

IMPROVING THE PREDICTION OF DIFFERENTIAL ITEM FUNCTIONING: A
COMPARISON OF THE USE OF AN EFFECT SIZE FOR LOGISTIC REGRESSION
DIF AND MANTEL-HAENSZEL DIF METHODS

A Dissertation

by

SUSAN CROMWELL DUNCAN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2006

Major Subject: Educational Psychology

IMPROVING THE PREDICTION OF DIFFERENTIAL ITEM FUNCTIONING: A
COMPARISON OF THE USE OF AN EFFECT SIZE FOR LOGISTIC REGRESSION
DIF AND MANTEL-HAENSZEL DIF METHODS

A Dissertation

by

SUSAN CROMWELL DUNCAN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,
Committee Members,

Head of Department,

Victor L. Willson
Robert Hall
David Martin
Michael Speed
Michael Benz

May 2006

Major Subject: Educational Psychology

ABSTRACT

Improving the Prediction of Differential Item Functioning: A Comparison of the Use of an Effect Size for Logistic Regression DIF and Mantel-Haenszel DIF Methods.

(May 2006)

Susan Cromwell Duncan, B.S., Sam Houston State University;

M.S., Abilene Christian University

Chair of Advisory Committee: Dr. Victor L. Willson

Psychometricians and test developers use DIF analysis to determine if there is possible bias in a given test item. This study examines the conditions under which two predominant methods for determining differential item function compare with each other in item bias detection using an effect size statistic as the basis for comparison. The main focus of the present research was to test whether or not incorporating an effect size for LR DIF will more accurately detect DIF and to compare the utility of an effect size index across MH DIF and LR DIF methods. A simulation study was used to compare the accuracy of MH DIF and LR DIF methods using a p value or supplemented with an effect size. Effect sizes were found to increase the accuracy of DIF and the possibility of the detection of DIF across varying ability distributions, population distributions, and sample size combinations. Varying ability distributions and sample size combinations affected the detection of DIF, while population distributions did not seem to affect the detection of DIF.

DEDICATION

Isabel and James. The reason I began this journey. *Mom.* You are the reason I am capable. You taught me about Jesus. *Bryan.* You were an *awesome* find during this journey, a journey worth finishing now because of what you have given me. *Dave, Robbie, and James.* Your support and guidance kept me focused on what was most important - finishing. *Greg and Narda.* You began the journey with me and kept me from throwing in the towel many times.

ACKNOWLEDGEMENTS

To God I give the glory of my completing this degree. Every time I hit a dead end with this huge study I would be silent and find the peace of understanding, the understanding God would help me finish. I thank my husband for his support and help with manning 40 computers and a tired 8-month pregnant woman. I thank my mom for taking care of my babies so I could get my last two chapters done and giving me the foundation to go after my dream. I thank God again for giving me Isabel and James so that I would remain motivated to finish my dissertation so I could provide for them!

Thank you to Dr. Willson for the guidance and long distance advising. His patience and support helped me realize what kind of advisor I want to be soon. Dr. Hall gets a big thanks for helping me get started with this big huge paper and stepping in when I needed him near the end. He kept me accountable and made sure I had nourishment through animal crackers. Thank you to Dr. Speed for taking the time to get me past many huge software and formula bumps. Thanks to Dr. Martin for teaching IRT to begin my dissertation idea.

To Bruce – I did enjoy learning from you for five years. Thanks for helping me find my new career.

Lastly, I want to thank Carol – the one who helps all the graduate students keep it together. She knows what you want and need before you know to ask. Plus, she knows when to invite you in to her office so you can take a big breath and then get back out there!

TABLE OF CONTENTS

	Page
ABSTRACT	iii
DEDICATION	iv
ACKNOWLEDGEMENTS	v
TABLE OF CONTENTS	vi
LIST OF FIGURES	ix
LIST OF TABLES.....	xi
CHAPTER	
I INTRODUCTION.....	1
Statement of Problem	5
Research Questions	7
II LITERATURE REVIEW	9
Item Bias Historical Overview	10
Differential Item Functioning.....	13
Early Versions of Item Bias Techniques	14
IRT-Based DIF Methods	15
Chi-square Type DIF Methods	23
Logistic Regression as a DIF Method	30
Limitations of DIF Methodology	31
Significance Testing versus Effect Sizes.....	35
The Use of an Effect Size in DIF Methodology	37
$R^2\Delta$: An Effect Size for LR DIF.....	38
Conclusions about Current Research	40
III METHODOLOGY	42
ACT Test.....	42
Population Parameter Estimation	43
Calculation of Raju's Formula to Detect DIF	48
Uniform DIF versus Non-uniform DIF	57

CHAPTER	Page
Simulation Study	58
Series of Sample Size Combinations	58
Congruent and Incongruent Ability Distributions	59
Shape of Population Distributions.....	60
Statistical Analyses.....	61
Logistic Regression Analysis	61
Repeated Measures ANOVA	62
 IV RESULTS.....	 65
Preliminary Analyses.....	66
Logistic Regressions.....	66
Repeated Measures ANOVA	67
Conclusions of Preliminary Analyses.....	72
Research Question 1	73
Research Question 2.....	77
Research Question 3.....	81
Research Question 4.....	84
Research Question 5.....	87
Research Question 6.....	90
Research Question 7	103
Research Question 8.....	106
Research Question 9.....	108
Research Question 10.....	112
 V DISCUSSION AND CONCLUSIONS	 124
Research Question 1	124
Research Question 2.....	125
Research Question 3.....	125
Research Question 4.....	125
Research Question 5.....	126
Research Question 6.....	126
Research Question 7	127
Research Question 8.....	128
Research Question 9.....	128
Research Question 10.....	129
Conclusions	130
Strengths and Limitations.....	131

	Page
REFERENCES	134
APPENDIX A	144
APPENDIX B.....	145
APPENDIX C.....	146
APPENDIX D	147
APPENDIX E.....	148
APPENDIX F	149
APPENDIX G	150
APPENDIX H	151
APPENDIX I.....	152
APPENDIX J.....	153
APPENDIX K	154
APPENDIX L.....	155
APPENDIX M.....	161
APPENDIX N	167
APPENDIX O	169
VITA	173

LIST OF FIGURES

FIGURE		Page
1	Historical Overview of DIF Methodology, 1960-2000	10
2	Regression Lines for Males versus Females on Item from GRE.....	27
3	Histogram of Original Area Calculations for Math Subtest for Items 1-60.....	49
4	Histogram of Manipulated Area Calculations for Math Subtest for Items 1-60	49
5	Scatterplot of a and b Parameters for Forty-Four Non-DIF Items from ACT Test	54
6	Scatterplot of a and b Parameters for Sixteen DIF Items from ACT Test	55
7	Estimated Marginal Means for MH DIF and LR DIF Ability Distributions (p Values)	83
8	Estimated Marginal Means for MH DIF and LR DIF Population Distributions (p Values)	87
9	Estimated Marginal Means for MH DIF and LR DIF Sample Size Combination (p Values)	89
10	Estimated Marginal Means for Normal Ability Distributions (Method by Sample Size) (p Values)	94
11	Estimated Marginal Means for Moderate Ability Distributions (Method by Sample Size) (p Values)	95
12	Estimated Marginal Means for Severe Ability Distributions (Method by Sample Size) (p Values)	96
13	Estimated Marginal Means for Sample Size Combination 1000/100 (Method by Ability Distribution) (p Values).....	97
14	Estimated Marginal Means for Sample Size Combination 500/100 (Method by Ability Distribution) (p Values).....	98

FIGURE	Page
15 Estimated Marginal Means for Sample Size Combination 300/300 (Method by Ability Distribution) (p Values).....	99
16 Estimated Marginal Means for Sample Size Combination 1000/300 (Method by Ability Distribution) (p Values).....	100
17 Estimated Marginal Means for MH DIF and LR DIF Ability Distributions (Effect Sizes)	104
18 Estimated Marginal Means for MH DIF and LR DIF Population Distributions (Effect Sizes)	108
19 Estimated Marginal Means for MH DIF and LR DIF Sample Size Combination (Effect Sizes)	110
20 Estimated Marginal Means for Normal Ability Distributions (Method by Sample Size) (Effect Sizes)	115
21 Estimated Marginal Means for Moderate Ability Distributions (Method by Sample Size) (Effect Sizes)	116
22 Estimated Marginal Means for Severe Ability Distributions (Method by Sample Size) (Effect Sizes)	117
23 Estimated Marginal Means for Sample Size Combination 1000/100 (Method by Ability Distribution) (Effect Sizes)	118
24 Estimated Marginal Means for Sample Size Combination 500/100 (Method by Ability Distribution) (Effect Sizes)	119
25 Estimated Marginal Means for Sample Size Combination 300/300 (Method by Ability Distribution) (Effect Sizes)	120
26 Estimated Marginal Means for Sample Size Combination 1000/300 (Method by Ability Distribution) (Effect Sizes)	121

LIST OF TABLES

TABLE		Page
1	Simpson's Paradox Contingency Table for Two Groups on a Single Item	25
2	Estimated Item Parameters of Reference and Focal Groups	45
3	Original and Manipulated Female Item <i>b</i> Parameters and Area Calculations for DIF	50
4	Item Parameters for Male and Female Non-DIF Items	56
5	Item Parameters for Male and Female DIF Items.....	57
6	Categorical Levels for WLS R^2 and Log Odds Ratio.....	68
7	Descriptive Statistics of RM-ANOVA for Comparison of p Values.....	69
8	Descriptive Statistics of RM-ANOVA for Comparison of Effect Sizes.....	69
9	Levene's Test of Equality of Error Variances for p Values	71
10	Levene's Test of Equality of Error Variances for Effect Sizes	72
11	Classification Table for the Null Model for LR DIF.....	74
12	Classification Table for Model with p Value Predictor Variable for LR DIF.....	74
13	Classification Table for the Full Model for LR DIF	75
14	Statistics for Block 1 and Block 2 Models for LR DIF.....	76
15	Statistics for Predictor Variables in Full Model for LR DIF.....	76
16	Classification Table for Null Model for MH DIF	78
17	Classification Table for Model with p Value Predictor Variable for MH DIF.....	78

TABLE	Page
18	Classification Table for the Full Model for MH DIF 79
19	Statistics for Block 1 and Block 2 Models for MH DIF..... 80
20	Statistics for Predictor Variables in Full Model for MH DIF 80
21	Estimated Marginal Means for Method by Ability Distribution (p Values) 82
22	Linear and Quadratic Trends for Method by Ability Distribution (p Values) 84
23	Estimated Marginal Means for Method by Population Distribution (p Values) 86
24	Estimated Marginal Means for Method by Sample Size Combination (p Values) 88
25	Linear and Quadratic Trends for Method by Sample Size Combination (p Values) 90
26	Estimated Marginal Means for Method by Ability Distribution by Sample Size Combinations (p Values)..... 92
27	Contrast Trends for Method by Ability Distribution by Sample Size Combinations (p Values)..... 102
28	Estimated Marginal Means for Method by Ability Distribution (Effect Sizes) 104
29	Linear and Quadratic Trends for Method by Ability Distribution (Effect Sizes) 105
30	Estimated Marginal Means for Method by Population Distribution (Effect Sizes) 107
31	Estimated Marginal Means for Method by Sample Size Combination (Effect Sizes) 109
32	Linear and Quadratic Trends for Method by Sample Size Combination (Effect Sizes) 111

TABLE		Page
33	Estimated Marginal Means for Method by Ability Distribution by Sample Size Combinations (Effect Sizes).....	113
34	Contrast Trends for Method by Ability Distribution by Sample Size Combination (Effect Sizes)	123

CHAPTER I

INTRODUCTION

Standardized tests and measurements are used primarily to distinguish between specific skill or ability levels of examinees (i.e., academics, vocational tasks, or personality characteristics). As part of the determination of validity for these tests differential item analysis is employed to evaluate the degree to which measurements distinguish true abilities among examinees in an unbiased manner. Psychometricians and test developers use DIF analysis to determine if there is possible bias in a given test item. This study examines the conditions under which two predominant methods for determining differential item function compare with each other in item bias detection using an effect size statistic as the basis for comparison.

Test scores are inevitably affected by sources of variation other than the ability measured by the test. If tests invariably measured perfectly what researchers wanted to measure, all scores would be perfectly reliable and valid. However, irrelevant sources of variation cannot be completely controlled; therefore, steps should be taken to avoid giving any unfair advantage to any subpopulations taking the test. This unfair advantage will exist if within two subpopulations both have equal standing on the ability of interest, the irrelevant sources of variation is differentially distributed for the two subpopulations.

This dissertation follows the style of *Educational and Psychological Measurement*.

The statistical methodology for determining if item bias creates an unfair advantage is termed differential item functioning (DIF). DIF is defined as the observation of statistical properties of an item across two subpopulations that are assumed to have the same ability level (Holland & Wainer, 1993). DIF is determined in a two-step process. The first step is the comparison of two subpopulations' outcome on an item and determining the presence of DIF. The second step includes a decision of whether there is a large enough difference between subpopulations to eliminate or change the item of interest. The second step sometimes includes a formal test of the statistical significance of DIF, when available.

A serious limitation in both steps of determining DIF is that large sample size may cause a false positive, or Type I error, for an unbiased item when statistical significance tests are used (Cohen, 1990, 1994; Finch, Cumming, & Thomason, 2001; Thompson, 1998, 1999, 2002; Traub, 1983). Several DIF methods employ formal statistical significance tests to evaluate DIF; however, the use of probability (p) values and chi-square (χ^2) tests are not robust to varying sample sizes that are not comparable across methods or studies. The inability to compare across studies is a weakness because lack of comparability hinders generalization of a given study (Cohen, 1990; Kirk, 2001; Thompson, 1998, 1999). The use of an effect size to quantify DIF takes into account not only the magnitude of DIF, but generalizability and replicability (Huberty, 2002; Thompson, 1998, 1999, 2002). There is a need for an effect size measure as a supplemental statistic to control for false positives (Cohen, 1990, 1994; Kirk, 1996,

Thompson, 2002) and quantify the *amount* of DIF when DIF is detected (Potenza & Dorans, 1995).

Two other limitations in DIF methodology that affect the detection of DIF through DIF analyses are the varying types of ability and population distributions of the data. The assumption of normality in the population made by many researchers and psychometricians is unwarranted, and Pearson (1895) raised the question of prevalence of normality among real-world distributions (Micceri, 1989), who reported that normal distributions were rare (3.2%) in 440 sets of real data that were examined for distributional characteristics. Pommerich, Spray, and Parshall (1994) and Sweeney (1996) both reported that incongruence in ability distributions created instability in the detection of DIF. This can be expected due to the assumption of DIF methodology that ability distributions for reference groups and focal groups are congruent.

While the development of DIF methods has been a part of measurement research since the 1960's (Angoff, 1972, 1993; Cardall & Coffman, 1964; Cleary & Hilton, 1968; Holland & Wainer, 1993; Lord, 1980; Raju, 1988; Scheuneman, 1979; Shepard, Camilli, & Averill, 1981; Thissen, Steinberg, & Wainer, 1988), researchers recognize that there remain serious statistical weaknesses in the process of determining DIF. False positive errors are common due to the large samples that are necessary for most DIF methods. This is especially true for all IRT-based DIF methods, chi-square type DIF methods, and more importantly for the formal significance testing that determines if the difference is large enough to take action. To improve the accuracy of detecting DIF a formal statistical method that is robust to sample size must be used.

Numerous DIF methods have been developed in the past 40 years, but only a few have been studied widely throughout the literature. As DIF methodology progressed, earlier versions of methods became obsolete or evolved into new methods through research and the discovery of flaws and limitations. The prevalent DIF methods studied in the literature are IRT-based models, chi-square type methods (standardized and Mantel-Haenszel), and logistic regression. IRT-based, the Mantel-Haenszel (MH), and the logistic regression DIF methods have been the most promising for DIF research (Swaminathan & Rogers, 1990). Of these methods, the MH and the logistic regression DIF are currently seen as a practical means of determining DIF because of their simplicity and ease of use, at the same time providing an effect size statistic to determine if the DIF found is damaging. More importantly, simulation studies demonstrated that an effect size could be incorporated with the MH (Roussos & Stout, 1996) and logistic regression DIF methods (Jodoin & Gierl, 2001).

Swaminathan and Rogers (1990) introduced logistic regression as a DIF method, and it has been compared to the MH method with promising conclusions (Mazor, Kanjee, & Clauser, 1995; Swaminathan & Rogers, 1990). Mazor, Kanjee, and Clauser (1995) evaluated the logistic regression analysis for DIF as “a viable procedure for detecting differential item functioning (DIF)” (p. 131).

DIF can be considered for two conditions: uniform and non-uniform. Uniform DIF occurs when there is no interaction between the ability level and group membership. Nonuniform DIF occurs when there is an interaction between ability level and group membership (Swaminathan & Rogers, 1990). Logistic regression DIF (LR DIF) has

been compared to MH and has been found to be a better detector of both uniform and nonuniform DIF (Swaminathan & Rogers, 1990).

