH) test developed by Holland and Thayer (1988). This test matches examinees by their total test scores and then computes the log-odds ratio of obtaining a correct response on an item. Another example of a CTT method is standardization developed by Dorans and Kulick (1983) where nonparametric item test regressions are developed for the focal and reference groups, then compared. The third CTT method is known as logistic regression, which predicts the probability of obtaining a correct response depending on the examinee's estimated ability level and whether or not the examinee is in the focal group and the interaction between these two variables. In contrast, Lord (1980) developed the Wald test based on the IRT framework, which determines the statistical significance of the difference between the item parameters for the reference and focal groups. Bechger and Maris (2015) extended the Wald test to compare pairs of items by calculating the difference in estimated item difficulties. Another IRT method was developed by Raju (1988), which calculates the area between the two item characteristic curves (ICCs) when calibrating two separate IRT models for the focal and reference groups. All of these methods require that there be an overlap in the ability distribution between the two groups and a minimum sample size of 100. Thus, in this study, a new method, relative item performance, is proposed to address the issue of small focal group sample sizes.

The relative item performance method does not require that there be an overlap in ability distribution between the two groups. The index is computed by first obtaining the difference in item difficulty estimates between the two groups, then transforming this difference into an effect size. The effect size is then compared against the standard

normal distribution to identify if the item is performing differently from other items on the assessment. This method does not indicate which group the item is easier or harder for as that would require the need to have an overlap in the ability distribution. To evaluate how well this method addressed the issue of small sample sizes, a simulation study was run with the following factors manipulated: the sample size of the focal group, the ability distribution of both groups representing the amount of overlap in the ability levels, the amount of difference in item difficulty between the two groups, the number of items on the assessment, and the number of items that have different item difficulties. To compare the proposed method with current methods used to detect items that performed differently, the Type I error rates and power were calculated for each method and evaluated using graphs, descriptive statistics, and logistic regression. The five methods used in the simulation study are discussed in Chapter 2, and the simulation design is discussed in Chapter 3.

Chapter 2 – Literature Review

As the current methods used to detect items that perform differently require an overlap in the ability distribution between the reference and the focal group for ability matching as well as a minimum sample size of 100, I propose a method that would not need these requirements. Before discussing the proposed method, a review of empirical studies where the focal group sample size is small as well as the current methods used to detect differential item performance is given.

For this literature review, items scored in several categories, such as those assigned partial credit or those assigned points based on a rubric (e.g. a Likert scale), are not examined. These items, known as polytomous items, are not as frequently used in standard assessments as those with a binary response (e.g. correct versus incorrect, yes versus no, etc.). These binary items, also known as dichotomous items, are evaluated in this study.

Empirical Studies with Small Focal Group Sample Sizes

There have been several empirical studies where the focal group sample size is small. Maller (1997) analyzed the Wechsler Intelligence Scale for Children – Third Edition (WISC-III) test used to measure intelligence in children between the ages of 6 and 16 to determine if there were any differences in performance between hearing and deaf peers. The author used an IRT method known as the Rasch model with a sample of 110 deaf students compared to a sample of 110 hearing students with similar age and IQ. The author concluded that several items favored the deaf group or favored the hearing group. However, many of the items showed a lack of good fit meaning that the

assumptions of the Rasch model may not have been valid for deaf children with similar ability to their hearing peers. The author recommended the use of a two-parameter logistic model instead of the Rasch model to evaluate how difficult the item was, as well as how well the item discriminated between those who had a lower ability and those who had a higher ability to answer the item. However, using a more complex model requires a larger sample size due to the increase in the number of parameters to be estimated.

Another example of an empirical study with small sample sizes for the focal group is an analysis of the SAT verbal test analyzed by Bennett, Rock, and Kaplan (1985), where the exam performance of several groups of students with disabilities (deaf, learning, physical, and visual) was compared against that of students without disabilities. The sample size for the focal groups ranged from 98 to 2983. The authors evaluated the level of performance for all groups and the differential item performance using logistic regression. For the level of performance, the authors found that deaf students scored lower than their hearing peers. For differential item performance, items were grouped into logical clusters based on characteristics that may perform differently for students with disabilities. If a cluster performed differently, then items within that cluster were evaluated for differential performance. The authors found that items focusing on sentence completion and algebra performed differently for deaf students and their hearing peers with the former item being more difficult and the latter item being easier.

Steinberg, Cline, Ling, Cook, and Tognatta (2014) evaluated the validity and fairness of a test measuring English comprehension at a state level, comparing deaf or hard of hearing students with hearing students in fourth and eighth grades. The focal group sizes were 236 for grade 4 and 289 for grade 8 for students classified as non-ESL

(not an English as a Second Language student). For ESL students, the sample sizes were smaller: 174 for grade 4 and 165 for grade 8. The author also evaluated the performance of deaf students who took the test with accommodations, such as extended testing time, interpreters for the directions, or a separate room for test examination. The sample sizes for these groups were even smaller: 104 deaf students in grade 4, 113 hard-of-hearing students in grade 4, and 130 deaf/hard-of-hearing students in grade 8. Using a log-odds ratio analysis known as the Mantel-Haenszel test, there was only one item flagged for differential performance on the test for grade 4 and one for grade 8 when deaf/hard-of-hearing students did not use accommodations. When using accommodations, two items were found to perform differently for deaf students in grade 4, while only one item was identified for deaf/hard-of-hearing students in grade 8.

Currently, there is no method that can be used for sample sizes less than 50 such as when Martin (2005) conducted an expert review of the New York State English test to determine which items performed differently, then compared the results with an empirical analysis from 44 deaf students. Eight experts (three of whom were deaf) were used to rate the items and found that 18% of the multiple-choice items failed to pass item review. The test data of 44 deaf students who had taken the exam revealed that six out of 25 multiple-choice items could be correctly answered by 50% of students. The author concluded that the expert review and the empirical study revealed the same results regarding item performance. However, without the use of statistical tests due to the small sample size, the results from this analysis could not be generalized to all deaf students that take the New York State English test.

With sample sizes ranging from 44 to 236 for the focal group in these empirical studies, a variety of methods were used to detect items that perform differently. When the sample sizes are small, there is the concern that the identification of items as performing differently may not be accurate for two reasons. First, the items may be misidentified as performing differently when they are actually performing similarly for both groups because a small sample size increases the risk of a Type I error. The second reason is that the items may be identified as performing differently when in fact it is the difference in the ability distribution between the two groups. A small sample size for the focal group means that the overlap in the ability distribution between the reference and focal group would be sparse and result in different ability distributions for the two groups. Before discussing the current methods used to detect differences in item performance, differential item functioning is defined and discussed.

Explanation of Differential Item Functioning

The primary statistical method for identifying potential bias in assessments is known as differential item functioning (DIF) analysis, where DIF refers to differences in performance on a specific test item between two groups assuming that the two groups have an overlap in their ability distribution. This overlap is also known as common support (Rosenbaum & Rubin, 1983) and the reason for this requirement is due to Simpson's paradox.

Simpson's paradox. This paradox was first identified by Simpson (1951) to address how the interaction in contingency tables are interpreted. Simpson found that it was possible to have an item favor one group when looking at the overall performance and have it favor the other group when looking at specific levels of the ability

distribution. An example of this is shown in Table 1 where there are four ability levels and 100 examinees in both the reference and the focal groups.

Table 1

Example of Simpson's Paradox

Ability Level	Reference			Focal			Direction
	N	C	P	N	C	P	
1	20	15	0.75	50	35	0.70	>
2	10	8	0.80	25	18	0.72	>
3	50	25	0.50	15	7	0.47	>
4	20	15	0.75	10	7	0.70	>
Total	100	63	0.63	100	67	0.67	<

As seen in the table, the proportion of correct response (P) is equal to the number of correct responses (C) divided by the number of examinees (N). For each of the four ability levels, the proportion of correct response is higher for the reference group than that of the focal group. However, when looking at the overall proportions, the focal group answered the item correctly more than that of the reference group. For this reason, methods that identify which group the item favors must ensure that there is common support in the ability distribution between the reference and focal groups, which may not be possible if the focal group sample size is small. These methods can be grouped into two approaches to evaluating DIF: Classical Test Theory (CTT) and Item Response Theory (IRT).

Classical test theory. The first framework for analyzing differences in item performance uses observed data instead of a specified model distribution to determine if an item has DIF. CTT focuses on the observed score, the actual information obtained from the examinee, which can be represented by the following formula:

$$X_i = T_X + E_i, \quad (1)$$

where X_i is the score on item i used to measure an overall variable such as reading comprehension, T_X is the true level of reading comprehension for the specific examinee, and E_i is the error associated with item i . This equation holds as long as we assume that the score on the item does not measure another factor such as test anxiety. Thus, the observed score on the item is a mix of both the true score and error (DeVellis, 2006). The error in the equation is assumed to be randomly distributed with a mean of zero. Another assumption is that the items in the assessment are unidimensional, which means that the items are designed to measure only one underlying ability such as critical thinking or mathematical ability (Crocker & Algina, 2006).

To evaluate the performance of an item, two measures are used: item difficulty and item discrimination. Item difficulty is simply the proportion of the number of correct responses across all test takers, where a high proportion indicates that the item is easy and a low proportion indicates that the item is difficult. Item discrimination measures how well the item distinguishes between high-scoring or low-scoring examinees. This measure can be calculated using two approaches: an index of discrimination or a point-biserial correlation coefficient. Before calculating the index of discrimination, examinees must first be classified as high-scorers or low-scorers based on their total test score. The index is then computed as the difference between the proportions of correct responses between the two groups. This index can range from -1 to 1 with a positive value indicating that the high-scorers correctly answered the item more than the low-scorers, while a negative value indicates that the low-scorers correctly answered the item more than the high-scorers. The point-biserial correlation coefficient is calculated as follows:

$$\hat{\rho} = \frac{(\hat{\mu}_1 - \hat{\mu})}{\hat{\sigma}} \sqrt{\frac{\hat{p}}{1 - \hat{p}'}} \quad (2)$$

where $\hat{\mu}_1$ is the mean total score for examinees who responded correctly to the item, $\hat{\mu}$ is the mean total score for all examinees, $\hat{\sigma}$ is the population standard deviation of the total score for all examinees, and \hat{p} is the item difficulty. The coefficient also can range from -1 to 1, with a positive value indicating that the high-scorers answered the item correctly more than the low-scorers and vice versa (Crocker & Algina, 2006).

Methods to detect DIF within the CTT framework focus on either the difference between the proportions of correct responses or the odds ratio of the probabilities of obtaining a correct response for the focal (*F*) and reference (*R*) groups. In these methods, the examinee's ability estimate is obtained by summing the item responses to determine the total test score. For DIF analysis, the percentage of correct responses to an item is compared between the reference and the focal groups after conditioning on the overall test scores, with the interpretation of a significant difference between the percentages of correct responses being that the item favors one of the groups. A few CTT methods used to detect DIF are explained in the section "Methods to Detect DIF within the CTT Framework". Next, I describe a different approach to evaluating DIF using item response theory.

Item response theory. The second framework for analyzing differences in item performance is item response theory (IRT) which assumes that the responses to an item can be mathematically modeled as a relation between the probability of obtaining a correct response and the examinee's ability. This relation can be shown visually as a normal ogive curve shown in Figure 1 with the ability distribution on the x-axis and the

probability of obtaining a correct response on the y-axis. To define the curve, there are four item parameters: the difficulty parameter (b_i) which indicates the level of ability needed by an examinee to have a .50 probability (for the IRT models without the upper and lower asymptotes) of responding with a correct response on the item, the discrimination parameter (a_i) which indicates how well the item distinguishes between respondents on the lower end of the ability spectrum versus those on the higher end of the spectrum, the lower asymptote (c_i) which indicates the probability of getting the item correct by guessing, and the upper asymptote (d_i) that takes into consideration that those with higher abilities may get easy questions wrong for various reasons such as illness or carelessness (Barton & Lord, 1981).

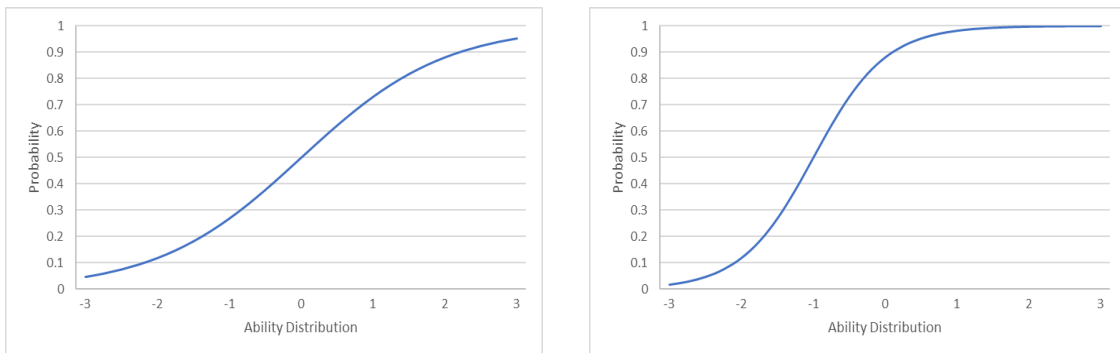


Figure 1. Two example item characteristic curves. The first curve has an item difficulty of 0 and item discrimination of 1, while the second curve has an item difficulty of -1 and item discrimination of 2.

Figure 1 shows the concept of item difficulty and item discrimination. As seen in the first graph, the item difficulty is zero with a probability of obtaining a correct response equal to 50% and the item discrimination is equal to 1. Thus, this item is neither difficult nor easy, and does not clearly distinguish between those on the upper or

lower end of the ability spectrum. The second graph has an item difficulty of -1 and an item discrimination of 2, which means that the item is considered to be easy, and does discriminate more strongly between those who are on the upper or lower end of the ability spectrum.

Using these item parameters, an item is said to not exhibit DIF if the following conditional probability formula (Zwick, 1990) holds:

$$P(Y = 1|\theta, G = R) = P(Y = 1|\theta, G = F), \quad (3)$$

where the probability of getting a correct response ($Y = 1$) is the same for the reference ($G = R$) and the focal ($G = F$) groups given the same ability (θ).

There are four different types of IRT models used for DIF analysis and the key difference is the parameters included in the model. The one-parameter logistic model, also known as the Rasch model, incorporates only the difficulty parameter (Hambleton, Swaminathan, & Rogers, 1991) as shown in Equation 4.

$$P(x_{ij} = 1|\theta_j) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}, \quad (4)$$

where θ_j represents the ability level of the j^{th} person, and x_{ij} is the response to the i^{th} item from the j^{th} person. The two-parameter logistic model includes the difficulty parameter and adds on the discrimination parameter as shown in Equation 5.

$$P(x_{ij} = 1|\theta_j) = \frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]} \quad (5)$$

This model can also be extended to include the lower and upper asymptote parameters. Once the estimates of these item parameters (a_i, b_i) are obtained for these models separately for the reference and the focal groups, they are then used to evaluate if an item has DIF.

Methods to detect DIF based on IRT have the intent of explicitly estimating parameters related to items and examinees during data analysis. For these methods, the researcher must specify a formal model for the item response pattern and estimate the parameters for comparison between the focal and reference groups (Penfield & Camilli, 2007). A test item is identified as having DIF if the item parameters or response patterns differ between the focal and reference group more than what is expected due to sampling or estimation errors (Lord, 1980).

Summary. Methods based on the CTT framework have the advantage over IRT methods in that the sample size requirements are smaller given that DIF is identified using the item and the test scores from the examinees rather than parameter estimates that come from specific assumed distributions (Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001). If the model used to develop the parameter estimates is not specified correctly, then DIF items could be misidentified as having no DIF or vice versa (Penfield & Camilli, 2007). CTT methods are not affected by this issue (Meijer & Baneke, 2004). However, a limitation of CTT methods is that the statistics are sample-dependent. An example is the item difficulty that reflects not just the difficulty of the item but also the ability of the examinees. Thus, a value of .80 for a group of examinees with low abilities is not the same as a value of .80 for a group of examinees with high abilities (De Champlain, 2009).

Methods to Detect DIF within the CTT Framework

In this section, three methods used to detect DIF within the CTT framework are discussed. The Mantel-Haenszel test is a method that uses a log-odds ratio to determine

which items perform differently. Another method is known as standardization that uses comparisons of nonparametric item test regressions between the focal and reference groups. The third CTT method is known as logistic regression, which predicts the probability of obtaining a correct response depending on the examinee's estimated ability level and whether or not the examinee is in the focal group and the interaction between these two variables. After discussing these three methods, factors that affect the performance of these methods are reviewed.

Mantel-Haenszel test. One of the methods used to detect DIF based on the CTT framework is known as the Mantel-Haenszel (MH) test developed by Mantel and Haenszel (1959) for medical research purposes. This test was later modified for DIF analysis by Holland and Thayer when studying diseases (Holland & Thayer, 1988). This method is an extension of the chi-square contingency table that compares the reference and focal groups in terms of the proportions of correct responses to a specific test item. First, those in the reference and focal groups are matched based on either an internal or external criterion (Clauser, Mazor, & Hambleton, 1993). However, many studies do not have a valid external criterion to use as their matching variable, so an internal criterion such as the overall test score is used to form k score levels (Uttaro & Millsap, 1994; Zwick, 1990). Once the matching strategy is determined, a 2×2 contingency table is tabulated for each score interval (k) using the layout shown in

Table 2.

Table 2

2×2 Contingency Table for Score Interval k

Group	Performance on item <i>i</i>		Total
	1	0	
<i>R</i>	a_k	b_k	N_{Rk}
<i>F</i>	c_k	d_k	N_{Fk}
Total	N_{Ik}	N_{Ok}	N_k

The MH statistic is then computed by aggregating the odds ratio of the reference group obtaining a correct response on the test item over the focal group across all the score intervals as given in the following formula.

$$\hat{\alpha}_{MH} = \frac{\sum_k (a_k d_k / N_k)}{\sum_k (b_k c_k / N_k)}, \quad (6)$$

where N_k represents the total number of responses for the test item from both groups in interval k . For ease of interpretability, the MH statistic is standardized (MH D-DIF) so that a value of zero would reflect the presence of no DIF in the test item. Also, MH D-DIF is negative when the item is more difficult for members of the focal group than similar members of the reference group. A positive MH D-DIF implies that the item is more difficult for members of the reference group than similar members of the focal group. This standardization is done by taking the log of the MH statistic and multiplying it by -2.35 (Holland & Thayer, 1988).

Educational Testing Service (ETS) has a system for categorizing the severity of DIF based on both the magnitude of the MH D-DIF index and its statistical significance (Zieky, 1993; Zwick, 2012). In this system, an item is classified as "A" (no DIF) when the absolute value of MH D-DIF is less than 1 and is not statistically significant with the

significance level equal to .05. A "C" item would have a moderate to large DIF when the absolute value of MH D-DIF is greater than 1.5 and statistically significantly greater than 1 with alpha equal to .05. Otherwise the items are labeled as having "B" DIF or slightly to moderate DIF. "C" items should be examined further for potential bias (Zwick, 2012). To indicate the direction of the DIF, a negative sign is added to the labels (e.g. "-C") if the DIF shows bias against the focal group.

The Mantel-Haenszel method can also be evaluated using either a chi-square test statistic or a log odds test statistic. The chi-square test statistic is calculated based on the following equation:

$$\chi_{MH}^2 = \frac{[|\sum_k a_k - \sum_k E(a_k)| - .5]^2}{\sum_k \text{Var}(a_k)}, \quad (7)$$

where

$$E(a_k) = N_{Rk}N_{1k}/N_k \quad (8)$$

$$\text{Var}(a_k) = \frac{N_{Rk}N_{1k}N_{Fk}N_{0k}}{N_k^2(N_k - 1)} \quad (9)$$

The chi-square test statistic is distributed as a chi-square distribution with 1 degree of freedom. The log odds test statistic is calculated by taking the log of the MH statistic and dividing it by a variance estimator developed by Phillips and Holland (1987). The variance estimator is given as follows:

$$\widehat{\text{Var}}(\log(MH)) = \frac{1}{2(\sum_k (a_k d_k / N_k))^2} * \sum_k \frac{1}{N_k^2} (a_k d_k + \hat{\alpha}_{MH} b_k c_k) [a_k + d_k + \hat{\alpha}_{MH} (b_k + c_k)] \quad (10)$$

This statistic is then evaluated to see if it is statistically significantly different from 1 with a significance level equal to .05.

Thick versus thin matching. There is one main concern that must be addressed when conducting the MH test: the overlap of the test scores between the reference and the focal groups. As the MH test uses the total test score for the score levels, there is the risk of not having enough data at each score level when the focal group sample size is smaller than that of the reference group. This concern is addressed by modifying the MH test to increase the width of the score levels, known as thick matching (Donoghue & Allen, 1993). Typically, when matching on the total test score, there are $k + 1$ possible score groups with k being the number of items. For example, if a test has 50 items, a respondent could obtain a score ranging from 0 to 50 thus resulting in 51 possible scores. However, depending on the common support of the ability distributions for the reference and focal groups, some score groups could have few or none from the focal group. Donoghue and Allen (1993) identifies the use of all score groups as thin matching, while the creation of score group intervals is known as thick matching. Several strategies can be used to create these intervals such as dividing the test score scale into intervals of equal widths (known as equal interval matching; Raju, Bode, & Larsen, 1989), dividing intervals based on percentages in the focal group (known as total percentage matching; Donoghue & Allen, 1993), dividing such that each interval has an equal number of focal examinees (known as focal percentage matching; Donoghue & Allen, 1993), and pooling extreme test scores into the nearest intervals to ensure a minimum number of examinees in each interval (known as censored matching; Zwick & Ercikan, 1989). Clauser, Mazor, and Hambleton (1994) conducted a simulation study to compare the equal interval matching technique with that of thin matching by varying the sample size (2000, 1000, 500, 200, 100 per group) and the score group intervals (81, 20, 10, 5, 2). They found that

if the sample was large and/or the ability distribution is the same for both groups, then the two matching techniques did not make a difference in DIF detection. However, when the sample size was small and the score group intervals were reduced then the group means were no longer equal, leading to potential misidentification of items with DIF. The authors recommended that if the sample size could not be increased, then the equal interval matching technique could be used, but cautiously. This is especially a concern as the authors mentioned that if a score level contained examinees from only one group, then it was dropped from the calculations. Donoghue and Allen (1993) also conducted a simulation study comparing the various thick matching techniques with the thin matching technique. Using a test with 5, 10, 20 or 40 items and a sample size of 300/100, 600/200, and 1200/400 (Reference/Focal), the authors found that thick matching can improve results but not for small tests (5 or 10 items) and even with longer tests the thin matching performed best when the sample size was large. Thick matching was developed to alleviate concern about the common support in ability distribution between the focal and the reference groups. However, there could still be potential loss of information when a score group interval only contains examinees from one of the groups and not the other. A possible solution for this concern is to disregard the requirement of common support in the ability distributions between the two groups and instead look at how the item performs relative to other items on the assessment.

Standardization. The second method used for DIF detection was developed by Dorans and Kulick (1983) to control for group differences in ability before making comparisons between the groups on a test item. Using the same table as the Mantel-Haenszel test (

Table 2) to group examinees by their total scores, the differences in proportions of correct response between the reference and focal group is calculated as follows:

$$D_k = P_{Fk} - P_{Rk}, \quad (11)$$

where

$$P_{Fk} = c_k/N_{Fk}; P_{Rk} = a_k/N_{Rk}. \quad (12)$$

There are two ways to flag an item as exhibiting DIF: the standardized p-difference (STD P-DIF) or the root mean weighted squared difference (RMWSD). Both of these flags require the use of a function to weight the differences at each score level prior to aggregating these differences to the item level. The calculations of the standardized p-difference and the root mean weighted squared difference are given in Equations 13 and 14, respectively.

$$\text{STD P-DIF} = \frac{\sum_{k=1}^K (W_k * D_k)}{\sum_{k=1}^K W_k}, \quad (13)$$

$$\text{RMWSD} = \left[\frac{\sum_{k=1}^K W_k * (D_k)^2}{\sum_{k=1}^K W_k} \right]^{.5}. \quad (14)$$

For the weights at each score level k , the researcher can choose from several values such as the total number of examinees from both groups, the number of examinees from the focal group, the number of examinees from the reference group, or the relative number of examinees from a norming population. The authors recommend the use of the number of examinees from the focal group as it gave the greatest weight to difference in the proportions of correct response at the score levels.

The authors recommend the use of the standardized p-difference instead of the RMWSD due to an article by Wright (1987) which found that the RMWSD is biased as it

contains the sampling error at each score level. Using the standardized p-difference, an item is flagged as exhibiting moderate DIF if the absolute value is between 0.05 and 0.10 and exhibiting substantial DIF if the absolute value exceeds .10.

Logistic regression. Another method for DIF detection is the use of logistic regression to predict the probability of obtaining a correct response depending on the examinee's estimated ability level and whether or not the examinee is in the focal group and the interaction between these two variables (Swaminathan & Rogers, 1990). The dependent variable is the probability of getting a correct response on the test item. The model is formulated below:

$$P(Y=1) = \frac{e^z}{1 + e^z}, \quad (15)$$

where

$$z = \beta_0 + \beta_1\theta + \beta_2G + \beta_3\theta G, \quad (16)$$

where $P(Y=1)$ is the probability of obtaining a correct response on the test item, θ is the ability level of the examinee, and G is the group membership. Once the beta parameters are estimated, then the statistical significance needs to be evaluated. If the group membership parameter estimate (β_2) is significantly different from zero, then the presence of DIF can be inferred for the test item. If the interaction is significant, then the ability distribution also has an effect on how the item performs differently for the reference and focal groups.

To obtain these estimates, maximum likelihood is used to evaluate the values that have the highest probability of creating the observed data. The $\hat{\beta}_2$ and $\hat{\beta}_3$ estimates are then compared simultaneously with the following hypotheses:

$$H_0: C\beta \text{ is equal to } 0$$

H₁: $C\beta$ is not equal to 0,

where

$$\mathbf{c} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (17)$$

The test statistic is calculated as shown in Equation 18 and compared against a chi-square distribution with 2 degrees of freedom.

$$\chi^2 = \hat{\beta}' \mathbf{c}' (\mathbf{c} \hat{\Sigma} \mathbf{c}')^{-1} \mathbf{c} \hat{\beta}, \quad (18)$$

where $\hat{\Sigma}$ is the estimated variance-covariance matrix of the parameter estimates.

Performance of the CTT methods. Through both simulation studies and empirical research, researchers have found that the performance of the Mantel-Haenszel, standardization, and logistic regression methods depend on several factors. These factors include the sample sizes for the reference and the focal groups, the type of items being studied, the number of test items, and how contaminated the assessment is with DIF. A review of these factors and their effect on the effectiveness of the three methods is given below.

Sample size requirements. As the MH statistic is comparing a reference group and a focal group, there needs to be an overlap between the two groups in their overall test scores. As the sample size decreases, there is the potential that a score interval could have no examinees from the focal group that matched on their test scores with the reference group (Dorans, 1989; Dorans & Holland, 1992). Initially, studies recommended a sample size of 1000 in both groups (Ackerman & Evans, 1992). Specific studies of DIF detection methods including the MH statistic have reported the effect of group (both focal and reference) size on the power and Type I error on detecting DIF for a single item. Some of these studies have examined the MH statistic when the sample

sizes are small (Fidalgo, Ferreres, & Muñiz, 2004; Mazor, Clauser, & Hambleton, 1992; Muñiz, Hambleton, & Xing, 2001; Rogers & Swaminathan, 1993; Zieky, 1993), while others have compared the effect of different group sizes on several detection procedures (Miller & Oshima, 1992; Narayanan & Swaminathan, 1996; Roussos & Stout, 1996). For example, Fidalgo, Ferreres, and Muñiz (2004) conducted a simulation study with the following sample sizes (reference/focal): 50/50, 100/50, 200/50, 100/100. Across all sample sizes, the power of the DIF detection using MH ranged between 0.08 and 0.44, but there were no substantial differences in Type I error for items with DIF or without DIF. Based on the study, the recommendation was to consider increasing the significance level from 5% to 20% as falsely identifying a test item as having DIF is not as severe a problem as not detecting a test item that does have DIF. However, a good number of the reviewed studies focus on situations in which the focal and the reference groups were of the same size, or at least in which the difference in size between the groups is not manipulated systematically. Herrera and Gómez (2008) examined the effect of different reference and focal group sizes on Type I error of the MH statistic and found that a reference group of 500 and a focal group of between 100 and 500 would control the Type I error. In terms of power, the MH statistic performed better than logistic regression with a small reference group of 500 and a focal group size between 125 and 200.

For the standardization method, there has been no simulation study that used sample size as a factor. However, Dorans and Kulick (1983) state that a large sample size is required to ensure that the probabilities of a correct response is stable across each score level. An empirical study was conducted by using the standardization method on the SAT exam where the sample size was 21,285 and 21,209 for males and females,

respectively, with only one item performing differently for the two groups (Dorans & Kulick, 1983). Another study by Kulick and Dorans (1983a) had sample sizes of 7,053, 27,382, and 24,910 depending on the father's level of education and found no differences between the three groups. The third study shows an effect of the sample size on the ability to detect DIF when analyzing the performance of 2,616 Asian-American students versus 65,942 white students. Out of 195 items on the assessment, 51 items were identified as having moderate or substantial DIF and the authors state sample size as a possible cause by not providing accurate estimates of the proportions of correct response (Kulick & Dorans, 1983b). Thus, large sample sizes for both reference and focal groups should be at least 3,000.

As the MH test is a simpler case of logistic regression when the ability distribution is discrete and does not interact with group membership, these two methods have often been compared to each other in terms of being able to detect DIF over various sample sizes. Swaminathan and Rogers (1990) conducted a simulation study comparing these two methods by manipulating the sample size (250 per group or 500 per group), the test length (40 items, 60 items, 80 items), and the nature of the DIF (20% of the items contained DIF – half with uniform DIF and the other half with non-uniform DIF). The authors found that both methods were comparable in terms of Type I error when detecting uniform DIF with 75% accuracy in a sample size of 250 and with 100% accuracy in a sample size of 500. For non-uniform DIF, the MH could not detect any DIF, while the logistic regression method detected DIF with 50% accuracy in a sample size of 250 and with 75% accuracy in a sample size of 500. In terms of power, the effect of sample size was similar to those in terms of Type I error rates. Herrera and Gómez

(2008) found that the MH statistic performed better in terms of Type I error rates in detecting DIF in small sample situations (reference = 500, focal = 125 to 200), while the logistic regression method performed best as compared with MH with 1500 examinees in the reference group. Narayanan and Swaminathan (1996) discussed the issue of detecting the interaction between the ability distribution and the group membership, which logistic regression has been shown to detect while the MH statistic cannot. Using sample sizes of 500/200, 500/500, 1000/200, 1000/500, they found that the ability to detect DIF was largely impacted by the size of the focal group and recommended a sample size of 500. Rogers and Swaminathan (1993) conducted a power study using sample sizes of 250 or 500, and found that detection rates increased by approximately 15% when the sample size was increased from 250 to 500. Whitmore and Schumacker (1999) compared the logistic regression method with analysis of variance (ANOVA) DIF detection methods using sample sizes of 200, 400, and 600. As proposed by Tang (1994), the IRT ANOVA method involves calibrating the item and ability parameters using the appropriate IRT model, computing the probability of a correct response for each examinee, computing the expected score of the examinee for that test item, computing the residual by subtracting the expected score from the actual item score, then running ANOVA with the residuals as the dependent variable and the group characteristic (gender, race, etc.) as the independent variable. She found that the logistic regression method performed better than the ANOVA methods regardless of sample size and that the detection rate increased as the sample size increased.

Type of item being studied. Mazor, Clauser, and Hambleton (1992) also studied the Type I error of detecting DIF using MH at sample sizes ranging from 100 to 2000,

and found that there were specific types of items that were not being detected. Specifically, items with low discrimination parameters or items with very high or very low difficulty parameters were the least likely to be identified as possibly affected by DIF regardless of the sample size. The authors recommended that if researchers are only concerned about the items with large DIF then a sample size of 200 in each group would be enough. Roussos and Stout (1996) evaluated the MH method across different item difficulty, discrimination, and guessing parameters using sample sizes of 500, 1000, and 3000 for both groups. The authors found that MH had inflated Type I error rates for items with high discrimination and low difficulty parameters ($a = 1.0, b = -1.5$; $a = 2.5, b = -1.5$; $a = 2.5, b = -0.5$). Rogers and Swaminathan (1993) also modified the item difficulty and discrimination parameters for five out of 40 test items as follows: ($a = 0.6, b = -1.5$), ($a = 1.0, b = 0.0$), ($a = 1.6, b = 1.5$), ($a = 0.6, b = 1.5$), and ($a = 1.6, b = -1.5$). The authors found that the MH method did not detect items with low difficulty parameters and non-uniform DIF as well as the logistic regression method. For standardization, Donoghue, Holland, and Thayer (1993) conducted a Monte Carlo study of factors where the difficulty ranged from $-.5$ to $.5$ and the discrimination parameter was either $.3, 1, \text{ or } 1.5$. The results of the simulation showed that the standardization detected items at the same rate as that of the MH test. However, the standardization method was more sensitive to the amount of DIF in the item. For all three methods, DIF detection depends on type of item being evaluated with extreme item difficulties being the least likely to be detected.

Number of test items and test contamination. Another factor to consider when evaluating the performance of the MH procedure is how many test items are being evaluated and the degree of the DIF contamination. Swaminathan and Rogers (1990)

evaluated the MH and logistic regression procedures using 40, 60, or 80 test items with the degree of DIF contamination equal to 20%. The authors found that as the number of test items increased, both procedures correctly identified the items with DIF, but only evaluated the power with 80 test items, so there is no knowledge of how much change in power the procedures would have as number of test items increases. Rogers and Swaminathan (1993) continued this research using 40 or 80 test items with either 0% or 15% DIF contamination and found that neither the test length nor the degree of DIF contamination affected the Type I error rates. Donoghue and Allen (1993) evaluated the MH procedure using 5, 10, 20, or 40 test items with one item exhibiting DIF. The result was that the MH procedure falsely identified DIF when there were few items on the test. Fidalgo, Mellenbergh, and Muñiz (2000) evaluated the effect of test length and the amount of DIF by using 20, 40, or 60 test items and 10%, 15%, or 30% DIF contamination. The authors found that test length did increase the power of the MH procedure but not by a large factor for tests with 10% contamination. They also found that for the higher degrees of contamination, that the Type I error rates were inflated and that the procedure did not have sufficient power to detect DIF. There has been no simulation study conducted to evaluate the effect of test length and item contamination on the standardization procedure. For the MH and the logistic regression methods, the longer test lengths allow for a decrease in Type I error rates and an increase in power.

Summary. For these three methods based on the CTT framework, researchers need to take into consideration the sample size, the type of item being studied, the amount of item contamination, and the number of items on the assessment. Given prior findings, it has been found that the minimum sample size is that of 125 for the MH

procedure while standardization requiring the largest sample size. The type of item is also a concern when evaluating DIF, as the MH and logistic regression tests could not detect DIF with extreme item difficulties. Also, the more items on the assessment, the more accurate the methods are in detecting DIF items. However, when the percentage of contaminated items increases, the ability of the methods to detect DIF items decrease.

Of these factors, the sample size of the focal group and the test length have an effect on the common support between the reference and the focal group. As the sample size for the focal group decreases, there could be a few score levels where there are examinees from the reference group but none from the focal group. As the test length increases, especially if the number of items is more than the number of examinees in the focal group, there could be a lack of common support for some of the score levels. Next, I discuss methods based on the IRT framework.

Methods to Detect DIF within the IRT Framework

In this section, three methods used to detect DIF using the IRT framework are discussed. The Wald test is a method that calculates the difference between the estimated item difficulties for each item, while the differential item pair functioning is an extension of the Wald test that evaluates pairs of item difficulties. The third IRT method is known as the unsigned/signed areas, which calculates the area between the two item characteristic curves for the focal and the reference groups. After discussing these three methods, the sample size requirements for these methods are reviewed.

Wald test. The Wald test is a method of DIF detection based on the IRT framework, which determines the statistical significance of the difference between the item parameters for the reference and focal groups. Lord (1980) developed a *t*-test

comparing the difficulty parameters when the 1PL model was used in developing the test item. First, the 1PL model is calibrated separately for the reference and the focal groups to obtain separate difficulty parameters. The null hypothesis of the test is that the two difficulty parameters for the item from each group are the same ($H_0: b_{iR} = b_{iF}$), and is tested by calculating the following statistic, which follows a standard normal distribution:

$$d = \frac{b_{iR} - b_{iF}}{SE(b_{iR} - b_{iF})}, \quad (19)$$

where the standard error of the difference is calculated as:

$$SE(b_{iR} - b_{iF}) = \sqrt{SE(b_{iR})^2 + SE(b_{iF})^2} \quad (20)$$

The d statistic is then compared with the student t distribution with $N_r + N_f - 2$ degrees of freedom. For items developed using the 2PL or 3PL models, the Wald statistic is used instead. In the case of the 2PL model, the null hypothesis would be that both the item difficulty and discrimination parameter are equal for both groups ($H_0: a_{iR} = a_{iF}$ and $b_{iR} = b_{iF}$). To calculate the statistic, the first step is to estimate the 2×2 variance-covariance matrix of the differences between the item parameter estimates for the reference group and the focal group ($\mathbf{S}_{2 \times 2}$) for the specific item of interest. The second step is to create a vector containing the differences between the item parameters from each group [$\mathbf{V}' = (a_{iR} - a_{iF}, b_{iR} - b_{iF})$]. The Wald statistic is then calculated as:

$$\chi^2 = \mathbf{V}'\mathbf{S}^{-1}\mathbf{V}, \quad (21)$$

which is evaluated using a chi-square distribution with 2 degrees of freedom. This method can be extended to include the lower asymptote from the 3PL model using a $3 \times$

3 \mathbf{S} matrix and the vector set equal to the differences between the item parameters from each group [$\mathbf{V}' = (a_{iR} - a_{iF}, b_{iR} - b_{iF}, c_{iR} - c_{iF})$]. However, since the c parameter is often poorly estimated, it is usually ignored and the Wald statistic is run with just the a and b parameters (Lord, 1980). There is a small sample adjustment to the Wald test to account for small sample sizes. This adjustment uses a Taylor series approximation in order to adjust for the bias from the sandwich estimator of the variance, and compares the test statistic to either an F or t distribution with an estimated degree of freedom depending on the number of parameters estimated (Fay & Graubard, 2001; Fitzmaurice, Laird, & Ware, 2011).