The use of an effect size, weighted-least-squares R^2 (WLS R^2), was introduced by Zumbo and Thomas (1996) and empirically tested through simulations by Jodoin and Gierl (2001) and Roussos and Stout (1996). Jodoin and Gierl (2001) focused on testing the effect size for LR DIF and empirically generating a classification guideline for negligible, moderate, and large DIF effect sizes, which is scaled similarly to Cohen's (1992) small, medium, and large effect size guidelines. Roussos and Stout (1996) did a comparison study of small sample sizes for MH DIF and the SIBTEST, and used effect sizes along with statistical tests, which reduced the Type I or false positives error rate. The MH DIF and the LR DIF methods can both incorporate effect sizes into DIF analyses. The MH DIF method has been compared to other methods (i.e., SIBTEST) when testing for Type I error with the use of an effect size; however, the LR DIF method and the inclusion of the WLS R^2 has only been empirically compared to the MH DIF method by Hidalgo and Lopez-Pina (2004) and no other DIF methods.

Statement of Problem

The main focus of the present research was a) to test whether or not incorporating an effect size for LR DIF will more accurately detect DIF and b) to compare the utility of an effect size index across MH DIF and LR DIF methods. A secondary focus of the present research was how various conditions, such as sample size, ability distributions, and population distributions, affect the detection of DIF by MH DIF and LR DIF methods.

Presently, the majority of DIF methods rely on p values to determine if DIF is present in an item. Both MH DIF and LR DIF methods introduced an effect size to supplement the formal statistical test of a p value (Holland, 1985; Holland & Thayer, 1988; Holland & Wainer, 1993; Zumbo & Thomas, 1996). The inclusion of an effect size increases the accuracy of DIF determination because an effect size is more stable than the formal test of statistical significance when exposed to varying sample sizes. Statistical tests, such as p values, are not robust to sample size. The main hypothesis is that the use of an effect size creates a more accurate outcome for determining if DIF is present in an item (Cohen, 1990, 1994; Finch, Cumming, and Thomason, 2001; Kirk, 1996, Thompson, 1996, 2002).

Currently, one study (Hidalgo & Lopez-Pina, 2004) compared MH and LR DIF methods and concluded that the LR DIF method detected more DIF items than MH method and was insensitive to specified DIF conditions. However, specified conditions consisted of a focal group of 1000 and a reference group of 1000 and normally distributed ability along with varying types of DIF (i.e., uniform, nonuniform). When Zumbo and Thomas (1996) introduced the WLS R^2 for LR DIF they suggested further research to test the accuracy of including an effect size in the determination of DIF. There have not been any simulation studies run to determine which DIF method, MH or LR, is more stable across varying conditions and whether the respective effect sizes of each method increase the accuracy of the determination of DIF.

Research Questions

Specifically, the following research questions are addressed.

- 1) Is the detection of DIF in an item more accurate when weighted-least-squares R^2 (effect size for LR DIF) supplements the use of a statistical significance test (p value) in the LR DIF analyses?
- 2) Is the detection of DIF in an item more accurate when a log odds ratio (effect size for MH DIF) supplements the use of a statistical significance (p value) test in the MH DIF analyses?
- 3) Is the detection of DIF more likely to occur when using a statistical significance test (p value) for LR DIF or a statistical significance test (p value) for MH DIF with varying ability distributions?
- 4) Is the detection of DIF more likely to occur when using a statistical significance test (p value) for LR DIF or a statistical significance test (p value) for MH DIF with varying population distributions?
- 5) Is the detection of DIF more likely to occur when using a statistical significance test (p value) for LR DIF or a statistical significance test (p value) for MH DIF with varying sample size combinations?
- 6) Is the detection of DIF more likely to occur when using a statistical significance test (p value) for LR DIF or a statistical significance test (p value) for MH DIF with varying ability distributions, population distributions, and sample size combinations?

- 7) Is the detection of DIF more likely to occur when using weighted-least-squares (WLS) R^2 (effect size for LR DIF) or log odds ratio (effect size for MH DIF) with varying ability distributions?
- 8) Is the detection of DIF more likely to occur when using weighted-least-squares (WLS) R^2 (effect size for LR DIF) or log odds ratio (effect size for MH DIF) with varying population distributions?
- 9) Is the detection of DIF more likely to occur when using weighted-least-squares (WLS) R^2 (effect size for LR DIF) or log odds ratio (effect size for MH DIF) with varying sample size combinations?
- 10) Is the detection of DIF more likely to occur when using weighted-least-squares (WLS) R^2 (effect size for LR DIF) or log odds ratio (effect size for MH DIF) with varying ability distributions, population distributions, and sample size combinations?

CHAPTER II

LITERATURE REVIEW

The term differential item functioning (DIF) made its debut in the 1990's. However the term has been included in studies that target the development of test fairness and alleviating the disparity in test performance between subpopulations (i.e., Black and Hispanics) since the 1960's, under the term *item bias* (Angoff, 1993). Many of these studies were conducted to better understand the cultural differences between Blacks and Hispanics, and specifically to demonstrate that the disparity in test scores had more to do with the bias found in test items than with the level of ability of either subpopulation. The overall focus of item bias studies is to identify items within the tests that might show bias towards one of two groups. Item bias is defined as (Angoff, 1993):

An item is biased if equally able (or proficient) individuals, from different groups, do not have equal probabilities of answering the item correctly. (p. 4)

Shepard et al. (1981) characterized bias as “a kind of invalidity that harms one group more than another” (p. 318). Angoff (1993) pointed out that the definitions used to define bias intimate there is a performance, an evaluation of the performance, and an unfair effect as a result. This unfair effect, the main focus of testing for bias, due to controversy regarding whether *any* difference in item performance across subpopulations, always reflects bias. It is inherently difficult to distinguish differential performance across subpopulations as item artifacts (e.g., content, wording) as opposed to real differences between subpopulations (Angoff, 1993).

Item Bias Historical Overview

Much of the conflict in determining the presence of bias has to do with the timing of item bias research and historical events. Figure 1 includes a timeline of published DIF methods that were influential in the conception of test/item bias.

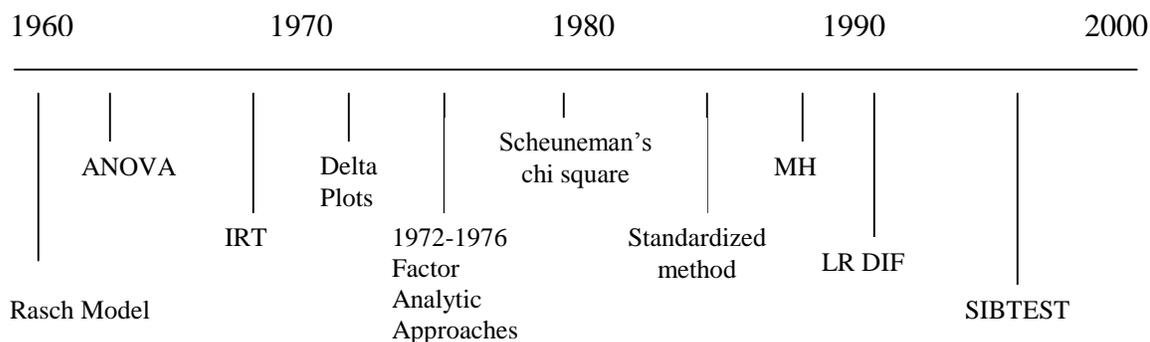


Figure 1. Historical Overview of DIF Methodology, 1960-2000

Cole (1993) compared item bias studies and historical events that were occurring simultaneously. In the 1960's when item bias detection was first used in research, the civil rights era had just begun. Concerns for bias grew out of this era, and became an established part of the test and measurement enterprise. The civil rights era was a time of establishing equality in education, employment, social services, and many other areas. Eventually, within areas of employment the term affirmative action was a popular legal concept. Prior to the civil right era, schools experienced major differences in educational resources and suffered tremendously from segregation. There were large differences in test scores across subpopulations reflecting how segregation and other

unfair treatment affected students in school, and people trying to get employment or acceptance into college.

Cole (1993) noted the testing community's response to these societal dynamics:

In this milieu, the technical testing community was enormously confused and even hurt by the attacks on its tests as biased. It viewed the tests as neutral tools that, granted, could be used for either good or bad, but were inherently neither. (p. 26)

However, Jenson (1980), in his *egalitarian fallacy*, argued that it is unreasonable to assume that all subpopulations are exactly equal in ability:

...the gratuitous assumption that all human populations are essentially identical or equal in whatever ability or trait the test purports to measure. Therefore, any difference between populations in the distribution of tests...is taken as evidence that the test is biased....with respect to scholastic performance, there is now general agreement that group differences in *achievement test scores* are not wholly due to test bias. The achievement differences between groups are more often attributed primarily to inequalities in schooling and home background. (p. 370)

Holland and Thayer (1986) introduced the term *impact*, which they defined as the real difference between groups in a test performance caused by a true difference across subpopulations in a valid ability (Ackerman, 1992).

Angoff (1993) pointed out an important characteristic of item bias and a disadvantage of item bias analysis – the difference in statistical versus social interpretation. In the 1960's when test score differences between groups were targeted as a sign of item bias a *social* perspective was used to understand the analyses.

Hypotheses were geared toward investigating differences in scores with the intention of assuming that the difference in minority test scores and majority test scores were due to item bias, and not impact. If there was a difference, item bias was assumed, and if there was not a difference, item bias was not assumed. But *statistically* item bias admits the possibility that item performance could deviate across subpopulations due to real group differences. Both interpretations should be taken into consideration before a test item is declared biased.

Psychometricians developed various item bias methodologies to determine whether tests included aberrant items reflecting real subpopulation differences, or biased items. Angoff (1993) noted that some aberrant items might have been biased, while other aberrant items might be quite fair and showing actual educational outcomes. It became obvious that the word bias was being used simultaneously, statistically and socially. This confusion in nomenclature of the word bias gave rise to the term differential item functioning (DIF), which refers to the observation that an item displays different statistical properties in different group settings. An important addition to this

definition is the term *statistical*, because the term infers that bias is determined by empirical means that come to the conclusion of bias based on an informed decision about the item under consideration.

Differential Item Functioning

DIF is the observation of an item that displays different statistical properties between two subpopulations that are assumed to have the same ability level (Holland & Wainer, 1993). Methods for investigating DIF are numerous in the literature, and methods are still being developed today. Researchers have classified DIF methods in various ways.

Some classifications focus on the conditioning variable, which is defined as a matching variable that subpopulations are assumed to be equal. The most common conditioning variable in DIF is ability level. For example, Millsap and Everson (1993) distinguished DIF methods based on the conditioning variable and whether it is observed or unobserved. Dorans and Potenza (1994) used categories based on the conditioning variable and the item response function (IRF) that described the relationship between the conditioning variable and item score as parametric or nonparametric. Researchers classify DIF methods into categories based on types of models and statistics used in the formulas. This is to give a better understanding of the various kinds of DIF methods and the origins of the statistics used in DIF detection.

The DIF methods presented here are further grouped together based on *formulaic similarities*. Unidimensional models were developed first and multidimensional models have become more prevalent in the past two decades, as noted in Figure 1. Early

research used the terminology “item bias”, while the later research uses “DIF”. The following research is reported with terminology consistent with its own literature. The more widely used DIF methods are discussed: early versions of item bias techniques (i.e., ANOVA, delta plots), IRT models, chi-square types of DIF, (i.e., MH, standardization), and logistic regression.

As seen in Figure 1, numerous DIF methods have been developed in the past 40 years, yet only a few have been used widely throughout the literature. As DIF methodology progressed, methods evolved due to the discovery of flaws and limitations. Some methods are corrected versions of previous methods (Rudner et. al, 1980). The following explanations of DIF methods focus on the most popular methods and those that are similar in computation. The DIF methods that are discussed are the ANOVA procedures, delta plots, Rasch model in conjunction with IRT models, Scheuneman’s chi square, the standardized and Mantel-Haenzsel method, and logistic regression.

Early Versions of Item Bias Techniques

In 1964, Cardall and Coffman (1964) developed the first formal procedure for DIF by applying the *analysis of variance* procedures to test two-way item performance (right versus wrong) by race (Black versus White) interactions for SAT data. Cleary and Hilton (1968) and Angoff and Sharon (1974) also followed the analysis of variance method, although this method did not seem to catch on (Holland & Wainer, 1993).

Angoff (1972) offered the *delta-plot* or transformed item-difficulty (TID) method, which looked at cultural differences. The Delta plot (DIT) method was very similar to Thurstone’s (1925) absolute scaling method and provided plots of item-by-

group interaction effects (Rudner et al., 1980). Indices of item difficulty (p values) were converted into a normal deviate and plotted to create the delta plot. Both groups had a point for each item chosen, and points were respectively plotted to form a graph that looked similar to the regression equation fit line (Angoff, 1972). The delta plot forms an ellipse, which represents the degree to which the two groups are similar (or not). Angoff (1972) recommended studying the specific items that are most aberrant by measuring the distance between each point and the major axis.

This method became quite popular due to its ease and simplistic nature. However, delta plots were flawed due to inconsistent discriminating power for items. Items could have been incorrectly determined as biased (or biased items were not found to be biased) because the items under consideration might not have had similar discriminating power. Although delta plots seemed simplistic and easy to use, the limitations involved left too many confounding variables present to trust the results (Angoff, 1982; Cole, 1978; Holland & Wainer, 1993).

IRT-Based DIF Methods

The use of item response models with item bias was actually introduced as early as the 1960's through the Rasch (1960) model, which is the one-parameter IRT model (McKinley & Mills, 1989). However, some recent Monte Carlo studies suggest that IRT methods may not have unique advantages over classical test theory (CTT) when it comes to DIF applications (Courville, 2004; Fan Xitao, 1998; MacDonald & Paunonen, 2002).

Lord (1952) and Lord and Novick (1968) introduced the theory's strong capacity for measuring differential item functioning. The field of DIF methodology literature

then exploded with IRT based DIF methods (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991; Linn & Harnisch, 1981; Paek, 1998; Rudner et al., 1980).

The Rasch model, or one-parameter model, is limited to measuring differences in item difficulty, while the two-parameter model is capable of measuring differences in both item discrimination and item difficulty. The most extensive model, the three-parameter model, is capable of measuring discrimination, difficulty, and the guessing parameter. The Rasch model, without knowledge of the a - and c - parameters, led to incorrect decisions regarding bias or lack of bias. The two- and three- parameter models are the two models used most frequently in DIF studies. Once the model is chosen, the type of IRT-based DIF method must also be chosen (Embretson & Reise, 2000; Hambleton et al., 1991). The main function of the IRT based DIF method is to determine if there is a difference in item parameters between the focal group (minority) and the reference (majority) group.

There are several definitions of DIF found in IRT literature. An inaccurate definition of DIF is “an item shows DIF if the majority and minority groups differ in their mean performances on the item” (Hambleton et al., 1991, p. 109). Hambleton et al.’s argument is that this definition does not take into account the other variables that could have influenced the real between-group difference in ability, which may be responsible for the difference in p -values. An accurate definition of DIF is “an item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item right” (Hambleton et al., 1991, p. 110). Yet, one

could make this accurate definition somewhat simpler by stating it as follows: “An item shows DIF if the item response functions across different subgroups are not identical. Conversely, an item does not show DIF if the item characteristic functions across different subgroups are identical” (Hambleton et al., p. 110). When IRT is used to detect possible DIF in items, the item characteristic curve that represents the item characteristic function is used.

There are also several ways in which item characteristic functions can be compared through DIF. The most popular, and simplistic, is the *comparison of item parameters*. This is based on the null hypothesis that the item response functions will be the same (subscripts are group identification) and is stated

$$H_0: \quad b_1=b_2; a_1=a_2; c_1=c_2,$$

where b is the item difficulty parameter, a is the discrimination parameter, and c is the guessing parameter (Embretson & Reise, 2000; Hambleton et al., 1991).

There are several criticisms of the comparison of item parameters. Rudner et al. (1980) were displeased with statistical significance testing within DIF, and developed the plot method which is based in IRT:

The development of the plot method which is an item response theory (IRT) based method was our attempt to by-pass the prevalent trend at the time to use statistical significance tests to flag potentially biased test items (or DIF). Our concern was that with a large enough sample size (and large sample sizes were being encouraged with

the IRT-based DIF method) even the most trivial differences between majority and minority groups would be identified as statistically significant (p. 2)

All test items could be considered biased given enough statistical power (i.e., sample size) (Traub, 1983). Another criticism involves the asymptotic distribution of the test statistic, and the statistic only being applicable when item parameters are estimated and the ability parameters are known.

Just as these limitations of the comparison of item parameters led to the development of the plot method, other alternative IRT-based DIF methods soon emerged. Two of these are the *area between ICCs* and the *fit evaluation in minority group through total group estimations*.

Area between ICCs

Taking the comparisons of item parameters a step further, methods computing the areas between ICCs across subpopulations were developed. In this DIF analysis, the ICCs are the focus, not the parameters. Granted the ICCs are created from the parameters, but this approach requires a much more thorough look at how different the parameters are. The area between the ICCs is the actual focus, and when the area is zero the conclusion is that DIF is not present. Conversely, when the area is not zero, DIF is present to some degree. Until technology caught up with these complex statistical procedures numerical procedures were used to compare ICCs. Several ways of doing this were dividing the ability range into k intervals and constructing rectangles centered around the midpoint of each interval. For examples see Hambleton et al. (1991).

Hambleton et al. (1991) and Raju (1988) both expressed this procedure in a formula. Hambleton's formula is a symbolic illustration for heuristic purposes, while Raju derived the exact expression that will compute the area between the ICCs for all parameter models (one, two, and three). Below is Hambleton et al.'s symbolic illustration:

$$A_i = \Sigma \left| P_{i1}(\theta) - P_{i2}(\theta) \right| \Delta\theta.$$

For this formula, the quantity of $\Delta\theta$ is the interval width and should be as small as possible. The terms r (below Σ and equal to θ) and s (above Σ) constitute the ability range of the area that is to be calculated to measure DIF. The range is arbitrary and chosen by the researcher, but typically is three standard deviations above and below the group mean ability.

Raju's (1988) formulas all take into account what is being estimated by the specific model, and uses the term "exact" to illustrate the difference in his formulas and the procedures listed in Hambleton et al. (1991):

$$\text{Area for 3-parameter} = (1 - c) \left| [2(a-a_1)/D_{a_1a_2}] \ln[1 + e^{D_{a_1a_2} (b_2-b_1)/(a_2-a_1)}] - (b_2-b_1) \right|;$$

$$\text{Area for 2-parameter} = \left| [2(a-a_1)/D_{a_1a_2}] \ln[1 + e^{D_{a_1a_2} (b_2-b_1)/(a_2-a_1)}] - (b_2-b_1) \right|;$$

$$\text{Area for 1-parameter} = \left| (b_2-b_1) \right|.$$

Once again, the a is item difficulty, b is discrimination, c is the guessing parameter, and D is a scaling constant that is usually set at 1.7 (Lord, 1980; Raju, 1990). Notice that the formula for the three-parameter model includes terms c , b , and a . In the two-parameter model the c term is missing, and in the one-parameter model the only term used is b . This is because of what parameters each model takes into account during estimation.

Raju (1988) established that all three of his formulas assume a common metric for group comparisons, and this can be accomplished with procedures from Hambleton et al. (1991). Estimations used in the formulas will vary from sample to sample, and small sample sizes can cause severe variations between groups. Also, equality of the lower asymptotes (c parameter) is evidence that the area between the ICCs is finite, which means Raju's formulas are very useful. However, when area estimates are based on finite intervals (i.e., $-3, +3$) of integration, the area estimates are much smaller than when the c parameters do not have to be estimated (Hambleton et al., 1993; Raju, 1988).

This dilemma of deflated area estimates between ICCs caused Linn and Harnisch (1981) to question the appropriate score interval for computing the area between two ICCs. The bottom line is that there needs to be equality in the lower asymptote in order to trust area estimates in the three-parameter model, and when estimating parameters there especially needs to be an appropriate sample size with a large ability range. Importantly, there might also be a problem with having a small minority sample size, which will lead to more erroneous decisions about the potential for DIF (Hambleton et al., 1991).

Fit Evaluation in Minority Group through Total Group Estimations

Linn and Harnisch (1981) developed the idea of using a goodness-of-fit test in the minority group through the utilization of the IRT model fit for the total group. Estimates of item and ability parameters for the combined group are obtained. The estimates obtained for the minority group from the total group analysis are used to see if the minority group fits the model from the data for the total group. If the model fits the

data for the minority group and no DIF is found, the total group ICC should fit the minority group data. The positive aspect of this procedure is that the parameters for the minority group are not estimated, which leads to less deflation of area estimates because minority groups are usually relatively small in sample size (Hambleton et al., 1991).

Linn and Harnisch (1981) described step by step how to use the total group estimates of a , b , c , and θ parameters to detect possible DIF between minority and majority groups. The basic steps taken to obtain DIF results once the item discriminating power, item difficulty, lower asymptote, and each participant's location on the θ scale are estimated are included below.

Step One: Three-Parameter Model Estimation

Step Two: $P_{ij} = c_i + 1 - c_i / 1 + \exp[-1.7a_i(\theta_j - b_i)]$

Step Three: $P_{ig} = 1/n_g \sum P_{ij}$

Step Four: $P_{i.} = \sum n_g P_{ig} / \sum n_g$

Step Five: $O_{i.} = \sum n_g O_{ig} / \sum n_g$

Step Six: $Di. = O_{i.} - P_{i.}$

Step one is obtaining total group estimates for a , b , c , and θ . Step two is the formula for the probability that a person j would answer item i correctly. Linn and Harnisch (1981) used the terminology of the target group (total group) and the subgroup (minority group) of the target group. In step three j is the participant in subgroup g who is expected to answer item i correctly based on the model. Step four is the formula for the proportion of people in the complete target group (total group). In step five the formula solves for the observed proportion correct on item i for subgroup g . Last, step

six finds the difference, which is an index of the degree to which members performed better or worse than expected on item i . Differences D_{ig} , are also used to note differences for each region on the θ scale. More specifically, Linn and Harnisch (1981) reported using D_i , D_{ig} , and corresponding differences between observed and expected performance of participants within subgroups to flag easy or difficult items, and these items are then compared in terms of item content and format (Linn & Harnisch, 1981).

The motive for using the procedure illustrated by Linn and Harnisch (1981) was the ability to use of the three-parameter model when only modest data was available, and more specifically the minority group was much smaller once separated from the total data. Also beneficial is the use of the difference that is calculated in step six, which quantifies differences.

Summary of IRT DIF Methods

While there are many IRT-based DIF methods, the comparison of item parameters is one of the most frequently used (Hambleton et al., 1991). Limitations of IRT DIF methods are the necessity of a unidimensional model, uniformity in the item characteristic curve, and the necessity of a large sample size. A limitation in conjunction with the need for a large sample size is the inflated Type I error actually caused by the large sample size.

The main factors in choosing a DIF method is usually based on resources available, ease of interpretation, and the researcher's ability to interpret and understand the analysis used. For ease in computation and interpretation less computationally-intensive DIF methods are a better choice. Thissen, Steinberg, and Wainer (1993)

reported that the process of using the general IRT-based DIF method (IRT-LR) is both labor and computationally intensive. The next DIF methods discussed here provides this ease of computation and interpretation that researchers sometimes look for when choosing a DIF method.

Chi Square Type DIF Methods

Analogous to, yet independent of the item characteristic curve, is Scheuneman's (1979) *modified chi-square* DIF method. The ability dimension is divided into discrete categories with the probability of correct responses in each category assumed constant, while discrimination among items vary and the lower asymptote is typically not zero. Scheuneman (1979) stated that "item characteristic curves for different ethnic groups can be very roughly approximated using relatively small samples..." (p. 145).

Scheuneman's version of the chi square method is concerned not only with frequencies of persons in each category as the usual chi square is, but with the number of correct responses made by persons in each ethnic group (or subpopulation) of interest. This is evident in the degrees of freedom for this method, which is $(k - 1)(r - 1)$ where k is number of subpopulations and r is the number of score groups, or categories.

Scheuneman's (1979) modified χ^2 formula is:

$$\chi^2 = \Sigma [(B_e - B_o)^2]/B_e + \Sigma [(W_e - W_o)^2]/W_e,$$

where B stands for subpopulation one and W stands for subpopulation two. For comparison purposes the usual χ^2 formula is:

$$\chi^2 = \Sigma[(O - E)^2]/E,$$

where O is the observed frequency in a given category and E is the expected frequency in a given category.

When establishing ability intervals on the total score scale, several criteria need to be met. The probability of a correct response within each ability interval must be less than one, and intervals are made larger or smaller to insure that there are some incorrect responses included in each interval. Expected frequencies must be at least five and all other cells must have somewhat large counts, a minimum of ten to twenty observed correct responses, due to small cells producing spurious results (Scheuneman, 1979). Scheuneman's (1979) χ^2 method was criticized because the values were too easily affected by sample size and did not have a chi square sampling distribution (Holland & Thayer, 1993; Rudner et al., 1980; Scheuneman, 1979).

Two contemporary chi square type methods used for DIF detection are the standardization method (Dorans, 1989) and the Mantel-Haenszel (1959) method. The MH method was introduced in 1959 by Mantel and Haenszel, but adapted to DIF procedures by Holland (1985) and Holland and Thayer (1988).

Dorans and Holland (1993) illustrated the standardization and MH method with Simpson's Paradox through the terms *impact* and *DIF*. Impact is defined as "the difference in performance between two intact groups" (p. 36) and DIF is "the differences in item functioning *after* groups have been matched with respect to the ability or attribute that the item purportedly measures" (p. 37). Simpson's Paradox compares results in a table that shows the performance of examinees for one item across two

groups. Table 1 is a contingency table used to explain Simpson's Paradox adapted from Dorans and Holland (1993).

Included in the contingency table is the number of examinees at each of three different ability levels (N_m), the number of correct responses from examinees at each ability level (N_{cm}), and the proportion of those who answered correctly over the number of examinees at that given ability level (N_{cm} / N_m). Impact is found when the *total* proportions for group *a* are divided by the *total* proportions for group *b*, which equals .10 (i.e., $.60 - .50 = .10$). DIF examines the proportions at the *individual ability levels* (e.g., .10 versus .20 for the first ability). When comparing the comparable statistics at each level in Table 1 the item favors Group B over Group A (i.e., .20 vs. 10; .60 vs. .50; 1.00 vs. .90), not Group A over Group B, as suggested by impact. Noticeably, the impact results are different from the DIF results due to the unequal distributions of item ability.

Table 1 Simpson's Paradox Contingency Table for Two Groups on a Single Item

Ability Level	Group A			Group B		
	N_m	N_{cm}	N_{cm}/N_m	N_m	N_{cm}	N_{cm}/N_m
1	400	40	.10	1000	200	.20
2	1000	500	.50	1000	600	.60
3	1000	900	.90	400	400	1.00
Total	2400	1440	.60	2400	1200	.50

Note. Adapted from Dorans and Holland (1993).

The *standardization* method was introduced by Dorans and Kulick (1983) after Dorans reviewed numerous item bias studies from the late 1970's. After dismissing the delta plots due to the exclusion of a discrimination parameter, and the IRT models for possible model misfit, Dorans and Kulick chose to use a method that was similar to IRT in theory, but where a total score used as an ability estimate was turned into an empirical item response curve. The definition of DIF for the standardization method is "when an expected performance on an item differs for examinees of equal ability from different groups. Expected performance on an item can be operationalized by nonparametric item test regressions. Differences in empirical item test regressions are indicative of DIF" (Dorans & Holland, 1993, p. 44). Important in the standardization method is to use all available data when estimating the conditional item performance of each group at each level of the matching variable (Dorans & Holland, 1993).

There are two steps in the standardization DIF method. First, all available data is used to estimate the nonparametric item test regression separately for the reference (*r*) group and focal (*f*) group. The standardization approach employs the definition $E_f(I | M) = E_r(I | M)$ where E is the item test regression, I is the item score variable, and M is the matching variable. A hypothetical data set of one item from the GRE for males and females is used to illustrate two regression lines plotted for visual analysis between the focal (male) and reference (female) groups in Figure 2.

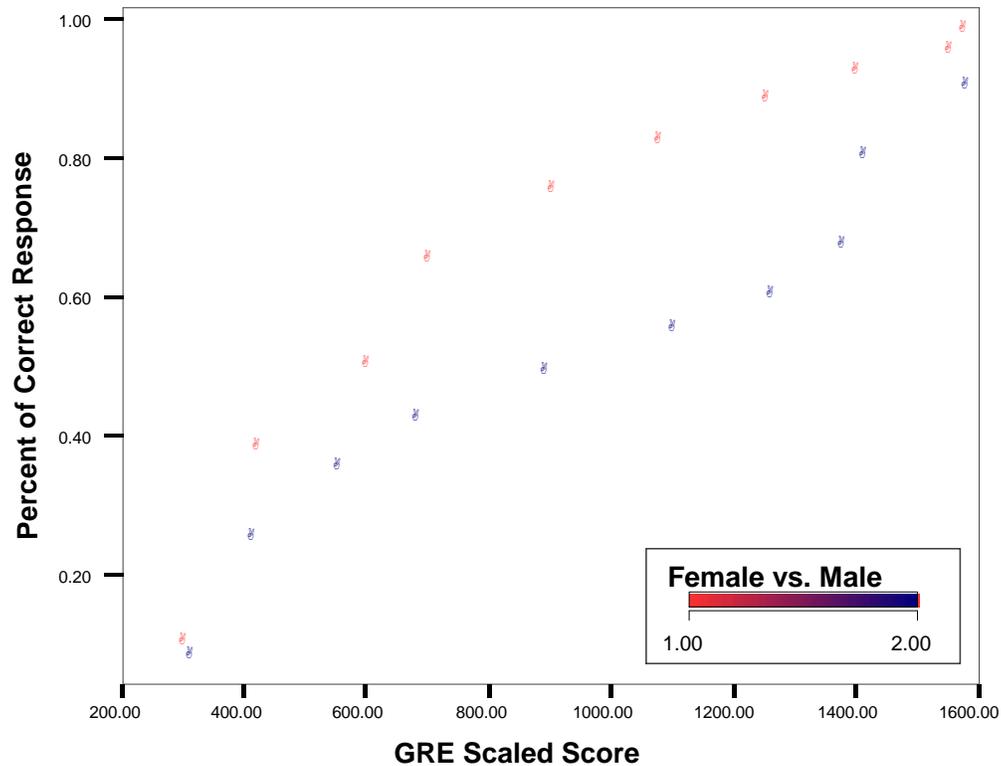


Figure 2. Regression Lines for Males versus Females on Item from GRE

Notice the area between the male regression line and female regression line.

There is substantial negative DIF for females in this hypothetical data set on this specific item.

While visual proof is not strong enough to make a decision of DIF, the standardization method includes a formal significance test. In this next step, a standardization group is used to create weights to use as a weight for each individual D_m

at each level. Weighted differences (D_m^2) are then accumulated across score levels to establish a summary item discrepancy index.

The *Standardized P-Difference test* is used to detect DIF statistically, as opposed to only visually. The formula for this is

$$\text{STD P-DIF} = \frac{\sum W_m(E_{fm} - E_{rm})}{\sum W_m} = \frac{\sum W_m D_m}{\sum W_m},$$

where $W_m/\sum W_m$ is the weighting factor at score level m from the standardization group to weight differences between focal and reference groups. Dorans and Holland (1993) explained that use of the same weight on E_{fm} and E_{rm} “is the essence of the standardization approach” (p. 49). The weights employed for standardization purposes depend on the reason for the research study.

The second contemporary chi square type method is the *MH DIF* method. The MH DIF method is a procedure for matched groups and data are placed in a $K \times 2$ table. Holland (1985) and Holland and Thayer (1988) took this chi square test and made it a more suitable procedure than previous chi square methods because their approach measures “the size of the departure of the data from H_0 ” (p. 133), which quantifies how much DIF is detected. The MH chi square statistic is

$$\text{MH-CHISQ} = \frac{(|\sum_j A_j - \sum_j E(A_j)| - 1/2)^2}{\sum_j \text{Var}(A_j)}.$$

Notice the addition of $- 1/2$ that is a continuity correction for accuracy. Holland and Thayer (1986) stated the MH-CHISQ is “the uniformly most powerful unbiased test of H_0 versus H_1 ” (p. 134), unless there is a violation of size constraint on the H_0 or lower power on H_1 for the other test.

Important to the MH method, Mantel and Haenszel (1959) also developed the constant odds ratio α_m and an estimate for α_{MH} . The constant odds ratio is

$$\alpha_m = (R_{rm}/W_{rm})/(R_{fm}/W_{fm}) = (R_{rm}/W_{fm})/(R_{fm}/W_{rm}),$$

where R is a correct response, W is an incorrect response, r is reference group, f is focal group, and m is the ability level of the studied item (Dorans & Holland, 1993).

The estimate for the constant odds ratio is derived from

$$\alpha_{MH} = (\sum_m R_{rm} W_{fm} / N_{tm}) / (\sum_m R_{fm} W_{rm} / N_{tm})$$

and is on a scale of 1 to ∞ . Dorans and Holland (1993) and Holland and Thayer (1989) provided formulas that will convert this estimate to a symmetrical scale for comparison purposes.

$$\Delta_{\alpha (MH)} = -2.35 \ln(\alpha_{MH})$$

This conversion formula that transforms the constant odds ratio into a log odds ratio improves the interpretative ability of the estimate (Hidalgo & Lopez-Pina, 2004).