Differential item pair functioning. When analyzing items for DIF in the IRT framework, one issue is that the difficulty of an item is not identifiable from the mean of the observations. Specifically, there are two parameters being estimated in the Rasch model: the difficulty of an item and the latent ability of the respondents (Gustafsson, 1980). As the center of the ability distribution is arbitrary, the difficulty parameter cannot be estimated. However, by multiplying the item difficulties with weights that sum to 1, the scale is normalized and the model is identifiable. One issue with normalization is that one of the items on the test must be used as a reference item, but if there is no knowledge of which items do or do not exhibit DIF, then there is the possibility of using a DIF item to create the normalization scale, which would lead to further misidentification of DIF in the other items. Also, if comparing groups, two normalizations would be needed, one for each group. Thus, Bechger and Maris (2015) proposed to do a pairwise comparison of all items to determine relative DIF and which item should be used as the reference item. The authors defined two goals of DIF research:

(1) determining if the test exhibits overall DIF, and (2) if so, which specific test items exhibit DIF? The authors state that the first question is easily answered by determining if there are differences between the two groups' matrices of all pairwise item comparisons, but the second is not. The authors proposed an extension of the Wald test to use the relative difficulties from the item pairs.

The first step in Bechger and Maris' (2015) proposal using a 1PL model to detect differential item pair functioning is to create a separate matrix for each group with the elements being the difference between any two items' difficulties e.g. $b_{21} - b_{31}$ (the difference between difficulties for items 2 and 3 from group 1). Note that by using differences between the item parameters for all items, an item is not chosen as a reference item. An example of this is given below using three items:

Item difficulty parameters
(Focal)

$$\begin{bmatrix} 0.8 \\ 1.2 \\ 1.4 \end{bmatrix}$$

Differences between item parameters
(Focal)

$$\begin{bmatrix} - & 0.8 - 1.2 & 0.8 - 1.4 \\ 0.8 - 1.2 & - & 1.2 - 1.4 \\ 0.8 - 1.4 & 1.2 - 1.4 & - \end{bmatrix} = \begin{bmatrix} - & -0.4 & -0.6 \\ -0.4 & - & -0.2 \\ -0.6 & -0.2 & - \end{bmatrix}$$

Item difficulty parameters
(Reference)

$$\begin{bmatrix} 0.7 \\ 1.6 \\ 1.7 \end{bmatrix}$$

Differences between item parameters
(Reference)

$$\begin{bmatrix} - & 0.7 - 1.6 & 0.7 - 1.7 \\ 0.7 - 1.6 & - & 1.6 - 1.7 \\ 0.7 - 1.6 & 1.6 - 1.7 & - \end{bmatrix} = \begin{bmatrix} - & -0.9 & -1.0 \\ -0.9 & - & -0.1 \\ -1.0 & -0.1 & - \end{bmatrix}$$

When there is no differential item pair functioning and the ability distribution between the two groups are equal, there is no difference between the two matrices (e.g., $\mathbf{R}_{p \times p}^{(R)} = \mathbf{R}_{p \times p}^{(F)}$). The difference between the two matrices is then computed, which results in a skew-symmetric matrix (where the lower off-diagonal elements are a mirror to that of the upper off-diagonal elements) as shown below:

$$\Delta \mathbf{R} = \begin{bmatrix} - & 0.5 & 0.4 \\ 0.5 & - & -0.1 \\ 0.4 & -0.1 & - \end{bmatrix}$$

As any of the columns or rows can be used to reproduce the entire matrix, one of the columns is evaluated by calculating the Wald statistic (Equation 21) which follows a chi-square distribution with $p - 1$ degrees of freedom where p is the number of items. If the null hypothesis that there is no difference between the two matrices is rejected, then the test exhibits DIF.

The next step is to determine which particular pair of items have DIF, which means evaluating and testing the specific elements in the column as shown below:

$$D_{ii'} = R_{ii'}^{(R)} - R_{ii'}^{(F)}, \quad (22)$$

Where i is an item and i' represents a different item. The test statistic is computed as follows:

$$\hat{D}_{ii'} = \frac{\hat{R}_{ii'}^{(R)} - \hat{R}_{ii'}^{(F)}}{\sqrt{\text{Var}(\hat{R}_{ii'}^{(R)}) + \text{Var}(\hat{R}_{ii'}^{(F)})}} \quad (23)$$

where

$$\text{Var}(\hat{R}_{ii'}^{(g)}) = \text{Var}(\hat{b}_{ig} - \hat{b}_{i'g}) = \text{Var}(\hat{b}_{ig}) + \text{Var}(\hat{b}_{i'g}) - 2\text{cov}(\hat{b}_{ig}, \hat{b}_{i'g}) \quad (24)$$

Under the null hypothesis, this statistic is asymptotically standard normal, so the null hypothesis is rejected if the absolute value of the test statistic is larger than $z_{\alpha/2}$.

Another way to evaluate items is to examine a heat map of the D_{ij} matrix. In Figure 2, a heat map is shown for 20 items where the shade of the color represents the magnitude of the D_{ij} element for each item pair. Where the color shade is much deeper or lighter than other items shows that these item pairs have a large D_{ij} . In this heat map, most of the pairs including the 17th item are shown to perform differently than other item pairs.

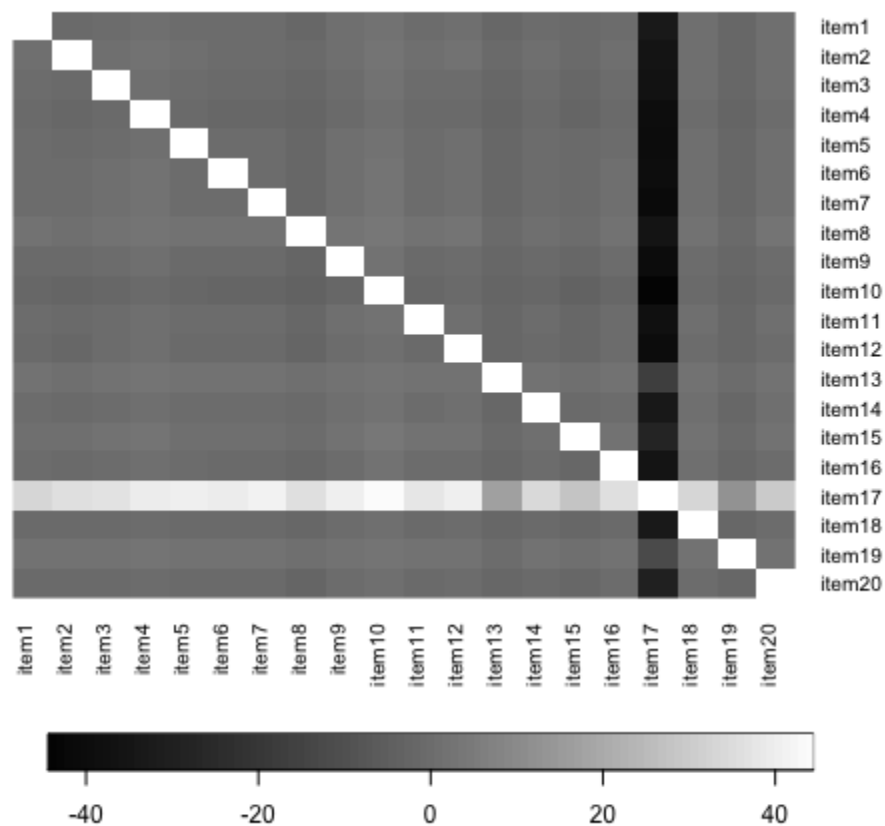


Figure 2. Heat map for differential item pair functioning. A heat map of item pairs is shown using 20 items with the test statistic ranging from -40 to 40.

Because this procedure relies on asymptotic distributions, a large sample is required. The authors conducted a simulation study using sample sizes between 100 and

10,000 in both groups, a test length of 30 items, the ability distribution being $N(0, 2)$ for the first group and $N(-0.2, 3)$ for the second group, and item difficulty parameters drawn from an uniform distribution with the mean between -1.5 and 0.5. The size of the DIF were simulated by taking the paired differences containing a specific item for the reference group and subtracting a factor ranging from 0 to 1 to obtain the paired differences for the focal group. As a result, samples below 500 in each group were insufficient to detect DIF. The Type I error rates were calculated by evaluating how often the difference of the difference between items 5 and 10 for the reference and focal groups were rejected using a significance level of 0.05 when DIF was not simulated. Power was calculated as how often the difference of the difference between items 8 and 10 for the reference and focal groups were correctly flagged as exhibiting DIF when DIF was simulated. The Type I error rates all ranged between 0.04 to 0.06 across all levels of sample sizes, while the power increased from an average of 20% for the smallest sample size of 500 to an average of 90% for the largest sample size of 10,000. If running this test with small samples, the authors recommend that an algorithm that utilizes an MCMC (Markov chain Monte Carlo) algorithm could be used to estimate the probability distribution of the parameter by first creating matrices that have similar marginal totals as that of the observed data used to estimate the difficulty parameters. The parameter estimates are then evaluated against the distribution to see how likely the resulting data would have occurred. Thus, if focus group sample sizes are small, then this procedure likely is not sufficient to detect DIF unless using MCMC simulation.

Raju's signed or unsigned areas. The final method using the IRT framework (Raju, 1988) involves two measures involving the calculation of the area between two

item characteristic curves (ICCs) when calibrating two separate IRT models, one for the focal group and the other for the reference group. The first measure is called the signed area (SA) and is calculated differently depending on which IRT model is used. If the 3PL model is used, then Equation 25 is used assuming that the lower asymptote parameter is the same for both groups. If either the 2PL or the 1PL model is used, then Equation 26 is used. Note that the discrimination parameter is not included in the calculation as the signed area measures are zero when the discrimination parameter differs across groups (Raju, 1988).

$$SA_i = (1 - c_i)(b_{iF} - b_{iR}), \quad (25)$$

$$SA_i = (b_{iF} - b_{iR}), \quad (26)$$

where c_i is the lower asymptote parameter for the item, b_{iF} is the item difficulty parameter for the focal group, and b_{iR} is the item difficulty parameter for the reference group (Raju, 1990). If this SA measure is zero, this could mean that there is no DIF for the item or the curves cross and the areas cancel each other out. To infer that DIF is present assuming the item was provided to all test-takers in a population, a z -test is calculated to determine if the SA value is significantly different from zero (Raju, 1990). One disadvantage with this method is that SA may show no DIF when there is bias because the lower abilities have a positive DIF while the higher abilities have a negative DIF and they cancel out or vice versa.

To address the disadvantage with SA, Raju (1988) proposed another method known as the unsigned area (UA), which is calculated differently depending on

assumptions about the item parameters. Equation 26 assumes that only the lower asymptote parameter is the same for both the reference and the focal group, while Equation 27 assumes that the lower asymptote parameter and the discrimination parameter is the same for both groups. Note that Equation 28 is the same as Equation 25, meaning that UA reduces to SA if the discrimination parameter is the same for the reference and the focal groups.

$$UA_i = (1 - c_i) * \left| \frac{2(a_{iF} - a_{iR})}{D a_{iF} a_{iR}} \ln \left[1 + \exp \left(\frac{D a_{iF} a_{iR} (b_{iF} - b_{iR})}{a_{iF} - a_{iR}} \right) \right] - (b_{iF} - b_{iR}) \right|, \quad (27)$$

$$UA_i = (1 - c_i)(b_{iF} - b_{iR}), \quad (28)$$

where a_{iF} is the item discrimination parameter for the focal group, a_{iR} is the item discrimination parameter for the reference group, D is a constant of 1.7 to equate the logistic curve with that of a normal curve (Lord, 1980). A large difference between UA and SA indicates that a test item may have non-uniform DIF. A limitation for both of these methods is that they do not take into consideration the ability distribution of the examinees and may lead to misinterpretation of the magnitude of DIF observed for a specific set of examinees (Penfield & Camilli, 2007).

Sample size requirements. For methods that use IRT to calibrate the parameters, the sample size requirements depend on the model used. For the one parameter model, Lord (1968) recommended a sample size of 1000 examinees with 50 test items. Morizot, Ainsworth, and Reise (2007) and Bond and Fox (2007) recommended a sample size of 100, while Lai, Teresi, and Gershon (2005) recommended a sample size of 200. For the two-parameter logistic model, studies recommend either 200 or 500 examinees (Bond &

Fox, 2007; Morizot, Ainsworth, & Reise, 2007) while Hulin, Lissak, and Drasgow (1982) recommended 500. However, these recommendations do not take into consideration the fact that there are two groups being compared when detecting DIF.

Kim, Cohen, and Kim (1994) evaluated Lord's Wald statistic using a Monte Carlo simulation study with sample sizes of 250 and 1000 total examinees using either the two-parameter logistic model or the three-parameter logistic model and found that Type I error rates were closer to the α level when the sample size was larger. For the differential item pair functioning method, a minimum sample size of 500 is required for the focal group.

Kim and Cohen (1995) compared the area measures with the Lord test and the likelihood ratio test using data from two forms of a university mathematics placement test with a sample size of 765 for the reference group and 725 for the focal group. The result of this study found that the area measures and the Lord statistic identified the same test items as having DIF. Raju (1990) evaluated these methods using two subsamples of a vocabulary test from students in grades 4 to 6, with a sample size of 1000 each for black and white students. The SA method identified 7 items as having DIF, while the UA method identified 13 items as having DIF. Donoghue, Holland, and Thayer (1993) conducted a simulation study comparing the area measures with the Lord statistic, and a closed interval measure proposed by Kim and Cohen (1991). The factors manipulated in this study were the total number of examinees (300/300, 600/600, or 1000/1000), the number of items (30 or 60), and the number of items with DIF (0, 3, 6, or 12). The result of this simulation study was that the UA method exhibited an inflated Type I error, rejecting the hypothesis of no DIF at more than twice the α level. Also, the SA method

was able to detect items with different difficulty parameters but not items with different discrimination parameters. Cohen and Kim (1993) conducted a simulation study comparing the area measures with the Lord statistic by manipulating the sample size (100/100, 500/500), the test length (20, 60 items), the ability distribution of the focal group, and the amount of DIF in the test items (0%, 10%, or 20%). The power of the three methods decreased as test length and significance level increased, and the results indicated that the Lord statistic was more effective than the area measures. For these studies, Type I error rates decreased and the power increased as the sample size increased.

Rationale for the Proposed Method

Both CTT and IRT methods provide a way to determine differences in item performance using comparisons of item difficulty between the reference and the focal group. For example, Holland and Thayer (1988) showed that for the 1PL model, the standardized MH statistic is equal to an estimate of the difference between the b parameters for the focal and reference groups multiplied by 2.35. However, the Mantel-Haenszel test is the CTT method with the smallest sample size requirement of 125 for the focal group while the IRT method (Differential item pair functioning) has a minimum sample size of 500 for the focal group. When the sample size for the focal group is much smaller than these minimums, there is an increased risk that the common support between the reference and the focal group no longer exist. For example, if there are 40 items on an assessment and only 25 examinees in the focal group, there will be at most 25 score intervals out of a possible 41 where examinees from both groups are matched. This could

lead to a loss of information as intervals without focal group examinees would not be included in the calculation for the Mantel-Haenszel test. Thus, a new method is proposed, titled relative item performance, to evaluate differences in item difficulties without the need to have common support, with the interpretation being that the item performs differently than other items on the assessment regardless of which group it favors.

The Proposed Method to Detect Relative Item Performance

For this proposed method, there is no requirement that common support must exist in the reference and focal groups. By not having this requirement, the sample size does not have to be large enough to ensure that there are examinees in both the reference and focal group at all score levels. The definition of DIF states that the probability of obtaining a correct response is the same for both the reference and the focal group assuming that the two groups have an overlap in their ability distribution. However, as this method assumes that there is no overlap between the two groups, DIF cannot be examined. The proposed method looks at how each item performs compared to other items on the assessment, assuming that differences between the reference and the focal groups on each item should be consistent across all items. The interpretation of this method is different from current methods in that this method only states that the item is performing differently than expected with no mention of which group the item favors.

To evaluate an item's performance relative to other items on the assessment, first the difference ($PDIFF_i$) between proportion of correct responses from the focal and reference groups must be calculated for all items as follows:

$$PDIFF_i = \frac{X_{Ri}}{N_{Ri}} - \frac{X_{Fi}}{N_{Fi}}, \quad (29)$$

where X_{Ri} and X_{Fi} are the number of correct responses for the item from the reference and focal groups, respectively, and N_{Ri} and N_{Fi} are the number of examinees for the reference and focal groups, respectively. This is how effect sizes are reported for both the Mantel-Haenszel and logistic regression methods, and is equivalent to comparing the b parameters for the reference and the focal groups (Steinberg & Thissen, 2006). The mean and population standard deviation are then calculated across all items as follows:

$$\overline{PDIFF} = \frac{\sum PDIFF_i}{N}, \quad (30)$$

$$SD(PDIFF) = \sqrt{\frac{\sum (PDIFF_i - \overline{PDIFF})^2}{N}}, \quad (31)$$

where N is the total number of items on the assessment. After obtaining the mean and standard deviation of the differences, an effect size for the item of interest is computed as follows:

$$Effect_i = \frac{PDIFF_i - \overline{PDIFF}}{SD(PDIFF)}. \quad (32)$$

The effect size is assumed to follow the standard normal distribution, as the proportion of a correct response follows a binomial distribution and the difference between two proportions follows a normal distribution if the sample is larger than 30 according to the Central Limit Theorem. Thus, an item is found to perform differently from other items in the assessment if the absolute value of $Effect_i$ is greater than 1.96, the standard normal value with $\alpha/2 = 0.05$.

Consider the situation where we have 10 items with the proportions of correct responses for the reference and focal groups as shown in Table 3:

Table 3

Example Effect Sizes for 10 Items

Item	Reference	Focal	PDIFF	Effect	MHStat
1	0.576	0.590	-0.014	-0.2617	0.9329
2	0.624	0.575	0.049	0.5925	1.2385
3	0.950	0.935	0.015	0.1315	1.2577
4	0.938	0.930	0.008	0.0366	1.0846
5	0.612	0.595	0.017	0.1586	1.0936
6	0.648	0.610	0.038	0.4433	1.1738
7	0.716	0.730	-0.014	-0.2617	0.8936
8	0.660	0.845	-0.185	-2.5800*	0.2932*
9	0.526	0.440	0.086	1.0941	1.4801*
10	0.508	0.455	0.053	0.6467	1.2374

After calculating the difference between the two proportions, the mean and population standard deviation is equal to 0.0053 and 0.0736, respectively. The effect size is then calculated and shown in the fifth column of Table 3. In column 6, the Mantel-Haenszel statistics are given for these items with items 8 and 9 being flagged as exhibiting DIF. As the absolute value of the effect size for item 8 is 2.5800, which is greater than 1.96, this item is performing differently than the other 9 items on the assessment as seen with a difference in proportion correct of 0.185 for this item and less than .10 for the other items. Although it appears that the item favors the reference group, this cannot be stated because there is no knowledge about the common support between the reference and the focal groups.

After determining which items perform differently, the items can then be evaluated through the use of cognitive or expert review to determine the potential impact of the item wording and/or content on the performance of the two groups of examinees (reference and focal).

Research Questions

Based on this literature review, the following research questions were answered in the simulation study:

- 1) Does the relative item performance method result in robust inference about item performance?
 - a. What is the Type I error rate of the relative item performance method? How does this error rate compare to other methods (Mantel-Haenszel with small intervals, Mantel-Haenszel with large intervals, Differential item pair functioning)?
 - b. For conditions in which the Type I error rate is controlled, what is the power of the relative item performance method when detecting an item that perform differently? How does this power compare to other methods (Mantel-Haenszel with small intervals, Mantel-Haenszel with large intervals, Differential item pair functioning)?
- 2) For selected factors (sample size for both the reference and focal groups, size of the DIF, the test length, and the proportion of items contaminated with DIF), what is their effect on the Type I error rate and power of the four methods?

Chapter 3 – Methods

In this proposed study, the relative item performance method is evaluated to detect differences in item performance when the focal group sample size is small. This method attempts to address the concern that there is a potential lack of common support in the ability distribution, which could lead to misidentification of which group the item favors. This method was evaluated by using a simulation design. By using simulated data, the research questions are addressed about the proposed method's Type I error rate and power as well as the effect of various factors on the proposed method.

Choosing DIF Detection Methods for Comparison

There are three CTT methods (Mantel-Haenszel, standardization, and logistic regression) and three IRT methods (Wald, differential item pair functioning, and signed/unsigned areas) that could be used in comparison to the relative item performance based on Type I error rates and power. In this study, only one CTT and one IRT method were used with Mantel-Haenszel chosen as the CTT method and differential item pair functioning chosen as the IRT method. The Mantel-Haenszel method was chosen as it has the smallest sample size requirement of the three methods and is currently used by ETS when conducting DIF analyses. The differential item pair functioning method was chosen because it compares pairs of items rather than individual items similar to how the relative item performance compares an item against the distribution of all items on the assessment.

For the Mantel-Haenszel test, prior research has found that the minimum sample size should be 125 for the focal group to allow for sufficient common support between both the reference and the focal group. As this study focuses on smaller sample sizes,

there needs to be a reduction in the amount of score levels used in the analysis. Donoghue and Allen (1993) mentioned thin versus thick matching where thin matching uses all possible score levels and thick matching breaks up the score levels into intervals. To allow this method to be used as comparison against the relative item performance method, two sets of intervals were used. For the first set of intervals, the score levels were grouped into 10 intervals. For example, if there are 20 items then there are 21 total score levels (0 to 20), and the levels would be broken into: 0-2, 3-4, 5-6, 7-8, 9-10, 11-12, 13-14, 15-16, 17-18, and 19-20. For this study, this method is known as MH with small intervals (MH1). The second set of intervals had the score levels grouped into 5 intervals with the method known as MH with large intervals (MH2). Using the same example as before, the levels would be broken into: 0-4, 5-8, 9-12, 13-16, and 17-20. By using these intervals, the Mantel-Haenszel test can be used as a comparison method even when the sample size is small.

Simulation Design

The simulation design first began with the generation of data following specific conditions. After the data were generated, several methods were then run using the simulated data: relative item performance, MH with small intervals, MH with large intervals, and differential item pair functioning. I analyzed the resulting statistics to determine whether each item is flagged as performing differently than expected.

Data generation. The first simulation factor is the number of examinees in each group; for this condition, the sample size for the focal group was chosen to be smaller than the minimum requirement of 125 for the Mantel-Haenszel test. Thus, the sample size factor was simulated in 4 different conditions: the focal group size being 25, 50, 100, or

200 and the reference group size being fixed at 500. Across the different simulation studies of the Mantel-Haenszel test, the sample sizes for both the reference group and the focal group have ranged from 50 to 3000. Some of these studies have examined the MH statistic when the sample sizes are small (Fidalgo, Ferreres, & Muñiz, 2004; Mazor, Clauser, & Hambleton, 1992; Muñiz, Hambleton, & Xing, 2001; Rogers & Swaminathan, 1993; Zieky, 1993), while others have compared the effect of different group sizes on several detection procedures (Miller & Oshima, 1992; Narayanan & Swaminathan, 1996; Roussos & Stout, 1996). Because I was interested in finding DIF methods that work with small sample sizes, I focused on evaluating DIF using small focal sample sizes ranging from 25 to 200. As the sample size increases, the power should increase with the Type I error rate consistently around 0.05.

To create the dataset of item responses, the first step was to generate the ability parameters of the examinees in the reference group by obtaining a value from a standard normal distribution with a mean of zero and a standard deviation of one, $N_r(0,1)$. For the focal group, the mean of the ability distribution ranged from 0 to -1 at -0.5 intervals. This should have an effect on the common support between the reference and the focal groups, given that as the ability mean decreases, the amount of overlap between the two ability distribution decreases. When the mean of the ability distribution for the focal group is one standard deviation lower than that of the reference group, the Type I error rate is expected to be inflated especially when other items on the assessment also exhibit DIF for the Mantel-Haenszel method (Wang & Su, 2004). For the proposed method, the mean of the ability distribution may not have as strong an effect because the expectation is that the differences in the percentages should remain consistent across items, except in

the case where other items exhibit DIF. Many of the simulations evaluating DIF detection using the Mantel-Haenszel have generated latent ability with the differences between the group means, with the majority using one standard deviation difference (e.g., Ackerman & Evans, 1992; Donoghue & Allen, 1993; Mazor, Clauser, & Hambleton, 1992) and a few using 0.5 standard deviation difference (e.g., Narayanan & Swaminathan, 1994).

For the assessment itself, data were manipulated such that the test length varied from what might be considered a short test to a long test. In previous simulation studies, the test length varied from 5 items up to 100 items, with 40 being the most common test length (Ackerman & Evans, 1992; Clauser, Mazor, & Hambleton, 1993; Donoghue & Allen, 1993; Fidalgo, Ferreres, & Muñiz, 2004; Fidalgo, Mellenbergh, & Muñiz, 2000; Miller & Oshima, 1992; Muñiz, Hambleton, & Xing, 2001; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Uttaro & Millsap, 1994). Thus, the assessment had either 20, 40, or 80 items similar to Clauser, Mazor, and Hambleton (1993). Based on these studies, increasing the number of test items could reduce the misidentification of items performing differently than other items for the relative item performance method.

Item responses (0/1) were generated based on the 2PL model (Equation 5) using the randomly generated person ability mentioned earlier. After grouping the number of items into five intervals, the difficulty item parameter was randomly drawn from a uniform distribution as follows: $-2.5 < b < -1.5$, $-1.5 < b < -0.5$, $-0.5 < b < 0.5$, $0.5 < b < 1.5$, $1.5 < b < 2.5$. For each of these intervals, the items were then split into intervals of two where the discrimination parameter was randomly drawn from a uniform distribution

ranging from either 0.2 to 1.0 or 1.0 to 2.0. Table 4 illustrates an example of how the parameters are created for 10 items.

Table 4

Difficulty and Discrimination Parameters

Items	Parameters		Stratum	
	Difficulty	Discrimination	Difficulty	Discrimination
1	-2.5 to -1.5	0.2 to 1.0	1	1
2		1.0 to 2.0		2
3	-1.5 to -0.5	0.2 to 1.0	2	1
4		1.0 to 2.0		2
5	-0.5 to 0.5	0.2 to 1.0	3	1
6		1.0 to 2.0		2
7	0.5 to 1.5	0.2 to 1.0	4	1
8		1.0 to 2.0		2
9	1.5 to 2.5	0.2 to 1.0	5	1
10		1.0 to 2.0		2

These item parameters were only drawn once from the corresponding uniform distribution for 20, 40, or 80 items, meaning that there were three sets of item parameters in total.

Along with the number of test items, the percentage of items exhibiting DIF has been shown to be a contributing factor to the performance of the Mantel-Haenszel method (Fidalgo, Mellenbergh, & Muñiz, 2000). To evaluate the Type I error rate and the power of the four methods, the percentage of DIF items ranged from 0% (none), 10% (small), 20% (moderate) and 30% (large). For percentages greater than 0, items were randomly drawn from the five intervals used to generate the item parameters. This ensures that the items chosen to exhibit DIF have different item difficulty and discrimination parameters. Many of the simulation designs ranged from 0% to 50% DIF

items (Narayanan & Swaminathan, 1996; Swaminathan & Rogers, 1990; Wang & Su, 2004). ETS found that there were 1 to 14 items out of 42 (approximately 2% to 33%) that were identified as having DIF when analyzing the TOEFL Junior Standard Test taken in October 2010 (Young, Morgan, Rybinski, Steinberg, & Wang, 2013). The proposed method could be affected by item contamination as the average and standard deviation of the differences includes these items, leading one to conclude that an item may not be performing differently when it is or vice versa.

The difference between proportions of correct response for the DIF items was simulated by adjusting the IRT difficulty parameter for the focal group by a specific factor. Simulation studies have adjusted the parameter by subtracting a number between 0 to 1.5, with either 3, 4 or 5 levels (Donoghue & Allen, 1993; Fidalgo, Ferreres, & Muñiz, 2004). Thus, for this study, the difficulty parameter for the items simulated to exhibit DIF for the focal group were adjusted by subtracting 0.25, 0.5, 0.75, and 1. By subtracting a factor from the difficulty parameter, this would mean that an item is easier for the focal group than the reference group.

These simulated conditions allowed me to evaluate the methods' ability to flag an item's performance: (a) the ability distribution (3 conditions), (b) number of examinees (4 conditions), (c) number of test items (3 conditions), (d) proportion of test items with DIF (4 conditions), and (e) size of the DIF (4 conditions). Note that when the proportion of test items with DIF was equal to 0%, the size of the DIF did not matter meaning that there were only 36 possible conditions ($3 \times 4 \times 3 \times 1$). For the remaining proportions, 432 possible conditions ($3 \times 4 \times 3 \times 3 \times 4$) were simulated. Within each condition, there

are 2 levels for the item discrimination parameters and 5 levels for the item difficulty parameter. The summary of the simulation factors is given in

Table 5.

Summary of Simulation Factors

Ability distribution	Focal group: $N(0, 1)$, $N(-0.5, 1)$, $N(-1, 1)$ Reference group: $N(0,1)$
Sample size	Focal group: 25, 50, 100, 200 Reference group: 500
Test length	20, 40, 80 items
% test contamination	0%, 10%, 20%, 30%
DIF size adjustment for item difficulty parameter	0.25, 0.5, 0.75, 1

To ensure that the data were generated correctly, distribution statistics were calculated to ensure that the mean and standard deviation of the ability distribution matched the specified conditions. Individual calculations were also conducted to ensure that the probability of obtaining a correct response was generated correctly using the specified ability of the examinees and the item parameters.

DIF detection methods. After generating the data, four methods were applied to the data: relative item performance (RI), the Mantel-Haenszel test with small (MH1) and large intervals (MH2), and the Differential Item Pair Functioning (DIPF) approach. For the two Mantel-Haenszel methods, the log-odds statistic (Equation 10) was used to determine if the item exhibited DIF. For the DIPF method, item difficulty parameters were calibrated using a package in R called plRasch (Li & Hong, 2014, R Core Team, 2016). For each item, all possible item pairs were then evaluated following Equations 22, 23, and 24. For a specific item to be flagged as significant, more than half of the item

pairs containing that item must be significantly larger than 0 as well. In the case of 20 items, a specific item has 19 possible pairs and is flagged if more than 10 of these pairs were significantly larger than 0. This cutoff was chosen arbitrarily to allow for the Type I error rates to be calculated for a single item rather than for pairs of items which was how the authors calculated their Type I error rates. For RI, the $PDIFF_i$ for each item, the mean and standard deviation for all items were calculated using Equations 29, 30, 31, and 32 resulting in a single RI effect size measure for each item. This effect size measure was then compared to a standard normal distribution with a critical value equal to $\alpha/2$ or 1.96.

Within each cell, the data generation and subsequent DIF analysis methods were conducted a thousand times. A pilot run with 20 items, an ability mean for the focal group equal to -1, item contamination equal to 30%, and a sample size for the focal group equal to 25 determined that the estimate for the Type I error rate stabilized at 900 runs. All of the simulation work was conducted using the statistical computing software *R* version 3.3.1 (R Core Team, 2016).

Analysis. After each replication, the RI, MH1, MH2, and DIPF test statistics were calculated and all items were flagged as either performing differently or not. Once all replications were complete, then the Type I error rate was calculated as the number of times the items were flagged as performing differently when there actually was no difference in the difficulty parameter and the power was calculated as the number of times the items were correctly flagged as performing differently when there actually was a difference in the difficulty parameter. There were three datasets created as a result of the simulation study. The first dataset contained the sample sizes, the focal ability mean parameter, the number of items, the difficulty parameter (both reference and focal) for

each item, the discrimination parameter for each item, the difference in the difficulty parameter between the reference and the focal group, the amount of item contamination, and the Type I error rate for the four methods (RI, MH1, MH2, DIPF) for items that did not exhibit DIF. The second dataset was similar to the first but instead of having the Type I error rate, it contained the power for each of the four methods for items that exhibited DIF. The third dataset contained the simulation factors, the effect size statistic, and the flags which indicated if an item was identified as exhibiting DIF for the four methods across all conditions and replications.

Addressing the Research Questions

For the first part of the first research question, the Type I error rates were evaluated using items that were flagged as exhibiting DIF when the item difficulty parameter is the same for both the focal group and the reference group. For each cell, the Type I error rate for each method is calculated as the number of times the item is flagged divided by the number of replications (1000). Graphs and descriptive statistics are provided to compare Type I error rates across the different methods.

For the second part of the first research question, power was calculated using items that were flagged as exhibiting DIF when the item difficulty parameter was different for the focal group and the reference group. For each cell where the Type I error rate is controlled, the power for each method was calculated as the number of times an item was flagged divided by the number of replications (1000). Graphs and descriptive statistics are provided to compare power across the different methods.

The second research question addressed what factors affected the Type I error rates and power for the detection methods considered in this simulation. To answer this

question, descriptive statistics and graphs were used to evaluate the effect from the sample size for both the reference and focal groups, size of the DIF, the test length, the proportion of items contaminated with DIF, the item parameters, and the mean of the focal ability distribution. In addition, logistic regression was used to determine the magnitude and direction of the effect from the simulation factors on the various methods where the outcome variable was whether or not an item was correctly flagged. Specifically, for the relative item performance method, descriptive statistics for the effect size was evaluated to determine the effect of the simulation factors on the ability to correctly flag items.

After the research questions are addressed, the limitations and implications of the results are discussed. Based on these limitations, recommendations are made for researchers as well as directions for future research.

Chapter 4 – Results

This chapter includes the results of the simulation study described in Chapter 3. Logistic regression is used to determine the effect of the simulation factors on the flagging of items, and descriptive statistics for the effect size across all simulation factors are provided. Based on the results of the logistic regression, descriptive statistics and graphical illustrations of the effect on Type I error rates and power from the significant simulation factors for each of the four methods are provided. Using these results, the research questions at the end of Chapter 2 are then addressed.

Type I Error Rates Without Item Contamination

In this section, descriptive statistics of the Type I error rates for all methods across simulation conditions without item contamination are reported. The reason that the Type I error rates are evaluated with or without item contamination is that the Type I error rates without item contamination does not include the size of DIF. If a method's Type I error rate obtained for each condition is not close to 0.05, then the method is not recommended for practical use. In Table 6, the summary statistics of the Type I error rates for each method are given. Using the median, the Type I error rate is near 0.05 for both Mantel-Haenszel methods, but not for the differential item pair functioning or the relative item performance method.

Table 6

Summary Statistics of Type I Error Rates Across Cells (N=360)

	MH1	MH2	DIPF	RI
Minimum	0.014	0.001	0.002	0.000
1st Quartile	0.041	0.042	0.007	0.016
Median	0.047	0.049	0.014	0.040
Mean	0.049	0.056	0.032	0.042
3rd Quartile	0.053	0.059	0.031	0.061
Maximum	0.147	0.271	0.384	0.173

To determine what caused the difference in Type I error rates for each of the four methods, logistic regression was conducted with the outcome variable being whether or not the item was flagged inappropriately when there was no item contamination. The independent variables are the main effects, two-way interaction effects, and the three-interaction effects of the following factors: the number of items (ni), the ability mean of the focal group ($theta$), the sample size for the focal group ($fsize$), the item difficulty parameter ($bref$), and the item discrimination parameter ($aref$). Note that any interaction between the number of items and the item parameters are excluded because the number of items within a specific interval for the difficulty parameter is determined by the total number of items. The equation used is given as follows:

$$\begin{aligned}
 Y_i = & ni + theta + fsize + aref + bref + ni * theta + ni * fsize \\
 & + theta * fsize + theta * aref + theta * bref \\
 & + fsize * aref + fsize * bref + aref * bref + ni \\
 & * theta * fsize + theta * fsize * aref + theta * fsize \\
 & * bref + theta * aref * bref + fsize * aref * bref
 \end{aligned} \tag{33}$$

Note that there are 140 items (20 + 40 + 80) times 12 conditions (3 ability mean levels and 4 sample size levels) times 1,000 replications, which gives a total of 1,680,000 observations. The results of the logistic regression are given for each method in Table 7 where the effect size is equal to the odds ratio with a value above 1 indicating that the odds of the item being inappropriately flagged for exhibiting DIF is greater than the odds of the item being correctly flagged as not exhibiting DIF. Effect sizes that have a difference from 1.000 of at least 0.05 are bolded to represent substantial effects. For the MH1, MH2, and RI methods, the factors with the largest effect on the Type I error rates are the ability mean of the focal group, the item parameters, and their interactions. The sample size and the number of items do not have a large effect on the Type I error rates. For the DIPF method, the Type I error rate is not controlled for any of the simulation conditions with both the mean and median Type I error rate much lower than that of the other methods. This could be due to the fact that the determination of an item being flagged as exhibiting DIF is based on an arbitrary cutoff of the number of paired items being statistically different from 0.

Table 7

Logistic Regression Odds Ratios for Items with No Item Contamination

(N=1,680,000)

	MH1	MH2	DIPF	RI
(Intercept)	0.0469	0.0465	0.0015	0.0745
ni	1.0002	1.0004	1.0034	1.0085
theta	0.8629	0.9306	0.1135	1.9245
fsize	1.0004	1.0005	1.0095	1.0000
aref	0.8708	0.8645	5.7745	0.3298
bref	0.9920	1.0160	0.6924	0.9658
ni:theta	1.0016	1.0020	1.0033	1.0016

ni:fsize	1.0000	1.0000	1.0000	1.0000
theta:fsize	0.9982	1.0004	0.9913	1.0081
theta:aref	1.0920	0.9613	2.2129	0.4427
theta:bref	0.9013	0.9229	0.7012	0.9251
fsize:aref	1.0006	1.0005	0.9892	1.0004
fsize:bref	1.0000	0.9998	1.0062	1.0001
aref:bref	1.0014	0.9809	1.2674	0.9801
ni:theta:fsize	1.0000	1.0000	1.0000	1.0000
theta:fsize:aref	0.9986	0.9956	0.9998	0.9938
theta:fsize:bref	0.9998	0.9995	1.0050	0.9998
theta:aref:bref	1.2986	1.4203	0.6711	1.3038
fsize:aref:bref	1.0001	1.0004	0.9946	1.0002

Type I error rates are given for each combination of these factors in Table A4 in the Appendix. The median Type I error rate are shown for each simulation factor that have a substantial odds ratio. Specifically, due to their large effects, the size of the discrepancy between the focal group and reference group ability mean and the values of the item parameters are discussed in detail.

Mean of the focal ability distribution. For the different means of the focal group's ability distribution (0, -0.5, -1), the median Type I error rates are shown for each method in Figure 3. The Type I error rates for MH1 and MH2 are approximately 0.05 regardless of the focal group's ability distribution, which is unexpected as previous research found that the Type I error rates were inflated when the ability mean was one standard deviation apart (Wang & Su, 2004). However, the Type I error rates for the RI method increases from a Type I error rate of 0.019 when the ability mean is equal to -1 to a Type I error rate of 0.051 when the ability mean is equal to 0. As seen in Table 7, the ability distribution has a smaller effect on the Type I error rates for the MH1 and MH2

methods (odds ratios of 0.8659 and 0.9494) compared to the RI method (odds ratio of 1.9374).