Both the MH and standardization methods are powerful methods that can be used with ease and simplicity along with a statistical significance test to show DIF. One of the limitations of MH is its inability to test nonuniform items (Hambleton & Rogers, 1989; Swaminathan & Rogers, 1990). The standardization method and the MH method both use an internal criterion for matching, which is to some a circularity problem and both do not take into account the possibility of multidimensionality in an item (Dorans & Holland, 1993). A second limitation concerns the statistical significance tests used for both methods where results are heavily influenced by sample size. Unlike Scheuneman's chi square method and the standardization method, the MH method

includes an effect size that quantifies the amount of DIF (Dorans & Holland, 1993; Roussos & Stout, 1996). Therefore, the main limitation of the MH method is its inability to detect nonuniform DIF.

Logistic Regression as a DIF Method

The use of logistic regression to detect DIF is similar to the MH DIF method as regards ease of computation and interpretation and provides a measurement of magnitude for DIF. One area in which LR DIF is stronger than MH DIF is the ability to detect nonuniform DIF (Dojoin & Gierl, 2001; Swaminathan & Rogers, 1990; Zumbo, 1996).

The use of logistic regression as a DIF method was introduced by Swaminathan and Rogers (1990) and through simulation studies has shown to be more powerful than the MH method in detecting nonuniform DIF and just as powerful as MH in detecting uniform DIF (Swaminathan & Rogers, 1990). Bertrand and Boiteau (2003) described LR DIF as a two-step process. In the first step, the total test score is used in the regression equation. The second step is a regression with two variables related to the group and the group-by-score interaction. If these two models lead to a statistically significant difference, then DIF occurs.

The LR DIF is based on the logistic regression model in Swaminathan and Rogers (1990),

$$P(u = 1 | \theta) = e^{(\beta_0 + \beta_1 \theta)} / [1 + e^{(\beta_0 + \beta_1 \theta)}],$$

“where u is the response to the item, θ is the observed ability of an individual, β_0 is the intercept parameter, and β_1 is the slope parameter” (p. 362). Each group under

consideration uses the above formula to derive a separate probability of a correct response to an item known as LR DIF (Swaminathan & Rogers, 1990):

$$P(u_{ij} = 1 | \theta_{ij}) = e^{(\beta_{oj} + \beta_{1j} \theta_{ij})} / [1 + e^{(\beta_{oj} + \beta_{1j} \theta_{ij})}].$$

This formula is used for each group where i refers to the person in j group for an item. DIF is present if each group has different probabilities of success but the same ability.

Swaminathan and Rogers (1990) compared MH and LR DIF methods through simulation studies. Overall, LR DIF detected nonuniform items in three different test lengths, while MH did not detect any nonuniform DIF. LR DIF also detected more uniform DIF items. However, MH performed better in determining false positives (i.e., DIF that is detected but not actually present in the item). The LR DIF method seemed a better model than the MH method when nonuniformity is a concern. Jodoin and Gierl (2001) also reported possible inflated Type I error and noted the weakness of no effect size measure available in LR DIF.

Limitations of DIF Methodology

All the above DIF methods have been discussed along with the strengths and weaknesses of each method. Limitations common to accuracy of DIF methods are sample size, shape of population distributions, congruence of ability distributions, amount of DIF, type of DIF (uniform and non-uniform), and test length. The current study focused on three limitations that were important in the study of statistical significance testing and effect sizes for DIF methodology. These limitations are sample

size, congruent and incongruent ability distributions between the focal and reference groups, and shape of population distribution.

The one limitation that affects the accuracy of the detection of DIF for all DIF analyses is that larger sample sizes can increase false positives in the detection of DIF, yet the majority of DIF analyses require large sample sizes (Hambleton, Swaminathan, & Rogers, 1991; Schmitt, Hollan, & Dorans, 1993; Zieky, 1993). Zieky (1993) and Roussos and Stout (1996) both reported sizes of no less than 100 for the focal groups and reference groups that vary between 200 and 1000 are sample sizes found in practice. The reference group is generally larger than the focal group.

The need for large sample sizes is a problem because larger sample sizes can cause false positives in DIF through the use of formal statistical significance tests. Traub (1983) stated that all test items could be considered biased if given enough statistical power (i.e., sample size). The null hypothesis will always be rejected if the sample size is large enough, and all of the DIF methods use null hypothesis significance testing (Thompson, 1996). This topic is discussed in greater depth in the following section that discusses *Significance Testing versus Effect Sizes*.

Congruence of ability distributions is important for all DIF analyses. This assumption is based on the notion that in order to examine group differences the data must be conditioned on a criterion variable, which is known as the ability parameter (Angoff, 1993). While the ability parameter can be observed or unobserved, the distributions of this ability parameter for the focal group and the reference group are assumed congruent.

Pommerich, Spray, and Parshall (1994) and Sweeney (1996) both reported the absence of congruence in ability distributions created instability in the detection of DIF. When group ability distributions had group mean differences of three standard deviations the DIF statistic became unstable. Only under moderately congruent distributions was the DIF statistic stable. Zwick (1990) and Schulz, Perlman, Rice, and Wright (in press) both found similar conclusions that the more incongruent the ability distributions for groups, the less likely the null hypothesis of no difference would be satisfied for focal and reference groups.

Roussos and Stout (1996) analyzed DIF detection while setting the ability distributions as normal, both focal and reference with a variance of one, but with varying means. The differences in means used were 0.0, 0.5, and 1.0. Roussos and Stout (1996) chose these amounts of mean differences based on an examination of real data and discussions with test data specialists. Roussos and Stout (1996) reported that a “slight tendency toward increasing Type I error with increasing sample size and increasing d_T [difference in ability distribution means], with MH seeming to have very slightly lower Type I error rates for $d_T > 0$.” (p. 221).

Jodoin and Gierl (2001) simulated data for equal ability distributions and unequal ability distributions under small sample sizes (250/250) to larger sample sizes (1000/1000). When comparing the Type I error and power rates between equal and unequal ability distributions rates were slightly (i.e., 0.2 points to 4.0 points) higher for unequal ability distributions for smaller samples, but evened out as the sample sizes increased. Jodoin and Gierl (2001) set the unequal ability distributions with a difference

of .50 for the means of the reference and focal group with the same SD. The smaller difference in this study might attribute to not finding differences in Type I error and power rates for the larger sample sizes.

Another limitation for DIF methodology is the assumption of normality in the data. Unfortunately, real data are rarely normally distributed. Non-zero estimates of kurtosis and skewness reflect the assumption of normality being violated. Micceri (1989) researched the degree and frequency of various forms of skewed distributions in real data. Across 440 real data sets only 28.4% were considered “relatively symmetric” (p. 160). Micceri (1985) reported:

No distributions among those investigated passed all tests of normality, and very few seem to be even reasonably close approximations to the Gaussian. (p. 161)

Although Micceri (1989) did not use kurtosis as a classification in his study, he reported “kurtosis estimates were computed and ranged from -1.70 to 37.37 . Ninety-seven percent of those distributions exhibiting kurtosis beyond the double exponential also showed extreme or exponential asymmetry...” (p. 161). Micceri also reported that of 43 distributions, 72.7% exhibited positive or negative coefficient of skewness greater than .39. Out of 18 of the distributions, seven exhibited skewness at or greater than .94.

Fleishman (1978) noted that when performing statistical testing the rejection of the null is equal to rejecting one of the main assumptions, such as the assumption of a normal population distribution. Fleishman used a polynomial transformation called the

power method to simulate non-normal distributions, and he published a table with skewness and kurtosis coefficients. Fleishman (1978) stated:

In other words, many, if not most of the psychological variables seen are skewed and/or kurtotic to various degrees. (p. 521)

Specifically, sample size, ability distribution, and population distribution are limitations to DIF analyses and should be considered when studying accuracy of DIF methodology.

Significance Testing versus Effect Sizes

Important in DIF detection methods is whether or not the method has a means of testing the difference, size of DIF, between subpopulations to assist in making the decision that the difference is enough to delete or change the item of interest. Earlier versions of item bias detection methods do not have formal tests, while more current DIF methods, such as IRT-based, MH, and standardization methods use a p value or χ^2 statistic. The use of a statistical test is better than not using a formal test, but as seen in the literature (Cohen, 1990, 1994; Finch, Cumming, & Thomason, 2001; Huberty, 2002; Thompson, 1998, 1999, 2002; Traub, 1983), statistical tests are controlled by sample size. More importantly, pointed out in the previous overview of DIF methods, the use of statistical significance tests inflate Type I error, because DIF analyses inherently use large sample sizes.

The need for an alternative or supplement to significance testing has been evident for decades and conveyed in many journal articles (Cohen, 1990, 1992, 1994; Fidler,

2002; Finch, Cumming, & Thomason, 2001; Huberty, 2002; Kirk, 1996; Rosnow & Rosenthal, 1996; Schmidt, 1996; Thompson, 1996, 2002). Huberty (2002) elaborated on the controversy of the practice of significance testing by pointing out that in the past several decades "there has been an exponential increase in the frequency of publications criticizing uses of statistical testing..." (p. 227). Harlow stated the true purpose of significance testing in *What if There Were No Significance Tests* (1997), which gives proof of the necessity for a supplement to significance testing from its birth:

NHST [significance testing] was intended to provide a method for ruling out chance, thus helping to build strong evidence in favor of one or more alternative hypotheses, rather than provide an indication of the proof or probability of these hypotheses.... (p. 2)

Kirk (1996) pointed out several areas of criticism concerning classical null hypothesis significance testing. First, statistical significance tests do not tell the researcher what they want to know. The researcher wants to know the probability of the null hypothesis being true in the population, but instead testing the significance of the null hypothesis tells the researcher the probability of obtaining sample data that supports the null hypothesis if the null hypothesis is assumed true in the population. Second, statistical significance testing is a trivial exercise because there will always be some degree of difference between the two variables; therefore, statistical significance can always be achieved at some sample size (Thompson & Keiffer, 2000). The higher the sample size, the more likely the researcher will find statistical significance (Cromwell,

2001). Often overlooked is whether not the effect is useful or large enough to make a practical difference, regardless of the level of statistical significance. This led to researchers to following the rules of null hypothesis statistical testing to such a narrow degree that researchers focused on controlling the Type I error that cannot occur, because essentially all null hypotheses are false.

Jodoin and Gierl (2001) noted that the use of null hypothesis significance testing is dangerous to use for DIF detection methods when not accompanied by an effect size (Cohen, 1990, 1994; Kirk, 1996; Thompson, 2002). Practical significance is important in this area of research and including an effect size would increase the accuracy of the interpretation of DIF (Jodoin & Gierl, 2001; Thompson, 2002).

The Use of an Effect Size in DIF Methodology

The use of LR DIF and MH DIF and both statistical significance testing and effect sizes would increase the confidence of the DIF results and decrease the amount of false positives. The MH DIF effect size measure, log odds ratio, was presented earlier and the LR DIF will be presented below.

The LR DIF method is the youngest DIF methodology currently in the literature. While the use of logistic regression for DIF detection was introduced nearly a decade ago, the use of an effect size for LR DIF has not been extensively researched. LR DIF is one of the first procedures intended to measure uniform and nonuniform DIF. The MH DIF method has been used to measure nonuniform DIF, but has been found to be less powerful in detecting nonuniformity than LR DIF. The MH and LR DIF methods are both known for the simplicity and ease of computation, which is attractive.

The MH DIF method has been researched and compared to several other DIF methods (Dorans & Holland, 1993; Holland & Thayer, 1986; Swaminathan & Rogers, 1990; Roussos & Stout, 1996) and found to be a powerful method for uniform DIF (Swaminathan & Rogers, 1990). Important to note is the one limitation for LR DIF is an inflated Type I error rate (Jodoin & Gierl, 2001; Swaminathan & Rogers, 1990). This could be controlled by the use of an effect size measure. If LR DIF can detect nonuniform DIF better than the MH DIF method, and is as powerful at detecting uniform DIF as the MH DIF method, then the inclusion of an effect size would make LR DIF a very attractive choice as a DIF detection method.

Jodoin and Gierl (2001) performed a simulation study on evaluating Type I error and power rates using an effect size with LR DIF. Zumbo and Thomas's (1996) $R^2\Delta$, the weighted least squares measure, quantifies the magnitude of uniform and nonuniform DIF. Jodoin and Gierl (2001) noted that if there is an effect size measure for LR DIF, the 2-*df* chi-square test could be separated into two 1-*df* tests for uniform and nonuniform tests, respectively. They found that if the uniform and nonuniform tests are separated into two tests in conjunction with an effect size measure the procedure results in superior power in the detection of uniform and nonuniform DIF than when using the 2-*df* test, even while using smaller samples. Jodoin and Gierl (2001) recommended the use of two separate 1-*df* tests for uniform and nonuniform DIF over the 2-*df* test for practitioners and researchers who need power because this "enhances uniform DIF detection, allows for nonuniform DIF detection, and curtails Type I error rates" (p. 347). The effect size $R^2\Delta$ for LR DIF is illustrated below.

R²Δ: An Effect Size for LR DIF

The formula for LR DIF, as stated in Swaminathan and Rogers (1990) is,

$$P(u_{ij} = 1 | \theta_{ij}) = e^{(\beta_{oj} + \beta_{lj} \theta_{ij})} / [1 + e^{(\beta_{oj} + \beta_{lj} \theta_{ij})}],$$

for $i = 1, \dots, nj$ and $j = 1, 2$.

The first parameter, β_o , is the intercept parameter, β_l is the slope parameter, u is the response to the item, θ is the observed ability of an individual. This is not a linear model, but can be stated as a linear model with Zumbo and Thomas's (1996) revisions. The equivalently linear model is,

$$\ln(P/1-P) = \tau_0 + \tau_1 + \tau_2 g + \tau_3 (\theta g).$$

Based on θ and g (group membership), P is the probability of responding correctly. When you apply Pregibon's (1981) results of the vector of maximum likelihood estimators, τ , to the LR coefficients the above equation is a weighted least squares model.

The formula for the maximum likelihood estimator, τ , is

$$\tau = (X'VX)^{-1} X'Vz,$$

“where $z = X\tau + V^{-1}r$, $r=(u-P)$, V is an $N \times N$ diagonal matrix with elements $P_i(1-P_i)$, $i=1, \dots, N$, X is an $N \times 4$ data matrix with rows $[1, \theta_i, g_i, \theta_i g_i]$, P is an $N \times 1$ vector of the fitted values of the LR model, u is an $N \times 1$ vector of examinee responses, and N is the combined sample size of the reference and focal groups” (Jodoin & Gierl, 2001, p. 333).

Given the above explanation of the LR DIF model in terms of a weighted least squares model, Zumbo and Thomas (1996) demonstrated that through the geometry of least squares an additive partitioning of explanatory variables was a reasonable idea:

$$R^2\Delta = R_1^2 - R_2^2.$$

Jodoin and Gierl (2001) explained the R_1^2 and R_2^2 terms in the formula are sums of the products of the standardized regression coefficient for each explanatory variable and the correlation between the response and each explanatory variable” (p. 333).

For comparison purposes a classification system was suggested by Zumbo and Thomas (1996) for the weighted least squares effect size, $R^2\Delta$, and Zwick and Ercikan (1989) proposed a classification system for the log odds ratio, $\Delta_{\alpha (MH)}$. For $R^2\Delta$, negligible DIF is an estimate below .13, moderate DIF is .13 to .26, and large DIF is an estimate above .26. For $\Delta_{\alpha (MH)}$, negligible DIF is an estimate less than than |1|, moderate DIF is |1| through |1.5|, and large DIF is an estimate greater than |1.5|(Zwick and Ercikan, 1989; Zumbo and Thomas, 1996).

The effect size, $R^2\Delta$, has only been empirically studied through simulation studies and compared with SIBTEST. The SIBTEST and LR DIF effect sizes had a curvilinear relationship. No other studies of the LR DIF effect size have been performed. The new effect size for the LR DIF needs to be studied through simulation and compared with results from a similar DIF method, such as the MH DIF method.

Conclusions about Current Research

The weaknesses found in all DIF statistical analyses led Kim and Cohen (1995) and Fidalgo, Ferreres, and Muniz (2004) to suggest the use of multiple DIF statistics to

analyze possible DIF within empirical data to reduce false positives. The literature above has illustrated that DIF analyses unreliably detect DIF across DIF methods and especially with the use of only a statistical significance test. While the majority of DIF analyses only use a formal statistical significance test to measure DIF, effect sizes have been introduced in recent years (Zumbo & Thomas, 1996; Jodoin & Gierl, 2001; Hidalgo & Lopez-Pina, 2004). Specifically, effect sizes have been established and utilized with the LR DIF and MH DIF methods. Hidalgo and Lopez-Pina (2004) found the LR DIF and MH DIF methods to be “highly comparable” (p. 912).

Both the LR DIF and MH DIF methods have been promoted in the literature through simulation studies and empirical studies. The effect size, $R^2\Delta$, for LR DIF is analyzed in the current study. If the effect size, $R^2\Delta$, is an accurate and stable statistic for LR DIF, then the use of LR DIF and MH DIF methods would be beneficial for the researcher. The use of both DIF methods for the analysis of DIF detection could give the researcher four formal statistics to interpret and make decisions about the presence of DIF.

The purpose of the current study was two-fold. First, the accuracy and stability between the LR DIF and MH DIF are analyzed through conditioning sample size, ability distributions, and population distributions. Secondly, the LR DIF and MH DIF effect sizes and p values were analyzed through conditioning sample size, ability distributions, and population distributions for accuracy and stability of detecting DIF.

CHAPTER III

METHODOLOGY

Simulation of DIF conditions was conducted under a condition of 15 populations and four sample size combinations to analyze the accuracy of formal statistical significance tests and effect sizes for LR DIF and MH DIF. Each population was simulated 200 times. First, an empirical population of 40,000 examinee responses from the math subtest was obtained from the American College Testing Program (ACT). The math subtest was chosen based on content and past evidence that DIF might occur between subgroups of male and female. Male was designated the reference group and female was designated the focal group. The data were used to ensure similarity to real data for the simulations instead of creating population parameters to simulate data. Initial data sets from ACT containing original test items were requested because original items for ACT tests are not revised to control for DIF. This request ensured the possibility of DIF items and ecological validity for distributions of DIF parameters similar to those found in actual test development situations.

ACT Test

The 40,000 examinee responses from the math subtest are a simple random sample from those who took the same form of the ACT under standardized conditions on the same national test date. The target population is college-bound 11th and 12th graders. The sample of 40,000 examinee responses' consisted of 17,201 males and 22,799 females.

The ACT math subtest is a 60-item test that has a time limit of 60 minutes. This test is designed to assess mathematical reasoning skills acquired in courses from 1st to 12th grade. The items are five-option multiple-choice items. Data were provided in the scale of 0 (incorrect) and 1 (correct) for each response. The content areas included are pre-algebra, elementary algebra, intermediate algebra, coordinate geometry, plane geometry, and trigonometry.

Population Parameter Estimation

The population data from the ACT examinees' responses math subtest consisted of binary responses to items. The math subtest consisted of 60 questions. For gender, the math subtest contained 17,201 males and 22,799 females. DIF methods require an establishment of a reference group and focal group. The reference group was male and the focal group was female for the math subtest simulation.

The population parameters under the 3-parameter model were estimated with BILOG-MG 3[®] (Scientific Software International [SSI], 2003) for the 40,000 examinees' responses from the ACT data.

Data from achievement or aptitude tests measures maximal performance, rather than typical performance. Data measuring maximal performance with multiple choice tests often include a guessing parameter; therefore, the three-parameter IRT-based model is the most appropriate model for the ACT data (Reise & Waller, 1990). The three-parameter IRT-based model was used to estimate the a-parameter, b-parameter, and c-parameter for the simulation study.

$$P_i(\theta) = c_i + (1 - c_i) \left[\frac{e^{Dai(\theta - bi)}}{1 + e^{Dai(\theta - bi)}} \right]$$

The c- parameter estimated from the 40,000 ACT examinees' responses was not used in the simulation study, and was set to .20 as a constant for each item because if the c- parameter is not equal across reference and focal groups, the areas under the item characteristic curves are infinite due to the asymptotic tails and subgroups can not be estimated for DIF (Raju, 1990, 1988). Various studies have used .20 (Schnipske et al., 2000; Swaminathan & Rogers, 1990) for the c-parameter in similar studies. As Embretson and Reise (2000) explained, the actual pseudo-guessing parameter for a four-response multiple choice question of .25 is an overestimation and any overestimation creates inflated results. Notably, the average of the c parameter for the math subtest items was 0.19 (see Table 2).

A total of 60 sets of estimated a-, b-, and c- item parameters for the math subtest for male (reference) and female (focal) groups are listed in Table 2. The item parameters found in Table 2 for the math subtest were used to calculate Raju's area formula to detect the items that contained DIF.

Table 2 Estimated Item Parameters of Reference and Focal Groups

Item	Reference (Male) Group			Focal (Female) Group		
	<i>a</i>	<i>b</i>	<i>c</i>	<i>a</i>	<i>b</i>	<i>c</i>
1	0.94	-1.76	0.20	0.86	-2.19	0.20
2	1.99	-0.21	0.20	1.72	-0.55	0.20
3	1.24	-1.21	0.20	1.22	-1.05	0.20
4	1.47	-1.40	0.20	1.45	-1.73	0.20
5	2.22	-0.78	0.20	2.59	-0.70	0.20
6	1.21	-1.56	0.20	1.25	-1.43	0.20
7	1.14	-1.10	0.20	1.03	-1.55	0.20
8	1.51	-0.92	0.20	1.48	-1.49	0.20
9	1.56	-1.14	0.20	1.51	-0.91	0.20
10	2.28	-0.23	0.20	1.78	-0.65	0.20
11	2.16	-0.91	0.20	2.08	-0.75	0.20
12	1.60	-0.52	0.20	1.47	-0.82	0.20
13	1.89	0.26	0.20	1.66	-0.04	0.20
14	2.09	0.03	0.20	1.23	0.00	0.20
15	2.26	0.04	0.20	1.79	-0.21	0.20
16	1.40	-0.25	0.20	1.48	-0.29	0.20
17	2.50	-0.21	0.20	2.34	-0.09	0.20
18	1.76	-0.26	0.20	1.54	-0.57	0.20
19	1.78	-0.54	0.20	1.64	-0.27	0.20
20	2.42	-0.15	0.20	2.11	-0.68	0.20

Table 2 Continued

Item	Reference (Male) Group			Focal (Female) Group		
	<i>a</i>	<i>b</i>	<i>c</i>	<i>a</i>	<i>b</i>	<i>c</i>
21	1.12	-1.08	0.20	0.92	-1.21	0.20
22	0.60	0.84	0.20	0.47	1.15	0.20
23	2.17	-0.44	0.20	2.15	-0.52	0.20
24	1.55	0.30	0.20	1.37	0.26	0.20
25	1.32	-0.63	0.20	1.28	-0.42	0.20
26	2.32	0.31	0.20	2.45	0.46	0.20
27	2.11	-0.18	0.20	2.16	0.01	0.20
28	1.28	-0.02	0.20	1.07	-0.08	0.20
29	2.04	0.14	0.20	1.98	-0.10	0.20
30	2.92	0.08	0.20	2.92	0.12	0.20
31	1.76	0.47	0.20	1.63	0.40	0.20
32	1.86	0.30	0.20	1.82	0.29	0.20
33	1.20	0.37	0.20	1.05	0.48	0.20
34	1.76	-0.11	0.20	1.68	-0.24	0.20
35	2.09	0.34	0.20	1.88	0.15	0.20
36	1.41	-0.04	0.20	1.39	0.13	0.20
37	1.71	0.11	0.20	1.67	0.15	0.20
38	1.50	0.70	0.20	1.48	1.01	0.20
39	1.49	-0.18	0.20	1.60	0.14	0.20
40	1.76	-1.01	0.20	1.63	-0.88	0.20
41	1.13	2.24	0.20	1.14	2.19	0.20
42	2.59	0.30	0.20	2.21	0.14	0.20

Table 2 Continued

Item	Reference (Male) Group			Focal (Female) Group		
	<i>a</i>	<i>b</i>	<i>c</i>	<i>a</i>	<i>b</i>	<i>c</i>
43	1.70	0.87	0.20	1.49	0.79	0.20
44	2.67	0.26	0.20	2.35	0.35	0.20
45	0.61	0.36	0.20	0.72	0.30	0.20
46	1.29	0.07	0.20	1.24	0.42	0.20
47	2.03	0.88	0.20	1.80	0.82	0.20
48	2.50	0.82	0.20	2.26	0.71	0.20
49	2.02	0.80	0.20	1.68	0.80	0.20
50	2.04	0.48	0.20	1.86	0.63	0.20
51	1.91	1.57	0.20	1.45	1.69	0.20
52	1.80	1.39	0.20	1.75	1.60	0.20
53	2.03	1.03	0.20	2.08	1.23	0.20
54	2.44	1.42	0.20	1.82	1.78	0.20
55	1.16	1.58	0.20	1.38	1.73	0.20
56	3.07	1.43	0.20	3.57	1.54	0.20
57	1.80	1.33	0.20	1.63	1.35	0.20
58	2.25	1.05	0.20	2.21	1.21	0.20
59	2.71	1.53	0.20	2.57	1.88	0.20
60	2.47	2.26	0.20	2.34	2.61	0.20

Calculation of Raju's Area Formula to Detect DIF

To flag DIF items and later control for the amount of DIF in the simulation study the area between the item characteristic curves across groups were calculated with Raju's (1988) formula using the item parameters estimated from Table 2. The formula used to calculate DIF is

$$\text{Area}_{3\text{PL}} = (1 - c) \left| \left[\frac{2(a-a_1)}{D_{a_1 a_2}} \ln \left[1 + e^{D_{a_1 a_2} (b_2 - b_1) / (a_2 - a_1)} \right] - (b_2 - b_1) \right] \right|,$$

where D is a constant of 1.7.

Items with ICC area differences calculated from Raju's area for 3-PL model of 0.40 or higher were considered DIF. Due to the low number of DIF items, area calculations were manipulated by adding 0.2 to the difference in reference b parameters and focal b parameters. This was done by adding or subtracting .2 to the focal group b parameter in order to increase the area between focal and reference groups. Item parameters with areas of .40 and larger were determined as DIF items. Studies illustrated in Swaminathan and Rogers (1990) and Hambleton, Swaminathan, and Rogers (1991) cited area calculations of .6 and higher as DIF items. The area between two ICCs of at least .40 and larger were considered large enough to determine DIF in order to keep the data realistic for further analyses and interpretations of results.

A histogram of original item area calculations and manipulated item area calculations for the math subtest are in Figure 3. The area calculations in Figure 4 were used in the simulation.

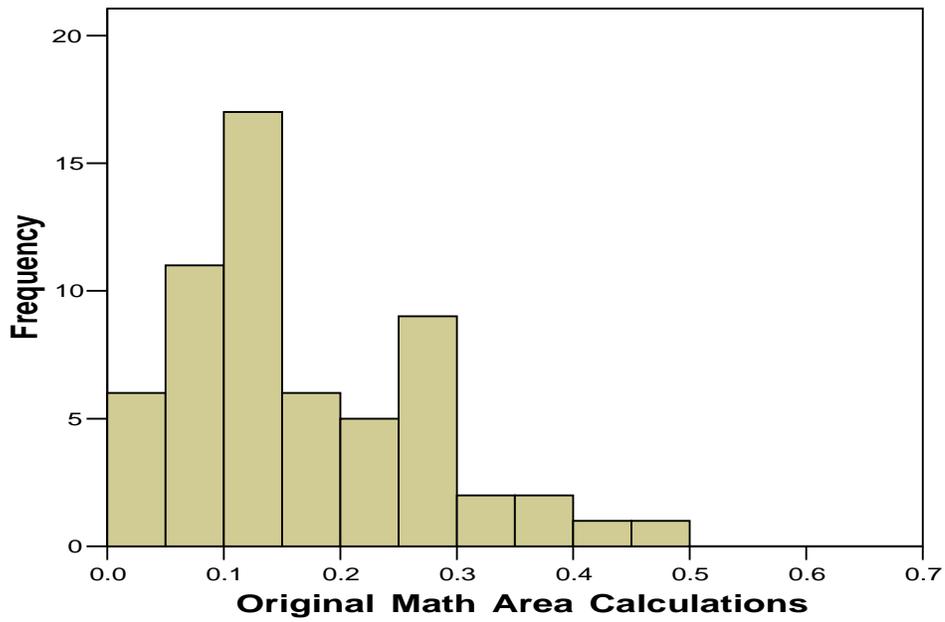


Figure 3. Histogram of Original Area Calculations for Math Subtest for Items 1-60

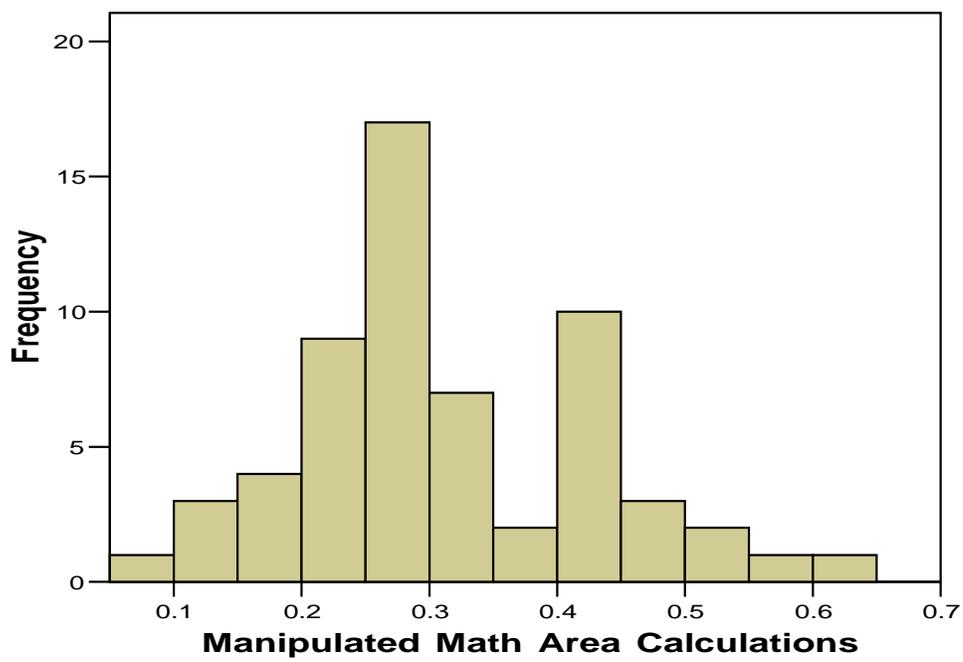


Figure 4. Histogram of Manipulated Area Calculations for Math Subtest for Items 1-60

After manipulation of item parameters for DIF as illustrated in Figure 4, the math subtest contained sixteen items with an area of 0.40 or higher. Previously, Table 2 listed the original item parameters for the reference and focal groups. Below, Table 3 lists the manipulated item b parameters for the reference and focal groups, manipulated area calculations, and original area calculations (for comparison purposes). The sixteen items determined to contain DIF (.040 or higher) are in italics.

Table 3 Original and Manipulated Female Item b Parameters and Area Calculations for DIF

Original b	Manipulated b	Original Area	Manipulated Area
-2.19	-2.39	0.35	0.51
-0.55	-0.75	0.28	0.44
-1.05	-0.85	0.13	0.29
-1.73	-1.93	0.26	0.42
-0.70	-0.91	0.07	0.10
-1.43	-1.23	0.11	0.27
-1.55	-1.75	0.36	0.52
-1.49	-1.69	0.46	0.62
-0.91	-0.71	0.18	0.34
-0.65	-0.85	0.33	0.49
-0.75	-0.55	0.13	0.29

Table 3 Continued

Original b	Manipulated b	Original Area	Manipulated Area
-0.82	-1.02	0.24	0.40
-0.04	-0.24	0.24	0.40
0.00	-0.20	0.22	0.27
-0.21	-0.41	0.20	0.36
-0.29	-0.50	0.04	0.20
-0.09	0.11	0.09	0.25
-0.57	-0.77	0.25	0.41
-0.27	-0.07	0.21	0.37
-0.68	-0.88	0.42	0.58
-1.21	-1.41	0.15	0.27
1.15	1.35	0.37	0.47
-0.52	-0.72	0.06	0.22
0.26	0.06	0.06	0.20
-0.42	-0.22	0.17	0.33
0.46	0.66	0.12	0.28
0.01	0.21	0.15	0.31
-0.08	-0.28	0.11	0.21
-0.10	-0.30	0.19	0.35
0.12	0.32	0.03	0.19
0.40	0.20	0.05	0.21
0.29	0.09	0.01	0.17
0.48	0.68	0.11	0.25

Table 3 Continued

Original b	Manipulated b	Original Area	Manipulated Area
-0.24	-0.44	0.10	0.26
0.15	-0.05	0.15	0.31
0.13	0.33	0.13	0.29
0.15	0.35	0.04	0.20
1.01	1.21	0.25	0.41
0.14	0.34	0.25	0.41
-0.88	-0.68	0.11	0.27
2.19	1.99	0.04	0.20
0.14	-0.06	0.13	0.29
0.79	0.99	0.07	0.11
0.35	0.55	0.08	0.23
0.30	0.50	0.18	0.20
0.42	0.62	0.28	0.44
0.82	1.03	0.06	0.12
0.71	0.91	0.09	0.07
0.80	0.60	0.06	0.17
0.63	0.83	0.12	0.28
1.69	1.89	0.14	0.26
1.60	1.80	0.17	0.33
1.23	1.43	0.17	0.33
1.78	1.98	0.29	0.45
1.73	1.93	0.14	0.28

Table 3 Continued

Original b	Manipulated b	Original Area	Manipulated Area
1.54	1.74	0.09	0.25
1.35	1.55	0.04	0.18
1.21	1.41	0.13	0.29
1.88	2.08	0.28	0.44
2.61	2.81	0.28	0.44

Note. $b = b$ parameter.