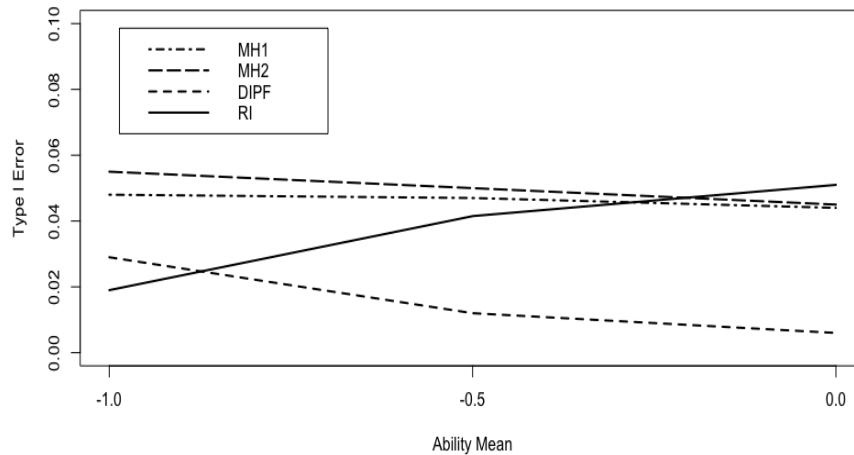


Figure 3. Median Type I error rates across focal ability distribution. The Type I error rates for MH1, MH2, DIPF, and RI methods are shown across the focal ability means (-1, -0.5, and 0).

Item parameters. As found from the logistic regression results, there are two item parameters that can affect the Type I error rates: the difficulty parameter and the discrimination parameter. First, in Figure 4, the effect of the difficulty parameter on the Type I error rate is shown.

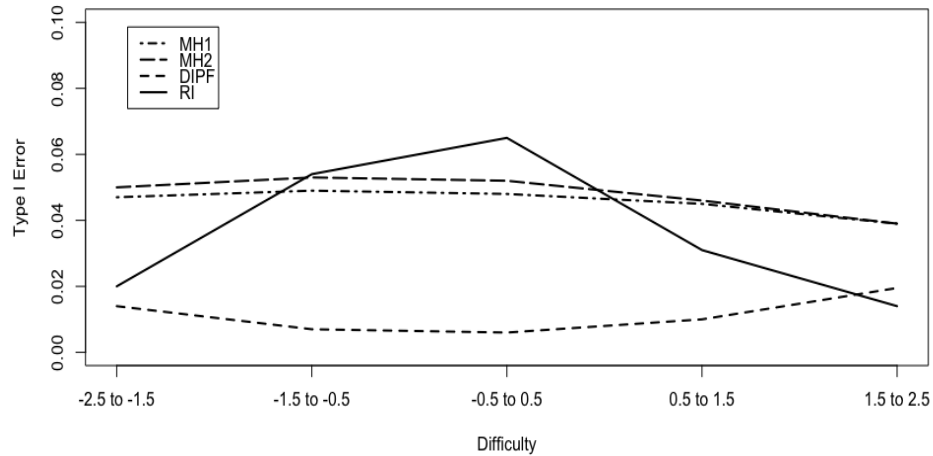


Figure 4. Median Type I error rates across difficulty parameters. The Type I error rates for MH1, MH2, DIPF, and RI are shown across the item difficulty parameters ranging from -2.5 to 2.5.

Before discussing the results of the difficulty parameter, the Type I error rates at different levels of the discrimination parameter need to be reviewed with the Type I error rates shown in Figure 5.

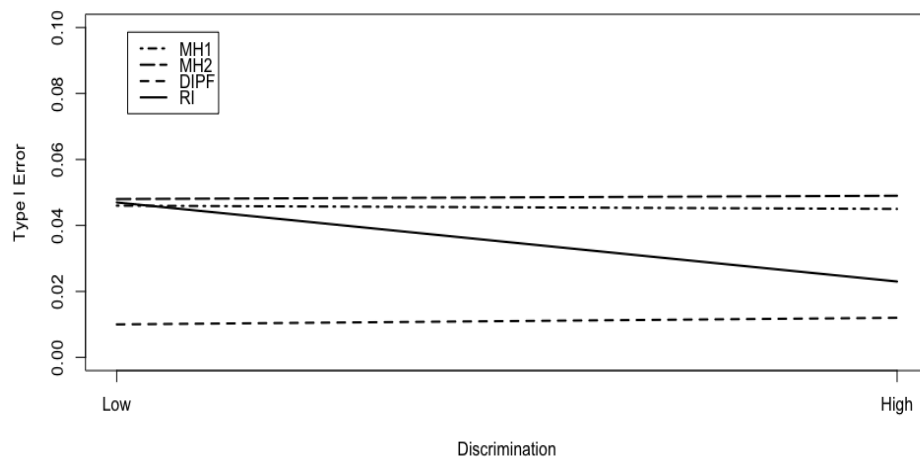


Figure 5. Median Type I error rates across discrimination parameters. The Type I error rates for MH1, MH2, DIPF, and RI are shown across the item discrimination

parameters where low represents a range of 0.2 to 1 and high represents a range of 1 to 2.

As the item parameters work together to determine the probability of obtaining a correct response on an item, the interaction between the two has an impact on the Type I error rates. This is shown in Figure 6.

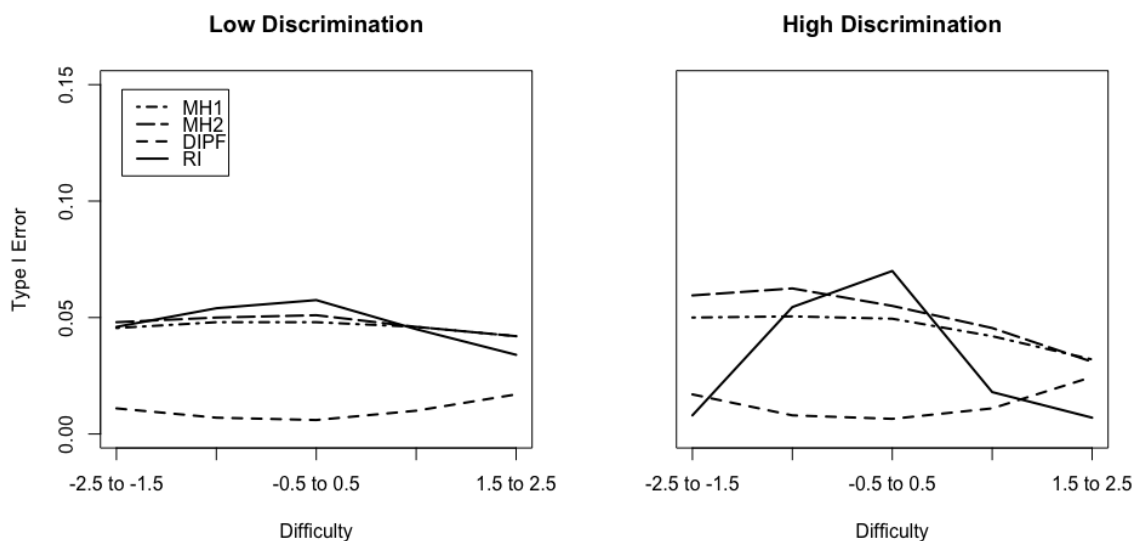


Figure 6. Median Type I error rates across item parameters. The left graph shows the Type I error rates for the four methods across the item difficulty parameters when the discrimination parameter is low, while the right graph shows the Type I error rates for the four methods across the item difficulty parameters when the discrimination parameter is high.

When looking at both the item difficulty and the item discrimination parameters, it is clear that items with a low discrimination parameter allow for consistent Type I error rates, while the pattern varies when the discrimination parameter is high. As seen in the

left graph, the Type I error rates for the MH1, MH2, and RI methods range from 0.034 to 0.058 across the difficulty parameters. In the right graph, the Type I error rates for the MH1 and MH2 exhibit a downward trend, while the Type I error rates for the RI method are much lower (0.007, 0.008) when the difficulty parameters are at the extreme ends, and equal to 0.070 when the difficulty parameter is between -0.5 and 0.5. This is also shown in the effect sizes for the item parameters. For the item difficulty parameters, the effect size for the RI method is slightly larger (odds ratio of 0.9539) than the MH1 and MH2 methods (odds ratio of 0.9795 and 0.9785), but the effect size for the RI method (odds ratio of 0.3305) is much stronger than the MH1 and MH2 methods (odds ratios of 0.8715 and 0.8663). One possible reason for this is the fact that the RI method is calculated using the percentage of correct responses and does not incorporate item discrimination. The item discrimination distinguishes between examinees with higher and lower ability in terms of how much higher their probability of obtaining a correct response. As the RI method disregards the ability distribution for both groups, this is not accounted for in the calculation.

Interaction between ability mean and item parameters. For all three methods, the effect of the two-way and three-way interactions between the ability mean and item parameters were shown to have an effect on the Type I error rates in the logistic regression. The first two-way interaction is between the ability mean and the difficulty parameter as shown in Figure 7. For the MH methods, the Type I error rates decrease for the extreme difficulty parameters as the ability mean decreases. For the RI method, the up-and-down pattern shifts to the left as the ability mean decreases. However, all the

effect sizes are approximately the same for all three methods (MH1 – 0.9021, MH2 – 0.9308, RI – 0.9283).

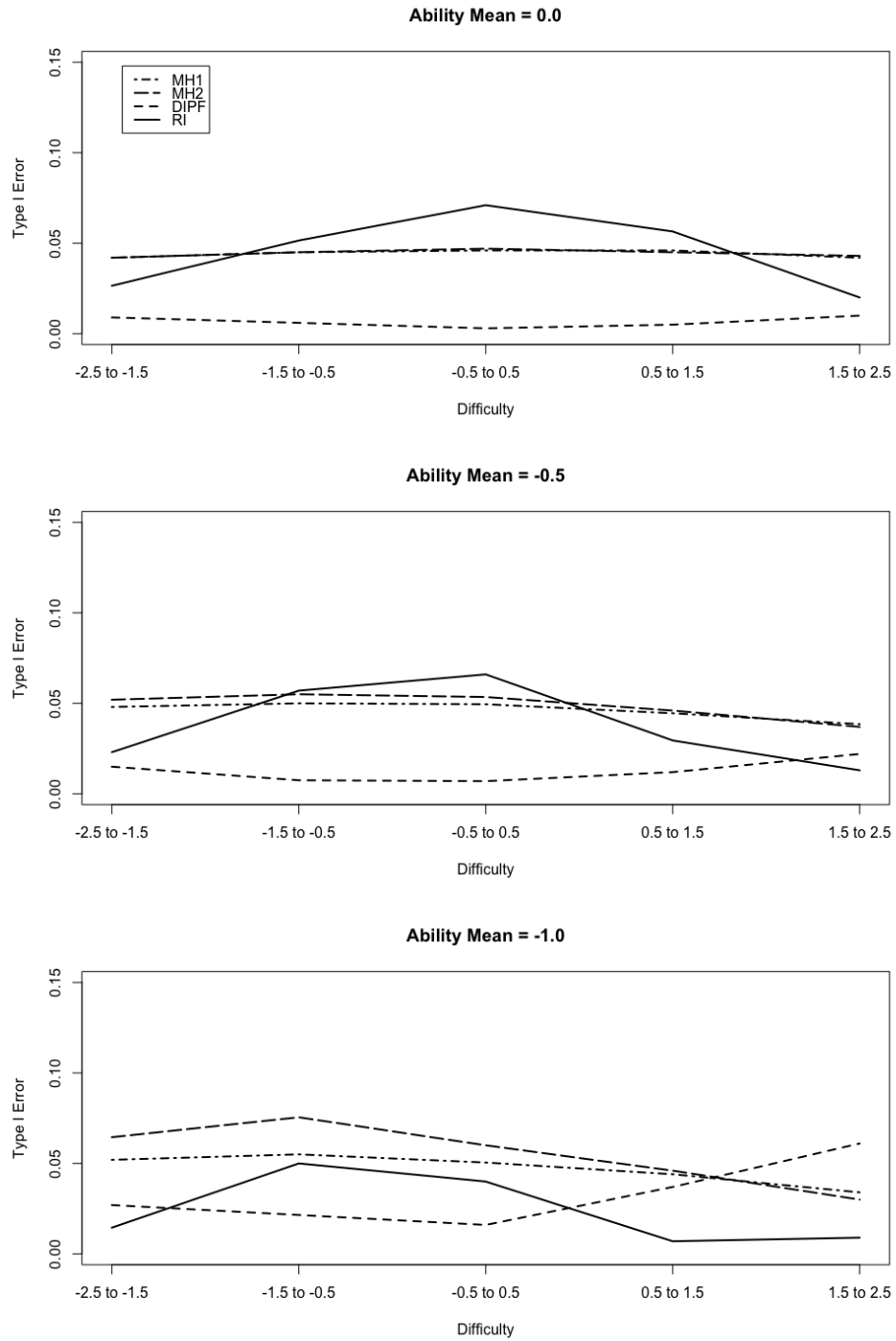
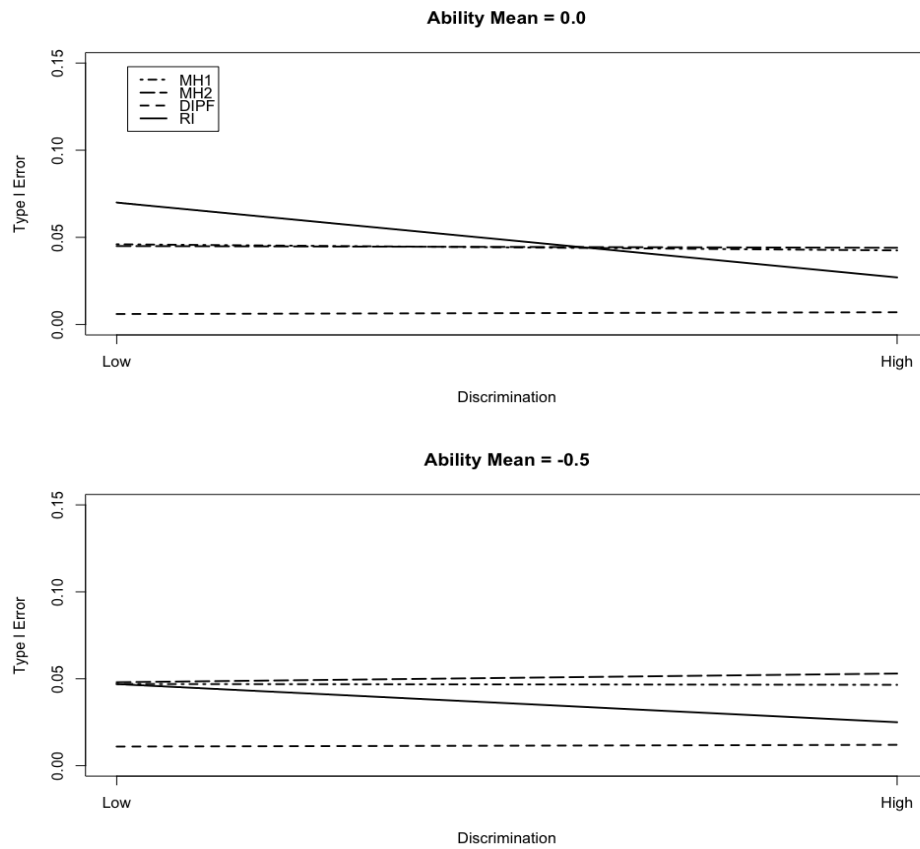


Figure 7. Median Type I error rates across ability mean and item difficulty. The first graph shows the Type I error rates for the four methods across the item

difficulty parameters when the ability mean is equal to 0, the second graph shows the Type I error rates for the four methods across the item difficulty parameters when the ability mean is equal to -0.5 and the third graph shows the Type I error rates for the four methods across the item difficulty parameters when the ability mean is equal to -1.0.

The second two-way interaction is between the ability mean and the discrimination parameter, shown in Figure 8.



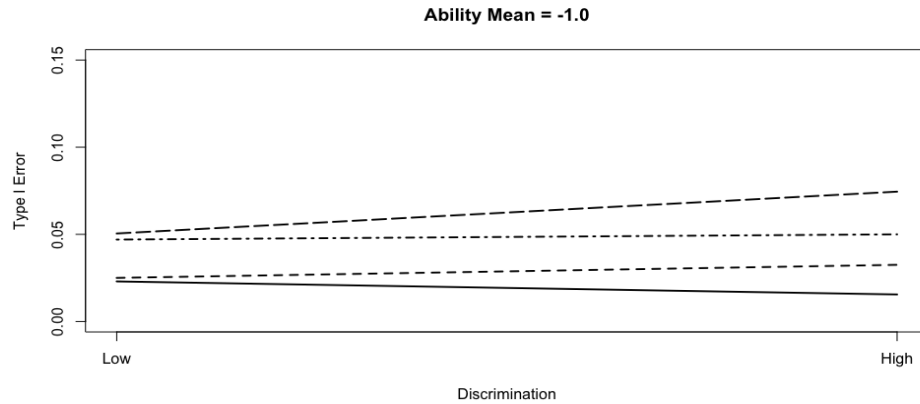


Figure 8. Median Type I error rates across ability mean and item discrimination.

The first graph shows the Type I error rates for the four methods across the item discrimination parameters when the ability mean is equal to 0, the second graph shows the Type I error rates for the four methods across the item discrimination parameters when the ability mean is equal to -0.5 and the third graph shows the Type I error rates for the four methods across the item discrimination parameters when the ability mean is equal to -1.0.

For the MH1 method, the Type I error rates stay consistently around 0.05 regardless of the item discrimination or the ability mean. For the MH2 method, the Type I error rates increase across the discrimination parameters as the ability mean also increases. The RI method shows the opposite pattern from that of the MH2 method with a downward trend. The effect on the Type I error rates for the RI method (odds ratio of 0.4401) is stronger than on the MH1 and MH2 methods (odds ratios of 1.0872 and 0.9411).

The three-way interaction between the ability mean and the item parameters are shown in Figure 9.

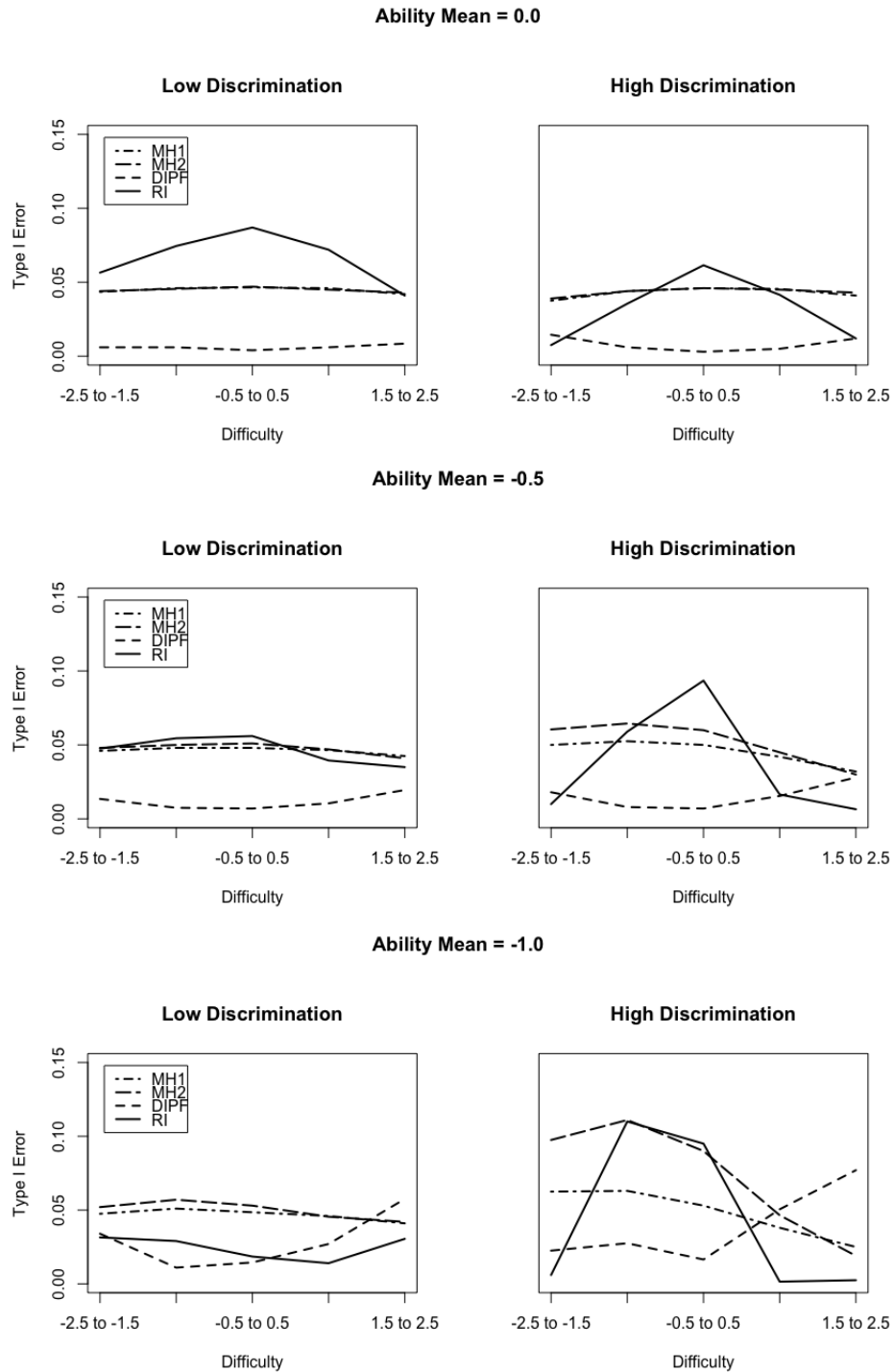


Figure 9. Median Type I error rates across ability mean and item parameters. The first graph shows the Type I error rates for the four methods across the item parameters when the ability mean is equal to 0, the second graph shows the Type I

error rates for the four methods across the item parameters when the ability mean is equal to -0.5 and the third graph shows the Type I error rates for the four methods across the item parameters when the ability mean is equal to -1.0.

What is interesting to note is that the MH1 and MH2 methods are consistent across the item parameters and ability mean except when the ability mean is equal to -1 and the discrimination parameter is high. For the RI method, the same up-and-down pattern is seen across all combinations except for two: ability mean equal to -0.5 with low discrimination and ability mean equal to -1.0 with low discrimination. For the first case, the RI method shows a consistent Type I error rate equal to 0.05. However, the second case shows a Type I error rate much lower than the MH1 and MH2 methods across all difficulty parameters. The effect on the Type I error rates is similar across all three methods (MH1 – 1.2975, MH2 - 1.4126, RI – 1.3048).

Summary. For all three methods, the main factors that affect the Type I error rates are the ability mean for the focal group and the item parameters. The sample size and the number of items were found not to have an effect on the Type I error rates for all methods. Overall, the Type I error rates without item contamination for the MH1 and MH2 methods are shown to increase as the focal group ability mean deviates from the reference group ability mean, especially when the discrimination parameter is high. For the RI method, the Type I error rates are shown to decrease as the focal group ability mean deviates from the reference group ability mean, especially with high discrimination and extreme difficulty parameters.

Type I Error Rates with Item Contamination

In this section, descriptive statistics of the Type I error rates for all methods across simulation conditions with item contamination are reported. In Table 8, the summary statistics of the Type I error rates for each method are given. Using the median, the Type I error rate is near 0.05 for both Mantel-Haenszel methods, but not for the differential item pair functioning or the relative item performance method.

Table 8

Summary Statistics of Type I Error Rates Across Cells (N=4,320)

	MH1	MH2	DIPF	RI
Minimum	0.004	0.002	0.000	0.000
1st Quartile	0.047	0.048	0.009	0.003
Median	0.059	0.063	0.015	0.015
Mean	0.085	0.097	0.041	0.031
3rd Quartile	0.090	0.106	0.039	0.040
Maximum	0.921	0.956	0.885	0.589

To determine what is causing the difference in Type I error rates for each of the four methods, logistic regression was conducted with the outcome variable being whether or not the item is flagged inappropriately when there is item contamination. Logistic regression was run separately for 20 items and 40 items, but could not be run for 80 items as the dataset was large ($N=9,216,000$). The independent variables are the main effects, two-way interaction effects, and the three-interaction effects of the following factors: the ability mean of the focal group (θ), the sample size for the focal group ($fsize$), the amount of item contamination ($itemcont$), the size of the DIF ($bdiff$), the item difficulty parameter ($bref$), and the item discrimination parameter ($aref$). The equation is given as follows:

$$\begin{aligned}
Y_i = & \theta + fsize + aref + bref + itemcont + bdiff + \\
& \theta * fsize + \theta * itemcont + \theta * bdiff + \\
& \theta * aref + \theta * bref + fsize * itemcont + \\
& fsize * bdiff + fsize * aref + fsize * bref + \\
& itemcont * bdiff + itemcont * aref + itemcont * bdiff \\
& + bdiff * aref + bdiff * bref + aref * bref + \\
& \theta * fsize * itemcont + \theta * fsize * bdiff + \\
& \theta * fsize * aref + \theta * fsize * bref + \theta * \quad (34) \\
& itemcont * bdiff + \theta * itemcont * aref + \theta * \\
& itemcont * bref + \theta * bdiff * aref + \theta * bdiff * \\
& bref + \theta * aref * bref + fsize * itemcont * bdiff + \\
& fsize * itemcont * aref + fsize * itemcont * bref + fsize * \\
& bdiff * aref + fsize * bdiff * bref + fsize * aref * bref + \\
& itemcont * bdiff * aref + itemcont * bdiff * bref + \\
& itemcont * aref * bref + bdiff * aref * bref
\end{aligned}$$

The results of the logistic regression are given for each method in Table 9 and Table 10 where effect sizes that have a difference from 1.0 of at least 0.05 are bolded to show substantial effect on the Type I error rates. For the MH1, MH2, and RI methods, the factors with the largest effect on the Type I error rates are the amount of item contamination, the ability mean of the focal group, the size of the DIF, the item parameters, and their interactions. The sample size alone or any of its interaction did not have a large effect, while the effect of the number of items could not be determined as the

logistic regression was run separately for 20 and 40 items. For the DIPF method, the Type I error rate is also not controlled for any of the simulation conditions with both the mean and median Type I error rate much lower than that of the other methods.

Table 9

Logistic Regression Odds Ratios for 20 Items with Item Contamination

($N=2,304,000$)

	MH1	MH2	DIPF	RI
(Intercept)	0.0311	0.0366	0.0007	0.0356
theta	0.4176	0.6785	0.0164	0.4081
Fsize	1.0034	1.0026	1.0082	1.0146
Itemcont	5.2709	3.4812	38207.3147	747.5472
Bdiff	1.4841	1.2909	14.4996	4.7996
Aref	1.2796	1.0697	7.9766	1.1602
Bref	0.9233	0.8993	0.8472	1.1303
theta:fsize	1.0009	1.0022	0.9918	1.0128
theta:itemcont	20.6559	11.7563	61595.0175	857.2813
theta:bdiff	1.9490	1.4418	11.1009	2.7485
theta:aref	2.1472	1.1737	6.6246	2.6778
theta:bref	0.7132	0.7574	0.5100	0.9074
fsize:itemcont	0.9851	0.9879	0.9601	0.9548
fsize:bdiff	0.9966	0.9973	0.9941	0.9733
fsize:aref	0.9976	0.9986	0.9949	0.9894
fsize:bref	0.9997	1.0000	1.0027	0.9994
itemcont:bdiff	0.2497	0.3072	0.0000	0.0000
itemcont:aref	0.1289	0.2091	0.0002	0.0024
itemcont:bref	1.8461	1.9285	0.4566	2.0132
bdiff:aref	0.5789	0.6998	0.1047	0.1487
bdiff:bref	0.9768	0.9899	1.1463	0.7904
aref:bref	1.0177	1.0234	1.0634	0.7787
theta:fsize:itemcont	1.0027	1.0003	0.9913	1.0113
theta:fsize:bdiff	1.0012	1.0004	0.9949	0.9934
theta:fsize:aref	0.9952	0.9937	1.0032	0.9892
theta:fsize:bref	0.9997	0.9998	1.0029	0.9993

	MH1	MH2	DIPF	RI
theta:itemcont:bdiff	2.1197	1.1487	1.8704	4.9097
theta:itemcont:aref	0.0477	0.0990	0.0002	0.0002
theta:itemcont:bref	1.3429	1.2003	1.3866	1.1015
theta:bdiff:aref	0.5115	0.7095	0.1507	0.2267
theta:bdiff:bref	1.0339	1.0304	1.0738	0.9679
theta:aref:bref	1.6001	1.5862	1.2769	1.2350
fsize:itemcont:bdiff	1.0254	1.0231	1.0256	0.9783
fsize:itemcont:aref	1.0133	1.0109	1.0294	1.0508
fsize:itemcont:bref	0.9997	0.9994	1.0022	0.9977
fsize:bdiff:aref	1.0036	1.0026	1.0001	1.0126
fsize:bdiff:bref	1.0002	1.0001	1.0004	1.0004
fsize:aref:bref	1.0008	1.0007	0.9973	1.0011
itemcont:bdiff:aref	98.8441	51.4311	33085.4356	6295.9518
itemcont:bdiff:bref	0.7809	0.7805	0.4315	0.5966
itemcont:aref:bref	0.5366	0.5125	2.8950	0.6449
bdiff:aref:bref	0.9991	1.0107	0.9784	1.3157

Table 10

Logistic Regression Odds Ratios for 40 Items with Item Contamination

(*N=4,608,000*)

	MH1	MH2	DIPF	RI
(Intercept)	0.0560	0.0621	0.0011	0.1903
theta	0.9094	1.4947	0.2742	1.9065
fsize	1.0013	1.0001	1.0080	1.0014
itemcont	0.3887	0.2689	25.3490	1.5985
bdiff	0.6570	0.6505	1.4839	1.0625
aref	0.8463	0.7604	7.2991	0.2827
bref	1.0805	1.0973	0.8546	1.2441
theta:fsize	1.0028	1.0008	0.9906	1.0117
theta:itemcont	0.7203	0.1843	524.2618	1.7639
theta:bdiff	1.0407	0.7817	2.8540	2.9879
theta:aref	1.2395	0.8202	1.4150	0.4407
theta:bref	0.7439	0.8468	0.2776	0.8634

	MH1	MH2	DIPF	RI
fsize:itemcont	0.9935	0.9979	0.9712	1.0022
fsize:bdiff	0.9977	0.9985	0.9980	0.9847
fsize:aref	0.9997	1.0005	0.9897	1.0026
fsize:bref	0.9995	0.9997	1.0017	1.0001
itemcont:bdiff	9.2396	8.1888	3.0622	0.0078
itemcont:aref	0.6520	1.0106	0.1085	1.5022
itemcont:bref	0.7250	0.6775	1.4520	0.3755
bdiff:aref	0.9792	1.0251	0.7240	0.9024
bdiff:bref	0.9312	0.9235	1.1080	0.8808
aref:bref	0.8944	0.8839	1.0075	0.7938
theta:fsize:itemcont	1.0006	1.0026	1.0084	0.9821
theta:fsize:bdiff	0.9994	0.9999	0.9998	0.9868
theta:fsize:aref	0.9963	0.9956	0.9986	0.9957
theta:fsize:bref	0.9998	0.9998	1.0052	0.9997
theta:itemcont:bdiff	3.0913	3.1298	0.6027	0.0170
theta:itemcont:aref	0.9353	1.9344	0.0133	1.6707
theta:itemcont:bref	1.2907	1.0660	1.6605	1.0969
theta:bdiff:aref	0.8528	0.9559	0.4807	0.6242
theta:bdiff:bref	1.1371	1.0806	1.1980	1.0237
theta:aref:bref	1.4008	1.4166	1.1343	1.3403
fsize:itemcont:bdiff	1.0276	1.0235	1.0104	0.9815
fsize:itemcont:aref	1.0067	1.0038	1.0279	0.9955
fsize:itemcont:bref	1.0013	1.0008	1.0059	0.9994
fsize:bdiff:aref	1.0013	1.0010	1.0009	1.0006
fsize:bdiff:bref	1.0001	1.0000	1.0005	0.9998
fsize:aref:bref	1.0007	1.0007	0.9985	1.0001
itemcont:bdiff:aref	9.5482	6.8400	0.9236	3.2555
itemcont:bdiff:bref	0.8294	0.8938	0.4577	0.7566
itemcont:aref:bref	1.1385	1.1928	0.6079	2.6135
bdiff:aref:bref	1.0577	1.0571	1.0233	1.1261

Type I error rates are given for each combination of these factors in Table A5 in the Appendix. The median Type I error rate are shown for each simulation factor shown to have a large effect in the next paragraphs. Specifically, due to their large effects, the

size of the discrepancy between the focal group and reference group ability mean, the amount of item contamination, the size of the DIF, and the values of the item parameters are discussed in detail.

Item contamination and the size of the DIF. The first significant factor is the amount of item contamination (10%, 20%, 30%), and Figure 10 shows the Type I error rates for the four methods.

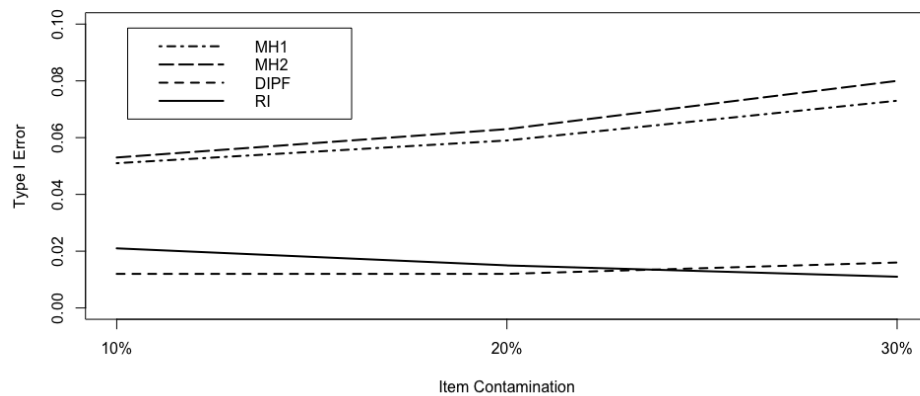


Figure 10. Median Type I error rates across item contamination. The Type I error rates for MH1, MH2, DIF, and RI methods are shown across the amount of item contamination (10%, 20%, and 30%).

For the MH1 and MH2 methods, increasing the amount of item contamination causes the Type I error rates to increase. However, the RI method shows a decrease in the Type I error rates as the amount of item contamination increases. This could be due to the fact that as more items exhibit DIF, the distribution of the percent differences becomes wider, making it less likely for an item to be identified as exhibiting DIF. When

there are 40 items, the effects for the three methods are approximately the same though in different directions (MH1 – 0.3887, MH2 – 0.2689, RI – 1.5985).

Along with the amount of item contamination, the size of the DIF must also be taken into consideration. Figure 11 shows the Type I error rates for the four methods across the size of the DIF (0.25, 0.50, 0.75, and 1.00).

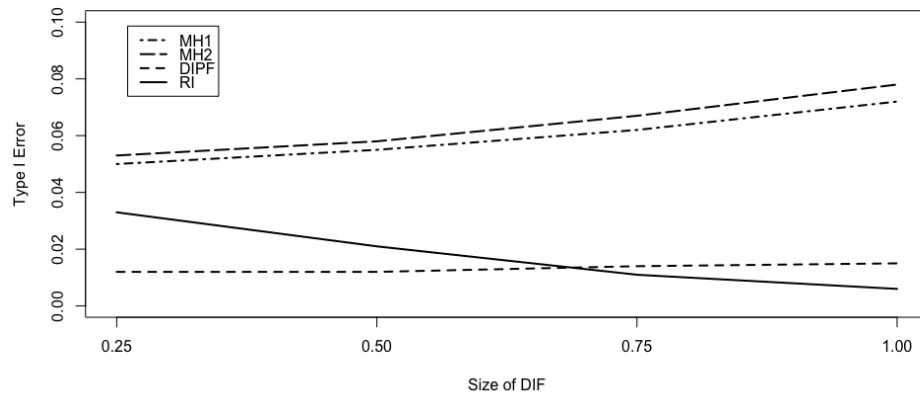


Figure 11. Median Type I error rates across the size of the DIF. The Type I error rates for MH1, MH2, DIPF, and RI methods are shown across the size of the DIF (0.25, 0.50, 0.75, 1.00).

For the size of the DIF, the same trend is seen in the Type I error rates for the methods as for the amount of item contamination with the effect sizes being 0.6750 for MH1, 0.6505 for MH2, and 1.0625 for RI. Figure 12 gives the interaction between the two factors. For the MH1 and MH2 methods, the Type I error rates increase across the size of the DIF along with an increase in the amount of item contamination. There appears to be no effect from the interaction between the amount of item contamination and the size of the DIF on the Type I error rates for the RI method, as it exhibits the same

downward trend across all amounts of item contamination even though the effect size is 0.0078.

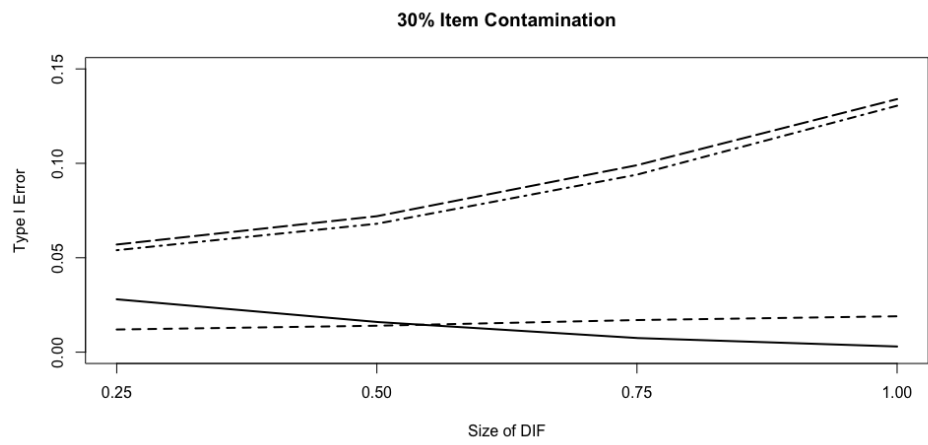
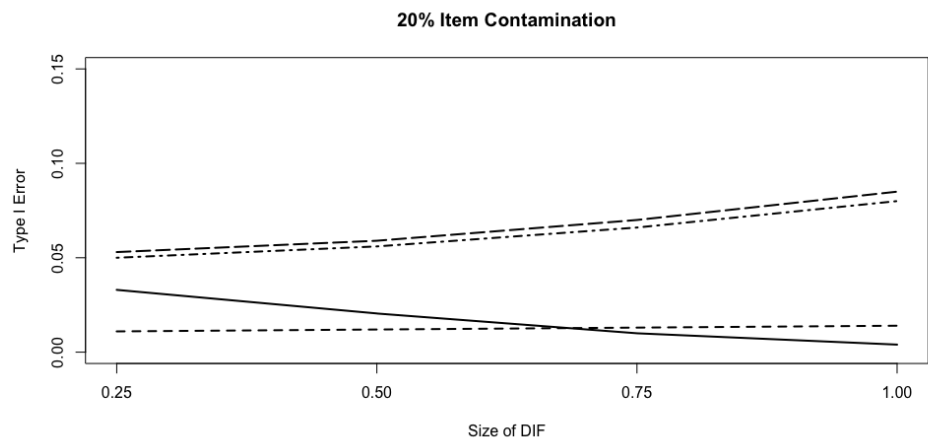
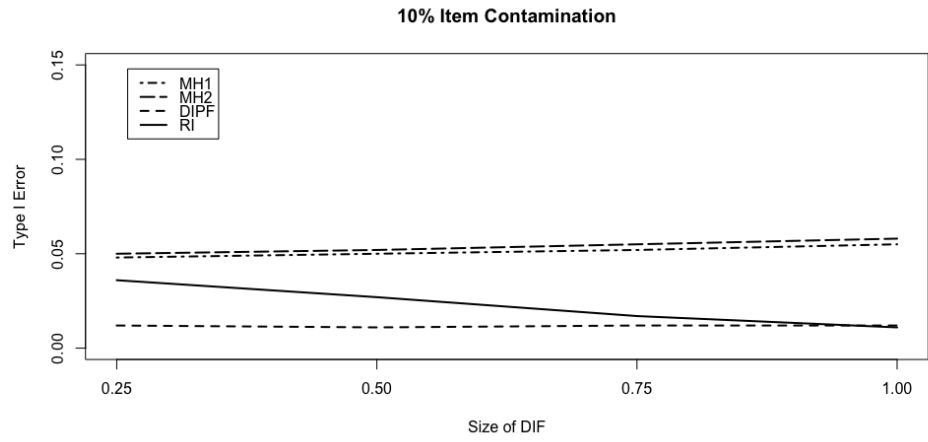


Figure 12. Median Type I error rates across the size of the DIF and amount of item contamination. The first graph shows the Type I error rates for the four methods across the size of the DIF when item contamination is 10%. The second graph shows the Type I error rates for the four methods across the size of the DIF when item contamination is 20%. The last graph shows the Type I error rates for the four methods across the size of the DIF when item contamination is 30%.