After the revision there were 44 non-DIF items and 16 DIF items. A sample of non-DIF items was systematically selected from the 44 non-DIF items and a sample of DIF items was systematically selected from the 16 DIF items; both were systematic samples that were drawn to uniformly span the range of parameter values within each distribution of items. The intent of the sampling was to investigate DIF properties across the ranges represented in the two populations. It is assumed that any variation observed will be smooth for unobserved items within the ranges. In the current study five non-DIF items and five DIF items were analyzed. The item parameters estimated from the 40,000 examinees' responses from the ACT math subtest were used to generate a sample of non-DIF and DIF items.

Five non-DIF items were systematically selected by examining the scatterplot of the a parameter and b parameter of the 44 non-DIF items. Four quadrants were created based on the mean of each parameter, or the intercept of the a and b parameters. The items that clustered around the a and b intercept were considered one group of items and

one item was selected to represent this cluster. Next, one item was selected from those items that fell outside of the middle cluster of items from each of the four quadrants. A total of five non-DIF items were taken to create a systematic sample of non-DIF items from the total of 44 non-DIF items from the ACT test. Item 6 was originally selected for the simulation; however item 5 was simulated. Item 6 was later simulated and added to the data, while item 5 was kept for analysis purposes. There were 6 non-DIF items in the sample. Figure 5 illustrates the scatter of the non-DIF items based on the a and b parameter, the cluster of items around the intercept, and the four quadrants used to select the sample of non-DIF items. Figure 5 also illustrates that the correlation between the a and b parameters of the non-DIF items has a weak correlation coefficient of 0.16. The five non-DIF items that were selected are tagged in Figure 5 with the item number.

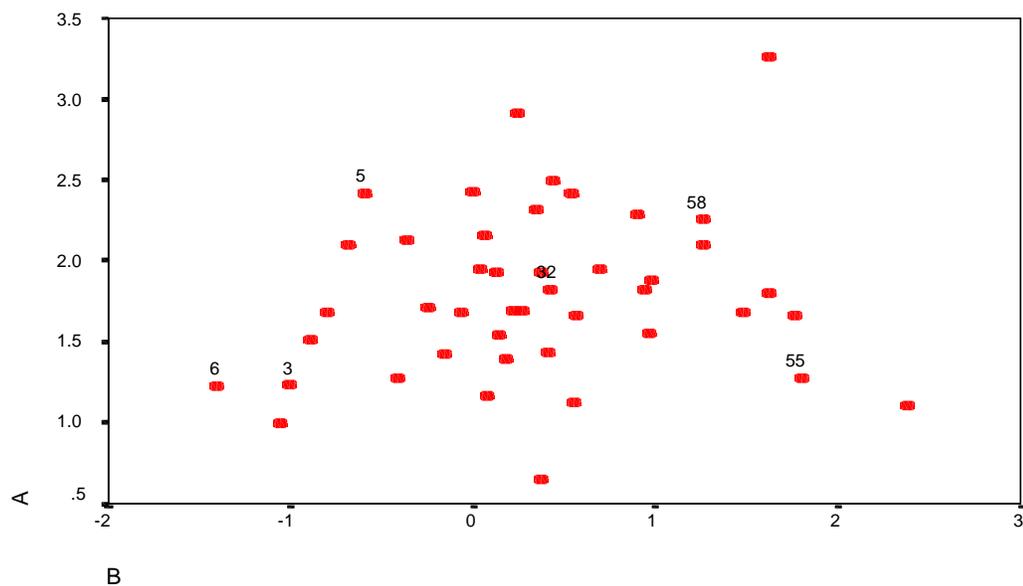


Figure 5. Scatterplot of a and b Parameters for Forty-Four Non-DIF Items from ACT Test

DIF items were systematically selected by running a correlation between the a parameter and b parameter of the 16 DIF items. Figure 6 illustrates the correlation between the a and b parameters of the DIF items and there is a strong correlation coefficient of 0.56. The DIF items found on the outer extremes of the correlation were selected for the current study. The five DIF items that were selected are tagged in Figure 6 with the item number.

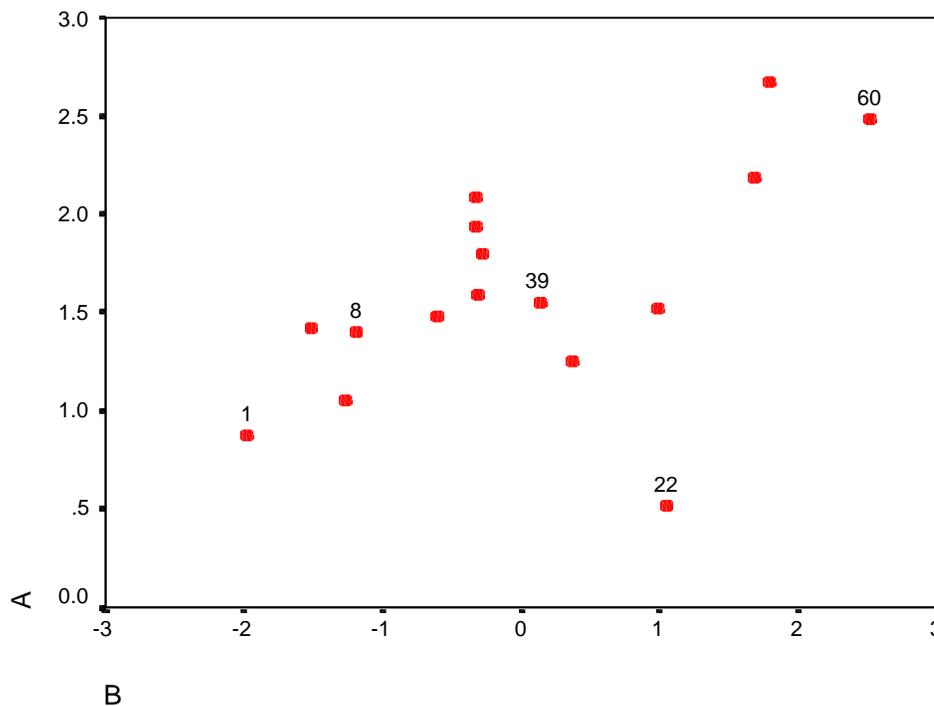


Figure 6. Scatterplot of a and b Parameters for Sixteen DIF Items from ACT Test

A total of eleven items, six non-DIF and five DIF, were included in the analysis. The a , b , and c item parameters for male (reference) and female (focal) groups

for Non-DIF items are listed in Table 4. The a, b, and c item parameters for male and female groups for DIF items are listed in Table 5.

Table 4 Item Parameters for Male and Female Non-DIF Items

Item	Male			Female			DIF
	a	b	c	A	b	c	
3	1.24	-1.21	0.20	1.22	-0.84	0.20	0.29
5	2.22	-0.78	0.20	2.59	-0.91	0.20	0.10
6	1.21	-1.56	0.20	1.25	-1.22	0.20	0.27
32	1.86	0.30	0.20	1.82	0.09	0.20	0.17
55	1.16	1.59	0.20	1.38	1.93	0.20	0.28
58	2.25	1.05	0.20	2.21	1.41	0.20	0.29

Note. a = a parameter, b = b parameter, c = c parameter.

Table 5 Item Parameters for Male and Female DIF Items

Item	Male			Female			DIF
	a	b	c	a	b	c	
1	0.94	-1.76	0.20	0.86	-2.39	0.20	0.51
8	1.51	-0.92	0.20	1.48	-1.69	0.20	0.62
22	0.60	0.84	0.20	0.47	1.35	0.20	0.47
39	1.49	-0.18	0.20	1.60	0.34	0.20	0.41
60	2.47	2.26	0.20	2.34	2.81	0.20	0.44

Note. a = a parameter, b = b parameter, c = c parameter.

Uniform DIF vs. Non-uniform DIF

The math subtest contained uniform DIF, and non-uniform DIF was not determined. Past research has illustrated that MH DIF does not distinguish between uniform and non-uniform DIF (Hambleton & Rogers, 1989; Swaminathan & Rogers, 1990), while recent research cautions against assuming MH DIF cannot distinguish non-uniform DIF. Mazor, Clauser, and Hambleton (1994) modified the MH statistic and reported that this modification improves the detection of non-uniform DIF (Hidalgo & Lopez-Pina, 2004). Hidalgo and Lopez-Pina (2004) suggested that non-uniform DIF is relatively rare in most test item data (Holland & Thayer, 1988; Dorans & Holland, 1993). The present study focused on uniform data and ICCs were created for observation of possible non uniform DIF. Item 22 was the only item that had possible non-uniform DIF. Several graduate students were asked to observe the ICC curve and

give an expert opinion and all were unsure of non-uniform DIF. The presence of the possible non-uniform DIF item was noted while interpreting the data analysis of the simulated data. Appendix A through Appendix K contain ICC plots of sample items used in the simulation.

Simulation Study

Using the SAS software, data generated from the sixty item parameters for male and female groups from Table 3 was used to simulate 200 samples for two populations with five sample size conditions. The two populations were 1) congruent and incongruent ability distributions, and 2) shape of population distributions. Sample size, ability distributions, and population distributions were the three main factors in the present study due to the effect these factors can have on false positives (Type I error), the power of DIF procedures, and effect sizes. The two population conditions and the five sample size conditions were selected to simulate realistic data samples.

Series of Sample Size Combinations

Five sample size conditions were established in this factor. Sample size affects false positives and the statistical power of any statistical analysis. The larger the sample size, the more likely it is for false positives to occur (Cohen, 1990, 1994; Thompson, 1999; Traub, 1983). Samples sizes used in DIF methodology require separate subsample sizes for the reference group and the focal group. In real data samples, usually the reference group is larger than the focal group (Roussos & Stout, 1996). An example of this is the comparison of Caucasians (reference) versus African Americans (focal). However, sometimes samples have equal size reference and focal groups, such as

gender. Sample size conditions consisted of five levels. The reference group had three sample sizes: $n_r = 100, 500, 1000$. The focal group had three sample sizes; $n_f = 100, 300, 1000$. The five sample size combinations of this study were $1000_r/100_f$, $500_r/100_f$, $300_r/300_f$, $1000_r/300_f$ and $1000_r/1000_f$. These give a realistic proportion to sample sizes between reference groups and focal groups of 10%, 5%, 100% (small sample size), and 30% and 100% (large sample size), respectively. The sample combinations included three combinations of unbalanced sample sizes and two balanced sample sizes for the reference and focal groups to model a variety of research situations.

The sample size combination of 1000/1000 was withheld from the simulation process due to the simulation program taking over 12 hours to complete for smaller sample size combinations (100/1000) and resources not available to run for more than that period of time. The sample combination was selected for omission due to larger balanced sample sizes containing more statistical power and that balanced sample sizes of 1000 are less likely in real data sets. The four remaining sample size combinations were expected to have more instability and accuracy.

Congruent and Incongruent Ability Distributions

The congruence of ability distributions was manipulated. Penny and Johnson (1999) and Monahan (2000) reported the effect on DIF analyses of differences in ability distributions. This difference was found to inflate Type I error and inflate effect sizes (Monahan, 2000; Penny & Johnson, 1999). Monahan (2000) found that only varying means across ability distributions did not influence Type I error rate, while varying both means and standard deviations inflated Type I error. Narayanan and Swaminathan

(1994) and Rogers and Swaminathan (1993) found that ability differences as large as 1 SD between reference and focal groups did not affect type I error rates (false positives). However, Jodoin and Zumbo (2001) reported that ability differences were common in real data sets and can create interaction effects with other variables. In this study three ability distribution conditions were modeled. The first condition was set with means of 0.0 for the focal and reference groups and standard deviations 1.0 for focal and reference groups. The second condition was set with equal means and unequal standard deviations across distributions, in which the focal group had a mean of 1.0 and a standard deviation of 1.0, and the reference group had a mean of 1.0 and standard deviation of 2.0. The third condition was set with unequal means and unequal standard deviations, in which the focal group had a mean of 0 and a standard deviation of 1 and the reference group had a mean of 1 and a standard deviation of 2.

Shape of Population Distributions

Test score population distribution shape was manipulated to test the robustness of the LR DIF and MH DIF methodologies, since a normal population distribution is assumed by standard IRT estimation methods. Four population distributions were manipulated. Schiel and King (1999) reported the skewness of the predictor variable did not have an effect on outcome in their simulation study of DIF. However, Johnson (1993) concluded/commented/ that there is actually a tolerable range. The present simulation study tested the limits of skewness and kurtosis by setting the coefficient of skewness and coefficient of kurtosis to an extreme limit. The extreme limits were based

on Johnson (1993) and Fleishman (1978) and compared to simulated distributions from Miccerri (1989).

Levels of skewness and kurtosis were determined using Fleishman's table of *Power Method Weights*, which include coefficients of skewness and kurtosis needed to simulate non-normal population distributions. Srivastava's (2002) explanation that kurtosis is not normally distributed if it is greater + 3.0 was taken into consideration along with Johnson's parameters for skewness. The four different populations distributions are normally distributed (coefficient of skewness = 0; coefficient of kurtosis = 0), moderately skewed and moderately leptokurtic (coefficient of skewness = 0.5; coefficient of kurtosis = 0.5), skewed and leptokurtic (coefficient of skewness = 1; coefficient of kurtosis = 0.5), skewed and extremely leptokurtic (coefficient of skewness = 0; coefficient of kurtosis = 3), and platykurtic (coefficient of skewness = 0.0; coefficient of kurtosis = -1).

Statistical Analyses

Logistic Regression Analysis

The first statistical analysis was a logistic regression with two independent variables. The focus of this analysis was on research question one and two which addressed the accuracy of the p value when supplemented by the effect size or log odds ratio when using the LR DIF and MH DIF methods, respectively. Two separate logistic regressions were run. The first logistic regression was run on the LR DIF method and the dependent variable was DIF/Non-DIF and the independent variables were the log transformation of the p value and the WLS R^2 . The second logistic regression was run

on the MH DIF method and the dependent variable was DIF/Non-DIF and the independent variables were the log transformation of the p value and the log odds ratio.

The first logistic regression analysis was used for research question one, which addressed the accuracy of the detection of DIF through LR DIF when WLS R^2 (effect size for LR DIF) supplements a statistical significance test (p value). The second logistic regression analysis was used for research question two, which addresses the accuracy of the detection of DIF through MH DIF when log odds ratio (effect size for MH DIF) supplements a statistical significance test (p value).

Repeated Measures ANOVA

The second type of statistical analysis is a repeated measures analysis of variance (RM-ANOVA) with a 4 x 3 x 5 multivariate design with three possible main effects, three two-way interaction effects, and one three-way interaction effect. RM-ANOVA was used to test for main effects and interaction effects of sample size, ability distributions, and population distributions. The RM-ANOVA is a mixed model design where the independent variables are fixed and the dependent variables are random. The two-way interaction effects between sample size and ability distribution, ability distribution and population distribution, and sample size and population distribution were tested. The three-way interaction effect of sample size, ability distribution, and population distribution was tested. Preliminary analyses were performed to verify if assumptions for multivariate normal distribution and independence of observation were met.

Two separate RM-ANOVAs were performed. The first RM-ANOVA addressed research questions three through six, which addressed whether the detection of DIF in an item is more likely with a statistical significance test (p value) for LR DIF method or a statistical significance test (p value) for MH DIF method with varying ability distributions, population distributions, and sample size combinations. Research question three addressed whether the detection of DIF in an item is more likely with a statistical significance test (p value) for LR DIF method or a statistical significance test (p value) for MH DIF method with varying ability distributions. Research question four addressed whether the detection of DIF in an item is more likely with a statistical significance test (p value) for LR DIF method or a statistical significance test (p value) for MH DIF method with varying population distributions. Research question five addressed whether the detection of DIF in an item is more likely with a statistical significance test (p value) for LR DIF method or a statistical significance test (p value) for MH DIF method with varying sample size combinations. Research question six addressed whether the detection of DIF in an item is more likely with a statistical significance test (p value) for LR DIF method or a statistical significance test (p value) for MH DIF method with varying ability distributions, population distributions, and sample size combinations. The second RM-ANOVA addressed research questions seven through ten, which addressed whether the detection of DIF in an item is more likely with weighted-least-squares R^2 (effect size for LR DIF) or log odds ratio (effect size for MH DIF) with varying ability distributions, population distributions, and sample size combinations. Research question seven addressed whether the detection of DIF in an item is more

likely with weighted-least-squares R^2 (effect size for LR DIF) or log odds ratio (effect size for MH DIF) with varying ability distributions. Research question eight addressed whether the detection of DIF in an item is more likely with weighted-least-squares R^2 (effect size for LR DIF) or log odds ratio (effect size for MH DIF) with varying population distributions. Research question nine addressed whether the detection of DIF in an item is more likely with weighted-least-squares R^2 (effect size for LR DIF) or log odds ratio (effect size for MH DIF) with varying sample size combinations. Research question ten addressed whether the detection of DIF in an item is more likely with weighted-least-squares R^2 (effect size for LR DIF) or log odds ratio (effect size for MH DIF) with varying ability distributions, population distributions, and sample size combinations.

CHAPTER IV

RESULTS

Chapter IV discusses the results of the ten research questions. Each research question is listed below and addressed.

- 1) Is the detection of DIF in an item more accurate when weighted-least-squares R^2 (effect size for LR DIF) supplements the use of a statistical significance test (p value) in the LR DIF analyses?
- 2) Is the detection of DIF in an item more accurate when a log odds ratio (effect size for MH DIF) supplements the use of a statistical significance (p value) test in the MH DIF analyses?
- 3) Is the detection of DIF more likely to occur when using a statistical significance test (p value) for LR DIF or a statistical significance test (p value) for MH DIF with varying ability distributions?
- 4) Is the detection of DIF more likely to occur when using a statistical significance test (p value) for LR DIF or a statistical significance test (p value) for MH DIF with varying population distributions?
- 5) Is the detection of DIF more likely to occur when using a statistical significance test (p value) for LR DIF or a statistical significance test (p value) for MH DIF with varying sample size combinations?
- 6) Is the detection of DIF more likely to occur when using a statistical significance test (p value) for LR DIF or a statistical significance test (p value) for MH DIF

with varying ability distributions, population distributions, and sample size combinations?

- 7) Is the detection of DIF more likely to occur when using weighted-least-squares (WLS) R^2 (effect size for LR DIF) or log odds ratio (effect size for MH DIF) with varying ability distributions?
- 8) Is the detection of DIF more likely to occur when using weighted-least-squares (WLS) R^2 (effect size for LR DIF) or log odds ratio (effect size for MH DIF) with varying population distributions?
- 9) Is the detection of DIF more likely to occur when using weighted-least-squares (WLS) R^2 (effect size for LR DIF) or log odds ratio (effect size for MH DIF) with varying sample size combinations?
- 10) Is the detection of DIF more likely to occur when using weighted-least-squares (WLS) R^2 (effect size for LR DIF) or log odds ratio (effect size for MH DIF) with varying ability distributions, population distributions, and sample size combinations?

Preliminary Analyses

Logistic Regressions

Research questions one and two were addressed with separate logistic regression analyses. Research question one was a logistic regression with DIF/Non-DIF as the dependent variable and the independent variables were the p value of the LR DIF method and the WLS R^2 of the LR DIF method. Research question two was a logistic regression with DIF/Non-DIF as the dependent variable and the independent variables

were the p value of the MH DIF method and the log odds ratio of the MH DIF method. Dependent variables were dichotomous (DIF/Non-DIF) and independent variables were continuous (p values, WLS R^2 , and log odds ratio). All logistic regression assumptions were met: 1) cases were independent, 2) independent variables were not linear combinations of each other for LR DIF logistic regression (pearson $r = -.22$) or MH DIF logistic regression (pearson $r = -.28$) and 3) the model was correctly specified.

Repeated Measures ANOVA

The dependent variables log odds ratio and WLS R^2 (effect sizes) in the RM-ANOVAs were transformed to control for large amounts of skewness and kurtosis. Analyses were run before and after transformation and results were similar. Effect sizes were transformed into categorical values with four levels, with the negligible level split into level one and level two. Initially, effect sizes were transformed into their respective categorical levels of negligible, moderate, and large effect sizes; however, the negligible levels for WLS R^2 and log odds ratio were both split into two levels due to high kurtosis. Category and respective effect size (dependent variables) value ranges are listed in Table 6.

Table 6 Categorical Levels for WLS R^2 and Log Odds Ratio

Category	WLS R^2 value range	Log odds ratio range
1	$\Delta R^2 < 0.035$	$\Delta_{\alpha MH} < 1 $
2	$0.0351 \leq \Delta R^2 \leq 0.0041$	$ 1.001 \leq \Delta_{\alpha MH} \leq 1.2999 $
3	$0.00411 \leq \Delta R^2 \leq 0.070$	$ 1.3 \leq \Delta_{\alpha MH} \leq 1.5 $
4	$\Delta R^2 > 0.070$	$\Delta_{\alpha MH} > 1.5 $

Note. $\Delta R^2 = \text{WLS } R^2$, $\Delta_{\alpha MH} = \text{Log odds ratio}$

Analyses were once again run before and after negligible categories were split into two levels and results were similar. The main difference in comparison of analyses before and after transformations was that the three-way interaction between-subjects effect of ability, sample size combination, and population distributions was statistically significant ($p, .001$) in all non-transformed analyses and not statistically significant in all other analyses. The partial eta squared statistic was .000 to three decimal places for this particular F ratio.

The descriptive statistics for the dependent variables for the RM-ANOVAs are listed in Table 7 and Table 8. Descriptive statistics for the independent variables across dependent variables p values and effect sizes for the RM-ANOVAs are in Appendix L and M, respectively.

Table 7 contains the statistics for the RM-ANOVA for the p value dependent variables and Table 8 contains the statistics for the RM-ANOVA for the effect size dependent variables. Skewness, kurtosis, and variance are included in the descriptive

statistics. All dependent variables meet the general rule of less than +3/-3 skewness and kurtosis. Missing data are found in dependent variables for LR DIF due to convergence difficulties in the simulation procedure. Glass and Hopkins (1996) reported if the variance of the dependent variables is equal to or less than a 4:1 ratio with the bigger sample in the numerator then the missing data will not affect the results. The ratio of the log odds ratio over the WLS R^2 is only slightly larger than the ratio of 4:1.

Table 7 Descriptive Statistics of RM-ANOVA for Comparison of p Values

DV	N	Mean	Std. Deviation	Variance	Skewness	Kurtosis
LR p value	131598	.231	.287	.082	1.155	.066
MH p value	132000	.276	.300	.090	.913	-.483

Note. DV = Dependent Variable, N = sample size.

Table 8 Descriptive Statistics of RM-ANOVA for Comparison of Effect Sizes

DV	N	Mean	Std. Deviation	Variance	Skewness	Kurtosis
Log odds ratio (categorical)	132000	2.766	1.131	1.279	-.346	-1.291
WLS R^2 (categorical)	131598	1.543	.557	.309	.437	-.390

Note. N = sample size.

Specific assumptions for RM-ANOVA are equality of covariance matrices, sphericity, and equal variances for dependent variables at each time point. All assumptions were rejected, which might be due to the large sample size of 132,000. Steps were taken to check for severe violations of assumptions in order to trust the results of the RM-ANOVAs. Two RM-ANOVAs were run, one to compare p values of MH DIF and LR DIF and one to compare log odds ratio for MH DIF and WLS R^2 for LR DIF.

RM-ANOVA for p Values

The equality of covariance matrices assumption was tested with Box's Test of Equality of Covariance Matrices in SPSS. Box's M was 33762.05 ($F = 190.682$, $df_1 = 177$, $df_2 = 620$, $p < .001$). Covariance matrices for each method of the dependent variable were MH DIF = .088 and LR DIF = .079. The bivariate correlation was run on the residuals for MH DIF and LR DIF methods ($r = .183$). The assumption of equality of covariance matrices was rejected, which could also be due to a sample size of 131,598. Areas not robust to the violation of equal variances are unequal sample sizes and large skewness and kurtosis. Skewness and kurtosis are low as seen in Table 7. Glass and Hopkins (1996) reported if the variance of the dependent variables (methods of dependent variable in this case) is equal to or less than a 4:1 ratio with the bigger sample in the numerator then the missing data (unbalanced design) will not affect the results. The ratio of the p value for MH DIF/ p value for LR DIF is .090/.082, which is less than a 4:1 ratio. Table 9 illustrates the results for the Levene's test of equality of variances

for MH DIF and LR DIF log p values across groups and overall variances of MH DIF and LR DIF methods.

Table 9 Levene's Test of Equality of Error Variances for p Values

Dependent Variable	F	df1	df2	Sig.	Variance
p LR	65.428	59	131538	.001	.082
p MH	13.306	59	131538	.001	.090

Note. F = F ratio, df1 = degrees of freedom numerator, df2 = degrees of freedom denominator, Sig. = significance.

RM-ANOVA for Effect Sizes

The equality of covariance matrices assumption was tested with Box's Test of Equality of Covariance Matrices in SPSS. Box's M was 7547.398 ($F = 42.626$, $df1 = 177$, $df2 = 620$, $p < .001$). Covariance matrices for each method of effect sizes was 1.279 for MH DIF and .297 for LR DIF. The bivariate correlation was run on the residuals for MH DIF and LR DIF methods ($r = .422$). The assumption of equality of covariance matrices was rejected, which could be due to a sample size of 131,598. The assumption of equal variances between groups for both methods of the dependent variable was rejected, which could also be due to the large sample size of 131,598. Areas not robust to the violation of equal variances are unequal sample sizes and large skewness and kurtosis. Skewness and kurtosis are low as seen in Table 8. Glass and Hopkins (1996) reported if the variance of the dependent variables (methods of dependent variable in this case) is equal to or less than a 4:1 ratio with the bigger sample

in the numerator then the missing data (unbalanced design) will not affect the results. The ratio of the log odds ratio for MH DIF/WLS R^2 for LR DIF is 1.279/.309, which is slightly higher than a 4:1 ratio. Table 10 illustrates the results for the Levene's test of equality of variances for MH DIF and LR DIF effect sizes across groups and overall variances of MH DIF and LR DIF methods.

Table 10 Levene's Test of Equality of Error Variances for Effect Sizes

	F	df1	df2	Sig.	Variance
Log odds ratio	49.161	59	131538	.001	1.279
WLS R^2	52.088	59	131538	.001	.309

Note. F = F ratio, df1 = degrees of freedom numerator, df2 = degrees of freedom denominator, Sig. = significance.

Conclusions of Preliminary Analyses

Preliminary analyses suggest results for logistic regressions are acceptable and unlikely due to unstable data and interpretation can be trusted. The RM-ANOVAs need to be interpreted with caution due to all assumptions being violated. Violations of assumptions are less severe for the RM-ANOVA for the p values than for the assumptions of the RM-ANOVA for the effect sizes. The strong assumption violations for the effect sizes RM-ANOVA are due to the different covariance matrices and high bivariate correlation of the dependent variable residuals. However, preliminary analyses of these violations suggest results can be trusted to cautiously interpret and make

decisions with discretion. Notably, assumptions will always be rejected with such a large sample size ($n = 132,000$).

Research Question 1

Research question one addresses whether the detection of DIF in an item is more accurate when WLS R^2 (effect size for LR DIF) supplements the use of a statistical significance test (p value) in the LR DIF analyses. Logistic regression was used to analyze the accuracy of the detection of DIF when the WLS R^2 is included with a formal statistical significance test (p value). The dependent variable was DIF/Non-DIF (0 was Non-DIF and 1 was DIF). The independent variables were p value for LR DIF and WLS R^2 for LR DIF. The total sample size was 132,000 with 402 missing cases. The total number of cases used in the analysis was 131,598.

The overall percent of cases predicted correctly by the null model (without predictor variables) was 54.5, the model with the p value predictor variable also predicted 54.5 of the cases, and the full model with the predictor variables p value and WLS R^2 predicted 65.2 percent of the cases. Notably, the p value predictor variable does have a higher corrected prediction percentage, but due to the high number of cases this change does not show in the overall percentages. The full model seems to be the best model of fit. When the model added the WLS R^2 predictor variable to the model the percentage of cases predicted increased 10.7 percent. Table 11 illustrates the proportion of predicted cases versus observed cases for the model without predictors (null model), Table 12 illustrates the model with the p value predictor variable, and

Table 13 illustrates the model with the p value and WLS R^2 predictor variables (full model).

Table 11 Classification Table for the Null Model for LR DIF

Observed	Predicted		Percentage
	Non-DIF	DIF	
Non-DIF	71702	0	100
DIF	59896	0	0
Overall Percentage			54.5

Table 12 Classification Table for Model with p Value Predictor Variable for LR DIF

Observed	Predicted		Percentage
	Non-DIF	DIF	
Non-DIF	70183	1519	97.9
DIF	58358	1538	2.6
Overall Percentage			54.5

Table 13 Classification Table for the Full Model for LR DIF

Observed	Predicted		Percentage
	Non-DIF	DIF	
Non-DIF	60278	11424	84.1
DIF	34432	25464	42.5
Overall Percentage			65.2

In Block 1 the model including only the p value LR had a chi-square of 177.675 ($p < .001$) and was statistically significant, but not meaningful. In Block 2 the model added WLS R^2 to the model and the chi-square was 17264.244 ($p < .001$), an increase of 17086.569. The Nagelkerke R-square increased from .002 to .164 with the addition of the WLS R^2 . The increase in chi-square and Nagelkerke R-square indicates meaningful improvement and statistical significance for the WLS R^2 predictor variable. Table 14 lists the chi-square statistic, degrees of freedom, Nagelkerke R-square, and significance level for the model in Block 1 (predictor variable p value only) and the model in Block 2 (the p value and the WLS R^2).

Table 14 Statistics for Block 1 and Block 2 Models for LR DIF

Model	Chi-square	df	Nagelkirke R-square	Sig.
Block 1	177.675	1	.002	< .001
Block 2	17264.244	2	.164	< .001

Note. df = degrees of freedom, Sig. = significance.

For the full model in Block 2 the omnibus test had a chi-square of 17264.244 (df=2, $p < .001$) and was statistically significant. The predictor variables in the model were the p value and WLS R^2 for LR DIF. Table 15 lists the statistics for the predictor variables for the full model in Block 2.

Table 15 Statistics for Predictor Variables in Full Model for LR DIF

Variable	B	S.E.	Wald	df	Sig.	Exp(B)
p value	.984	.021	2160.438	1	< .001	2.675
WLS R^2	96.368	.891	11692.602	1	< .001	7.114

Note. B = Beta weight, S.E. = standard error, df = degrees of freedom, Sig. = significance, Exp(B) = odds ratio.

The predictor p value is statistically significant with a B weight (log odds ratio) of .984 and an odds ratio of 2.675. The predictor variable p value was 2% less likely to detect DIF if the item was a DIF item, all other variables held constant. The desired outcome for the predictor variable is a negative B weight due to the scale of the variable (increases for Non-DIF and decreases for DIF). The predictor variable WLS R^2 is

statistically significant with a B weight (log odds ratio) of 96.368 and an odds ratio of 7.115. WLS R^2 was more likely to detect DIF if the item was a DIF item, all other variables held constant. The WLS R^2 predictor variable is a stronger predictor than the p value predictor variable in the detection of DIF when the item is predetermined to contain DIF.

Research Question 2

Research question two addresses whether the detection of DIF in an item is more accurate when log odds ratio (effect size for MH DIF) supplements the use of a statistical significance test (p value) in the MH DIF analyses. Logistic regression was used to analyze the accuracy of the detection of DIF when the log odds ratio is included with a formal statistical significance test (p value). The dependent variable was DIF/Non-DIF (0 was Non-DIF and 1 was DIF). The independent variables were p value for MH DIF and log odds ratio for MH DIF. There were no missing cases, which gave a total sample size of 132,000 cases included in the logistic regression.

The overall percent of cases predicted by the null model (without predictor variables) was 54.5, the model with the p value predictor variable predicted 61.0 of the cases, and the full model with the predictor variables p value and log odds ratio predicted 60.6 percent of the cases. When the model including only the p value added the log odds ratio predictor variable to the model the percentage of cases predicted decreased .04 percent in the full model; however, the chi square increased for the full model and the Nagelkirke R-square increased from .067 to .069. Table 16 illustrates the proportion of predicted cases versus observed cases for the model without predictors

(null model), Table 17 illustrates the model with the p value predictor variable, and Table 18 illustrates the model with the p value and log odds ratio predictor variables.

Table 16 Classification Table for Null Model for MH DIF

Observed	Predicted		Percentage
	Non-DIF	DIF	
Non-DIF	72000	0	100
DIF	60000	0	0
Overall Percentage			54.5

Table 17 Classification Table for Model with p Value Predictor Variable for MH DIF

Observed	Predicted		Percentage
	Non-DIF	DIF	
Non-DIF	42999	29001	59.7
DIF	22482	37518	62.5
Overall Percentage			61.0

Table 18 Classification Table for the Full Model for MH DIF

Observed	Predicted		Percentage
	Non-DIF	DIF	
Non-DIF	42859	29141	59.5
DIF	22892	37108	61.8
Overall Percentage			60.6

In Block 1 the model including only the p value had a chi-square of 6784.854 ($p < .001$) and was statistically significant, but not meaningful. In Block 2 the model added log odds ratio to the model and the chi-square was 6967.834 ($p < .001$), an increase of 182.980. The Nagelkerke R-square increased from .067 to .069 with the addition of the log odds ratio predictor variable. The increase in chi-square and Nagelkerke R-square indicates improvement and statistical significance for the log odds ratio predictor variable; however, improvement might not be meaningful based on the percentage of cases predicted in Table 17 and Table 18. Table 19 lists the chi-square statistic, degrees of freedom, Nagelkerke R-square, and significance level for the model in Block 1 (predictor variable p value only) and the model in Block 2 (the p value and the log odds ratio).