Interaction of amount of item contamination with other simulation factors.

Besides the size of the DIF, the amount of item contamination also has an interaction effect with the ability mean of the focal group and the item parameters. These effects are shown in Figure 13 and Figure 14.

For the interaction between the ability mean and the amount of item contamination, the Type I error rate for the MH1 method is consistent across all levels while the Type I error rate for the MH2 method increases as the ability mean decreases and the amount of item contamination increases. The Type I error rate for the RI method increases as the ability mean increases and the amount of item contamination decreases. This is reflected in the effect sizes of 0.7203 for the MH1 method, 0.1843 for the MH2 method, and 1.7639 for the RI method when there are 40 items.

For the interaction between the item parameters and the amount of item contamination, the Type I error rates for the MH1 and MH2 methods show an upward trend across the item difficulty parameter, but much more severely for high item discrimination parameters as the amount of item contamination increases. The Type I error rate for the RI method shows an increase as the amount of item contamination

increases but is extremely low for items with high discrimination and extreme difficulty parameters. The effect sizes for the interaction between the amount of item contamination and the item parameters are given in Table 11.

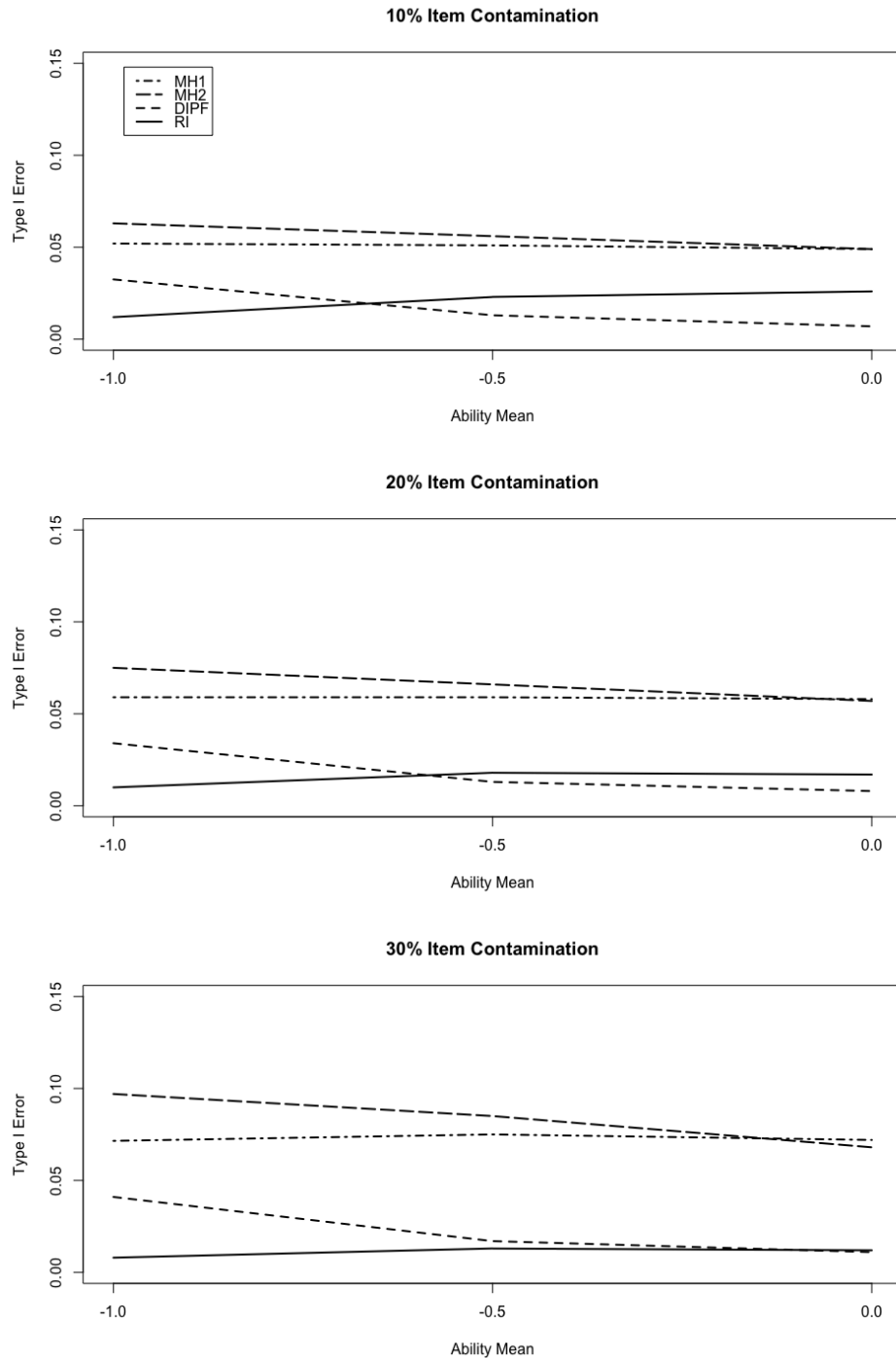
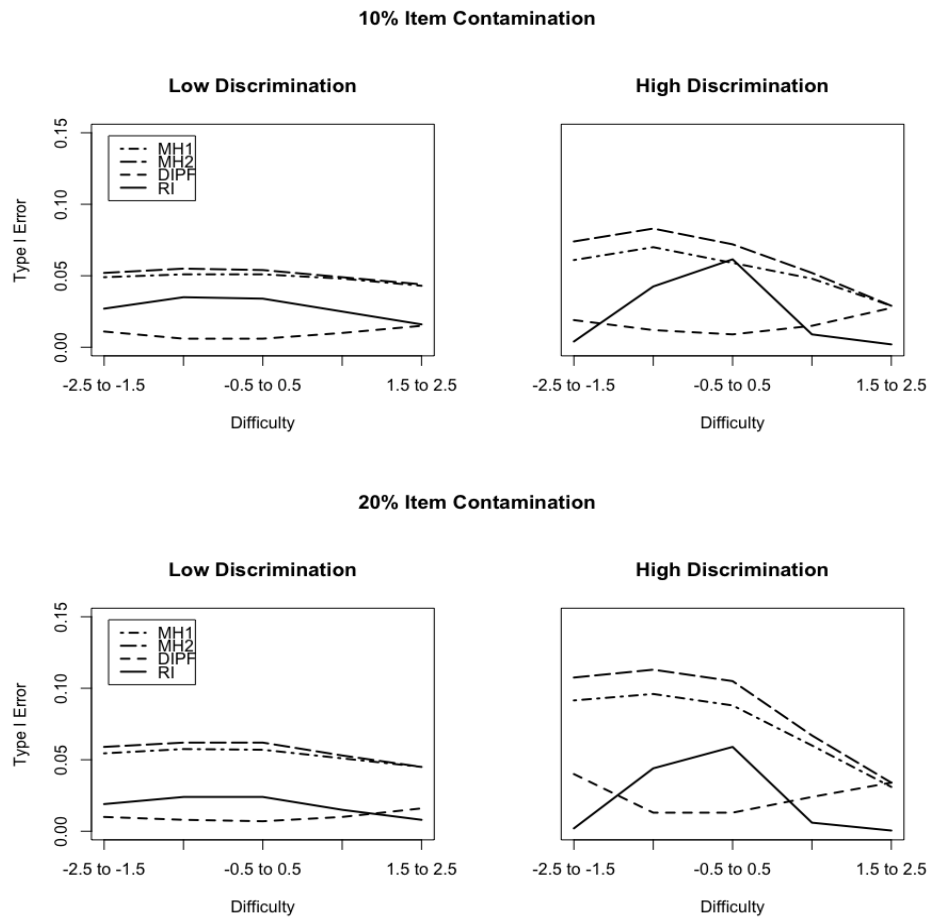


Figure 13. Median Type I error rates across the ability mean and amount of item contamination. The first graph shows the Type I error rates for the four methods across the ability mean when item contamination is 10%. The second graph shows the Type I error rates for the four methods across the ability mean when item contamination is 20%. The last graph shows the Type I error rates for the four methods across the ability mean when item contamination is 30%.



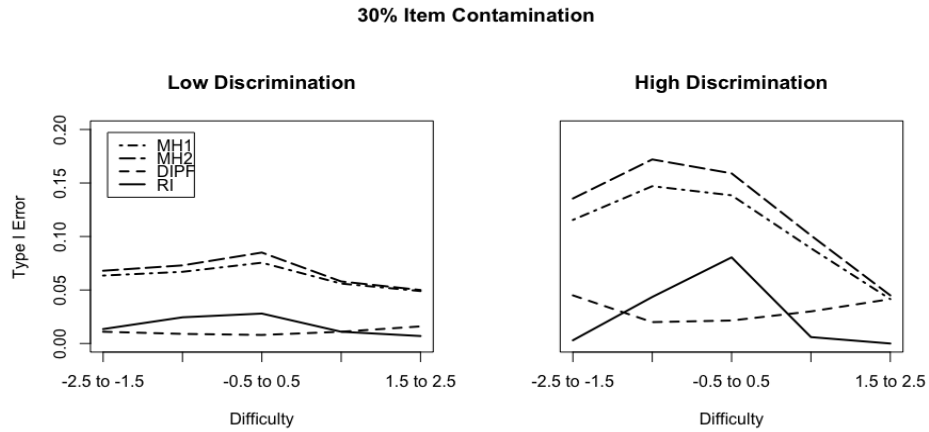


Figure 14. Median Type I error rates across the item parameters and amount of item contamination. The first graph shows the Type I error rates for the four methods across the item parameters when item contamination is 10%. The second graph shows the Type I error rates for the four methods across the item parameters when item contamination is 20%. The last graph shows the Type I error rates for the four methods across the item parameters when item contamination is 30%.

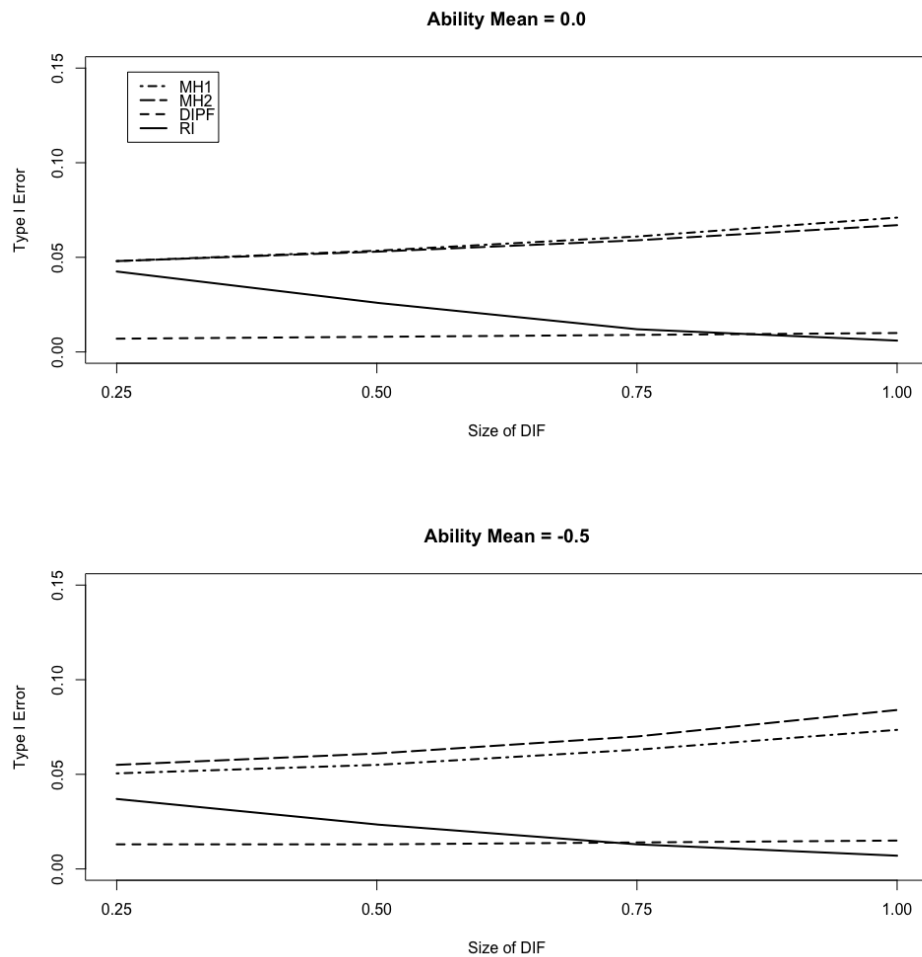
Table 11

Odds Ratios for Interaction between Item Contamination and Item Parameters

	MH1	MH2	RI
Item Contamination * Discrimination	0.6520	1.0106	1.5022
Item Contamination * Difficulty	0.7250	0.6775	0.3755
Item Contamination * Diff * Disc	1.1385	1.1928	2.6135

Interaction of the size of the DIF with other factors. The size of the DIF has been shown to have an interaction effect with the ability mean for the focal group and the

item parameters on the Type I error rates for the MH1, MH2, and RI methods. The interaction between the size of the DIF and the ability mean is shown in Figure 15, the interaction between the size of the DIF and the item difficulty parameter is shown in Figure 16. And the interaction between the size of the DIF and the item discrimination parameter is shown in Figure 17.



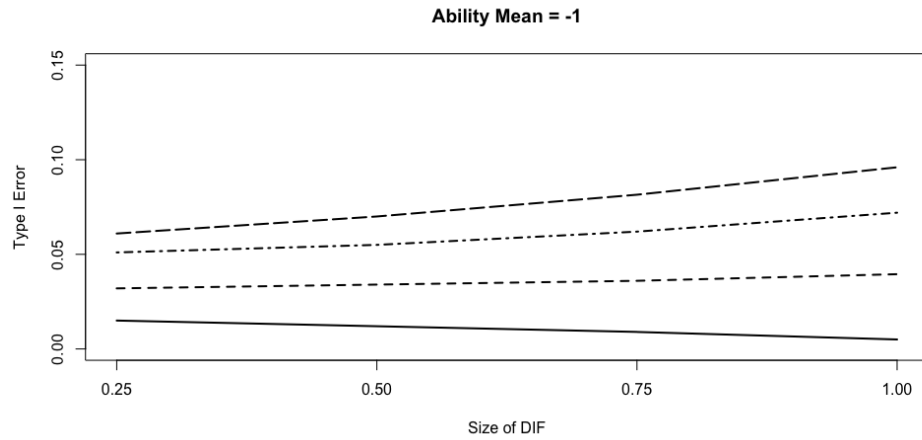
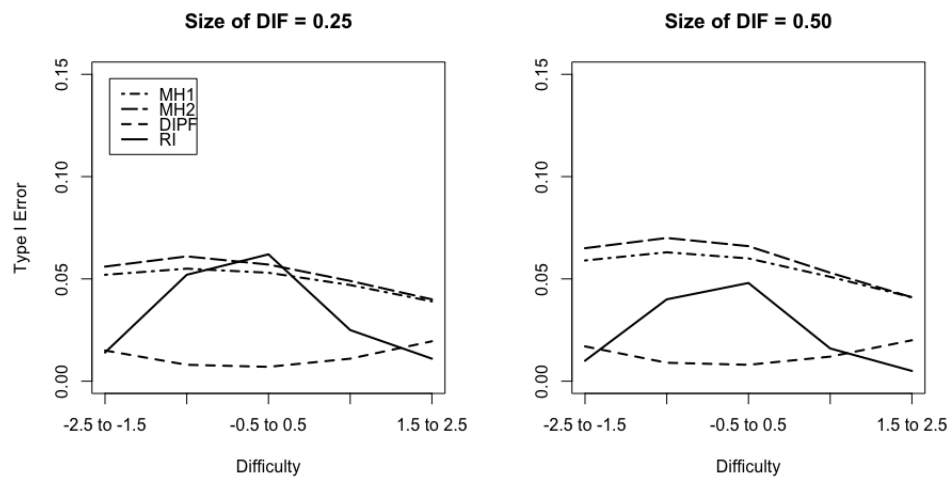


Figure 15. Median Type I error rates across the size of the DIF and the ability mean. The first graph shows the Type I error rates for the four methods across the item parameters when item contamination is 10%. The second graph shows the Type I error rates for the four methods across the item parameters when item contamination is 20%. The last graph shows the Type I error rates for the four methods across the item parameters when item contamination is 30%.



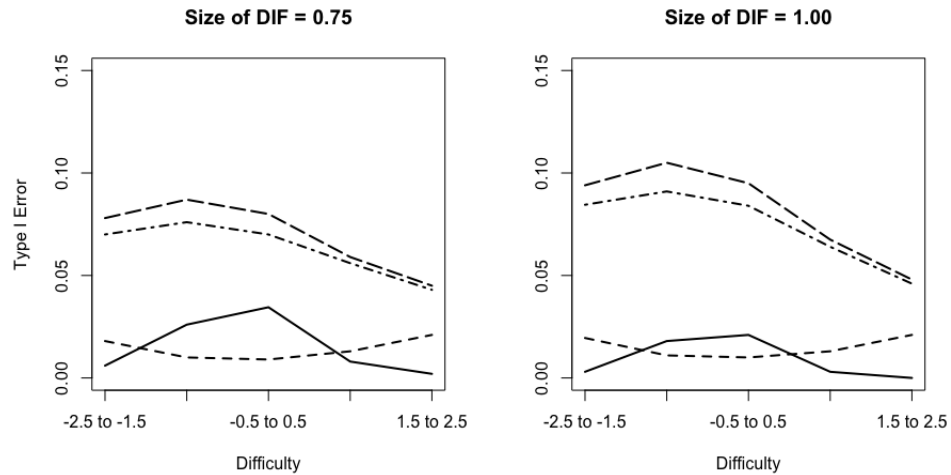


Figure 16. Median Type I error rates across the size of the DIF and the item difficulty parameter. Four graphs show the Type I error rates for the four methods across the item difficulty parameter separated by the size of the DIF.

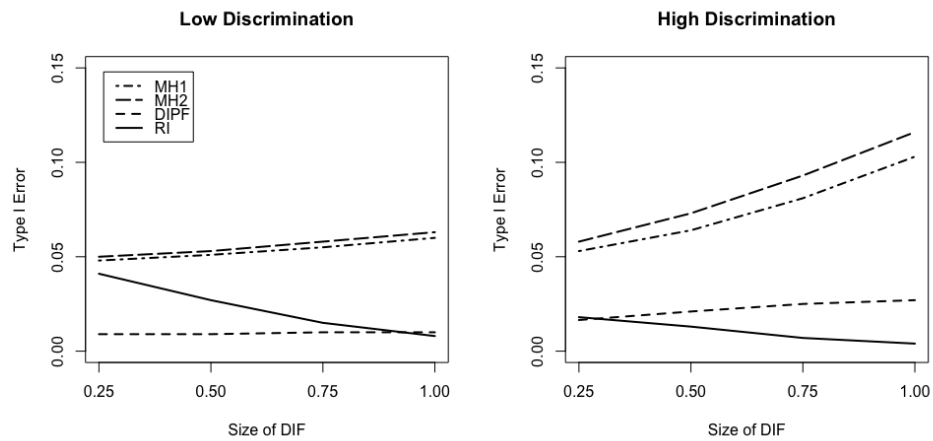


Figure 17. Median Type I error rates across the size of the DIF and the item discrimination parameter. The left graph shows the Type I error rates for the four methods across the size of the DIF when item discrimination is low while the right graph shows the Type I error rates across the size of the DIF when item discrimination is high.

For the interaction between the size of the DIF and the ability mean for the focal group, the Type I error rates for the MH1 and MH2 methods increase across the size of the DIF along with a decrease in the ability mean. There appears to be no effect from the interaction between the amount of item contamination and the size of the DIF on the Type I error rates for the RI method, as it exhibits the same downward trend across all amounts of item contamination even though the effect size for 40 items is 2.9879.

For the interaction between the size of the DIF and the item parameters, the Type I error rates for the MH1 and MH2 methods increase as the size of the DIF increases, but decrease as the item difficulty parameter increases. The Type I error rates for the RI method decrease as the size of the DIF increases, but is much lower for extreme difficulty parameters. The effect for the RI method (.8808) is stronger than the effects for the MH1 and MH2 methods (.9312 and .9235). For the item discrimination parameter, the Type I error rates for the MH1 and MH2 methods increase across the size of the DIF but is much larger when the item discrimination is high, while the Type I error rates for the RI method decrease across the size of the DIF and is much smaller when the item discrimination is high. The effect for the RI method (odds ratio of .9024) is stronger than the effects for the MH1 and MH2 methods (odds ratios of .9792 and 1.0251).

Interaction between the ability mean and the item parameters. For all three methods, the effect of the two-way and three-way interactions between the ability mean and item parameters were shown to have an effect on the Type I error rates in the logistic regression. The first two-way interaction is between the ability mean and the difficulty parameter as shown in Figure 18. For the MH methods, the Type I error rates decrease for the extreme difficulty parameters as the ability mean decreases. For the RI method, the

up-and-down pattern shifts to the left as the ability mean decreases. However, all the effect sizes are approximately the same for all three methods (MH1 – 0.7439, MH2 – 0.8468, RI – 0.8634).

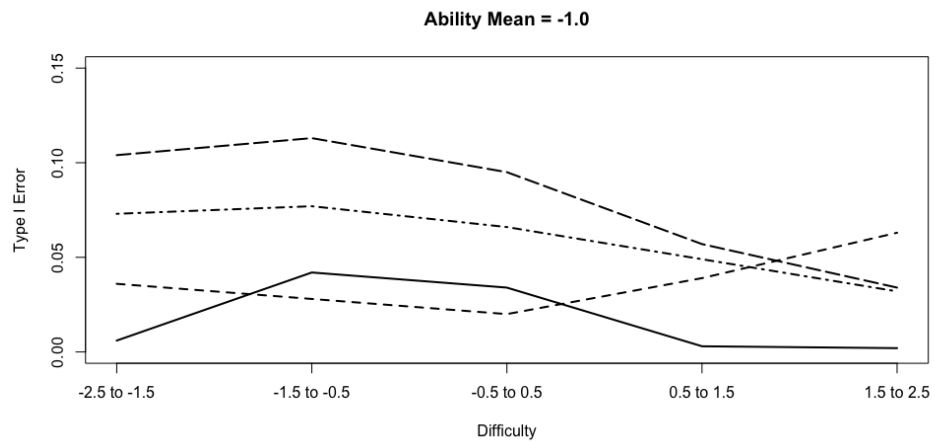
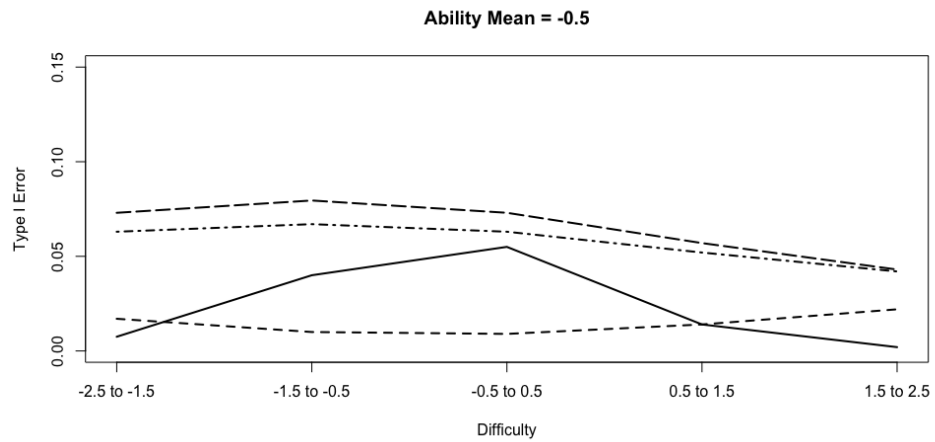
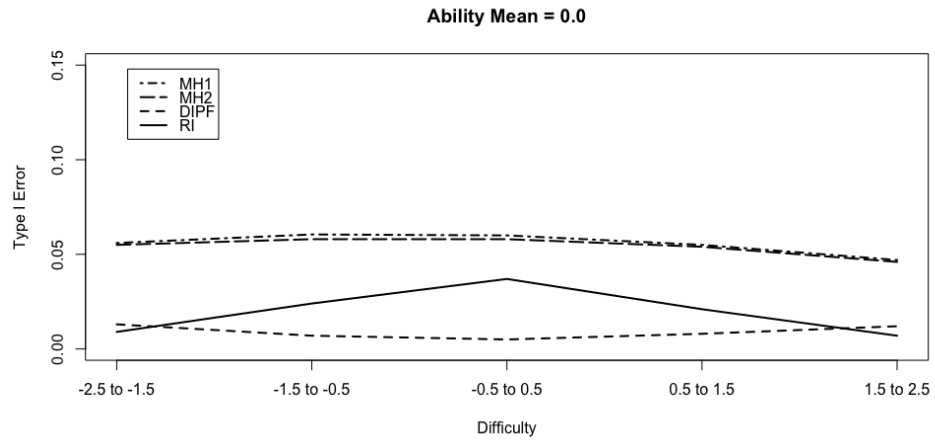
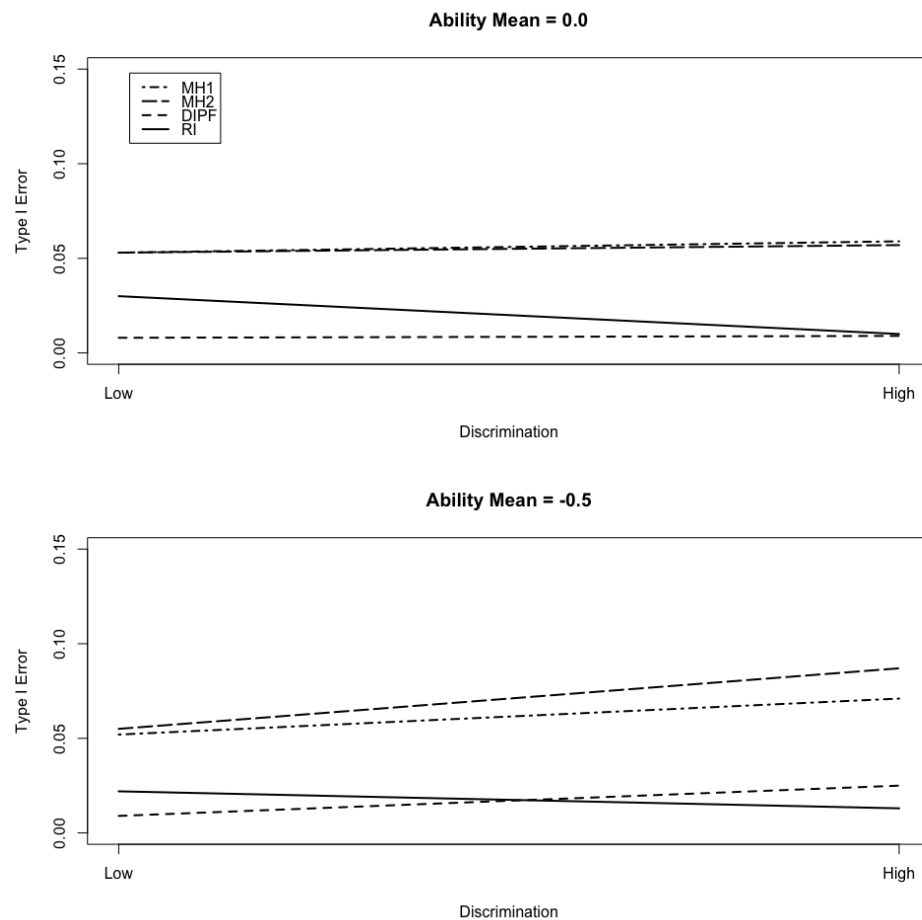


Figure 18. Median Type I error rates across ability mean and item difficulty. The first graph shows the Type I error rates for the four methods across the item difficulty parameters when the ability mean is equal to 0, the second graph shows the Type I error rates for the four methods across the item difficulty parameters when the ability mean is equal to -0.5 and the third graph shows the Type I error rates for the four methods across the item difficulty parameters when the ability mean is equal to -1.0.

The second two-way interaction is between the ability mean and the discrimination parameter, shown in Figure 19.



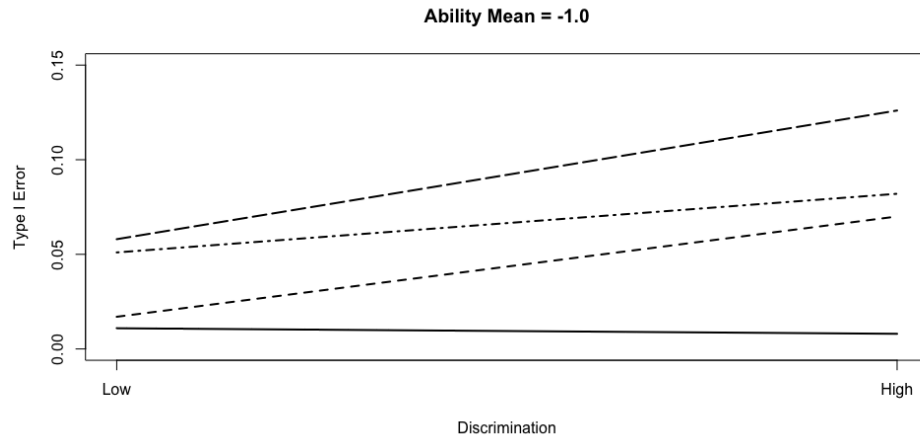


Figure 19. Median Type I error rates across ability mean and item discrimination.

The first graph shows the Type I error rates for the four methods across the item discrimination parameters when the ability mean is equal to 0, the second graph shows the Type I error rates for the four methods across the item discrimination parameters when the ability mean is equal to -0.5 and the third graph shows the Type I error rates for the four methods across the item discrimination parameters when the ability mean is equal to -1.0.

For the MH1 method, the Type I error rates stay consistently around 0.05 regardless of the item discrimination or the ability mean. For the MH2 method, the Type I error rates increase across the discrimination parameters as the ability mean also increases. The RI method shows the opposite pattern from that of the MH2 method with a downward trend. The effect on the Type I error rates for the RI method (odds ratio of 0.4407) is stronger than on the MH1 and MH2 methods (odds ratios of 1.2395 and 0.8202).

The three-way interaction between the ability mean and the item parameters are shown in Figure 20.

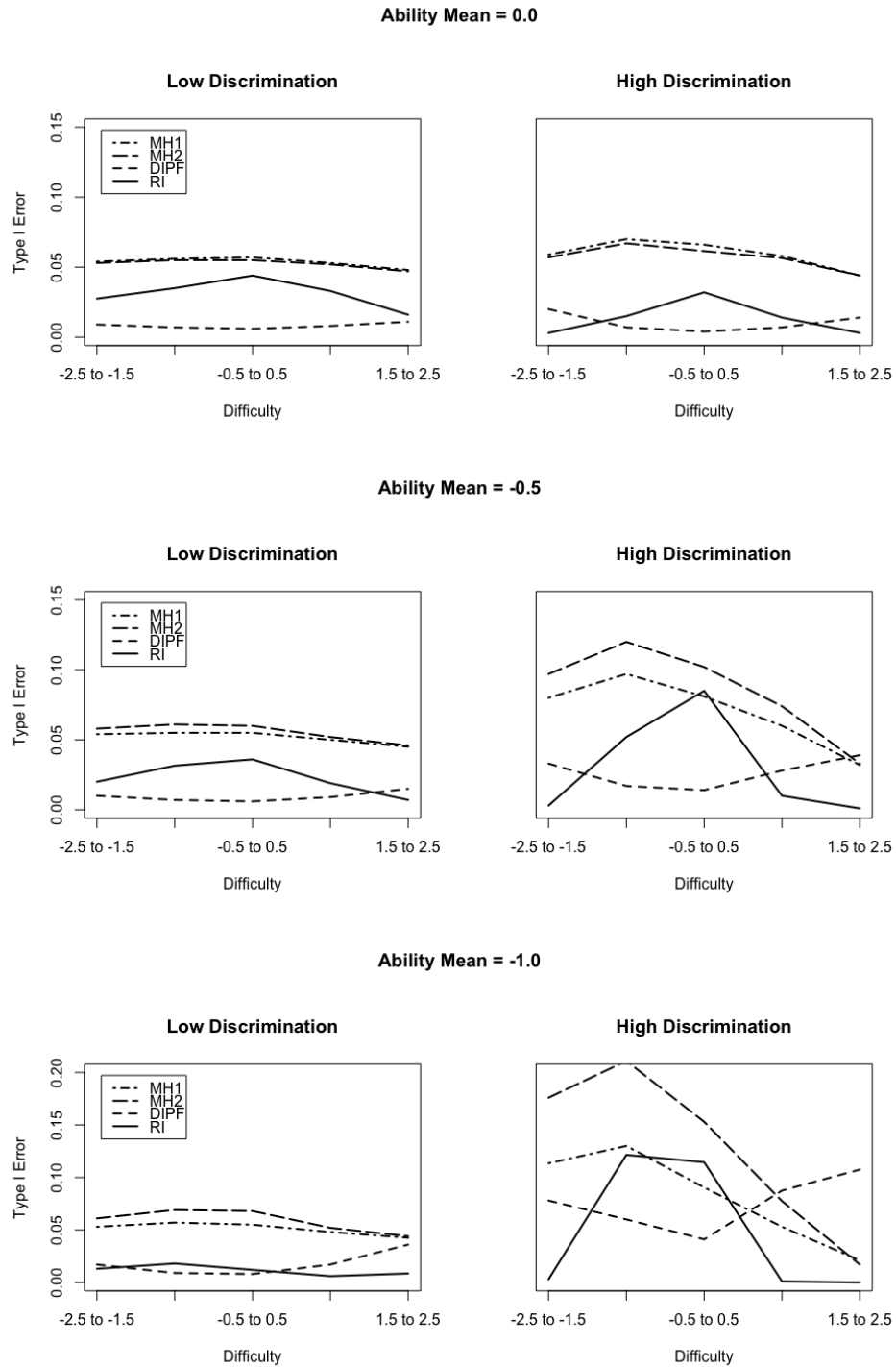


Figure 20. Median Type I error rates across ability mean and item parameters.

The first graph shows the Type I error rates for the four methods across the item parameters when the ability mean is equal to 0, the second graph shows the Type I

error rates for the four methods across the item parameters when the ability mean is equal to -0.5 and the third graph shows the Type I error rates for the four methods across the item parameters when the ability mean is equal to -1.0.

What is interesting to note is that the MH1 and MH2 methods are consistent across the item parameters and ability mean except when the ability mean is equal to -0.5 or -1 and the discrimination parameter is high. For the RI method, the same up-and-down pattern is seen across all combinations. The effect on the Type I error rates is similar across all three methods (MH1 – 1.4008, MH2 - 1.4166, RI – 1.3403).

Summary. For all three methods, the main factors that have an effect on the Type I error rates are the amount of item contamination, the size of the DIF, the ability mean for the focal group, and the item parameters. The sample size and the number of items were found not to have an effect on the Type I error rates for all methods. Overall, the Type I error rates with item contamination for the MH1 and MH2 methods are shown to increase as the amount of item contamination and the size of the DIF increases. Also, an increase in the amount of item contamination and a decrease in the ability mean leads to an increase in the Type I error rates for the MH1 and MH2 methods. For the RI method, the Type I error rate is shown to decrease as the amount of item contamination and the size of the DIF decreases, especially with high discrimination and extreme difficulty parameters. A summary of how the Type I error rates increased due to the simulation factors is given in Table 12 for the MH1, MH2, and RI methods. The DIPF method is not included in the table as the Type I error rate was not controlled for in any of the simulation conditions. This could be due to the fact that the determination of an item

being flagged as exhibiting DIF is based on an arbitrary cutoff of the number of paired items being statistically different from 0.

Table 12

Effect of Simulation Factors on Type I Error Rates

Interaction	MH1	MH2	RI
% item contamination * size of DIF	item contamination ↑ size of DIF ↑		item contamination ↓ size of DIF ↓
% item contamination * ability mean	no effect	ability mean ↓ item contamination ↑	ability mean ↑ item contamination ↓
% item contamination * item parameters	item difficulty ↗ item discrimination ↑ item contamination ↑		item difficulty: moderate item discrimination ↑ item contamination ↑
size of DIF * ability mean	size of DIF ↑ ability mean ↓		no effect
size of DIF * item parameters	size of DIF ↑ item difficulty ↓ item discrimination ↑		size of DIF ↓ item difficulty: moderate item discrimination ↑

Analysis of Effect Size for the RI Method

To determine why the relative item performance is not as effective as the MH methods in terms of Type I error rates, descriptive statistics of the effect size are given in

-Table 16 for items simulated without DIF and items simulated with DIF.

Table 13

Descriptive Statistics of Effect Size for Items with No DIF

	Overall	20 Items	40 Items	80 Items
Minimum	-5.0170	-3.7130	-4.3390	-5.0170
2.5 Percentile	-1.5465	-1.5612	-1.4965	-1.5678
1st Quartile	-0.3891	-0.3964	-0.3732	-0.3951
Median	0.2016	0.1931	0.2138	0.1977
Mean	0.2052	0.1958	0.2150	0.2027
3rd Quartile	0.7954	0.7909	0.7961	0.7961
97.5 Percentile	1.9713	1.9376	1.9520	1.9899
Maximum	5.0290	3.5880	4.3640	5.0290

Table 14

Descriptive Statistics of Effect Size for Items with No DIF and No Item Contamination

	Overall	20 Items	40 Items	80 Items
Minimum	-4.9510	-3.6240	-4.2560	-4.9510
2.5 Percentile	-1.8719	-1.8681	-1.8373	-1.8909
1st Quartile	-0.6845	-0.6779	-0.6935	-0.6817
Median	-0.0329	-0.0248	-0.0342	-0.0342
Mean	0.0000	0.0000	0.0000	0.0000
3rd Quartile	0.6675	0.6710	0.6693	0.6657
97.5 Percentile	1.9828	1.9187	1.9906	1.9964
Maximum	5.0290	3.4100	4.1760	5.0290

Table 15

Descriptive Statistics of Effect Size for Items with No DIF and Item Contamination

	Overall	20 Items	40 Items	80 Items
Minimum	-5.0170	-3.7130	-4.3390	-5.0170
2.5 Percentile	-1.4972	-1.5160	-1.4408	-1.5201
1st Quartile	-0.3579	-0.3671	-0.3397	-0.3647
Median	0.2220	0.2120	0.2348	0.2180
Mean	0.2266	0.2162	0.2374	0.2239
3rd Quartile	0.8061	0.8012	0.8064	0.8072
97.5 Percentile	1.9702	1.9393	1.9482	1.9892
Maximum	4.8860	3.5880	4.3640	4.8860

Table 16

Descriptive Statistics of Effect Size for Items with DIF

	Overall	20 Items	40 Items	80 Items
Minimum	-5.5060	-4.0190	-4.7770	-5.5060
2.5 Percentile	-2.6135	-2.5151	-2.7006	-2.5949
1st Quartile	-1.4790	-1.4430	-1.5140	-1.4690
Median	-0.9388	-0.8957	-0.9795	-0.9294
Mean	-0.9065	-0.8648	-0.9494	-0.8955
3rd Quartile	-0.3591	-0.3187	-0.3991	-0.3496
97.5 Percentile	0.9431	0.9551	0.8899	0.9650
Maximum	4.5150	3.2530	3.9210	4.5150

What is interesting to note about these tables is that the left side of the effect size empirical distribution for items simulated to not have DIF depends on whether or not item contamination is present. If the effect size statistic is modified to use an empirical distribution instead of the normal distribution when determining the critical values, the critical value for the left side of the distribution would be -1.9 without item contamination (Table 14) and -1.5 with item contamination (Table 15). For the right side

of the distribution, the critical value is equal to 1.98 regardless of whether or not item contamination is present. Figure 21 illustrates the difference between the empirical distribution and the standard normal distribution when looking at the critical values on the left side of the distribution.

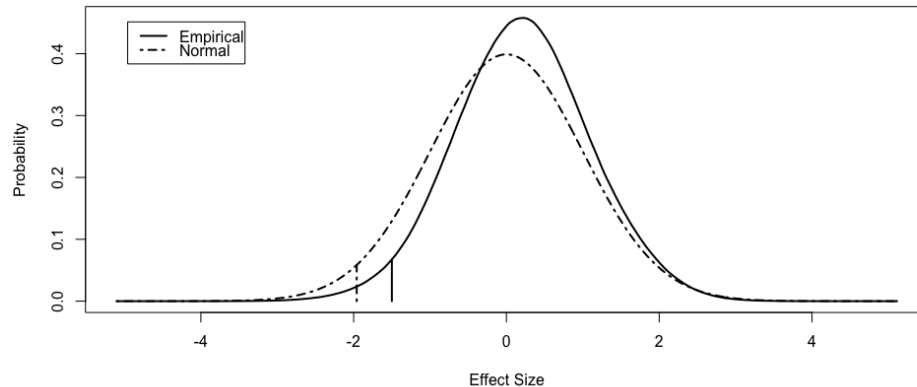


Figure 21. Comparison of Empirical and Standard Normal Distributions. The empirical distribution of the effect size is compared with the standard normal distribution.

If these critical values were used to evaluate items with DIF, the Type I error rate could be near 0.05. However, the potential power for this method would only be 26% given that the -1.5 value is the first quartile of the empirical distribution and 1.98 is larger than the 97.5 percentile of the empirical distribution.

Power

As the Type I error rate needs to be controlled to allow for an accurate measure of statistical power and this does not occur frequently for the relative item performance method, this section focuses on the Mantel-Haenszel methods instead. The criterion used

to determine if the Type I error rates were controlled was whether or not the Type I error rate was between 0.036 and 0.064, using a margin of error calculated from the square root of 0.05 times 1 – 0.05 divided by 1000 replications. For the RI method, Type I error rates were not consistently controlled across the various simulation factors. For example, the Type I error rate was controlled when the size of the DIF was equal to 0.25, except when the difficulty parameters were high, and the discrimination parameter was low. However, when the ability mean was equal to -1, none of the Type I error rates were controlled.

Power rates are given for each combination of conditions in Table A6 in the Appendix. Descriptive statistics for power over all conditions is given in Table 17.

Table 17

Summary Statistics for Power Across All Cells (N=2,640)

	MH1	MH2
Minimum	0.000	0.000
1st Quartile	0.090	0.077
Median	0.203	0.181
Mean	0.314	0.293
3rd Quartile	0.478	0.436
Maximum	1.000	1.000

Overall, MH1 has better power than MH2 due to the larger number of intervals. Using the median, power is .203 for MH1 and .181 for MH2. To determine what is causing the difference in power for these two methods, logistic regression is conducted with the outcome variable being whether or not the item is flagged appropriately as exhibiting DIF. The independent variables are the main effects, two-way interaction effects, and the three-interaction effects of the following factors: the number of items

(n_i), the ability mean of the focal group (θ), the sample size for the focal group ($fsize$), the amount of item contamination ($itemcont$), the size of the DIF ($bdiff$), the item difficulty parameter ($bref$), and the item discrimination parameter ($aref$). Note that any interaction between the number of items and the item parameters are excluded because the number of items within a specific interval for the difficulty parameter is determined by the total number of items. The equation used for the logistic regression is given as follows:

$$\begin{aligned}
Y_i = & n_i + \theta + fsize + aref + bref + itemcont + bdiff + \\
& n_i * \theta + n_i * fsize + n_i * itemcont + n_i * bdiff + \\
& \theta * fsize + \theta * itemcont + \theta * bdiff + \\
& \theta * aref + \theta * bref + fsize * itemcont + \\
& fsize * bdiff + fsize * aref + fsize * bref + \\
& itemcont * bdiff + itemcont * aref + itemcont * bdiff \\
& + bdiff * aref + bdiff * bref + aref * bref + \\
& n_i * \theta * fsize + n_i * \theta * itemcont + n_i * \theta \\
& * bdiff + n_i * fsize * itemcont + n_i * fsize * bdiff + \\
& n_i * itemcont * bdiff + \theta * fsize * itemcont + \theta \\
& * fsize * bdiff + \theta * fsize * aref + \theta * fsize \\
& * bref + \theta * itemcont * bdiff + \theta * itemcont * \\
& aref + \theta * itemcont * bref + \theta * bdiff * aref \\
& + \theta * bdiff * bref + \theta * aref * bref + fsize * \\
& itemcont * bdiff + fsize * itemcont * aref + fsize * \\
& itemcont * bref + fsize * bdiff * aref + fsize * bdiff *
\end{aligned} \tag{35}$$

$$\begin{aligned}
& bref + fsize * aref * bref + itemcont * bdiff * aref + \\
& itemcont * bdiff * bref + itemcont * aref * bref + \\
& bdiff * aref * bref
\end{aligned}$$

The results of the logistic regression are given for each method in Table 18 where the effect size is equal to the odds ratio with a value above 1 indicating that the odds of the item being appropriately flagged for exhibiting DIF is greater than the odds of the item being incorrectly flagged as not exhibiting DIF. Odds ratio where the difference from 1.0 is greater than 0.05 are bolded in the table to show substantial effects on the Type I error rates.

Table 18

Logistic Regression Odds Ratios for Power (N=4,032,000)

	MH1	MH2
(Intercept)	0.0571	0.0611
ni	0.9942	0.9949
theta	1.0536	1.3919
fsize	0.9982*	0.9981*
itemcont	0.1742	0.1871
bdiff	2.1360	2.0567
aref	1.0178	0.9251
bref	0.5117	0.5591
ni:theta	0.9976	0.9993
ni:fsize	1.0000	1.0000
ni:itemcont	1.0158	1.0141
ni:bdiff	1.0049	1.0037
theta:fsize	0.9934	0.9923
theta:itemcont	4.0453	3.2510
theta:bdiff	0.6225	0.4933
theta:aref	1.1847	1.1064

	MH1	MH2
theta:bref	0.5290	0.5647
fsize:itemcont	1.0121	1.0105
fsize:bdiff	1.0046	1.0045
fsize:aref	1.0001	1.0004
fsize:bref	1.0008	1.0005
itemcont:bdiff	3.5680	3.5314
itemcont:aref	1.6920	1.7819
itemcont:bref	6.9460	6.1521
bdiff:aref	3.2586	3.5081
bdiff:bref	1.4656	1.3479
aref:bref	1.7239	1.6412
ni:theta:fsize	1.0000	1.0000
ni:theta:itemcont	0.9842	0.9841
ni:theta:bdiff	1.0044	1.0033
ni:fsize:itemcont	0.9999	0.9999
ni:fsize:bdiff	1.0000	1.0000
ni:itemcont:bdiff	0.9843	0.9867
theta:fsize:itemcont	0.9971	0.9970
theta:fsize:bdiff	1.0033	1.0045
theta:fsize:aref	1.0064	1.0085
theta:fsize:bref	1.0009	1.0007
theta:itemcont:bdiff	0.7011	0.9704
theta:itemcont:aref	0.7564	0.8517
theta:itemcont:bref	0.6658	0.6939
theta:bdiff:aref	0.9499	1.0689
theta:bdiff:bref	1.8189	1.7453
theta:aref:bref	1.7546	1.6774
fsize:itemcont:bdiff	0.9781	0.9812
fsize:itemcont:aref	0.9910	0.9918
fsize:itemcont:bref	0.9974	0.9977
fsize:bdiff:aref	1.0227	1.0222
fsize:bdiff:bref	1.0004	1.0006
fsize:aref:bref	0.9995	0.9995
itemcont:bdiff:aref	0.1143	0.1226
itemcont:bdiff:bref	0.4462	0.4988

	MH1	MH2
itemcont:aref:bref	0.1937	0.2061
bdiff:aref:bref	1.4452	1.4819

Note. fsize has an asterisk for the odds ratio due to the fact that the odds ratios are calculated assuming only one unit change (one person added to the sample size)

For the MH1 and MH2 methods, the factors with the largest effect on power are the amount of item contamination, the ability mean of the focal group, the size of the DIF, the item parameters, and their interactions. The sample size does not have a large effect (0.9982 for MH1 and 0.9981 for MH2).

Sample size. Although sample size (at least those levels used in the simulation research here) does not have a large effect on power for the MH1 and MH2 methods, it is still worth evaluating what the minimum sample size should be for future research. For the sample sizes of 25, 50, 100, and 200, the power for MH1 and MH2 are given in

Table 19. For the odds ratio, the MH1 and MH2 being .9982 and .9981 is for every additional person added to the focal group sample size. However, if 25 people were added to the focal group sample size, then the odds ratios would be .9559 and .9536, respectively. For 50 people, the odds ratio would be .9139 and .9093. Thus, at least 50 people must be added to the focal group sample size for the odds ratio to have a large effect on the Type I error rates.

Table 19

Power for MH1 and MH2 Methods Across Sample Size

	Focal Group Sample Size			
	25	50	100	200
MH1	0.0910	0.1705	0.3015	0.4800
MH2	0.0810	0.1490	0.2570	0.3965

Not unexpectedly, for both the MH1 and MH2 methods, the power increases as the sample size increase. With the smallest sample size having approximately 10% in power, these methods may not be effective for small sample sizes.

Mean of the focal ability distribution. For the different means of the focal group’s ability distribution (0, -0.5, -1), the median power is shown for each method in Figure 22. The power for MH1 and MH2 methods show a slight upward trend as the ability mean increases with effect sizes of 1.0536 and 1.3919, respectively.

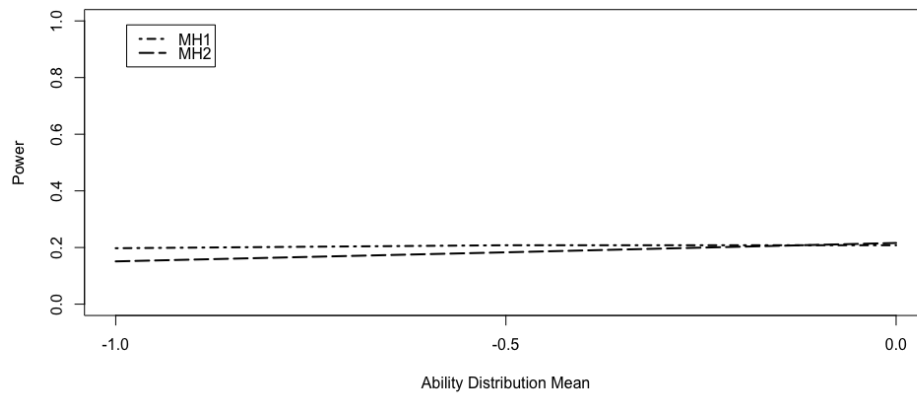


Figure 22. Median power rates across focal ability distribution. The power for MH1 and MH2 methods are shown across the focal ability means (-1, -0.5, and 0).

Item contamination and size of the DIF. In Figure 23, the power for each method is broken down by the amount of item contamination (10%, 20%, and 30%).

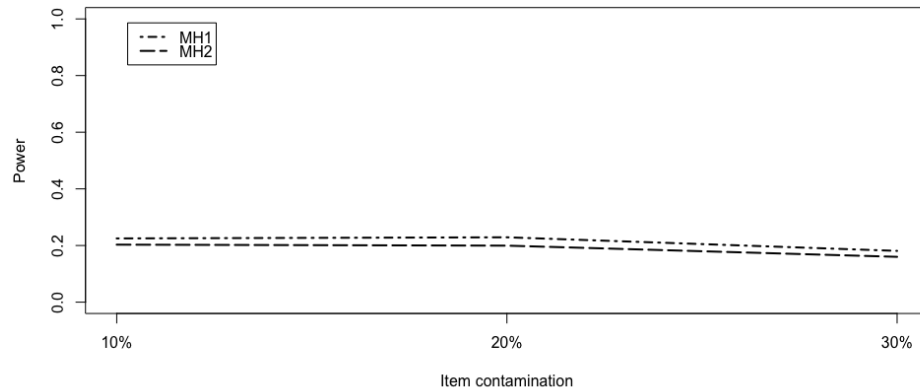


Figure 23. Median power across item contamination. The power for MH1 and MH2 methods are shown across the amount of item contamination (10%, 20%, and 30%).

The MH1 and MH2 methods show a small downward trend as the amount of item contamination increases and this is due to the fact that the MH methods only look at the individual items and not the item distribution. However, their effect sizes are large (odds ratios of 0.1742 and 0.1871).

The interaction between item contamination and the size of DIF has an effect on the methods' power (odds ratios of 3.5680 and 3.5314). This interaction is shown in Figure 24 below. As the size of DIF increases, the power for the Mantel-Haenszel methods increase, but decreases as the amount of item contamination increases.

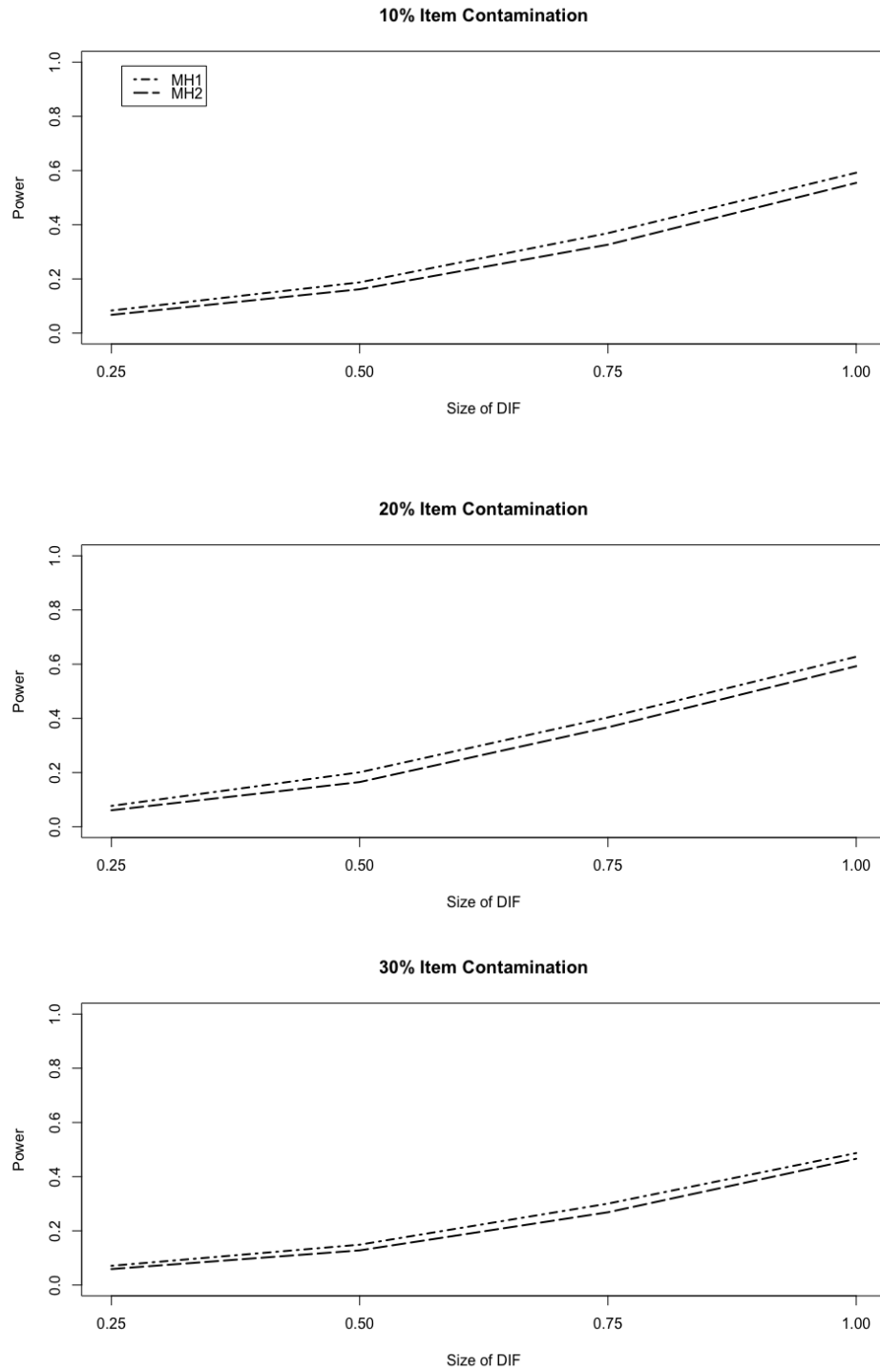


Figure 24. Median power rates across item contamination and size of DIF. The power for MH1 and MH2 methods are shown across the amount of DIF (0.25,

0.50, 0.75, 1.00) for each of the percentages of item contamination (10%, 20%, 30%).

Item parameters. There are two item parameters that affect power: the difficulty parameter (odds ratios of 0.5117 and 0.5591) and the discrimination parameter (odds ratios of 1.0178 and 0.9251). First, in Figure 25, the effect of the difficulty parameter on power is shown.

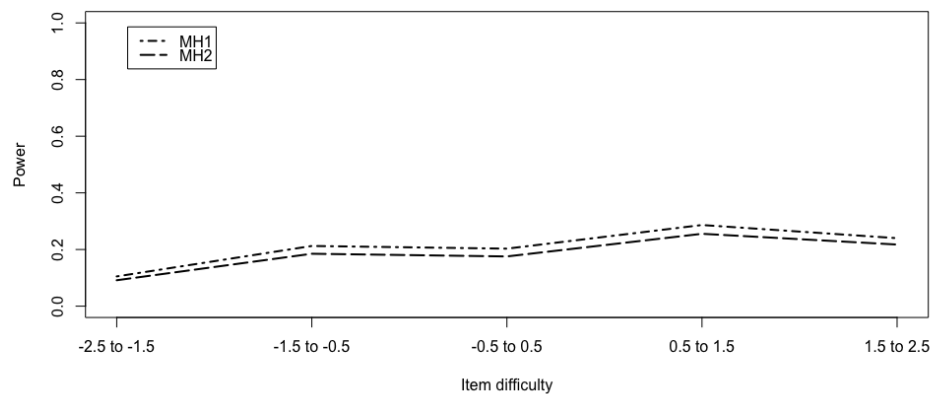


Figure 25. Median power across difficulty parameters. The power for MH1 and MH2 methods are shown across the item difficulty parameters ranging from -2.5 to 2.5.

Before discussing the results of the difficulty parameter, the power at different levels of the discrimination parameter need to be reviewed with the power shown in

Table 20.

Table 20

Median Power Across Discrimination Parameters

Methods	Discrimination	
	Low	High
MH1	0.1400	0.3145
MH2	0.1290	0.2715

The MH1 method has higher power than the MH2 method across the discrimination parameters. As the item parameters work together to determine the probability of obtaining a correct response on an item, the interaction between the two could have an impact on power. This is shown in Figure 26.

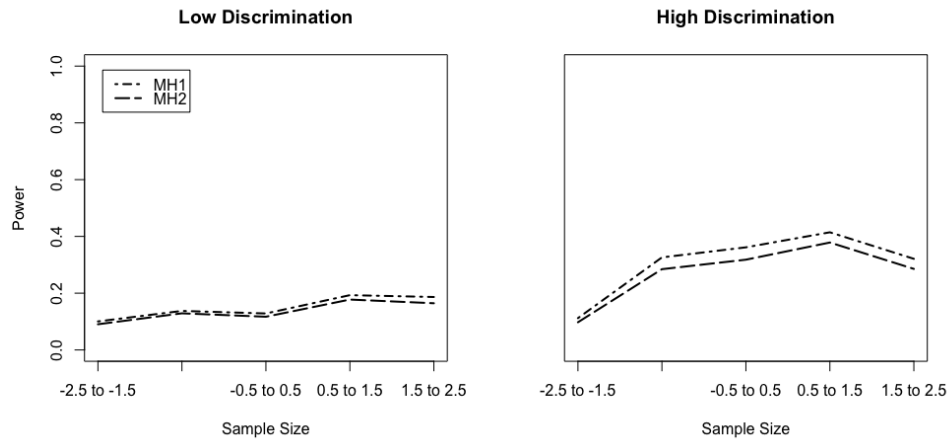


Figure 26. Median power across item parameters. The left graph shows the power for the two methods across the item difficulty parameters when the discrimination parameter is low, while the right graph shows the power for the two methods across the item difficulty parameters when the discrimination parameter is high.

When looking at both the item difficulty and the item discrimination parameters, it is clear that items with a low discrimination parameter have lower power than items

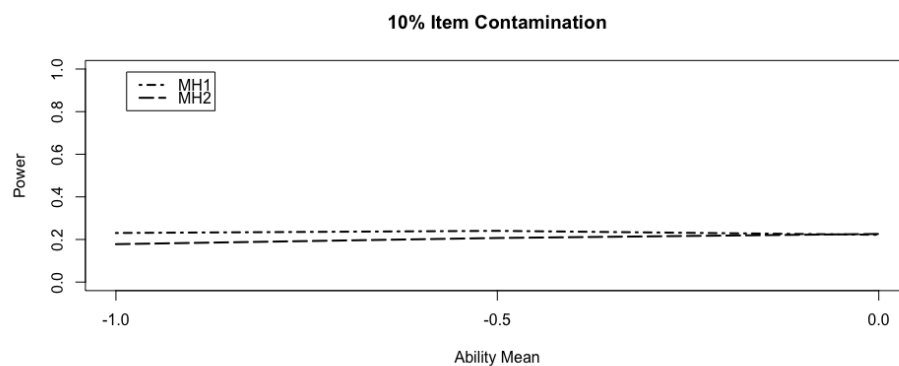
with a high discrimination parameter. Also, items with low difficulty parameters have lower power.

Interaction of amount of item contamination with other simulation factors.

Besides the size of the DIF, the amount of item contamination also has an interaction effect with the ability mean of the focal group and the item parameters. These effects are shown in Figure 27 and Figure 28.

For the interaction between the ability mean and the amount of item contamination, the power for the MH1 and MH2 methods increases as the ability mean increases but decreases as the amount of item contamination increases. This is reflected in the effect sizes of 4.0453 for the MH1 method and 3.2510 for the MH2 method.

For the interaction between the item parameters and the amount of item contamination, the power for the MH1 and MH2 methods show an upward trend across the item difficulty parameter, but much more severely for high item discrimination parameters as the amount of item contamination decreases. The effect sizes for the interaction between the amount of item contamination and the item parameters are given in Table 21.



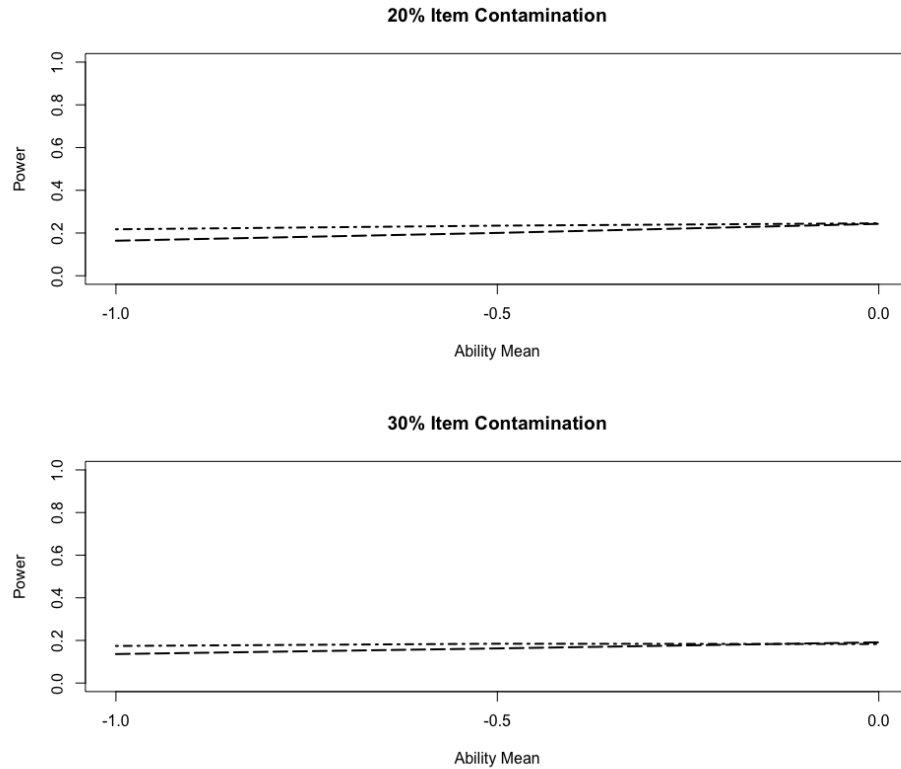


Figure 27. Median power rates across the ability mean and amount of item contamination. The first graph shows the power for the MH1 and MH2 methods across the ability mean when item contamination is 10%. The second graph shows the power for the MH1 and MH2 methods across the ability mean when item contamination is 20%. The last graph shows the power for the MH1 and MH2 methods across the ability mean when item contamination is 30%.

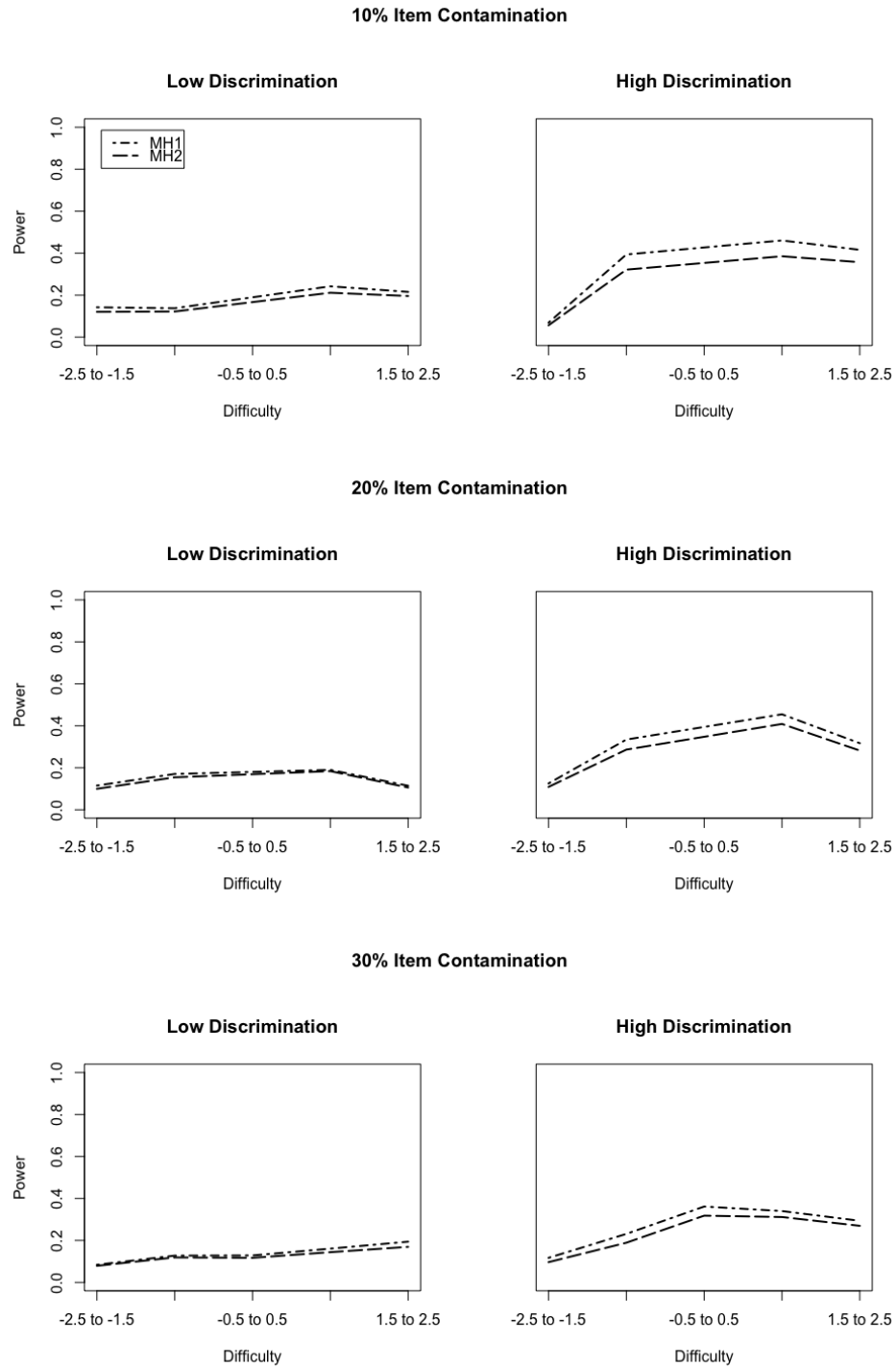


Figure 28. Median power rates across the item parameters and amount of item contamination. The first graph shows the power for the MH1 and MH2 methods across the item parameters when item contamination is 10%. The second graph

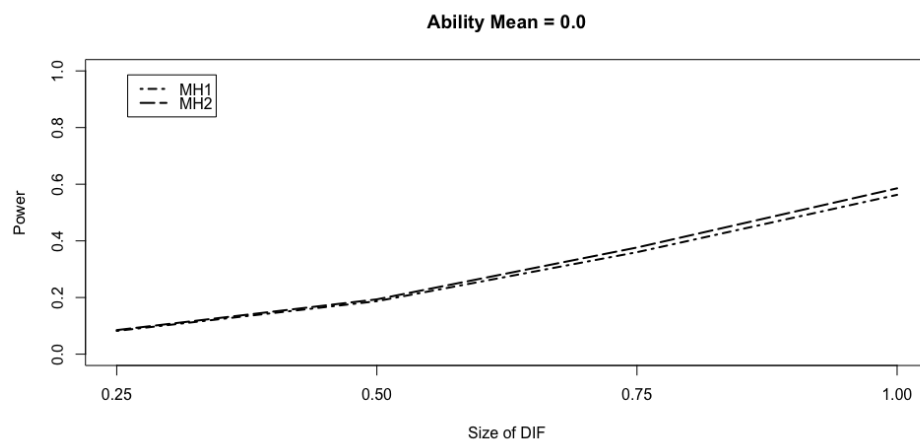
shows the power for the MH1 and MH2 methods across the item parameters when item contamination is 20%. The last graph shows the power for the MH1 and MH2 methods across the item parameters when item contamination is 30%.

Table 21

Odds Ratios for Interaction Between Ability Mean and Item Parameters

	MH1	MH2
Item Contamination * Discrimination	1.6920	1.7819
Item Contamination * Difficulty	6.9460	6.1521
Item Contamination * Diff * Disc	0.1937	0.2061

Interaction of the size of the DIF with other factors. The size of the DIF has been shown to have an interaction effect with the ability mean for the focal group and the item parameters on the power for the MH1 and MH2. The interaction between the size of the DIF and the ability mean is shown in Figure 29, the interaction between the size of the DIF and the item difficulty parameter is shown in Figure 30. And the interaction between the size of the DIF and the item discrimination parameter is shown in Figure 31.



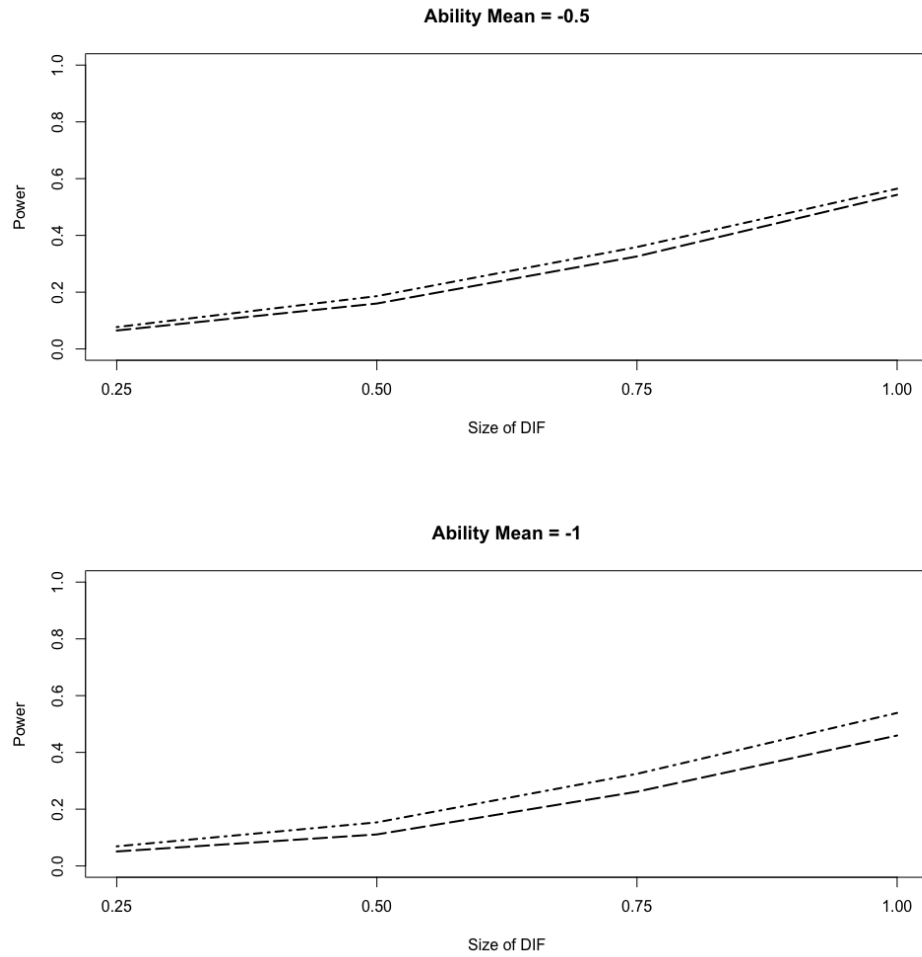


Figure 29. Median power rates across the size of the DIF and the ability mean.

The first graph shows the power for the MH1 and MH2 methods across the item parameters when item contamination is 10%. The second graph shows the power for the MH1 and MH2 methods across the item parameters when item contamination is 20%. The last graph shows the power for the MH1 and MH2 methods across the item parameters when item contamination is 30%.

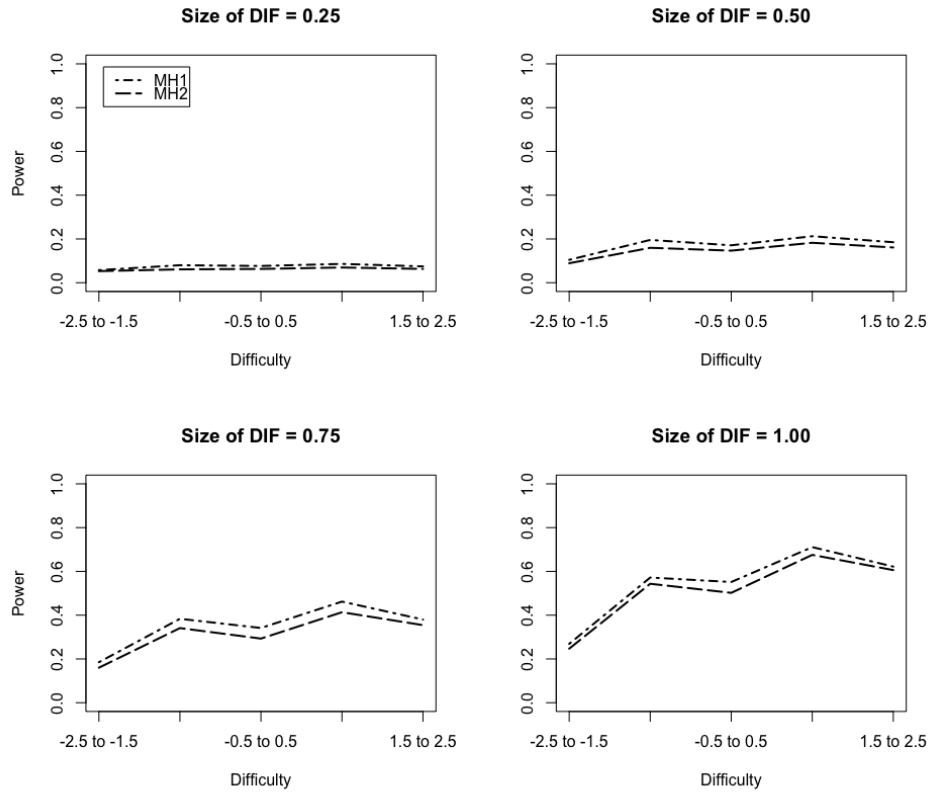


Figure 30. Median power rates across the size of the DIF and the item difficulty parameter. Four graphs show the power for the MH1 and MH2 methods across the item difficulty parameter separated by the size of the DIF.

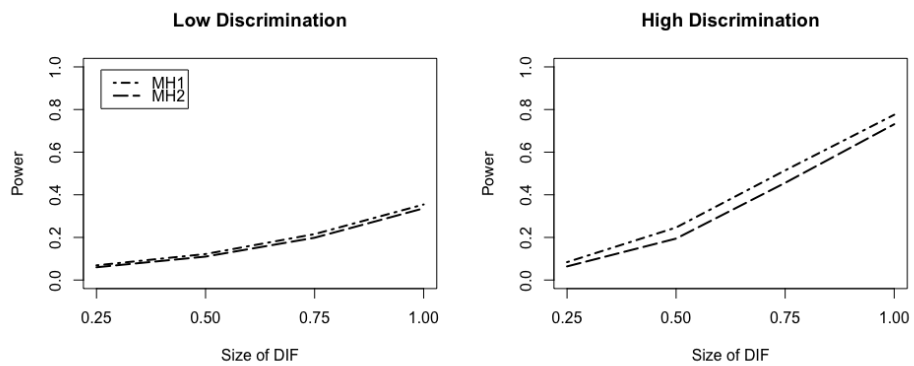


Figure 31. Median power across the size of the DIF and the item discrimination parameter. The left graph shows the power for the MH1 and MH2 across the size

of the DIF when item discrimination is low while the right graph shows the power across the size of the DIF when item discrimination is high.

For the interaction between the size of the DIF and the ability mean for the focal group, the power for the MH1 and MH2 methods increase across the size of the DIF and the ability mean. The effect on power is stronger for the MH2 method (odds ratio of 0.4933) than for the MH1 method (odds ratio of 0.6225).

For the interaction between the size of the DIF and the item parameters, the power for the MH1 and MH2 methods increase as the size of the DIF and the item difficulty parameter increases. The effect for the MH1 method (odds ratio of 1.4656) is stronger than the effect for MH2 method (odds ratio of 1.3479). For the item discrimination parameter, the power for the MH1 and MH2 methods increase across the size of the DIF but is much larger when the item discrimination is high. The effect for the MH1 method (odds ratio of 3.2586) is slightly weaker than the effects for the MH2 method (odds ratio of 3.5081).

Interaction between the ability mean and the item parameters. For the MH1 and MH2 methods, the effect of the two-way and three-way interactions between the ability mean and item parameters were shown to have an effect on power in the logistic regression. The first two-way interaction is between the ability mean and the difficulty parameter as shown in Figure 32. For the MH methods, the power increases as the difficulty parameter and the ability mean increases with a slightly stronger effect for the MH1 method (odds ratio of 0.5290) than the MH2 method (odds ratio of 0.5647).

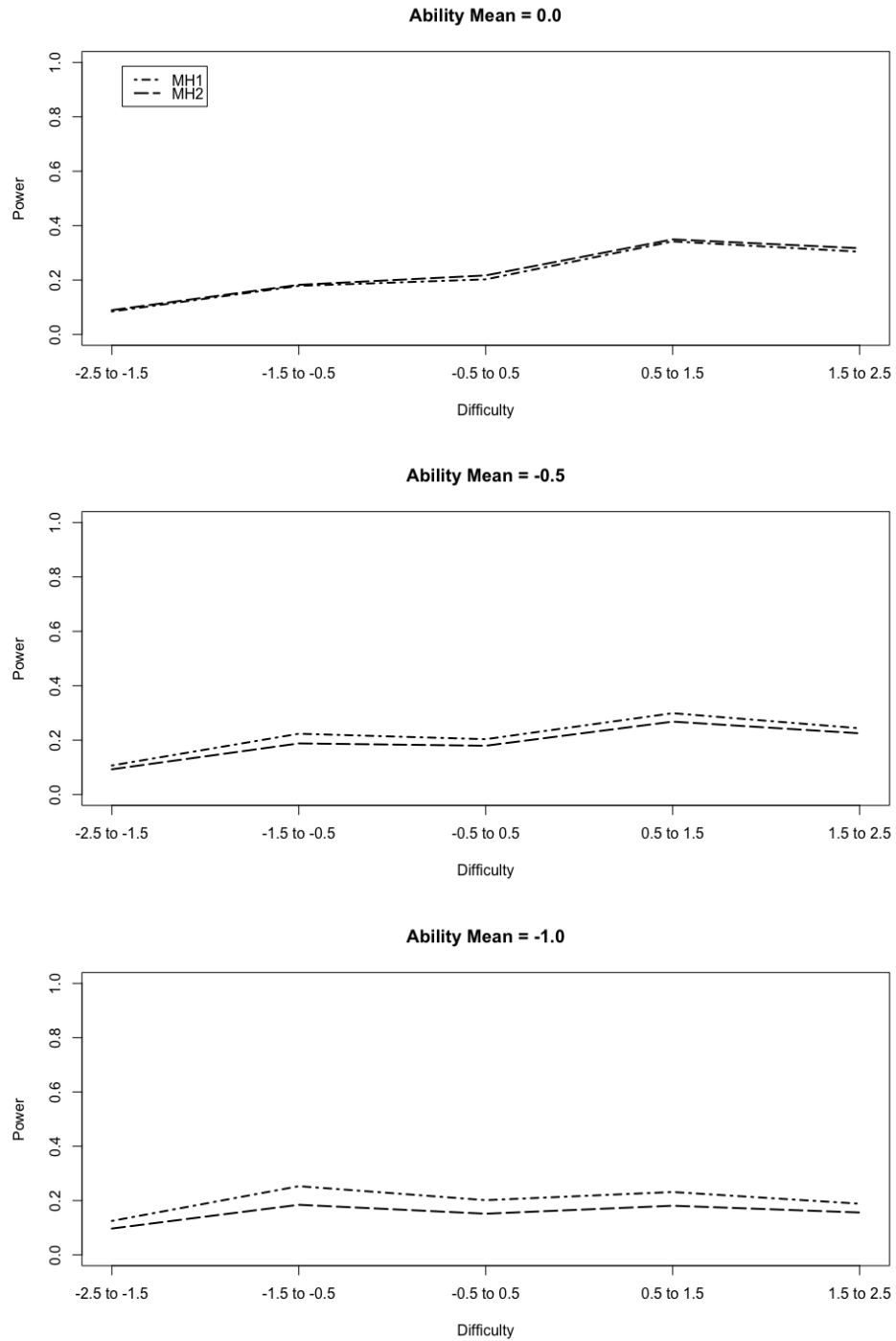
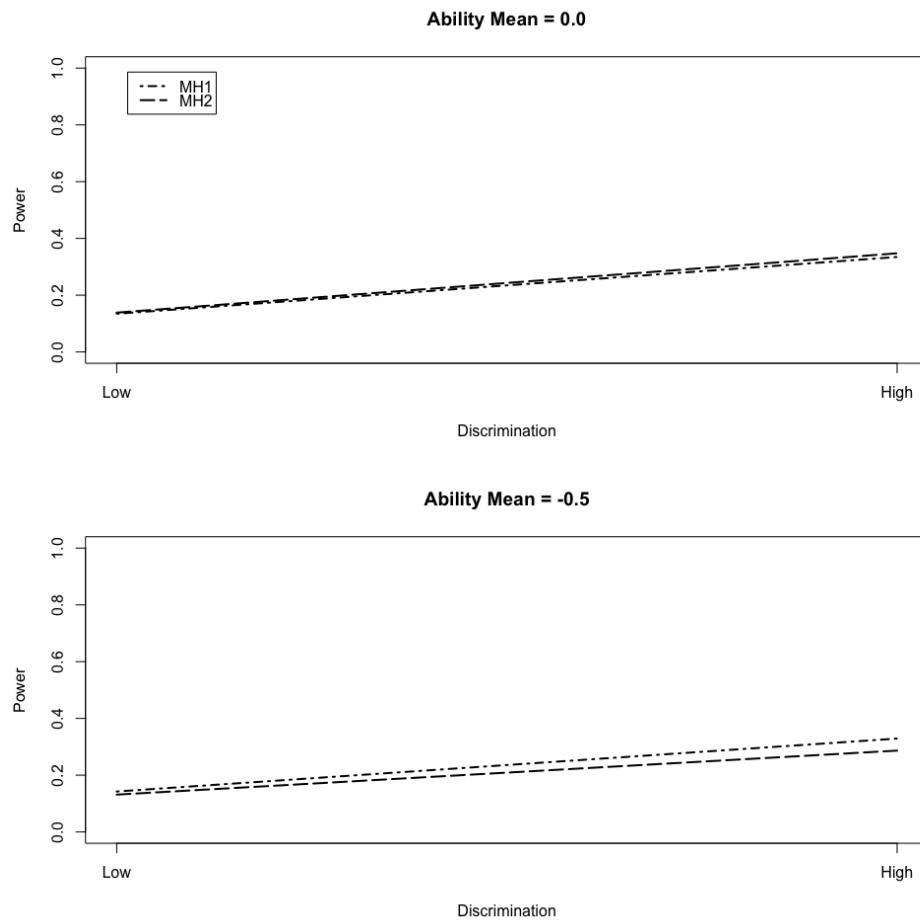


Figure 32. Median power rates across ability mean and item difficulty. The first graph shows the power for the MH1 and MH2 methods across the item difficulty parameters when the ability mean is equal to 0, the second graph shows the power for the MH1 and MH2 methods across the item difficulty parameters when the

ability mean is equal to -0.5 and the third graph shows the power for the MH1 and MH2 methods across the item difficulty parameters when the ability mean is equal to -1.0.

The second two-way interaction is between the ability mean and the discrimination parameter, shown in Figure 33.



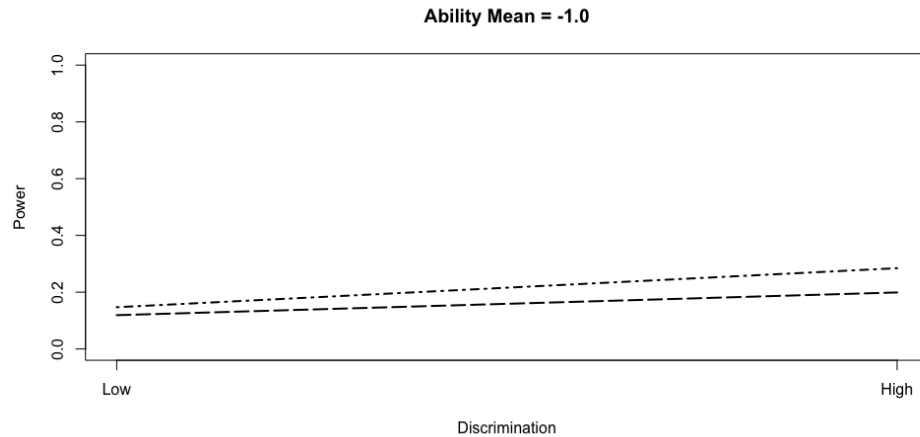


Figure 33. Median power rates across ability mean and item discrimination. The first graph shows the power for the MH1 and MH2 methods across the item discrimination parameters when the ability mean is equal to 0, the second graph shows the power for the MH1 and MH2 methods across the item discrimination parameters when the ability mean is equal to -0.5 and the third graph shows the power for the MH1 and MH2 methods across the item discrimination parameters when the ability mean is equal to -1.0.

For both methods, the power increases across the discrimination parameters as the ability mean also increases. The effect on power for the MH1 method (odds ratio of 1.1847) is stronger than on the MH2 method (odds ratio of 1.1064).

The three-way interaction between the ability mean and the item parameters are shown in Figure 34.

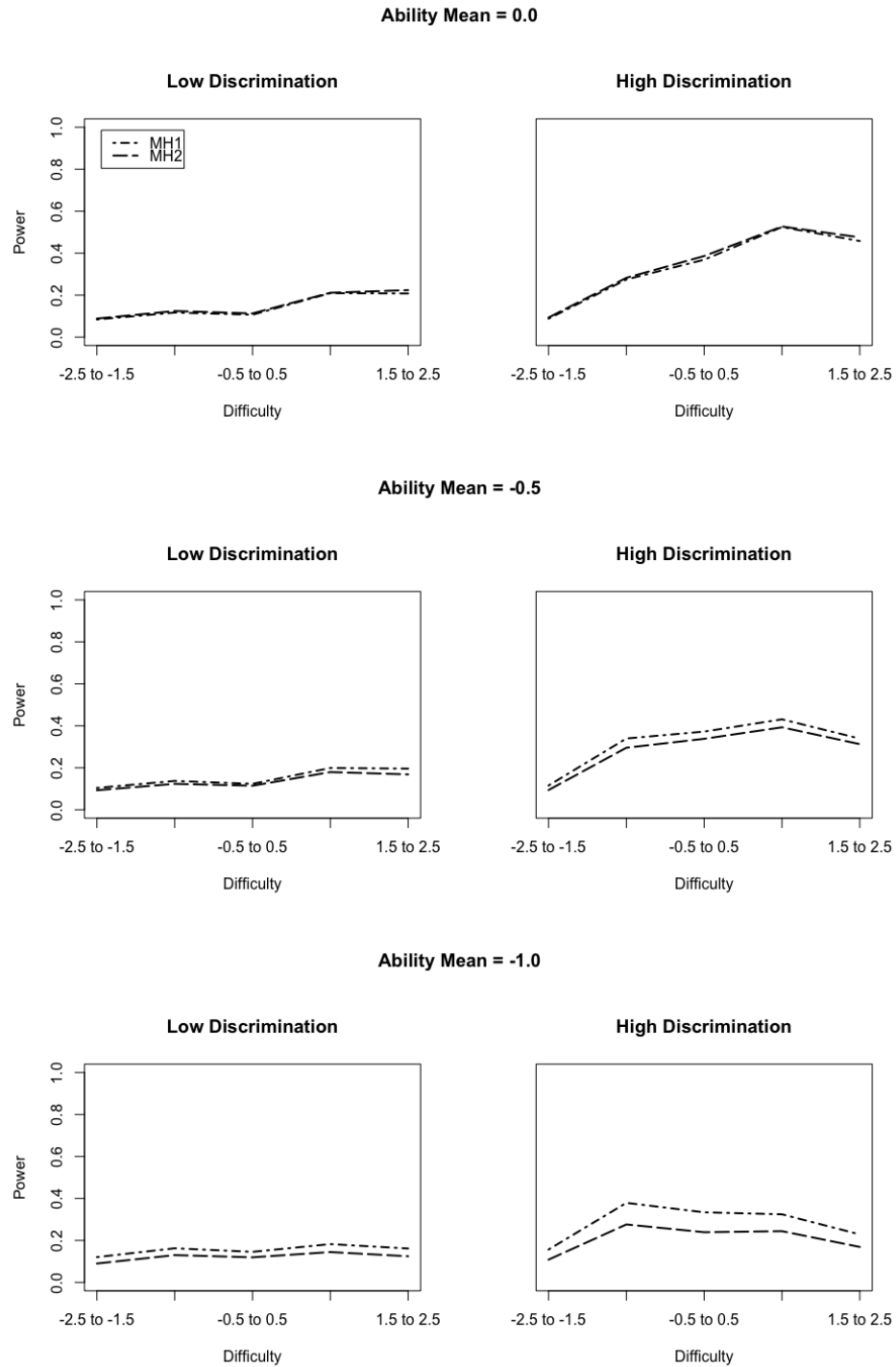


Figure 34. Median power across ability mean and item parameters. The first graph shows the power for the MH1 and MH2 methods across the item parameters when the ability mean is equal to 0, the second graph shows the power

for the MH1 and MH2 methods across the item parameters when the ability mean is equal to -0.5 and the third graph shows the power for the MH1 and MH2 methods across the item parameters when the ability mean is equal to -1.0.

The power for the MH1 and MH2 methods increases as the ability mean increases and when item discrimination is high. The highest power for both methods is 0.52 when the ability mean equals 0, the item discrimination is high, and the difficulty parameter ranges from 0.5 to 1.5. The effect on the power is similar for both methods (MH1 – 1.7546, MH2 – 1.6774).

Summary. Overall, the MH1 method has more power than that of the MH2 method across the different simulation factors even at reduced sample sizes. The strongest influence on power comes from the item parameters, the amount of item contamination, the size of the DIF, and the ability mean of the focal group. The highest power for both methods is found when the sample size is large, the ability mean is 0, the amount of item contamination is 10%, the size of the DIF is large, the item discrimination parameter is high, and the item difficulty parameter ranges from 0.5 to 1.5.

Answering the Research Questions

For the first part of the first research question, the overall Type I error rate for the RI method is much lower than the MH1 and MH2 methods. As the Type I error rate was not controlled for most of the simulation conditions, power was only analyzed for the Mantel-Haenszel methods in response to the second part of the first research question. The MH1 and MH2 performed similarly regardless of the simulation factors with the

MH1 method having slightly more power.

For the second research question, the effect of the simulation factors on the Type I error rates was evaluated with or without item contamination. Without item contamination, the influential factors were the ability mean for the focal group and the item parameters. For the RI method, the Type I error rates were much lower than that of the MH methods. With item contamination, the main factors that affect the Type I error rates were the amount of item contamination, the size of the DIF, and the ability mean for the focal group and the item parameters. For power, only the MH methods were evaluated as these methods had Type I error rates that were controlled across the various simulation factors. The following factors had an effect on the power of the MH methods: the item parameters, the amount of item contamination, the size of the DIF, and the ability mean for the focal group.

Chapter 5 – Discussion

Summary of Results

For the MH1, MH2, and RI methods, the main factors that affected the Type I error rates were the amount of item contamination, the size of the DIF, the ability mean for the focal group, and the item parameters. The sample size and the number of items were found not to have an effect on the Type I error rates for all methods. Overall, the Type I error rates with item contamination for the MH1 and MH2 methods were shown to increase as the amount of item contamination and the size of the DIF increases. Also, an increase in the amount of item contamination and a discrepancy in the ability mean led to an increase in the Type I error rates for the MH1 and MH2 methods. For the RI method, the Type I error rates were shown to decrease as the amount of item contamination and the size of the DIF decreases, especially with high discrimination and extreme difficulty parameters.

As the Type I error rate overall for the RI method was much lower than that of the MH1 and MH2 methods and not controlled across the simulation factors, power was only evaluated for the MH1 and MH2 methods. Overall, the MH1 method has more power than that of the MH2 method across the different simulation factors even at reduced sample sizes. The largest influences on power given a one-unit change came from the item parameters, the amount of item contamination, the size of the DIF, and the ability mean of the focal group. The highest power for both methods was found when the sample size was large, there was no discrepancy in the ability mean for the reference and focal groups, the amount of item contamination was 10%, the size of the DIF was large, the

item discrimination parameter was high, and the item difficulty parameter ranged from 0.5 to 1.5. The median power of these methods were .203 and .181, respectively.

Post Analysis

One possible reason for the RI and DIPF methods not being as effective as expected might be due to the use of the two-parameter logistic model instead of the one-parameter logistic model for data generation. For the RI method, only item difficulty was included in the effect size, meaning that the item discrimination was not taken into account when a 2PL IRT model was run. For the DIPF method, only differences in item difficulties were analyzed with the 1PL IRT model being used to obtain the estimated item difficulties. In order to examine if the performance of these methods might have been due to model specification, a post analysis was performed with the following conditions: 20 items, the ability mean of the focal group was set to -1 (with the reference at 0), the focal group sample size was set equal to 50 with 500 in the reference group, and no item contamination (i.e., no DIF in any items). The same item difficulty parameters were used in this simulation (see Table A1), with all the item discrimination parameters set to 1. Table 22 shows the Type I error rates for the DIPF and the RI methods from the two simulation studies: one with item responses generated using the 1PL IRT model and one with item responses generated using the 2PL IRT model. For the DIPF method, the Type I error rates are even larger when used on data generated from the 2PL IRT model than the 1PL IRT model which means that model misspecification was not the cause of the poor performance of the DIPF method. When the authors conducted a simulation study to evaluate the DIPF method, they manipulated the ability distribution by decreasing the ability mean from 0 for the reference group to -0.2 for the focal group and

increasing the standard deviation from 2 for the reference group to 3 for the focal group. As the simulation study conducted for this study did not change the standard deviation for the ability distribution for the two groups, this could be a reason why the DIPF method did not perform as expected. For the RI method, some of the Type I error rates approach 0.05 when applied to data from a one-parameter logistic model, but some are also further from 0.05. Thus, generating item responses from a one-parameter model instead of a two-parameter model does not necessarily mean that the RI method would be more effective than the MH1 and MH2 methods. As the interaction of the item difficulty and item discrimination parameters have a large effect on the Type I error rates for the RI method, one possible solution would be to modify the method to incorporate item discrimination.

Table 22

Type I Error Rates Across Methods for Data Generated from the 1PL and 2PL IRT Models

Item	Difficulty Parameter Stratum	DIPF		RI	
		1PL	2PL	1PL	2PL
1	1	0.004	0.009	0.027	0.019
2	1	0.008	0.069	0.020	0.042
3	1	0.005	0.100	0.029	0.018
4	1	0.010	0.018	0.027	0.015
5	2	0.002	0.008	0.054	0.031
6	2	0.005	0.115	0.067	0.004
7	2	0.001	0.066	0.059	0.009
8	2	0.005	0.014	0.069	0.034
9	3	0.005	0.006	0.052	0.061
10	3	0.005	0.200	0.011	0.024
11	3	0.005	0.062	0.048	0.051
12	3	0.007	0.009	0.054	0.075

13	4	0.010	0.022	0.008	0.034
14	4	0.014	0.137	0.010	0.043
15	4	0.006	0.003	0.014	0.044
16	4	0.009	0.037	0.007	0.064
17	5	0.045	0.059	0.057	0.022
18	5	0.033	0.017	0.035	0.039
19	5	0.023	0.033	0.024	0.018
20	5	0.013	0.088	0.018	0.017

Along with checking the model misspecification, the analysis also evaluated the overlap between the reference and focal groups across the MH score group widths to quantify the amount of overlap in the simulation conditions. Tables 23 and 24 show the score group widths (intervals), the mean number of reference group examinees in each interval (Reference), the mean number of focal group examinees in each interval (Focal), and the percentage of replications for which there were no focal group examinees in the interval (% No Focal).

Table 23

Overlap for MH1 Score Group Widths

Intervals	Reference	Focal	% No Focal
1	9.288	5.356	0
2	28.162	8.880	0
3	55.867	10.552	0
4	82.043	9.969	0
5	97.530	7.471	0
6	92.902	4.485	0
7	71.555	2.351	3.6
8	41.947	0.832	41.7
9	17.393	0.100	90.7
10	3.283	0.004	99.6

Table 24

Overlap for MH2 Score Group Widths

Intervals	Reference	Focal	% No Focal
1	37.45	14.200	0
2	137.940	20.533	0
3	190.432	11.950	0
4	113.502	3.114	3.2
5	20.676	0.203	82.0

Out of 1000 replications, there were 996 where the 10th interval do not have any focal group examinees for the MH1 method. For the MH2 method, only 820 out of 1000 replications had no focal group examinees in the fifth interval. Thus, reference examinees in the last intervals are not included when calculating the MH1 and MH2 statistics.

Although there were no focal group examinees in the last intervals, the Type I error rates for the MH1 and MH2 methods were consistently higher than that of the RI method.

Alignment with Literature Review

Clauser, Mazor, and Hambleton (1994) conducted a simulation study to compare the equal interval matching technique with that of thin matching by varying the sample size (2000, 1000, 500, 200, 100 per group) and the score group intervals (81, 20, 10, 5, 2). They found that if the sample was large and/or the ability distribution is the same for both groups, then the two matching techniques did not make a difference in DIF detection. However, when the sample size was small and the score group intervals were reduced then the group means were no longer equal, leading to potential misidentification of items with DIF. The authors recommended that if the sample size could not be increased, then the equal interval matching technique could be used, but cautiously. As

in this study, the authors found that items with higher discrimination, items with larger DIF size, and items with moderate difficulty parameters were the most likely to be correctly identified as exhibiting DIF. However, the authors did not examine the situation in which the reference group and the focal group sample sizes were unequal. Donoghue and Allen (1993) also conducted a simulation study comparing the various thick matching techniques with the thin matching technique. Using a test with 5, 10, 20 or 40 items and a sample size of 300/100, 600/200, and 1200/400 (Reference/Focal), the authors found that thick matching can improve results but not for small tests (5 or 10 items) and even with longer tests the thin matching performed best when the sample size was large. Both of these previous studies align with the results of this simulation study in terms of the effect of sample size on the power of the MH1 and MH2 methods.

Implications

With the MH1 and MH2 methods found to control Type I error rates regardless of the simulation factors, these methods can be used to evaluate items for DIF when the focal group sample size is small. However, with the median power being around .10 for the smallest sample size and .45 for the largest sample size, the minimum sample size recommended for these two methods is 100. These methods should not be the only approach used, but in tandem with cognitive and expert review. For example, in the Maller (1997) study, the 1PL IRT model was used to compare 110 deaf students with their hearing peers but many of the items showed a lack of good fit. Using the MH1 in this study would address the issue of model misspecification. However, neither of these methods should be used in the Martin (2005) study where 44 deaf students were

evaluated using the New York State English test, due to low power of .1705 and .1490 for the MH1 and MH2 method when the sample size was 50.

Limitations

For this simulation study, there are limitations that can affect the results such as the accuracy of the computer code, the type of factors used, and the methods chosen for comparison. These limitations are discussed in the next paragraphs, while the next section discusses how these limitations could provide directions for future research.

First, it is assumed that the computer code written for the simulation study was accurate. Although the code was debugged throughout the process of development, there is still the potential for miscalculations and misleading results. For example, the score group intervals developed for both Mantel-Haenszel methods had to be created then a new dataset of item responses was created based on these score group intervals. The resulting dataset was then processed by the Mantel-Haenszel test in the DIFR package. The reason for this was that the DIFR package does not provide the ability to match the reference and the focal group by another variable besides the individual total test scores. For transparency, the code is included in Appendix B.

Another limitation was the creation of the item difficulty and item discrimination parameters. Only one set of parameters was created for each number of items in the assessment (20, 40, and 80). By creating only one set, this does not allow for different combination of items such as all easy items versus a mix of easy and moderate items. See Tables A1-A3 in Appendix A for the item parameters used in this study. Along with the item parameters, a restriction was placed on the direction and variation of the DIF. For items exhibiting DIF, the difficulty parameter for the focal group was changed by

subtracting a specific value making the item easier for the focal group. However, there could be an assessment where one item was easier for the focal group, while another item was easier for the reference group. With the DIF for these items in opposite directions, there is the potential for cancelation meaning that these items could be found to not exhibit DIF. Also, the amount subtracted from the difficulty parameter was the same for all items simulated to have DIF. An assessment could have items with various amount of DIF, affecting the ability of the methods to detect these items.

Another factor that is a limitation in this simulation study is the sample size for the reference group being set to 500. There is no knowledge gained from this simulation study on how the sample size of the reference group could affect the four methods in terms of DIF detection. Both the Mantel-Haenszel method and the differential item pair functioning method has been shown to work with a sample size of 500 for the reference group. Decreasing the sample size for the reference group could lead to less power in detecting DIF items.

When evaluating items using the DIPF method, there is a concern that the DIPF method obtains the estimated item difficulty using a one-parameter logistic model. As the item parameters were created using a two-parameter logistic model, it is possible that the estimated item difficulties do not match the true item difficulties which leads to misidentification. Bechger and Maris (2015) mentioned that their method could be extended to a two-parameter logistic model, but the method would require the comparison of item ratios rather than item difficulties.

These limitations described in this section suggest ways that this simulation could be adapted for further research. The next section of this chapter discusses how this research could be extended.

Future Research

Based on the results of this study, the empirical distribution of the effect size should be reviewed to see if using a cutoff point equal to $\alpha/2$ would increase the Type I error rate. In addition, there needs to be a way to incorporate item discrimination into the method as the current method only uses the item difficulty parameters under the classical test theory framework.

Also, the MH1 and MH2 methods should be evaluated further using different sets of item parameters, changing directions in the size of the DIF, allowing items to have similar item parameters, and varying the size of the reference group. These methods should also be compared with the original MH method in terms of Type I error rates and power and possibly compared with other CTT methods such as standardization and logistic regression. As the differential item pair functioning was shown not to work for small sample sizes nor with the largest due to an arbitrary cutoff of half the item pairs being flagged as significant for the item to be flagged as significant, other IRT methods could be used such as the Wald test.

Also, the MH1 and MH2 methods should be applied to empirical data where the focal group sample sizes are small along with cognitive and expert review to determine if there is a practical use of these methods.

Conclusion

Based on this simulation study, the proposed method cannot be recommended for future studies due to the Type I error rates not being controlled and no consistency in the type of conditions where the RI method did have controlled Type I error rates. As there needs to be a common support between the reference and focal groups as shown by Simpson's paradox, the Mantel-Haenszel methods have been shown to exhibit controlled Type I error rates. However, the minimum sample size of 100 is recommended due to power being 10%-20% for the smaller sample sizes. It is recommended that these methods can be used along with cognitive and expert review given that the power of these methods range from 0.1 for the smallest sample size and 0.45 for the largest sample size.

Appendix A

Table A1

Item Parameters for 20 Items (Rounded to Four Places After the Decimal)

Item	Discrimination	Difficulty	Discrimination Stratum	Difficulty Stratum
1	0.8157	-1.5999	1	1
2	0.4001	-2.0703	1	1
3	1.8653	-1.5158	2	1
4	1.2175	-2.1315	2	1
5	0.8023	-0.7506	1	2
6	0.3258	-0.8086	1	2
7	1.6633	-0.8597	2	2
8	1.3287	-1.0849	2	2
9	0.7940	-0.2773	1	3
10	0.2027	0.4729	1	3
11	1.7138	-0.4171	2	3
12	1.1107	-0.0804	2	3
13	0.6018	1.2552	1	4
14	0.2665	0.8360	1	4
15	1.0536	0.5391	2	4
16	1.5723	1.0939	2	4
17	0.4754	2.0616	1	5
18	0.8326	1.9063	1	5
19	1.1475	1.7138	2	5
20	1.4822	1.6130	2	5

Table A2

Item Parameters for 40 Items (Rounded to Four Places After the Decimal)

Item	Discrimination	Difficulty	Discrimination Stratum	Difficulty Stratum
1	0.8476	-2.0411	1	1
2	0.5911	-2.1660	1	1
3	0.6996	-1.9450	1	1
4	0.7541	-1.9586	1	1
5	1.9265	-1.9714	2	1
6	1.0563	-2.4021	2	1
7	1.4934	-2.3481	2	1
8	1.1823	-1.9524	2	1
9	0.8715	-0.9912	1	2
10	0.9208	-1.4986	1	2
11	0.5689	-0.6602	1	2
12	0.9352	-0.7731	1	2
13	1.7656	-1.4700	2	2
14	1.4694	-1.1884	2	2
15	1.2249	-0.5362	2	2
16	1.2039	-0.6401	2	2
17	0.4255	0.1298	1	3
18	0.6758	-0.3022	1	3
19	0.7154	-0.2541	1	3
20	0.8239	0.0889	1	3

Item	Discrimination	Difficulty	Discrimination Stratum	Difficulty Stratum
21	1.4206	-0.4535	2	3
22	1.0502	0.3463	2	3
23	1.1741	-0.4028	2	3
24	1.3458	-0.4549	2	3
25	0.3748	1.0641	1	4
26	0.9177	0.7584	1	4
27	0.7985	1.2359	1	4
28	0.7708	1.4089	1	4
29	1.8707	1.0124	2	4
30	1.5634	0.7691	2	4
31	1.8542	1.2722	2	4
32	1.6909	0.6845	2	4
33	0.8781	1.6953	1	5
34	0.6768	2.4622	1	5
35	0.7346	2.2617	1	5
36	0.5778	2.1236	1	5
37	1.9892	1.7690	2	5
38	1.4666	1.6854	2	5
39	1.3953	1.8566	2	5
40	1.0821	2.0773	2	5

Table A3

Item Parameters for 80 Items (Rounded to Four Places After the Decimal)

Item	Discrimination	Difficulty	Discrimination Stratum	Difficulty Stratum
1	0.9990	-1.5110	1	1
2	0.5107	-2.2462	1	1
3	0.7165	-1.7783	1	1
4	0.4603	-2.0878	1	1
5	0.3202	-1.8912	1	1
6	0.2063	-1.8674	1	1
7	0.3466	-2.0174	1	1
8	0.5498	-1.6991	1	1
9	1.1794	-1.6107	2	1
10	1.1317	-2.3068	2	1
11	1.9300	-2.0528	2	1
12	1.2329	-2.1233	2	1
13	1.2872	-1.9830	2	1
14	1.1161	-1.8834	2	1
15	1.6552	-1.5254	2	1
16	1.9377	-2.2955	2	1
17	0.4281	-0.5432	1	2
18	0.3732	-1.1681	1	2
19	0.7284	-0.6986	1	2
20	0.8363	-1.3751	1	2

Item	Discrimination	Difficulty	Discrimination Stratum	Difficulty Stratum
21	0.3289	-0.9744	1	2
22	0.8539	-1.4900	1	2
23	0.8883	-0.9256	1	2
24	0.4743	-1.0197	1	2
25	1.0101	-1.2759	2	2
26	1.8614	-1.3035	2	2
27	1.0415	-0.9910	2	2
28	1.6699	-0.8179	2	2
29	1.9202	-1.3471	2	2
30	1.5260	-0.6299	2	2
31	1.4529	-1.1060	2	2
32	1.4260	-0.9096	2	2
33	0.7154	-0.1989	1	3
34	0.9677	-0.1326	1	3
35	0.7886	0.2644	1	3
36	0.7263	-0.3063	1	3
37	0.5405	-0.1771	1	3
38	0.8199	0.3272	1	3
39	0.6687	-0.0717	1	3
40	0.6220	0.0450	1	3

Item	Discrimination	Difficulty	Discrimination Stratum	Difficulty Stratum
41	1.7628	-0.0567	2	3
42	1.5937	-0.4726	2	3
43	1.3134	0.4465	2	3
44	1.2060	-0.3903	2	3
45	1.5067	0.2722	2	3
46	1.5193	0.0980	2	3
47	1.2536	-0.0815	2	3
48	1.3319	-0.2156	2	3
49	0.7725	1.1112	1	4
50	0.6452	1.2479	1	4
51	0.3879	1.3481	1	4
52	0.8514	1.4010	1	4
53	0.6013	1.4497	1	4
54	0.5352	0.9508	1	4
55	0.5332	1.2959	1	4
56	0.5319	0.8869	1	4
57	1.5576	0.6338	2	4
58	1.7098	0.9960	2	4
59	1.0983	1.1681	2	4
60	1.0558	1.0020	2	4

Item	Discrimination	Difficulty	Discrimination Stratum	Difficulty Stratum
61	1.0146	0.8674	2	4
62	1.4335	0.5277	2	4
63	1.7523	1.3636	2	4
64	1.7505	1.0546	2	4
65	0.9607	2.3387	1	5
66	0.2612	2.3128	1	5
67	0.8820	2.1195	1	5
68	0.9771	2.2945	1	5
69	0.2204	1.8449	1	5
70	0.9658	2.0010	1	5
71	0.4389	1.9565	1	5
72	0.5072	1.9412	1	5
73	1.3474	1.5097	2	5
74	1.3409	1.5650	2	5
75	1.6339	1.5481	2	5
76	1.0825	2.0971	2	5
77	1.4040	2.1751	2	5
78	1.9751	1.8488	2	5
79	1.4462	1.8565	2	5
80	1.1086	1.6635	2	5

Table A4

Median Type I Error Rates for All Conditions without Item Contamination

theta	a	b	MH1	MH2	DIPF	RI
0	1	1	0.044	0.044	0.006	0.057
0	1	2	0.046	0.046	0.006	0.075
0	1	3	0.047	0.047	0.004	0.087
0	1	4	0.046	0.045	0.006	0.072
0	1	5	0.042	0.043	0.009	0.041
0	2	1	0.038	0.039	0.015	0.008
0	2	2	0.044	0.044	0.006	0.036
0	2	3	0.046	0.046	0.003	0.062
0	2	4	0.046	0.045	0.005	0.042
0	2	5	0.041	0.043	0.012	0.012
-0.5	1	1	0.046	0.048	0.014	0.048
-0.5	1	2	0.048	0.050	0.008	0.055
-0.5	1	3	0.048	0.051	0.007	0.056
-0.5	1	4	0.047	0.047	0.011	0.040
-0.5	1	5	0.043	0.041	0.020	0.035
-0.5	2	1	0.050	0.061	0.018	0.010
-0.5	2	2	0.053	0.065	0.008	0.059
-0.5	2	3	0.050	0.060	0.007	0.094
-0.5	2	4	0.042	0.045	0.016	0.017
-0.5	2	5	0.032	0.030	0.028	0.007
-1	1	1	0.048	0.052	0.034	0.032
-1	1	2	0.051	0.057	0.011	0.029
-1	1	3	0.049	0.053	0.015	0.019
-1	1	4	0.046	0.046	0.027	0.014
-1	1	5	0.041	0.042	0.058	0.031
-1	2	1	0.063	0.098	0.023	0.006
-1	2	2	0.063	0.111	0.028	0.110
-1	2	3	0.053	0.090	0.017	0.095
-1	2	4	0.038	0.047	0.051	0.002
-1	2	5	0.025	0.019	0.077	0.003

Note: The columns are defined as follows: theta = ability mean, a = discrimination

stratum, b = difficulty stratum, MH1 = Mantel-Haenszel with small intervals,

MH2 = Mantel-Haenszel with large intervals, DIPF = differential item pair functioning, RI = relative item performance method

Table A5

Median Type I Error Rates with Item Contamination

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
0	0.1	0.25	1	1	0.0450	0.0465	0.0080	0.0500
0	0.2	0.25	1	1	0.0470	0.0475	0.0080	0.0410
0	0.3	0.25	1	1	0.0510	0.0500	0.0080	0.0435
0	0.1	0.5	1	1	0.0490	0.0485	0.0070	0.0380
0	0.2	0.5	1	1	0.0525	0.0490	0.0090	0.0285
0	0.3	0.5	1	1	0.0650	0.0625	0.0110	0.0265
0	0.1	0.75	1	1	0.0500	0.0490	0.0080	0.0255
0	0.2	0.75	1	1	0.0630	0.0600	0.0090	0.0150
0	0.3	0.75	1	1	0.0770	0.0720	0.0150	0.0150
0	0.1	1	1	1	0.0510	0.0520	0.0080	0.0155
0	0.2	1	1	1	0.0690	0.0640	0.0100	0.0085
0	0.3	1	1	1	0.0985	0.0890	0.0140	0.0045
0	0.1	0.25	1	2	0.0485	0.0485	0.0050	0.0670
0	0.2	0.25	1	2	0.0510	0.0510	0.0065	0.0595
0	0.3	0.25	1	2	0.0520	0.0505	0.0090	0.0610
0	0.1	0.5	1	2	0.0490	0.0475	0.0050	0.0515
0	0.2	0.5	1	2	0.0520	0.0525	0.0070	0.0350
0	0.3	0.5	1	2	0.0650	0.0620	0.0085	0.0360
0	0.1	0.75	1	2	0.0515	0.0510	0.0060	0.0370
0	0.2	0.75	1	2	0.0675	0.0635	0.0070	0.0190
0	0.3	0.75	1	2	0.0820	0.0735	0.0115	0.0195
0	0.1	1	1	2	0.0555	0.0540	0.0065	0.0235
0	0.2	1	1	2	0.0765	0.0725	0.0090	0.0070
0	0.3	1	1	2	0.1150	0.0985	0.0120	0.0095
0	0.1	0.25	1	3	0.0490	0.0490	0.0045	0.0785
0	0.2	0.25	1	3	0.0490	0.0495	0.0060	0.0680

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
0	0.3	0.25	1	3	0.0530	0.0525	0.0050	0.0690
0	0.1	0.5	1	3	0.0500	0.0500	0.0050	0.0625
0	0.2	0.5	1	3	0.0585	0.0575	0.0060	0.0475
0	0.3	0.5	1	3	0.0660	0.0660	0.0080	0.0340
0	0.1	0.75	1	3	0.0530	0.0510	0.0060	0.0455
0	0.2	0.75	1	3	0.0650	0.0625	0.0070	0.0220
0	0.3	0.75	1	3	0.1085	0.0995	0.0100	0.0180
0	0.1	1	1	3	0.0580	0.0550	0.0060	0.0310
0	0.2	1	1	3	0.0795	0.0765	0.0080	0.0105
0	0.3	1	1	3	0.1400	0.1235	0.0120	0.0090
0	0.1	0.25	1	4	0.0480	0.0480	0.0070	0.0625
0	0.2	0.25	1	4	0.0490	0.0470	0.0070	0.0600
0	0.3	0.25	1	4	0.0460	0.0465	0.0070	0.0575
0	0.1	0.5	1	4	0.0480	0.0490	0.0060	0.0495
0	0.2	0.5	1	4	0.0510	0.0500	0.0085	0.0360
0	0.3	0.5	1	4	0.0575	0.0555	0.0110	0.0315
0	0.1	0.75	1	4	0.0500	0.0485	0.0065	0.0365
0	0.2	0.75	1	4	0.0575	0.0580	0.0090	0.0175
0	0.3	0.75	1	4	0.0770	0.0710	0.0130	0.0110
0	0.1	1	1	4	0.0500	0.0510	0.0070	0.0230
0	0.2	1	1	4	0.0655	0.0615	0.0090	0.0095
0	0.3	1	1	4	0.0945	0.0865	0.0140	0.0075
0	0.1	0.25	1	5	0.0450	0.0425	0.0100	0.0345
0	0.2	0.25	1	5	0.0450	0.0430	0.0100	0.0340
0	0.3	0.25	1	5	0.0475	0.0455	0.0110	0.0370
0	0.1	0.5	1	5	0.0460	0.0460	0.0100	0.0295
0	0.2	0.5	1	5	0.0460	0.0450	0.0120	0.0165
0	0.3	0.5	1	5	0.0510	0.0500	0.0120	0.0185
0	0.1	0.75	1	5	0.0425	0.0440	0.0090	0.0160
0	0.2	0.75	1	5	0.0495	0.0470	0.0120	0.0070
0	0.3	0.75	1	5	0.0545	0.0535	0.0145	0.0055
0	0.1	1	1	5	0.0460	0.0445	0.0095	0.0125
0	0.2	1	1	5	0.0530	0.0535	0.0125	0.0030
0	0.3	1	1	5	0.0695	0.0675	0.0160	0.0015
0	0.1	0.25	2	1	0.0400	0.0400	0.0140	0.0080

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
0	0.2	0.25	2	1	0.0445	0.0460	0.0180	0.0055
0	0.3	0.25	2	1	0.0510	0.0510	0.0185	0.0060
0	0.1	0.5	2	1	0.0440	0.0450	0.0160	0.0060
0	0.2	0.5	2	1	0.0610	0.0570	0.0205	0.0030
0	0.3	0.5	2	1	0.0740	0.0690	0.0230	0.0030
0	0.1	0.75	2	1	0.0485	0.0470	0.0150	0.0030
0	0.2	0.75	2	1	0.0770	0.0730	0.0200	0.0015
0	0.3	0.75	2	1	0.1085	0.0985	0.0280	0.0010
0	0.1	1	2	1	0.0560	0.0550	0.0150	0.0020
0	0.2	1	2	1	0.1005	0.0915	0.0240	0.0000
0	0.3	1	2	1	0.1485	0.1345	0.0285	0.0000
0	0.1	0.25	2	2	0.0475	0.0485	0.0055	0.0355
0	0.2	0.25	2	2	0.0535	0.0550	0.0040	0.0405
0	0.3	0.25	2	2	0.0590	0.0580	0.0060	0.0335
0	0.1	0.5	2	2	0.0515	0.0515	0.0070	0.0270
0	0.2	0.5	2	2	0.0690	0.0645	0.0060	0.0195
0	0.3	0.5	2	2	0.0900	0.0860	0.0080	0.0170
0	0.1	0.75	2	2	0.0575	0.0540	0.0060	0.0110
0	0.2	0.75	2	2	0.0995	0.0870	0.0070	0.0080
0	0.3	0.75	2	2	0.1515	0.1280	0.0105	0.0075
0	0.1	1	2	2	0.0715	0.0690	0.0055	0.0095
0	0.2	1	2	2	0.1255	0.1115	0.0080	0.0035
0	0.3	1	2	2	0.2290	0.1940	0.0135	0.0030
0	0.1	0.25	2	3	0.0505	0.0480	0.0030	0.0595
0	0.2	0.25	2	3	0.0525	0.0515	0.0035	0.0530
0	0.3	0.25	2	3	0.0600	0.0555	0.0035	0.0510
0	0.1	0.5	2	3	0.0510	0.0510	0.0030	0.0455
0	0.2	0.5	2	3	0.0720	0.0685	0.0045	0.0335
0	0.3	0.5	2	3	0.0995	0.0885	0.0070	0.0300
0	0.1	0.75	2	3	0.0560	0.0555	0.0030	0.0315
0	0.2	0.75	2	3	0.0995	0.0895	0.0050	0.0170
0	0.3	0.75	2	3	0.1560	0.1395	0.0085	0.0185
0	0.1	1	2	3	0.0675	0.0590	0.0030	0.0200
0	0.2	1	2	3	0.1310	0.1180	0.0060	0.0055
0	0.3	1	2	3	0.2585	0.2050	0.0095	0.0090

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
0	0.1	0.25	2	4	0.0450	0.0445	0.0050	0.0350
0	0.2	0.25	2	4	0.0525	0.0525	0.0060	0.0300
0	0.3	0.25	2	4	0.0560	0.0555	0.0060	0.0260
0	0.1	0.5	2	4	0.0490	0.0495	0.0050	0.0255
0	0.2	0.5	2	4	0.0580	0.0555	0.0070	0.0130
0	0.3	0.5	2	4	0.0850	0.0790	0.0090	0.0155
0	0.1	0.75	2	4	0.0510	0.0490	0.0050	0.0180
0	0.2	0.75	2	4	0.0760	0.0705	0.0080	0.0060
0	0.3	0.75	2	4	0.1370	0.1175	0.0110	0.0055
0	0.1	1	2	4	0.0570	0.0540	0.0050	0.0080
0	0.2	1	2	4	0.1070	0.0950	0.0085	0.0010
0	0.3	1	2	4	0.1975	0.1760	0.0120	0.0020
0	0.1	0.25	2	5	0.0380	0.0390	0.0130	0.0100
0	0.2	0.25	2	5	0.0400	0.0395	0.0130	0.0090
0	0.3	0.25	2	5	0.0415	0.0380	0.0120	0.0090
0	0.1	0.5	2	5	0.0375	0.0395	0.0120	0.0060
0	0.2	0.5	2	5	0.0405	0.0390	0.0145	0.0035
0	0.3	0.5	2	5	0.0555	0.0555	0.0155	0.0025
0	0.1	0.75	2	5	0.0415	0.0415	0.0120	0.0030
0	0.2	0.75	2	5	0.0530	0.0520	0.0140	0.0010
0	0.3	0.75	2	5	0.0785	0.0675	0.0170	0.0000
0	0.1	1	2	5	0.0450	0.0420	0.0120	0.0005
0	0.2	1	2	5	0.0650	0.0595	0.0150	0.0000
0	0.3	1	2	5	0.1120	0.0995	0.0190	0.0000
-0.5	0.1	0.25	1	1	0.0460	0.0485	0.0125	0.0495
-0.5	0.2	0.25	1	1	0.0495	0.0510	0.0100	0.0405
-0.5	0.3	0.25	1	1	0.0540	0.0575	0.0095	0.0330
-0.5	0.1	0.5	1	1	0.0515	0.0530	0.0115	0.0340
-0.5	0.2	0.5	1	1	0.0510	0.0560	0.0090	0.0240
-0.5	0.3	0.5	1	1	0.0575	0.0615	0.0070	0.0175
-0.5	0.1	0.75	1	1	0.0500	0.0525	0.0105	0.0230
-0.5	0.2	0.75	1	1	0.0595	0.0655	0.0090	0.0125
-0.5	0.3	0.75	1	1	0.0780	0.0820	0.0080	0.0060
-0.5	0.1	1	1	1	0.0525	0.0550	0.0100	0.0125
-0.5	0.2	1	1	1	0.0700	0.0750	0.0110	0.0055

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
-0.5	0.3	1	1	1	0.0955	0.0990	0.0085	0.0045
-0.5	0.1	0.25	1	2	0.0500	0.0530	0.0060	0.0510
-0.5	0.2	0.25	1	2	0.0500	0.0535	0.0075	0.0470
-0.5	0.3	0.25	1	2	0.0525	0.0555	0.0065	0.0525
-0.5	0.1	0.5	1	2	0.0510	0.0525	0.0060	0.0450
-0.5	0.2	0.5	1	2	0.0520	0.0600	0.0070	0.0325
-0.5	0.3	0.5	1	2	0.0640	0.0685	0.0075	0.0350
-0.5	0.1	0.75	1	2	0.0505	0.0565	0.0060	0.0295
-0.5	0.2	0.75	1	2	0.0605	0.0685	0.0070	0.0185
-0.5	0.3	0.75	1	2	0.0855	0.0960	0.0080	0.0170
-0.5	0.1	1	1	2	0.0545	0.0585	0.0060	0.0190
-0.5	0.2	1	1	2	0.0750	0.0835	0.0090	0.0090
-0.5	0.3	1	1	2	0.1015	0.1050	0.0100	0.0150
-0.5	0.1	0.25	1	3	0.0460	0.0500	0.0070	0.0530
-0.5	0.2	0.25	1	3	0.0500	0.0555	0.0060	0.0575
-0.5	0.3	0.25	1	3	0.0555	0.0630	0.0045	0.0655
-0.5	0.1	0.5	1	3	0.0500	0.0540	0.0060	0.0455
-0.5	0.2	0.5	1	3	0.0550	0.0565	0.0060	0.0405
-0.5	0.3	0.5	1	3	0.0685	0.0770	0.0055	0.0515
-0.5	0.1	0.75	1	3	0.0505	0.0540	0.0050	0.0330
-0.5	0.2	0.75	1	3	0.0610	0.0690	0.0060	0.0290
-0.5	0.3	0.75	1	3	0.1090	0.1190	0.0090	0.0300
-0.5	0.1	1	1	3	0.0540	0.0565	0.0060	0.0185
-0.5	0.2	1	1	3	0.0695	0.0710	0.0070	0.0150
-0.5	0.3	1	1	3	0.1500	0.1590	0.0090	0.0135
-0.5	0.1	0.25	1	4	0.0435	0.0460	0.0105	0.0370
-0.5	0.2	0.25	1	4	0.0470	0.0485	0.0090	0.0370
-0.5	0.3	0.25	1	4	0.0470	0.0500	0.0090	0.0335
-0.5	0.1	0.5	1	4	0.0480	0.0490	0.0100	0.0320
-0.5	0.2	0.5	1	4	0.0500	0.0520	0.0075	0.0215
-0.5	0.3	0.5	1	4	0.0505	0.0530	0.0080	0.0170
-0.5	0.1	0.75	1	4	0.0470	0.0475	0.0100	0.0195
-0.5	0.2	0.75	1	4	0.0545	0.0555	0.0090	0.0105
-0.5	0.3	0.75	1	4	0.0620	0.0660	0.0090	0.0110
-0.5	0.1	1	1	4	0.0480	0.0505	0.0100	0.0100

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
-0.5	0.2	1	1	4	0.0565	0.0600	0.0090	0.0040
-0.5	0.3	1	1	4	0.0790	0.0850	0.0105	0.0015
-0.5	0.1	0.25	1	5	0.0430	0.0435	0.0150	0.0315
-0.5	0.2	0.25	1	5	0.0430	0.0420	0.0175	0.0185
-0.5	0.3	0.25	1	5	0.0440	0.0445	0.0150	0.0240
-0.5	0.1	0.5	1	5	0.0435	0.0425	0.0165	0.0175
-0.5	0.2	0.5	1	5	0.0475	0.0475	0.0155	0.0070
-0.5	0.3	0.5	1	5	0.0450	0.0490	0.0120	0.0075
-0.5	0.1	0.75	1	5	0.0435	0.0430	0.0155	0.0080
-0.5	0.2	0.75	1	5	0.0445	0.0455	0.0150	0.0035
-0.5	0.3	0.75	1	5	0.0530	0.0560	0.0135	0.0020
-0.5	0.1	1	1	5	0.0440	0.0435	0.0180	0.0060
-0.5	0.2	1	1	5	0.0505	0.0520	0.0155	0.0000
-0.5	0.3	1	1	5	0.0560	0.0575	0.0125	0.0000
-0.5	0.1	0.25	2	1	0.0530	0.0650	0.0175	0.0060
-0.5	0.2	0.25	2	1	0.0650	0.0795	0.0320	0.0040
-0.5	0.3	0.25	2	1	0.0665	0.0825	0.0280	0.0060
-0.5	0.1	0.5	2	1	0.0590	0.0735	0.0190	0.0070
-0.5	0.2	0.5	2	1	0.0790	0.0965	0.0430	0.0030
-0.5	0.3	0.5	2	1	0.1055	0.1250	0.0435	0.0030
-0.5	0.1	0.75	2	1	0.0710	0.0820	0.0185	0.0035
-0.5	0.2	0.75	2	1	0.1095	0.1280	0.0475	0.0015
-0.5	0.3	0.75	2	1	0.1570	0.1745	0.0535	0.0010
-0.5	0.1	1	2	1	0.0750	0.0925	0.0185	0.0025
-0.5	0.2	1	2	1	0.1530	0.1740	0.0480	0.0010
-0.5	0.3	1	2	1	0.2270	0.2400	0.0580	0.0010
-0.5	0.1	0.25	2	2	0.0595	0.0740	0.0115	0.0650
-0.5	0.2	0.25	2	2	0.0645	0.0835	0.0105	0.0795
-0.5	0.3	0.25	2	2	0.0750	0.0930	0.0135	0.0835
-0.5	0.1	0.5	2	2	0.0680	0.0835	0.0110	0.0545
-0.5	0.2	0.5	2	2	0.0870	0.1080	0.0150	0.0650
-0.5	0.3	0.5	2	2	0.1180	0.1405	0.0225	0.0655
-0.5	0.1	0.75	2	2	0.0775	0.1005	0.0140	0.0420
-0.5	0.2	0.75	2	2	0.1235	0.1430	0.0180	0.0465
-0.5	0.3	0.75	2	2	0.2115	0.2350	0.0310	0.0490

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
-0.5	0.1	1	2	2	0.0865	0.1135	0.0145	0.0345
-0.5	0.2	1	2	2	0.1665	0.1875	0.0230	0.0315
-0.5	0.3	1	2	2	0.3080	0.3160	0.0365	0.0295
-0.5	0.1	0.25	2	3	0.0555	0.0665	0.0090	0.1035
-0.5	0.2	0.25	2	3	0.0610	0.0775	0.0110	0.1210
-0.5	0.3	0.25	2	3	0.0805	0.1015	0.0145	0.1280
-0.5	0.1	0.5	2	3	0.0590	0.0770	0.0090	0.0930
-0.5	0.2	0.5	2	3	0.0845	0.1010	0.0150	0.1005
-0.5	0.3	0.5	2	3	0.1250	0.1545	0.0185	0.1185
-0.5	0.1	0.75	2	3	0.0650	0.0805	0.0110	0.0745
-0.5	0.2	0.75	2	3	0.1220	0.1435	0.0225	0.0740
-0.5	0.3	0.75	2	3	0.2195	0.2380	0.0345	0.0905
-0.5	0.1	1	2	3	0.0780	0.0970	0.0110	0.0510
-0.5	0.2	1	2	3	0.1635	0.1905	0.0210	0.0505
-0.5	0.3	1	2	3	0.3500	0.3725	0.0395	0.0525
-0.5	0.1	0.25	2	4	0.0465	0.0495	0.0200	0.0155
-0.5	0.2	0.25	2	4	0.0450	0.0525	0.0285	0.0135
-0.5	0.3	0.25	2	4	0.0580	0.0710	0.0230	0.0190
-0.5	0.1	0.5	2	4	0.0500	0.0550	0.0205	0.0130
-0.5	0.2	0.5	2	4	0.0600	0.0700	0.0315	0.0085
-0.5	0.3	0.5	2	4	0.0885	0.0995	0.0355	0.0150
-0.5	0.1	0.75	2	4	0.0535	0.0645	0.0200	0.0085
-0.5	0.2	0.75	2	4	0.0770	0.0905	0.0370	0.0060
-0.5	0.3	0.75	2	4	0.1270	0.1440	0.0425	0.0095
-0.5	0.1	1	2	4	0.0645	0.0740	0.0210	0.0060
-0.5	0.2	1	2	4	0.1055	0.1260	0.0400	0.0020
-0.5	0.3	1	2	4	0.2040	0.2120	0.0510	0.0020
-0.5	0.1	0.25	2	5	0.0310	0.0275	0.0320	0.0040
-0.5	0.2	0.25	2	5	0.0265	0.0300	0.0330	0.0025
-0.5	0.3	0.25	2	5	0.0285	0.0270	0.0380	0.0020
-0.5	0.1	0.5	2	5	0.0300	0.0255	0.0320	0.0020
-0.5	0.2	0.5	2	5	0.0310	0.0320	0.0455	0.0005
-0.5	0.3	0.5	2	5	0.0385	0.0420	0.0475	0.0000
-0.5	0.1	0.75	2	5	0.0290	0.0300	0.0290	0.0005
-0.5	0.2	0.75	2	5	0.0280	0.0315	0.0435	0.0000

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
-0.5	0.3	0.75	2	5	0.0555	0.0555	0.0555	0.0000
-0.5	0.1	1	2	5	0.0255	0.0270	0.0320	0.0000
-0.5	0.2	1	2	5	0.0425	0.0455	0.0435	0.0000
-0.5	0.3	1	2	5	0.0755	0.0820	0.0575	0.0000
-1	0.1	0.25	1	1	0.0460	0.0515	0.0255	0.0225
-1	0.2	0.25	1	1	0.0485	0.0550	0.0220	0.0265
-1	0.3	0.25	1	1	0.0500	0.0550	0.0225	0.0190
-1	0.1	0.5	1	1	0.0480	0.0575	0.0225	0.0175
-1	0.2	0.5	1	1	0.0510	0.0635	0.0160	0.0215
-1	0.3	0.5	1	1	0.0605	0.0740	0.0130	0.0090
-1	0.1	0.75	1	1	0.0485	0.0540	0.0220	0.0170
-1	0.2	0.75	1	1	0.0550	0.0660	0.0150	0.0150
-1	0.3	0.75	1	1	0.0685	0.0855	0.0115	0.0075
-1	0.1	1	1	1	0.0530	0.0600	0.0285	0.0110
-1	0.2	1	1	1	0.0660	0.0750	0.0160	0.0060
-1	0.3	1	1	1	0.0895	0.1100	0.0120	0.0020
-1	0.1	0.25	1	2	0.0485	0.0585	0.0075	0.0225
-1	0.2	0.25	1	2	0.0510	0.0605	0.0090	0.0290
-1	0.3	0.25	1	2	0.0545	0.0650	0.0060	0.0255
-1	0.1	0.5	1	2	0.0520	0.0595	0.0070	0.0235
-1	0.2	0.5	1	2	0.0550	0.0680	0.0080	0.0220
-1	0.3	0.5	1	2	0.0665	0.0770	0.0100	0.0250
-1	0.1	0.75	1	2	0.0535	0.0665	0.0075	0.0220
-1	0.2	0.75	1	2	0.0600	0.0730	0.0100	0.0150
-1	0.3	0.75	1	2	0.0805	0.1015	0.0130	0.0155
-1	0.1	1	1	2	0.0585	0.0675	0.0090	0.0175
-1	0.2	1	1	2	0.0660	0.0860	0.0105	0.0055
-1	0.3	1	1	2	0.0985	0.1340	0.0130	0.0090
-1	0.1	0.25	1	3	0.0500	0.0560	0.0115	0.0205
-1	0.2	0.25	1	3	0.0495	0.0600	0.0085	0.0130
-1	0.3	0.25	1	3	0.0560	0.0665	0.0070	0.0180
-1	0.1	0.5	1	3	0.0485	0.0580	0.0090	0.0140
-1	0.2	0.5	1	3	0.0520	0.0670	0.0080	0.0115
-1	0.3	0.5	1	3	0.0630	0.0915	0.0070	0.0235
-1	0.1	0.75	1	3	0.0530	0.0640	0.0095	0.0180

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
-1	0.2	0.75	1	3	0.0610	0.0760	0.0070	0.0070
-1	0.3	0.75	1	3	0.0930	0.1260	0.0085	0.0165
-1	0.1	1	1	3	0.0540	0.0690	0.0090	0.0075
-1	0.2	1	1	3	0.0695	0.0915	0.0090	0.0075
-1	0.3	1	1	3	0.1425	0.1840	0.0170	0.0090
-1	0.1	0.25	1	4	0.0470	0.0495	0.0275	0.0170
-1	0.2	0.25	1	4	0.0440	0.0495	0.0220	0.0115
-1	0.3	0.25	1	4	0.0455	0.0525	0.0250	0.0095
-1	0.1	0.5	1	4	0.0465	0.0495	0.0250	0.0115
-1	0.2	0.5	1	4	0.0470	0.0495	0.0145	0.0085
-1	0.3	0.5	1	4	0.0500	0.0550	0.0110	0.0080
-1	0.1	0.75	1	4	0.0480	0.0480	0.0230	0.0105
-1	0.2	0.75	1	4	0.0520	0.0540	0.0135	0.0035
-1	0.3	0.75	1	4	0.0590	0.0625	0.0110	0.0040
-1	0.1	1	1	4	0.0440	0.0495	0.0225	0.0035
-1	0.2	1	1	4	0.0520	0.0590	0.0140	0.0010
-1	0.3	1	1	4	0.0650	0.0775	0.0135	0.0000
-1	0.1	0.25	1	5	0.0400	0.0435	0.0470	0.0230
-1	0.2	0.25	1	5	0.0415	0.0415	0.0400	0.0165
-1	0.3	0.25	1	5	0.0465	0.0485	0.0535	0.0185
-1	0.1	0.5	1	5	0.0415	0.0425	0.0470	0.0190
-1	0.2	0.5	1	5	0.0410	0.0440	0.0370	0.0085
-1	0.3	0.5	1	5	0.0430	0.0395	0.0295	0.0095
-1	0.1	0.75	1	5	0.0430	0.0450	0.0420	0.0105
-1	0.2	0.75	1	5	0.0420	0.0455	0.0305	0.0045
-1	0.3	0.75	1	5	0.0435	0.0470	0.0225	0.0030
-1	0.1	1	1	5	0.0425	0.0435	0.0430	0.0040
-1	0.2	1	1	5	0.0415	0.0455	0.0290	0.0010
-1	0.3	1	1	5	0.0500	0.0570	0.0250	0.0000
-1	0.1	0.25	2	1	0.0640	0.1150	0.0235	0.0050
-1	0.2	0.25	2	1	0.0890	0.1455	0.0640	0.0025
-1	0.3	0.25	2	1	0.0930	0.1480	0.0465	0.0040
-1	0.1	0.5	2	1	0.0715	0.1190	0.0320	0.0040
-1	0.2	0.5	2	1	0.1225	0.1845	0.0930	0.0050
-1	0.3	0.5	2	1	0.1345	0.2075	0.0925	0.0045

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
-1	0.1	0.75	2	1	0.0865	0.1395	0.0335	0.0030
-1	0.2	0.75	2	1	0.1630	0.2460	0.1200	0.0020
-1	0.3	0.75	2	1	0.2160	0.3025	0.1450	0.0035
-1	0.1	1	2	1	0.1050	0.1640	0.0375	0.0030
-1	0.2	1	2	1	0.2105	0.2970	0.1315	0.0015
-1	0.3	1	2	1	0.3105	0.3855	0.1695	0.0010
-1	0.1	0.25	2	2	0.0820	0.1480	0.0450	0.1265
-1	0.2	0.25	2	2	0.0780	0.1380	0.0315	0.1410
-1	0.3	0.25	2	2	0.1030	0.1660	0.0440	0.1485
-1	0.1	0.5	2	2	0.0875	0.1685	0.0490	0.1225
-1	0.2	0.5	2	2	0.1070	0.1785	0.0440	0.1400
-1	0.3	0.5	2	2	0.1560	0.2350	0.0760	0.1645
-1	0.1	0.75	2	2	0.1050	0.1845	0.0535	0.1065
-1	0.2	0.75	2	2	0.1595	0.2250	0.0565	0.1145
-1	0.3	0.75	2	2	0.2515	0.3345	0.1120	0.1500
-1	0.1	1	2	2	0.1320	0.2125	0.0625	0.0835
-1	0.2	1	2	2	0.2130	0.2910	0.0660	0.0900
-1	0.3	1	2	2	0.3765	0.4700	0.1390	0.1195
-1	0.1	0.25	2	3	0.0580	0.1015	0.0225	0.1070
-1	0.2	0.25	2	3	0.0695	0.1140	0.0290	0.1185
-1	0.3	0.25	2	3	0.0785	0.1355	0.0385	0.1385
-1	0.1	0.5	2	3	0.0710	0.1165	0.0265	0.1045
-1	0.2	0.5	2	3	0.0950	0.1420	0.0450	0.1370
-1	0.3	0.5	2	3	0.1450	0.2295	0.0650	0.1785
-1	0.1	0.75	2	3	0.0790	0.1375	0.0290	0.0945
-1	0.2	0.75	2	3	0.1335	0.2060	0.0580	0.1235
-1	0.3	0.75	2	3	0.2310	0.3190	0.0985	0.1495
-1	0.1	1	2	3	0.0875	0.1480	0.0285	0.0740
-1	0.2	1	2	3	0.1770	0.2575	0.0710	0.0905
-1	0.3	1	2	3	0.3700	0.4595	0.1375	0.1360
-1	0.1	0.25	2	4	0.0395	0.0545	0.0690	0.0010
-1	0.2	0.25	2	4	0.0375	0.0475	0.0900	0.0005
-1	0.3	0.25	2	4	0.0500	0.0785	0.0540	0.0015
-1	0.1	0.5	2	4	0.0425	0.0615	0.0680	0.0010
-1	0.2	0.5	2	4	0.0505	0.0695	0.1015	0.0005

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
-1	0.3	0.5	2	4	0.0715	0.1025	0.0865	0.0010
-1	0.1	0.75	2	4	0.0445	0.0675	0.0770	0.0010
-1	0.2	0.75	2	4	0.0520	0.0795	0.1180	0.0005
-1	0.3	0.75	2	4	0.1105	0.1565	0.1205	0.0010
-1	0.1	1	2	4	0.0465	0.0685	0.0795	0.0000
-1	0.2	1	2	4	0.0880	0.1155	0.1245	0.0000
-1	0.3	1	2	4	0.1705	0.2255	0.1290	0.0000
-1	0.1	0.25	2	5	0.0240	0.0180	0.0855	0.0020
-1	0.2	0.25	2	5	0.0205	0.0180	0.0855	0.0015
-1	0.3	0.25	2	5	0.0220	0.0175	0.0925	0.0010
-1	0.1	0.5	2	5	0.0225	0.0165	0.1040	0.0010
-1	0.2	0.5	2	5	0.0195	0.0190	0.1125	0.0000
-1	0.3	0.5	2	5	0.0200	0.0165	0.1075	0.0000
-1	0.1	0.75	2	5	0.0200	0.0175	0.0915	0.0000
-1	0.2	0.75	2	5	0.0185	0.0145	0.1300	0.0000
-1	0.3	0.75	2	5	0.0150	0.0120	0.1420	0.0000
-1	0.1	1	2	5	0.0210	0.0170	0.0960	0.0000
-1	0.2	1	2	5	0.0155	0.0160	0.1310	0.0000
-1	0.3	1	2	5	0.0165	0.0195	0.1540	0.0000

Note: The columns are defined as follows: theta = ability mean, itemcont = amount of item contamination, bdiff = size of the DIF, a = discrimination stratum, b = difficulty stratum, MH1 = Mantel-Haenszel with small intervals, MH2 = Mantel-Haenszel with large intervals, DIPF = differential item pair functioning, RI = relative item performance method

Table A6

Median Power Across All Conditions

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
-1	0.1	0.25	1	1	0.0640	0.0680	0.0550	0.0440
-1	0.1	0.25	1	2	0.0700	0.1275	0.1010	0.2485
-1	0.1	0.25	1	4	0.0300	0.0850	0.0615	0.0540
-1	0.1	0.25	1	5	0.0535	0.0720	0.0605	0.0620
-1	0.1	0.25	2	1	0.0150	0.0560	0.0385	0.0140
-1	0.1	0.25	2	2	0.0085	0.0925	0.0585	0.0060
-1	0.1	0.25	2	4	0.0030	0.0780	0.0665	0.0150
-1	0.1	0.25	2	5	0.0030	0.0780	0.0555	0.0160
-1	0.1	0.5	1	1	0.0960	0.1365	0.0965	0.1310
-1	0.1	0.5	1	2	0.1040	0.2025	0.1625	0.3405
-1	0.1	0.5	1	4	0.0580	0.1675	0.1200	0.1120
-1	0.1	0.5	1	5	0.0890	0.1530	0.1230	0.1400
-1	0.1	0.5	2	1	0.0645	0.1500	0.0955	0.0480
-1	0.1	0.5	2	2	0.0185	0.3065	0.2010	0.0335
-1	0.1	0.5	2	4	0.0085	0.2675	0.1960	0.0725
-1	0.1	0.5	2	5	0.0195	0.2350	0.1820	0.0570
-1	0.1	0.75	1	1	0.1410	0.2375	0.1905	0.2510
-1	0.1	0.75	1	2	0.1345	0.2770	0.2335	0.4295
-1	0.1	0.75	1	4	0.1475	0.3415	0.2825	0.3055
-1	0.1	0.75	1	5	0.1245	0.2645	0.2230	0.2070
-1	0.1	0.75	2	1	0.1375	0.3110	0.2360	0.1245
-1	0.1	0.75	2	2	0.0780	0.6215	0.4915	0.2065
-1	0.1	0.75	2	4	0.0510	0.5745	0.4895	0.2710
-1	0.1	0.75	2	5	0.0605	0.5005	0.4085	0.1515
-1	0.1	1	1	1	0.2155	0.3580	0.2930	0.4860
-1	0.1	1	1	2	0.1740	0.3645	0.3195	0.5365
-1	0.1	1	1	4	0.2435	0.5880	0.5015	0.5510
-1	0.1	1	1	5	0.2000	0.4660	0.4180	0.4085
-1	0.1	1	2	1	0.1825	0.4880	0.4210	0.3035
-1	0.1	1	2	2	0.2305	0.8545	0.7700	0.5520
-1	0.1	1	2	4	0.1775	0.8300	0.7775	0.5750
-1	0.1	1	2	5	0.1865	0.7635	0.6970	0.4320

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
-1	0.2	0.25	1	1	0.0295	0.0700	0.0535	0.0705
-1	0.2	0.25	1	2	0.0040	0.0850	0.0570	0.0355
-1	0.2	0.25	1	4	0.0060	0.0695	0.0510	0.0525
-1	0.2	0.25	1	5	0.0730	0.0645	0.0525	0.1290
-1	0.2	0.25	2	1	0.0140	0.0755	0.0520	0.0075
-1	0.2	0.25	2	2	0.0260	0.0705	0.0490	0.0040
-1	0.2	0.25	2	4	0.0045	0.0770	0.0525	0.0085
-1	0.2	0.25	2	5	0.0075	0.0560	0.0415	0.0370
-1	0.2	0.5	1	1	0.0615	0.1180	0.0900	0.1235
-1	0.2	0.5	1	2	0.0140	0.1590	0.1200	0.0950
-1	0.2	0.5	1	4	0.0295	0.1445	0.1075	0.1160
-1	0.2	0.5	1	5	0.1130	0.1115	0.0955	0.1885
-1	0.2	0.5	2	1	0.0250	0.1760	0.1200	0.0345
-1	0.2	0.5	2	2	0.0105	0.2805	0.1770	0.0100
-1	0.2	0.5	2	4	0.0100	0.2335	0.1595	0.0205
-1	0.2	0.5	2	5	0.0195	0.1505	0.1180	0.0490
-1	0.2	0.75	1	1	0.0995	0.1795	0.1495	0.1940
-1	0.2	0.75	1	2	0.0815	0.2940	0.2305	0.2220
-1	0.2	0.75	1	4	0.0895	0.2795	0.2320	0.2390
-1	0.2	0.75	1	5	0.1070	0.1580	0.1305	0.2185
-1	0.2	0.75	2	1	0.0475	0.3325	0.2450	0.1245
-1	0.2	0.75	2	2	0.0245	0.5675	0.4355	0.0975
-1	0.2	0.75	2	4	0.0275	0.5350	0.4320	0.0925
-1	0.2	0.75	2	5	0.0260	0.3485	0.2870	0.1355
-1	0.2	1	1	1	0.0985	0.2930	0.2560	0.3675
-1	0.2	1	1	2	0.1095	0.4610	0.3980	0.4110
-1	0.2	1	1	4	0.1110	0.4300	0.3730	0.4015
-1	0.2	1	1	5	0.1010	0.2410	0.2085	0.3455
-1	0.2	1	2	1	0.0480	0.5075	0.4285	0.3080
-1	0.2	1	2	2	0.0655	0.8250	0.7330	0.3455
-1	0.2	1	2	4	0.0910	0.8160	0.7435	0.4365
-1	0.2	1	2	5	0.0490	0.5620	0.4935	0.3115
-1	0.3	0.25	1	1	0.0580	0.0615	0.0535	0.0485
-1	0.3	0.25	1	2	0.0240	0.0725	0.0515	0.0310
-1	0.3	0.25	1	3	0.0370	0.0740	0.0570	0.0610

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
-1	0.3	0.25	1	4	0.0135	0.0635	0.0500	0.0370
-1	0.3	0.25	1	5	0.0140	0.0635	0.0520	0.0360
-1	0.3	0.25	2	1	0.0090	0.0545	0.0545	0.0105
-1	0.3	0.25	2	2	0.0145	0.0755	0.0470	0.0085
-1	0.3	0.25	2	3	0.0225	0.0710	0.0495	0.0050
-1	0.3	0.25	2	4	0.0025	0.0625	0.0405	0.0180
-1	0.3	0.25	2	5	0.0030	0.0560	0.0345	0.0515
-1	0.3	0.5	1	1	0.0780	0.0990	0.0810	0.0850
-1	0.3	0.5	1	2	0.0360	0.1275	0.0985	0.0755
-1	0.3	0.5	1	3	0.0630	0.1340	0.1060	0.1030
-1	0.3	0.5	1	4	0.0365	0.1155	0.0935	0.0605
-1	0.3	0.5	1	5	0.0220	0.1135	0.0865	0.0565
-1	0.3	0.5	2	1	0.0105	0.1255	0.0880	0.0155
-1	0.3	0.5	2	2	0.0025	0.2370	0.1360	0.0060
-1	0.3	0.5	2	3	0.0050	0.2355	0.1520	0.0125
-1	0.3	0.5	2	4	0.0025	0.1620	0.1060	0.0150
-1	0.3	0.5	2	5	0.0075	0.1450	0.1065	0.0380
-1	0.3	0.75	1	1	0.0690	0.1595	0.1335	0.1395
-1	0.3	0.75	1	2	0.0540	0.2195	0.1755	0.1530
-1	0.3	0.75	1	3	0.0645	0.2120	0.1740	0.1630
-1	0.3	0.75	1	4	0.0510	0.2180	0.1810	0.1360
-1	0.3	0.75	1	5	0.0245	0.2330	0.1905	0.1100
-1	0.3	0.75	2	1	0.0175	0.2910	0.2195	0.0650
-1	0.3	0.75	2	2	0.0105	0.5305	0.3915	0.0440
-1	0.3	0.75	2	3	0.0160	0.5350	0.4265	0.0700
-1	0.3	0.75	2	4	0.0140	0.4250	0.3480	0.0655
-1	0.3	0.75	2	5	0.0090	0.3150	0.2555	0.0870
-1	0.3	1	1	1	0.0375	0.2235	0.1950	0.2760
-1	0.3	1	1	2	0.0510	0.3625	0.3170	0.3145
-1	0.3	1	1	3	0.0640	0.3365	0.2940	0.2995
-1	0.3	1	1	4	0.0560	0.3630	0.3210	0.3030
-1	0.3	1	1	5	0.0260	0.3805	0.3180	0.2410
-1	0.3	1	2	1	0.0080	0.4675	0.3815	0.2345
-1	0.3	1	2	2	0.0190	0.7620	0.6575	0.2290
-1	0.3	1	2	3	0.0505	0.7985	0.7225	0.3420

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
-1	0.3	1	2	4	0.0310	0.7055	0.6245	0.2765
-1	0.3	1	2	5	0.0105	0.5680	0.5040	0.2235
-0.5	0.1	0.25	1	1	0.0785	0.0690	0.0590	0.0320
-0.5	0.1	0.25	1	2	0.1255	0.0930	0.0850	0.0730
-0.5	0.1	0.25	1	4	0.0845	0.0810	0.0695	0.0270
-0.5	0.1	0.25	1	5	0.0605	0.0725	0.0650	0.0345
-0.5	0.1	0.25	2	1	0.0315	0.0430	0.0360	0.0230
-0.5	0.1	0.25	2	2	0.0435	0.0940	0.0735	0.0110
-0.5	0.1	0.25	2	4	0.0375	0.1110	0.0905	0.0175
-0.5	0.1	0.25	2	5	0.0235	0.1035	0.0835	0.0205
-0.5	0.1	0.5	1	1	0.1535	0.1300	0.1120	0.0840
-0.5	0.1	0.5	1	2	0.1410	0.1235	0.1140	0.1035
-0.5	0.1	0.5	1	4	0.1690	0.1875	0.1630	0.0935
-0.5	0.1	0.5	1	5	0.1330	0.1920	0.1630	0.0855
-0.5	0.1	0.5	2	1	0.0555	0.1065	0.0965	0.1370
-0.5	0.1	0.5	2	2	0.1370	0.3240	0.2695	0.0835
-0.5	0.1	0.5	2	4	0.1260	0.3495	0.2940	0.0990
-0.5	0.1	0.5	2	5	0.1160	0.3330	0.2860	0.0930
-0.5	0.1	0.75	1	1	0.2125	0.2335	0.2125	0.1930
-0.5	0.1	0.75	1	2	0.1640	0.1805	0.1680	0.1555
-0.5	0.1	0.75	1	4	0.2845	0.3705	0.3340	0.2280
-0.5	0.1	0.75	1	5	0.1890	0.3150	0.2915	0.2095
-0.5	0.1	0.75	2	1	0.0580	0.2055	0.1815	0.2595
-0.5	0.1	0.75	2	2	0.3550	0.6115	0.5640	0.3105
-0.5	0.1	0.75	2	4	0.3765	0.6830	0.6445	0.3620
-0.5	0.1	0.75	2	5	0.2715	0.6545	0.6110	0.3205
-0.5	0.1	1	1	1	0.2440	0.3335	0.3110	0.3300
-0.5	0.1	1	1	2	0.1895	0.2755	0.2560	0.2475
-0.5	0.1	1	1	4	0.4330	0.6230	0.5920	0.4585
-0.5	0.1	1	1	5	0.3165	0.5380	0.5065	0.4235
-0.5	0.1	1	2	1	0.0415	0.2750	0.2620	0.4150
-0.5	0.1	1	2	2	0.5435	0.8395	0.8055	0.6095
-0.5	0.1	1	2	4	0.6350	0.8720	0.8570	0.6395
-0.5	0.1	1	2	5	0.5470	0.8770	0.8540	0.6780
-0.5	0.2	0.25	1	1	0.0645	0.0635	0.0575	0.0280

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
-0.5	0.2	0.25	1	2	0.0330	0.0745	0.0630	0.0260
-0.5	0.2	0.25	1	4	0.0465	0.0675	0.0660	0.0260
-0.5	0.2	0.25	1	5	0.0930	0.0595	0.0550	0.0415
-0.5	0.2	0.25	2	1	0.0310	0.0580	0.0505	0.0150
-0.5	0.2	0.25	2	2	0.0270	0.0975	0.0770	0.0075
-0.5	0.2	0.25	2	4	0.0355	0.1020	0.0810	0.0085
-0.5	0.2	0.25	2	5	0.0180	0.0785	0.0665	0.0225
-0.5	0.2	0.5	1	1	0.0730	0.1115	0.0970	0.0690
-0.5	0.2	0.5	1	2	0.0725	0.1485	0.1260	0.0630
-0.5	0.2	0.5	1	4	0.1135	0.1660	0.1515	0.0705
-0.5	0.2	0.5	1	5	0.1090	0.0975	0.0850	0.0555
-0.5	0.2	0.5	2	1	0.0575	0.1375	0.1155	0.0655
-0.5	0.2	0.5	2	2	0.0745	0.3110	0.2560	0.0635
-0.5	0.2	0.5	2	4	0.0885	0.3330	0.2870	0.0610
-0.5	0.2	0.5	2	5	0.0440	0.2290	0.1940	0.0715
-0.5	0.2	0.75	1	1	0.0920	0.1740	0.1575	0.1265
-0.5	0.2	0.75	1	2	0.1470	0.2905	0.2625	0.1730
-0.5	0.2	0.75	1	4	0.1305	0.3115	0.2765	0.1650
-0.5	0.2	0.75	1	5	0.1035	0.1740	0.1575	0.1095
-0.5	0.2	0.75	2	1	0.0445	0.2890	0.2605	0.1765
-0.5	0.2	0.75	2	2	0.1365	0.5985	0.5375	0.2390
-0.5	0.2	0.75	2	4	0.2210	0.6645	0.6180	0.2860
-0.5	0.2	0.75	2	5	0.0850	0.4835	0.4505	0.2415
-0.5	0.2	1	1	1	0.0655	0.2835	0.2685	0.2265
-0.5	0.2	1	1	2	0.1375	0.4505	0.4280	0.3580
-0.5	0.2	1	1	4	0.1820	0.4890	0.4590	0.3655
-0.5	0.2	1	1	5	0.0995	0.2615	0.2430	0.2230
-0.5	0.2	1	2	1	0.0190	0.4540	0.4200	0.3395
-0.5	0.2	1	2	2	0.1720	0.8190	0.7805	0.5755
-0.5	0.2	1	2	4	0.4290	0.9090	0.8790	0.6825
-0.5	0.2	1	2	5	0.1530	0.7935	0.7630	0.5885
-0.5	0.3	0.25	1	1	0.0790	0.0550	0.0505	0.0215
-0.5	0.3	0.25	1	2	0.0575	0.0675	0.0580	0.0195
-0.5	0.3	0.25	1	3	0.0790	0.0690	0.0615	0.0220
-0.5	0.3	0.25	1	4	0.0395	0.0650	0.0600	0.0170

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
-0.5	0.3	0.25	1	5	0.0250	0.0695	0.0635	0.0225
-0.5	0.3	0.25	2	1	0.0180	0.0590	0.0505	0.0115
-0.5	0.3	0.25	2	2	0.0100	0.0700	0.0510	0.0040
-0.5	0.3	0.25	2	3	0.0360	0.1010	0.0795	0.0060
-0.5	0.3	0.25	2	4	0.0150	0.0920	0.0790	0.0085
-0.5	0.3	0.25	2	5	0.0105	0.0850	0.0630	0.0225
-0.5	0.3	0.5	1	1	0.0630	0.0850	0.0795	0.0505
-0.5	0.3	0.5	1	2	0.0805	0.1110	0.1030	0.0490
-0.5	0.3	0.5	1	3	0.0920	0.1175	0.1045	0.0435
-0.5	0.3	0.5	1	4	0.0720	0.1260	0.1150	0.0425
-0.5	0.3	0.5	1	5	0.0355	0.1470	0.1325	0.0500
-0.5	0.3	0.5	2	1	0.0185	0.1305	0.1195	0.0545
-0.5	0.3	0.5	2	2	0.0305	0.2515	0.2170	0.0445
-0.5	0.3	0.5	2	3	0.0725	0.3030	0.2525	0.0410
-0.5	0.3	0.5	2	4	0.0495	0.2655	0.2270	0.0450
-0.5	0.3	0.5	2	5	0.0250	0.2300	0.1955	0.0695
-0.5	0.3	0.75	1	1	0.0335	0.1180	0.1060	0.0875
-0.5	0.3	0.75	1	2	0.0705	0.2040	0.1875	0.0860
-0.5	0.3	0.75	1	3	0.0930	0.1810	0.1715	0.0900
-0.5	0.3	0.75	1	4	0.0695	0.2295	0.2105	0.1045
-0.5	0.3	0.75	1	5	0.0385	0.2755	0.2520	0.1325
-0.5	0.3	0.75	2	1	0.0100	0.2400	0.2180	0.1375
-0.5	0.3	0.75	2	2	0.0395	0.4925	0.4430	0.1990
-0.5	0.3	0.75	2	3	0.1580	0.5875	0.5580	0.2110
-0.5	0.3	0.75	2	4	0.0980	0.5465	0.5045	0.2125
-0.5	0.3	0.75	2	5	0.0315	0.4570	0.4310	0.2195
-0.5	0.3	1	1	1	0.0145	0.1895	0.1765	0.1595
-0.5	0.3	1	1	2	0.0470	0.3395	0.3265	0.2335
-0.5	0.3	1	1	3	0.0780	0.2925	0.2760	0.1685
-0.5	0.3	1	1	4	0.0745	0.3755	0.3630	0.2500
-0.5	0.3	1	1	5	0.0425	0.4770	0.4555	0.3215
-0.5	0.3	1	2	1	0.0025	0.3660	0.3465	0.3120
-0.5	0.3	1	2	2	0.0260	0.6920	0.6590	0.4790
-0.5	0.3	1	2	3	0.2260	0.8590	0.8335	0.5760
-0.5	0.3	1	2	4	0.1895	0.8225	0.7960	0.5735

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
-0.5	0.3	1	2	5	0.0490	0.7570	0.7355	0.5570
0	0.1	0.25	1	1	0.0765	0.0620	0.0635	0.0135
0	0.1	0.25	1	2	0.0905	0.0670	0.0650	0.0125
0	0.1	0.25	1	4	0.1115	0.0890	0.0920	0.0125
0	0.1	0.25	1	5	0.0600	0.0775	0.0790	0.0175
0	0.1	0.25	2	1	0.0005	0.0205	0.0230	0.0735
0	0.1	0.25	2	2	0.0965	0.1130	0.1075	0.0255
0	0.1	0.25	2	4	0.1455	0.1430	0.1465	0.0340
0	0.1	0.25	2	5	0.0755	0.1500	0.1520	0.0350
0	0.1	0.5	1	1	0.0950	0.1100	0.1090	0.0270
0	0.1	0.5	1	2	0.0840	0.0855	0.0845	0.0180
0	0.1	0.5	1	4	0.2170	0.2125	0.2185	0.0470
0	0.1	0.5	1	5	0.1030	0.1900	0.1955	0.0555
0	0.1	0.5	2	1	0.0005	0.0350	0.0350	0.2065
0	0.1	0.5	2	2	0.2675	0.3205	0.3210	0.1310
0	0.1	0.5	2	4	0.3920	0.4020	0.4065	0.1495
0	0.1	0.5	2	5	0.2740	0.4475	0.4500	0.1820
0	0.1	0.75	1	1	0.1175	0.2115	0.2125	0.0750
0	0.1	0.75	1	2	0.0830	0.1170	0.1235	0.0285
0	0.1	0.75	1	4	0.3655	0.4145	0.4195	0.1375
0	0.1	0.75	1	5	0.1990	0.3685	0.3685	0.1570
0	0.1	0.75	2	1	0.0000	0.0430	0.0460	0.4000
0	0.1	0.75	2	2	0.4405	0.5740	0.5760	0.3420
0	0.1	0.75	2	4	0.6855	0.7020	0.7035	0.4040
0	0.1	0.75	2	5	0.5950	0.7855	0.7855	0.5460
0	0.1	1	1	1	0.1010	0.2915	0.3000	0.1265
0	0.1	1	1	2	0.0695	0.1715	0.1770	0.0525
0	0.1	1	1	4	0.5215	0.6350	0.6440	0.3395
0	0.1	1	1	5	0.2950	0.5865	0.5890	0.3235
0	0.1	1	2	1	0.0000	0.0560	0.0660	0.4980
0	0.1	1	2	2	0.5800	0.7990	0.8025	0.6015
0	0.1	1	2	4	0.8915	0.9125	0.9115	0.7260
0	0.1	1	2	5	0.8390	0.9445	0.9450	0.8320
0	0.2	0.25	1	1	0.0490	0.0535	0.0565	0.0135
0	0.2	0.25	1	2	0.0410	0.0650	0.0655	0.0115

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
0	0.2	0.25	1	4	0.1150	0.0825	0.0870	0.0140
0	0.2	0.25	1	5	0.0815	0.0675	0.0705	0.0070
0	0.2	0.25	2	1	0.0240	0.0535	0.0565	0.0340
0	0.2	0.25	2	2	0.0800	0.0990	0.1050	0.0405
0	0.2	0.25	2	4	0.1120	0.1455	0.1515	0.0210
0	0.2	0.25	2	5	0.0350	0.1285	0.1270	0.0420
0	0.2	0.5	1	1	0.0360	0.0930	0.0960	0.0270
0	0.2	0.5	1	2	0.0495	0.1325	0.1370	0.0360
0	0.2	0.5	1	4	0.1935	0.1700	0.1725	0.0325
0	0.2	0.5	1	5	0.0660	0.0905	0.0950	0.0145
0	0.2	0.5	2	1	0.0245	0.1065	0.1085	0.0935
0	0.2	0.5	2	2	0.1165	0.2775	0.2835	0.1285
0	0.2	0.5	2	4	0.2545	0.4395	0.4355	0.1515
0	0.2	0.5	2	5	0.0890	0.3300	0.3370	0.1540
0	0.2	0.75	1	1	0.0230	0.1515	0.1610	0.0525
0	0.2	0.75	1	2	0.0950	0.2550	0.2625	0.0995
0	0.2	0.75	1	4	0.1385	0.3060	0.3155	0.0850
0	0.2	0.75	1	5	0.0630	0.1690	0.1740	0.0330
0	0.2	0.75	2	1	0.0045	0.1970	0.2145	0.2200
0	0.2	0.75	2	2	0.1260	0.5540	0.5665	0.3780
0	0.2	0.75	2	4	0.5875	0.7650	0.7680	0.4780
0	0.2	0.75	2	5	0.2010	0.6510	0.6510	0.4570
0	0.2	1	1	1	0.0115	0.2200	0.2300	0.1080
0	0.2	1	1	2	0.0700	0.3945	0.4040	0.2195
0	0.2	1	1	4	0.1720	0.5365	0.5440	0.2655
0	0.2	1	1	5	0.0555	0.2665	0.2750	0.0860
0	0.2	1	2	1	0.0000	0.2970	0.3095	0.3575
0	0.2	1	2	2	0.1170	0.7740	0.7825	0.6725
0	0.2	1	2	4	0.7645	0.9475	0.9535	0.8590
0	0.2	1	2	5	0.3495	0.8905	0.9010	0.8030
0	0.3	0.25	1	1	0.0530	0.0540	0.0545	0.0110
0	0.3	0.25	1	2	0.0830	0.0665	0.0690	0.0090
0	0.3	0.25	1	3	0.0920	0.0605	0.0615	0.0080
0	0.3	0.25	1	4	0.0805	0.0680	0.0715	0.0100
0	0.3	0.25	1	5	0.0280	0.0795	0.0810	0.0200

theta	itemcont	bdiff	a	b	MH1	MH2	DIPF	RI
0	0.3	0.25	2	1	0.0070	0.0455	0.0470	0.0300
0	0.3	0.25	2	2	0.0240	0.0920	0.1010	0.0365
0	0.3	0.25	2	3	0.1310	0.1175	0.1165	0.0170
0	0.3	0.25	2	4	0.0710	0.1145	0.1185	0.0205
0	0.3	0.25	2	5	0.0365	0.1095	0.1135	0.0445
0	0.3	0.5	1	1	0.0210	0.0685	0.0725	0.0150
0	0.3	0.5	1	2	0.0670	0.1065	0.1115	0.0275
0	0.3	0.5	1	3	0.0935	0.1075	0.1075	0.0180
0	0.3	0.5	1	4	0.0805	0.1465	0.1560	0.0235
0	0.3	0.5	1	5	0.0365	0.1760	0.1800	0.0495
0	0.3	0.5	2	1	0.0020	0.0780	0.0835	0.1225
0	0.3	0.5	2	2	0.0280	0.2365	0.2350	0.1385
0	0.3	0.5	2	3	0.2610	0.3420	0.3540	0.1135
0	0.3	0.5	2	4	0.2130	0.3730	0.3740	0.1210
0	0.3	0.5	2	5	0.0405	0.3115	0.3230	0.1660
0	0.3	0.75	1	1	0.0055	0.0965	0.1030	0.0390
0	0.3	0.75	1	2	0.0465	0.1745	0.1855	0.0640
0	0.3	0.75	1	3	0.0795	0.1710	0.1800	0.0375
0	0.3	0.75	1	4	0.0925	0.2575	0.2705	0.0660
0	0.3	0.75	1	5	0.0450	0.3435	0.3620	0.1510
0	0.3	0.75	2	1	0.0000	0.1330	0.1420	0.2590
0	0.3	0.75	2	2	0.0115	0.3870	0.4000	0.3475
0	0.3	0.75	2	3	0.3365	0.6230	0.6415	0.3795
0	0.3	0.75	2	4	0.3610	0.6865	0.6985	0.3925
0	0.3	0.75	2	5	0.0790	0.6315	0.6480	0.4855
0	0.3	1	1	1	0.0015	0.1370	0.1410	0.0985
0	0.3	1	1	2	0.0165	0.2700	0.2940	0.1545
0	0.3	1	1	3	0.0575	0.2785	0.2985	0.0945
0	0.3	1	1	4	0.1045	0.4210	0.4360	0.1760
0	0.3	1	1	5	0.0495	0.5675	0.5875	0.4035
0	0.3	1	2	1	0.0000	0.1785	0.1995	0.3775
0	0.3	1	2	2	0.0040	0.5405	0.5745	0.6020
0	0.3	1	2	3	0.3860	0.8690	0.8875	0.7780
0	0.3	1	2	4	0.5625	0.8860	0.8995	0.7505
0	0.3	1	2	5	0.1305	0.8885	0.8940	0.8355

Note: The columns are defined as follows: theta = ability mean, itemcont = amount of item contamination, bdiff = size of the DIF, a = discrimination stratum, b = difficulty stratum, MH1 = Mantel-Haenszel with small intervals, MH2 = Mantel-Haenszel with large intervals, DIPF = differential item pair functioning, RI = relative item performance method

Appendix B

Code for Simulation Study

```
library(difR)
library(plRasch)

replication <- 1000
results_type1 <- data.frame(ni=numeric(0),fsize=numeric(0),itemcont=numeric(0),
  theta=numeric(0), bdiff=numeric(0), aref=numeric(0), bref=numeric(0),
  bfoc=numeric(0), type1_mh1=numeric(0), type1_mh2=numeric(0),
  type1_dipf=numeric(0),type1_ri=numeric(0))

results_power <- data.frame(ni=numeric(0),fsize=numeric(0),itemcont=numeric(0),
  theta=numeric(0), bdiff=numeric(0),aref=numeric(0),bref=numeric(0),
  bfoc=numeric(0), power_mh1=numeric(0), power_mh2=numeric(0),
  power_dipf=numeric(0), power_ri=numeric(0))

results_flag <- data.frame(ni=numeric(0),fsize=numeric(0),itemcont=numeric(0),
  theta=numeric(0), bdiff=numeric(0), aref=numeric(0), bref=numeric(0),
  bfoc=numeric(0), flag_mh1=numeric(0), flag_mh2=numeric(0),
  flag_dipf=numeric(0),flag_ri=numeric(0),flag_DIF=numeric(0))

#set number of test items and item parameters
for (ni in c(20,40,80)){
  aref <- c(rep(NA, ni)) #discrimination parameter
  bref <- c(rep(NA, ni)) #difficulty parameter

  bref[1:(ni/5)] <- runif(ni/5,-2.5,-1.5)
  bref[((ni/5)+1):(2*ni/5)] <-runif(ni/5,-1.5,-0.5)
  bref[((2*ni/5)+1):(3*ni/5)] <-runif(ni/5,-0.5,0.5)
  bref[((3*ni/5)+1):(4*ni/5)] <-runif(ni/5,0.5,1.5)
  bref[((4*ni/5)+1):ni] <-runif(ni/5,1.5,2.5)

  aref[1:(ni/10)] <- runif(ni/10,0.2,1)
  aref[((ni/10)+1):(2*ni/10)] <- runif(ni/10,1,2)
  aref[((2*ni/10)+1):(3*ni/10)] <- runif(ni/10,0.2,1)
  aref[((3*ni/10)+1):(4*ni/10)] <- runif(ni/10,1,2)
  aref[((4*ni/10)+1):(5*ni/10)] <- runif(ni/10,0.2,1)
  aref[((5*ni/10)+1):(6*ni/10)] <- runif(ni/10,1,2)
  aref[((6*ni/10)+1):(7*ni/10)] <- runif(ni/10,0.2,1)
  aref[((7*ni/10)+1):(8*ni/10)] <- runif(ni/10,1,2)
  aref[((8*ni/10)+1):(9*ni/10)] <- runif(ni/10,0.2,1)
  aref[((9*ni/10)+1):ni] <- runif(ni/10,1,2)
```

```

#set percentage of items with DIF
for (itemcont in c(0,0.1,0.2,0.3)){

  itemrand <- c(rep(NA, ni*itemcont))
  #create DIF using the difference in the focal group b paramater

  if (itemcont==.10 & ni==20){
    item1 <- sample(1:8,1)
    item2 <- sample(13:20,1)
    itemrand <- c(item1, item2)
  }
  else if ((itemcont==.10 & ni==40) | (itemcont==.20 & ni==20)){
    item1 <- sample(1:(ni/5),1)
    item2 <- sample(((ni/5)+1):(2*ni/5),1)
    item3 <- sample(((3*ni/5)+1):(4*ni/5),1)
    item4 <- sample(((4*ni/5)+1):ni,1)
    itemrand <- c(item1,item2,item3,item4)
  }
  else if ((itemcont==.10 & ni==80) | (itemcont==.20 & ni==40)){
    item1 <- sample(1:(ni/5),2)
    item2 <- sample(((ni/5)+1):(2*ni/5),2)
    item3 <- sample(((3*ni/5)+1):(4*ni/5),2)
    item4 <- sample(((4*ni/5)+1):ni,2)
    itemrand <- c(item1,item2,item3,item4)
  }
  else if (itemcont==.20 & ni==80){
    item1 <- sample(1:(ni/5),4)
    item2 <- sample(((ni/5)+1):(2*ni/5),4)
    item3 <- sample(((3*ni/5)+1):(4*ni/5),4)
    item4 <- sample(((4*ni/5)+1):ni,4)
    itemrand <- c(item1,item2,item3,item4)
  }
  else if (itemcont==.30){
    if (ni==20){
      item1 <- sample(1:4,1)
      item2 <- sample(5:8,1)
      item3 <- sample(9:10,1)
      item4 <- sample(11:12,1)
      item5 <- sample(13:16,1)
      item6 <- sample(17:20,1)
      itemrand <- c(item1,item2,item3,item4,item5,item6)
    }
  }
}

```

```

else if (ni==40){
  item1 <- sample(1:8,2)
  item2 <- sample(9:16,2)
  item3 <- sample(17:20,2)
  item4 <- sample(21:24,2)
  item5 <- sample(25:32,2)
  item6 <- sample(33:40,2)
  itemrand <- c(item1,item2,item3,item4,item5,item6)
}
else if (ni==80){
  item1 <- sample(1:16,4)
  item2 <- sample(17:32,4)
  item3 <- sample(33:40,4)
  item4 <- sample(41:48,4)
  item5 <- sample(49:64,4)
  item6 <- sample(65:80,4)
  itemrand <- c(item1,item2,item3,item4,item5,item6)
}
}

```

```

#set amount of DIF
for (bdiff in c(0.25,0.5,0.75,1.0)){

  if (itemcont==0 & bdiff == 0.5){break}

  afoc <- aref
  bfoc <- bref

  if (itemcont !=0){
    bfoc[itemrand] <- bfoc[itemrand] - bdiff
  }
}

```

```

#Set ability mean for focal group
for (theta in c(0,-.5,-1)){

  #set sample size for reference and focal group
  for (f in c(25,50,100,200)){
    simfacts <- cbind(ni, itemcont, bdiff, theta, f)
    print(simfacts)
    fsize <- f
    rsize <- 500

    r_totmatrix<-rsize*ni
    f_totmatrix<-fsize*ni
  }
}

```



```

flag_mh1 <- matrix(rep(NA, replication*ni),replication,ni)
flag_mh2 <- matrix(rep(NA, replication*ni),replication,ni)
flag_dipf <- matrix(rep(NA, replication*ni),replication,ni)
flag_ri <- matrix(rep(NA, replication*ni),replication,ni)
effsize <- matrix(rep(NA, replication*ni),replication,ni)

for (repl in 1:replication){

  r_ability <- rnorm(rsize,0,1)
  f_ability <- rnorm(fsize,theta,1)

  r_data <- matrix(rep(NA, r_totmatrix), rsize, ni)
  group <- matrix(rep(0, r_totmatrix),rsize,1)
  r_data <- cbind(r_data, group)

  f_data <- matrix(rep(NA, f_totmatrix), fsize, ni)
  group <- matrix(rep(1, f_totmatrix),fsize,1)
  f_data <- cbind(f_data, group)

  amat_ref <- matrix(aref,ncol=ni,nrow=rsize,byrow=TRUE)
  amat_foc <- matrix(afoc,ncol=ni,nrow=fsize,byrow=TRUE)
  bmat_ref <- matrix(bref,ncol=ni,nrow=rsize,byrow=TRUE)
  bmat_foc <- matrix(bfoc,ncol=ni,nrow=fsize,byrow=TRUE)

  r_prob <- exp((r_ability-bmat_ref)*amat_ref)/(1+exp((r_ability-
    bmat_ref)*amat_ref))
  f_prob <- exp((f_ability-bmat_foc)*amat_foc)/(1+exp((f_ability-
    bmat_foc)*amat_foc))

  for (i in 1:ni){
    for (j in 1:rsize){
      rini<-runif(1)
      r_data[j,i] <- as.numeric(rini<r_prob[j,i])
    }
    for (j in 1:fsize){
      rini<-runif(1)
      f_data[j,i] <- as.numeric(rini<f_prob[j,i])
    }
  }

  g_data<- rbind(r_data,f_data)

  tot_score <- rowSums(g_data[,1:ni])

```

```

#MH with small intervals
matchsw1_1 <- as.numeric(tot_score <= (1/10)*ni)
matchsw1_2 <- as.numeric(tot_score > (1/10)*ni & tot_score <= (2/10)*ni)*2
matchsw1_3 <- as.numeric(tot_score > (2/10)*ni & tot_score <= (3/10)*ni)*3
matchsw1_4 <- as.numeric(tot_score > (3/10)*ni & tot_score <= (4/10)*ni)*4
matchsw1_5 <- as.numeric(tot_score > (4/10)*ni & tot_score <= (5/10)*ni)*5
matchsw1_6 <- as.numeric(tot_score > (5/10)*ni & tot_score <= (6/10)*ni)*6
matchsw1_7 <- as.numeric(tot_score > (6/10)*ni & tot_score <= (7/10)*ni)*7
matchsw1_8 <- as.numeric(tot_score > (7/10)*ni & tot_score <= (8/10)*ni)*8
matchsw1_9 <- as.numeric(tot_score > (8/10)*ni & tot_score <= (9/10)*ni)*9
matchsw1_10 <- as.numeric(tot_score > (9/10)*ni & tot_score <= ni)*10
matchsw_all <- cbind(matchsw1_10, matchsw1_9, matchsw1_8, matchsw1_7,
  matchsw1_6, matchsw1_5, matchsw1_4, matchsw1_3, matchsw1_2,
  matchsw1_1)
matchsw1 <- rowSums(matchsw_all)

matchsw1_1 <- as.numeric(tot_score <= (1/5)*ni)
matchsw1_2 <- as.numeric(tot_score > (1/5)*ni & tot_score <= (2/5)*ni)*2
matchsw1_3 <- as.numeric(tot_score > (2/5)*ni & tot_score <= (3/5)*ni)*3
matchsw1_4 <- as.numeric(tot_score > (3/5)*ni & tot_score <= (4/5)*ni)*4
matchsw1_5 <- as.numeric(tot_score > (4/5)*ni & tot_score <= ni)*5
matchsw_all <- cbind(matchsw1_5, matchsw1_4, matchsw1_3, matchsw1_2,
  matchsw1_1)
matchsw2 <- rowSums(matchsw_all)

#re-creation of dataset for all items
newdata <- matrix(rep(NA,dim(g_data)[1]*11),dim(g_data)[1],11)
for (i in 1:ni){
  newdata[,1] <- g_data[,i]
  for (j in 1:nrow(g_data)){
    newdata[j,2:11] <- c(rep(1,matchsw1[j]-newdata[j,1]),rep(0,10-(matchsw1[j]-
      newdata[j,1])))
  }

  MH1 <- dichoDif(newdata,group=g_data[,ni+1],focal.name=1,method=
    "MH",MHstat="logOR",correct=FALSE)

  flag_mh1[repl,i] <- as.numeric(1 %in% MH1$DIFitems)
}

#MH with large intervals

```

```

#re-creation of dataset for all items
newdata <- matrix(rep(NA,dim(g_data)[1]*6),dim(g_data)[1],6)
for (i in 1:ni){
  newdata[,1] <- g_data[,i]
  for (j in 1:nrow(g_data)){
    newdata[j,2:6] <- c(rep(1,matchsw2[j]-newdata[j,1]),rep(0,5-(matchsw2[j]-
      newdata[j,1])))
  }

  MH2 <- dichoDif(newdata,group=g_data[,ni+1],focal.name=1,method=
    "MH",MHstat="logOR",correct=FALSE)

  flag_mh2[repl,i] <- as.numeric(1 %in% MH2$DIFitems)
}

#DIPF
item.mix <- rep(1,ni)
trait.mix <- 1

ref_items <- RaschPLE(r_data[,1:ni],item.mix,trait.mix)$coefficients
ref_var <- RaschPLE(r_data[,1:ni],item.mix,trait.mix)$covb
ref_matrix <- outer(ref_items,ref_items,'-')
ref_var_matrix <- outer(diag(ref_var),diag(ref_var),'+') - 2*ref_var

foc_items <- RaschPLE(f_data[,1:ni],item.mix,trait.mix)$coefficients
foc_var <- RaschPLE(f_data[,1:ni],item.mix,trait.mix)$covb
foc_matrix <- outer(foc_items,foc_items,'-')
foc_var_matrix <- outer(diag(foc_var),diag(foc_var),'+') - 2*foc_var

Dij <- (ref_matrix-foc_matrix)/sqrt(ref_var_matrix+foc_var_matrix)

for (i in 1:ni){
  itemnbr <- c(1:ni)
  Dijcol <- cbind(itemnbr,-abs(Dij[i,]))
  Dijsort <- Dijcol[order(Dijcol[,2]),]
  pval <- round(2*pnorm(Dijsort[,2]),3)
  Dijpval <- cbind(Dijsort,pval,as.numeric(pval <= 0.05))
  flag_dipf[repl,i] <- as.numeric(sum(Dijpval[1:ni-1,4]) >= ni/2)
}

#RI
pref <- colSums(r_data[,1:ni])/rsize
pfoc <- colSums(f_data[,1:ni])/fsize
pdiff <- pref - pfoc
effsize[repl,] <- (pdiff - mean(pdiff))/sqrt(var(pdiff))
flag_ri[repl,] <- as.numeric(abs(effsize[repl,]) >= 1.96)

```

```

flag_DIF = c(rep(NA,ni))

if (itemcont==0){
  flag_DIF[1:ni] = 0
}else{
  flag_DIF[-itemrand]=0
  flag_DIF[itemrand]=1
}

flag_all <- cbind(ni,f,itemcont,theta,bdiff,aref,bref,bfoc,flag_mh1[repl,],
  flag_mh2[repl,],flag_dipf[repl,],flag_ri[repl,],flag_DIF,effsize[repl,])
colnames(flag_all) <- c("ni", "fsize", "itemcont", "theta", "bdiff", "aref", "bref",
  "bfoc", "flag_mh1", "flag_mh2", "flag_dipf", "flag_ri",
  "flag_DIF","effsize")
results_flag <- rbind(results_flag,flag_all)

}

if (itemcont==0){
  type1_mh1 <- colSums(flag_mh1)/replication
  type1_mh2 <- colSums(flag_mh2)/replication
  type1_dipf <- colSums(flag_dipf)/replication
  type1_ri <- colSums(flag_ri)/replication
}else{
  type1_mh1 <- colSums(flag_mh1[-itemrand])/replication
  type1_mh2 <- colSums(flag_mh2[-itemrand])/replication
  type1_dipf <- colSums(flag_dipf[-itemrand])/replication
  type1_ri <- colSums(flag_ri[-itemrand])/replication

  power_mh1 <- colSums(flag_mh1[,itemrand])/replication
  power_mh2 <- colSums(flag_mh2[,itemrand])/replication
  power_dipf <- colSums(flag_dipf[,itemrand])/replication
  power_ri <- colSums(flag_ri[,itemrand])/replication
}

if (itemcont==0){
  data_type1 <- cbind(ni,f,itemcont, theta, bdiff, aref, bref, bfoc, type1_mh1,
  type1_mh2, type1_dipf,type1_ri)
  colnames(data_type1) <- c("ni", "fsize", "itemcont", "theta", "bdiff", "aref",
  "bref", "bfoc","type1_mh1", "type1_mh2", "type1_dipf",
  "type1_ri")
  results_type1 <- rbind(results_type1,data_type1)
} else {

```


References

- Ackerman, T. A., & Evans, J. A. (1992). *An investigation of the relationship between reliability, power, and the Type I error rate of the Mantel-Haenszel and Simultaneous Item Bias Detection procedures*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- American Educational Research Association, A. P. (2014). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.
- Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three- parameter logistic item- response model. *ETS Research Report Series, 1981*(1).
- Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika, 80*(2), 317-340.
- Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1985). The psychometric characteristics of the SAT for nine handicapped groups. *ETS Research Report Series, 1985*(2).
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.
- Clauser, B., Mazor, K. M., & Hambleton, R. K. (1994). The effects of score group width on the Mantel- Haenszel procedure. *Journal of Educational Measurement, 31*(1), 67-78.
- Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6*(4), 269-279.

- Cohen, A. S., & Kim, S. (1993). A comparison of Lord's chi-square and Raju's area measures in detection of DIF. *Applied Psychological Measurement, 17*, 39-52.
- Crocker, L., & Algina, J. (2006). *Introduction to modern and classical test theory*. Fort Worth, TX: Cengage Learning.
- De Champlain, A. F. (2009). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44*(1), 109-117.
- DeVellis, R. F. (2006). Classical test theory. *Medical Care, 44*(11), S50-S59.
- Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational and Behavioral Statistics, 18*(2), 131-154.
- 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, & H. Wainer (Eds.), *Differential Item Functioning* (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel--Haenszel method. *Applied Measurement in Education, 2*(3), 217-233.
- Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel- Haenszel and standardization. *ETS Research Report Series, 1992*(1), i-40.
- Dorans, N. J., & Kulick, E. (1983). Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach. *ETS Research Report Series, 1983*(1), 83-89.

- Fay, M. P., & Graubard, B. I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics*, *57*, 1198-1206.
- Fidalgo, A. M., Ferreres, D., & Muñiz, J. (2004). Utility of the Mantel-Haenszel procedure for detecting differential item functioning in small samples. *Educational and Psychological Measurement*, *64*(6), 925-936.
- Fidalgo, A. M., Mellenbergh, G. J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online*, *5*(3), 43-53.
- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied Longitudinal Analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.
- Gallaudet. (2007). *Monitoring report to the Middle States Commission on Higher Education*. Retrieved from <http://ims.gallaudet.edu/pdf/20070914-0003.pdf>
- Gustafsson, J. E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *33*(2), 205-233.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.
- Herrera, A. N., & Gómez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel-Haenszel and logistic regression techniques. *Quality & Quantity*, *42*(6), 739-755.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

- Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two-and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*(3), 249-260.
- Jiao, H., & Chen, Y. F. (2014). Differential Item and Testlet Functioning Analysis. In A. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 1282-1300). Malden, MA: John Wiley & Sons, Inc.
- Kim, S. H., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement, 15*(3), 269-278.
- Kim, S. H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education, 8*(4), 291-312.
- Kim, S., Cohen, A. S., & Kim, H. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement, 18*, 217-228.
- Kulick, E., & Dorans, N. J. (1983a). Assessing unexpected differential item performance of candidates reporting different levels of father's education on SAT form CSA2 and TSWE form E29. *ETS Statistical Reports, SR-83-27*.
- Kulick, E., & Dorans, N. J. (1983b). Assessing unexpected differential item performance of Oriental candidates on SAT form CSA6 and TSWE form E33. *ETS Statistical Reports, SR-83-106*.

- Lai, J. S., Teresi, J., & Gershon, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Evaluation & the Health Professions*, 28(3), 283-294.
- Li, Z., & Hong, F. (2014). *plRasch: Log linear by linear association models and Rasch family models by pseudolikelihood estimation*. Retrieved from R package version 1.0.: <https://CRAN.R-project.org/package=plRasch>
- Lord, F. M. (1968). Some test theory for tailored testing. *ETS Research Bulletin Series*, 1968(2), i-62.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Maller, S. J. (1997). Deafness and WISC-III item difficulty: Invariance and fit. *Journal of School Psychology*, 35(3), 299-314.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4), 719-748.
- Marschark, M., & Hauser, P. C. (2008). Cognitive underpinnings of learning by deaf and hard-of-hearing students. In M. Marschark, & P. C. Hauser (Eds.), *Deaf cognition: Foundations and outcomes* (pp. 3-23). New York, N.Y.: Oxford University Press.
- Martin, P. (2005). An examination of the appropriateness of the New York state English Language Arts grade 8 test for deaf students. *Unpublished doctoral dissertation*, Gallaudet University, Washington, D.C., March 2005.

- Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*(2), 443-451.
- Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9*(3), 354-368.
- Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement, 16*(4), 381-388.
- Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology* (pp. 407-423). New York: Guilford.
- Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*(2), 115-135.
- Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*(4), 315-328.
- Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*(3), 257-274.
- Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. *Handbook of Statistics, 26*, 125-167.

- Phillips, A., & Holland, P. W. (1987). Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics*, 425-431.
- R Core Team. (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from <https://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, 197-207.
- Raju, N. S., Bode, R. K., & Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel for studying differential item performance. *Applied Measurement in Education*, 2(1), 1-13.
- Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, 17(2), 105-116.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 41-55.
- Roussos, L. A., & Stout, W. F. (1996). Simulation studies 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.
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 238-241.

- Stark, S., Chernyshenko, O. S., Chan, K. Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology, 86*(5), 943-953.
- Steinberg, J., Cline, F., Ling, G., Cook, L., & Tognatta, N. (2014). Examining the validity and fairness of a state standards-based assessment of English-language arts for deaf or hard of hearing students. *Journal of Applied Testing Technology, 10*(2), 1-33.
- Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychological Methods, 11*(4), 402-415.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.
- Tang, M. (1994). *A new IRT-based small sample DIF method*. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX.
- Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18*(1), 15-25.
- Valli, C. (2000). *Linguistics of American sign language: An introduction*. Washington, D.C.: Gallaudet University Press.

- Wang, W. C., & Su, Y. H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education, 17*(2), 113-144.
- 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*, 910-927.
- Wright, D. (1987). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. In A. P. Schmitt, & N. J. Dorans (Eds.), *Differential item functioning on the Scholastic Aptitude Test (Research Memorandum No. 87-1)*. Princeton, NJ: Educational Testing Service.
- Young, J. W., Morgan, R., Rybinski, P., Steinberg, J., & Wang, Y. (2013). Assessing the Test Information Function and Differential Item Functioning for the TOEFL Junior® Standard Test. *ETS Research Report Series, 2013*(1).
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational and Behavioral Statistics, 15*(3), 185-197.
- Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series, 2012*(1), i-30.

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