Table 19 Statistics of Block 1 and Block 2 Models for MH DIF

Model	Chi-square	df	Nagelkerke R-square	Sig.
Block 1	6784.854	1	.067	< .001
Block 2	6947.234	2	.069	< .001

Note. df = degrees of freedom, Sig. = significance.

For the full model in Block 2 the omnibus test had a chi-square of 6967.834 (df=2, $p < .001$) and was statistically significant. The predictor variables in the model were the p value and log odds ratio. Table 20 lists the statistics for the predictor variables for the full model in Block 2.

Table 20 Statistics for Predictor Variables in Full Model for MH DIF

Variable	B	S.E.	Wald	Df	Sig.	Exp(B)
p value	-1.516	.021	5213.536	1	< .001	.219
Log odds ratio	.078	.006	151.513	1	< .001	1.081

Note. B = Beta weight, S.E. = standard error, df = degrees of freedom, Sig. = significance, Exp(B) = odds ratio.

The predictor variable p value is statistically significant with a B weight (log odds ratio) of -1.516 and an odds ratio of .219. A negative B weight shows a decrease in the predictor variable for a decrease in the dependent variable; therefore, when items are predetermined DIF the p values overall decrease. The predictor variable p value was 78% more likely to detect DIF if the item was a DIF item, all other variables held

constant. The predictor variable log odds ratio is statistically significant with a B weight (log odds ratio) of .078 and an odds ratio of 1.081. The predictor variable log odds ratio was 93% more likely to detect DIF if the item was a DIF item, all other variables held constant. The log odds ratio predictor variable is a stronger predictor than the p value predictor variable in the detection of DIF when the item is predetermined to contain DIF.

Research Question 3

Research question 3 addresses whether the detection of DIF in an item is more likely to occur with a statistical significance test (p value) for LR DIF or a statistical significance test (p value) for MH DIF with varying ability distributions. RM-ANOVA with p values as the dependent variable and ability distributions as the independent variable was used to answer this research question. The independent variable's levels were *normal* (mM0/mF0/stM1/stF1), *moderate* (mM1/mF1/stM1/stF2), and *severe* (mM0/mF1/stM1/stF2) differences between focal and reference ability distributions. Within-subjects effect was the focus of the RM-ANOVA to reveal possible differences in methods across ability distributions. Descriptives for the RM-ANOVA including varying ability distributions as an independent variable are in appendix L.

RM-ANOVA revealed statistically significant differences in the within-subjects effects test across methods for ability distribution levels ($F(df=2, 131,538) = 809.883, p < .001, \text{effect size} = 1.2\%$). The overall p values for MH DIF were different than LR DIF indicating MH DIF and LR DIF do not similarly detect DIF across varying ability distributions. Table 21 lists descriptive statistics for estimated marginal means for LR

DIF and MH DIF for ability distribution levels. Figure 7 illustrates the method by ability distribution effect for the estimated marginal means from Table 21.

Table 21 Estimated Marginal Means for Method by Ability Distribution (p Values)

Ability	Method	Mean	Std. Error
Normal	LR DIF	.276	.001
	MH DIF	.269	.001
Moderate	LR DIF	.180	.001
	MH DIF	.273	.001
Severe	LR DIF	.237	.001
	MH DIF	.284	.001

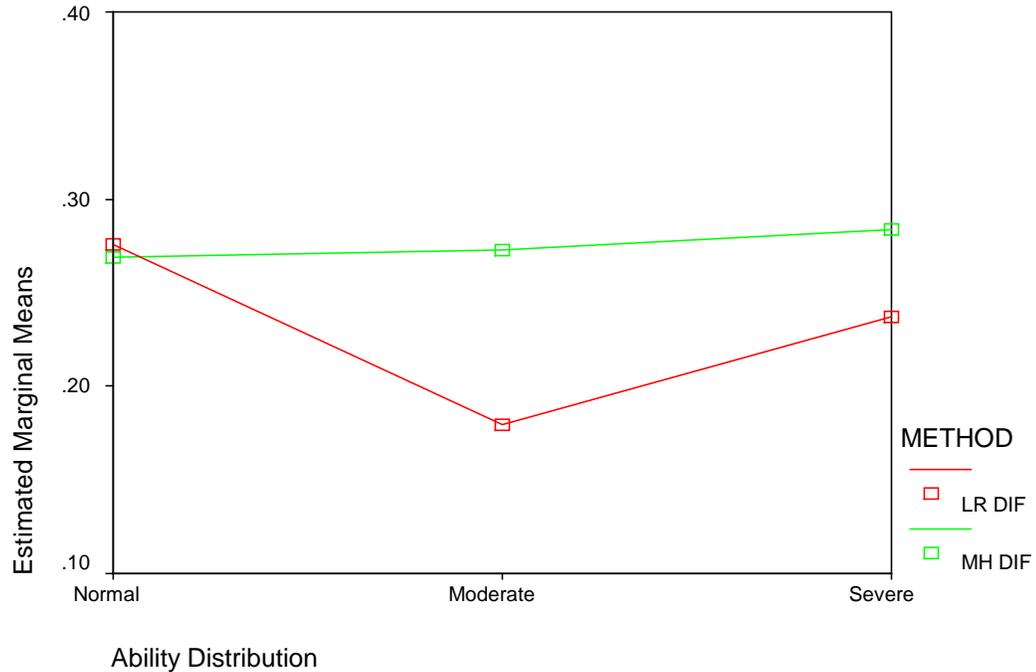


Figure 7. Estimated Marginal Means for MH DIF and LR DIF Ability Distributions (p Values)

Linear and quadratic contrasts were run for method by ability distribution. The interaction effect had a statistically significant linear trend ($F(df=1, 131538) = 1619.809, p < .001, \text{effect size} = 1.2\%$) and a statistically significant quadratic trend ($F(df=1, 131538) = 4.559, p = .032, \text{effect size} < .00\%$). The quadratic trend seems to be the best fit for the method by ability distribution interaction effect due to its smallest F statistic and effect size. Table 22 lists the statistics calculated for the linear and quadratic trends for method by ability distribution with corrected error sums of squares. Figure 7 illustrates the trends for each method and for the interaction of method by ability distribution.

Table 22 Linear and Quadratic Trends for Method by Ability Distribution (p Values)

Effect	Type I SOS	df	Mean Square	F	Sig.	Effect Size (%)
method	129.333	1	129.333	1901.956	.000	1.4
method * alin	110.147	1	110.147	1619.809	.000	1.2
method * aquad	.310	1	.310	4.559	.032	< .00
Error (method)	8970.543	131538	.068			
Total	9227.018					

Note. df = degrees of freedom, F = F ratio, Sig. = significance, alin = ability linear contrast, aquad = ability quadratic contrast.

Research Question 4

Research question 4 addresses whether the detection of DIF in an item is more likely to occur with a statistical significance test (p value) for LR DIF or a statistical significance test (p value) for MH DIF with varying population distributions. RM-ANOVA was run with p values as the dependent variable and population distribution as the independent variable. The independent variable's levels were *normally distributed* (coefficient of skewness = 0; coefficient of kurtosis = 0), *moderately skewed and moderately leptokurtic* (coefficient of skewness = 0.5; coefficient of kurtosis = 0.5),

skewed and leptokurtic (coefficient of skewness = 1; coefficient of kurtosis = 0.5), *skewed and extremely leptokurtic* (coefficient of skewness = 0; coefficient of kurtosis = 3), and *platykurtic* (coefficient of skewness = 0.0; coefficient of kurtosis = -1).

Within-subjects effect was the focus of the RM-ANOVA to reveal possible differences in methods across population distribution levels. Descriptives for the RM-ANOVA including varying population distributions as an independent variable are in appendix L.

RM-ANOVA revealed no statistically significant differences in the within-subjects effects test across methods for population distribution levels (F ($df=4, 131,538$) = .572, $p = .683$, effect size < .00%). The overall p values for MH DIF were not statistically significantly different than LR DIF indicating MH DIF and LR DIF detect DIF similarly across varying population distributions with LR DIF p values slightly lower across population distribution levels.

There was not an overall within-subjects effect; therefore, simple custom contrasts were run on method by population distribution. The overall simple contrast was not statistically significant (F ($df = 4, 131538$) = .628, $p < .643$, effect size < .00%) for population distribution. Level one, *normally distributed*, was the reference for the custom contrasts and not statistically significant versus level 2 ($p = .667$), level 3 ($p = .349$), level 4 ($p = .454$), and level 5 ($p = .900$). Table 23 lists descriptive statistics for estimated marginal means for LR DIF and MH DIF for population distribution levels.

Table 23 Estimated Marginal Means for Method by Population Distribution (p Values)

Population Levels	Method	Mean	Std. Error
Normally Dist.	LR DIF	.229	.002
	MH DIF	.276	.002
Moderately Dist.	LR DIF	.229	.002
	MH DIF	.274	.002
Skew/Lepto.	LR DIF	.233	.002
	MH DIF	.275	.002
Skew/Extremely Lepto	LR DIF	.231	.002
	MH DIF	.276	.002
Platykurtic	LR DIF	.231	.002
	MH DIF	.274	.002

Figure 8 illustrates the lack of linear, quadratic, or cubic trends for method by population distribution levels with estimated marginal means for population distribution levels.

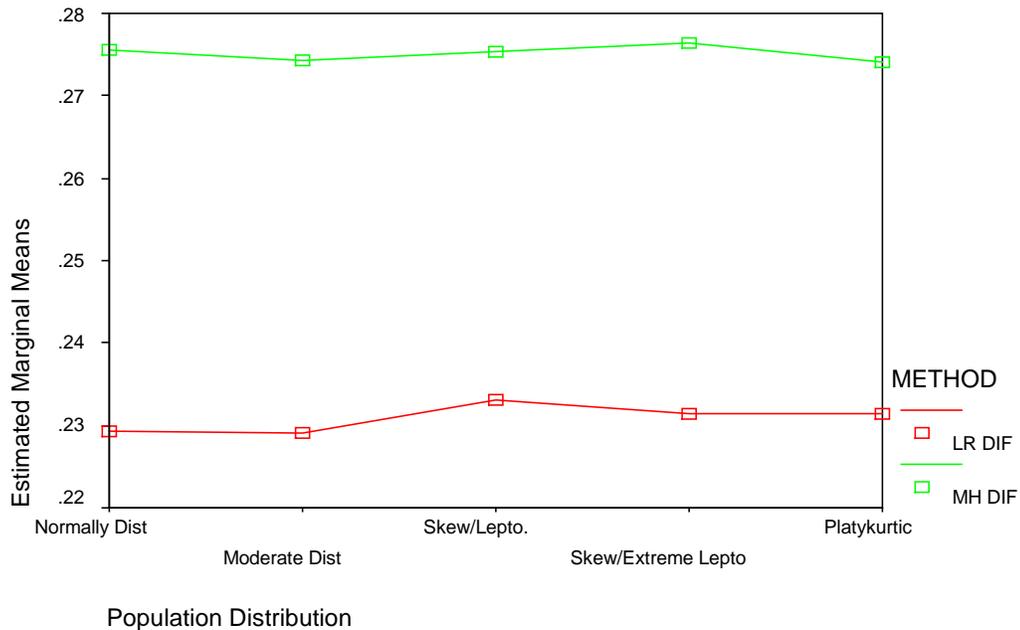


Figure 8. Estimated Marginal Means for MH DIF and LR DIF Population Distributions (p Values)

Research Question 5

Research question 5 addresses whether the detection of DIF in an item is more likely to occur with a statistical significance test (p value) for LR DIF or a statistical significance test (p value) for MH DIF with varying sample size combinations. RM-ANOVA was run with p values as the dependent variable and sample size combination as the independent variable. The four sample size combination levels were 1000_r/100_f, 500_r/100_f, 300_r/300_f, 1000_r/300_f.

Within-subjects effect was the focus of the RM-ANOVA to reveal possible differences in methods across sample size combination levels. Descriptives for the RM-ANOVA including the independent variable sample size combination are in appendix L.

RM-ANOVA revealed statistically significant differences in the within-subjects effects test across methods for sample size combinations ($F(df=3, 131,538) = 15.354, p < .001, \text{effect size} < .03\%$). The overall p values for MH DIF were statistically significantly different than LR DIF indicating MH DIF and LR DIF do not detect DIF similarly across sample size combinations. Figure 9 illustrates the method by ability distribution effect for the estimated marginal means from Table 24. Table 24 lists descriptive statistics for estimated marginal means for LR DIF and MH DIF for population distribution levels.

Table 24 Estimated Marginal Means for Method by Sample Size Combination (p Values)

Sample Size Levels (r/f)	Method	Mean	Std. Error
1000/100	LR DIF	.256	.002
	MH DIF	.309	.002
500/100	LR DIF	.276	.002
	MH DIF	.324	.002
300/300	LR DIF	.218	.002
	MH DIF	.235	.002
1000/300	LR DIF	.173	.002
	MH DIF	.212	.002

Note. r/f = reference/focal

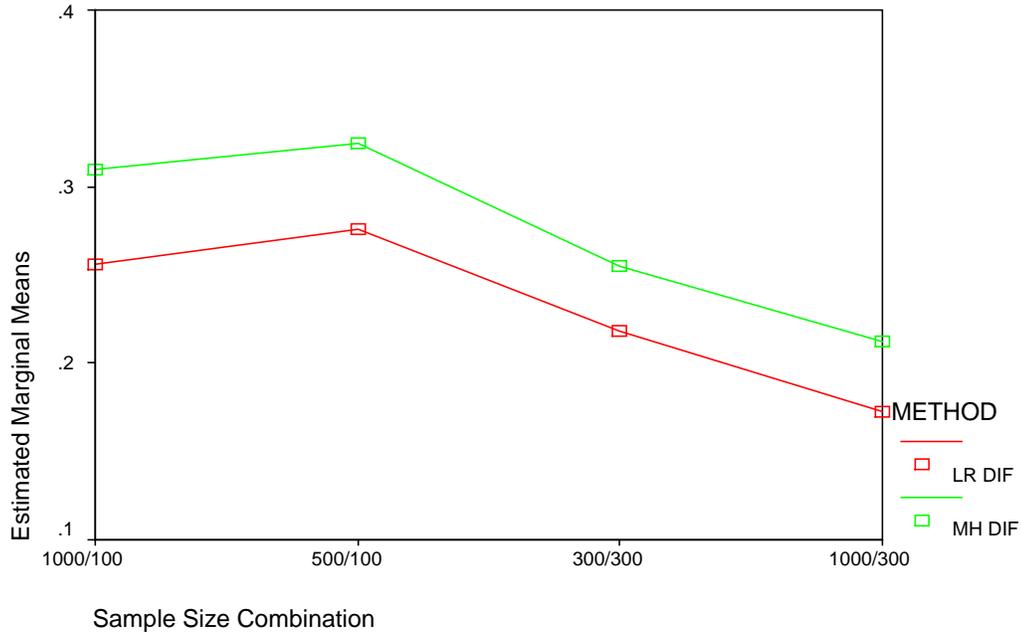


Figure 9. Estimated Marginal Means for MH DIF and LR DIF Sample Size Combination (p Values)

Linear, quadratic, and cubic contrasts were run for method by sample size combinations. The interaction effect had a statistically significant linear trend ($F(df=1, 131538) = 36.603, p < .001, \text{effect size} = 0.03\%$), a statistically significant quadratic trend ($F(df=1, 131538) = 4.118, p = .04, \text{effect size} < 0.00\%$), and a statistically significant cubic trend ($F(df=1, 131538) = 5.485, p = .02, \text{effect size} < 0.00\%$). The quadratic trend seems to be the best fit for the method by ability distribution interaction effect due to its smallest F statistic and effect size, with the cubic trend being a slightly better fit than the linear trend. Table 25 lists the statistics calculated for the linear and

quadratic trends for method by ability distribution. Figure 9 illustrates the trends for each method and for the interaction of methods by sample size combinations.

Table 25 Linear and Quadratic Trends for Method by Sample Size Combination (p Values)

Effect	Type I SOS	df	MS	F	Sig.	Effect Size (%)
Method	129.333	1	129.333	1901.956	< .00	1.4
method * slin	2.490	1	2.490	36.479	< .00	.03
method * squad	.280	1	.280	4.099	.04	< .00
method * scub	.373	1	.373	5.468	.02	< .00
Error (method)	8970.543	13153 8	.068			
Total	9227.018					

Note. df = degrees of freedom, MS = Mean Square, F = F ratio, Sig. = significance, slin = sample linear contrast, squad = sample quadratic contrast, scub = sample cubic contrast.

Research Question 6

Research question 6 addresses whether the detection of DIF in an item is more likely to occur with a statistical significance test (p value) for LR DIF or a statistical

significance test (p value) for MH DIF with varying ability distributions, population distributions, and sample size combinations. RM-ANOVA was run with p values as the dependent variable and ability distribution, population distribution, and sample size combination as the independent variables.

Within-subjects effect was the focus of the RM-ANOVA to reveal possible differences in methods across sample size combination levels. Descriptives for the RM-ANOVA including sample size combinations as an independent variable are in appendix L.

RM-ANOVA revealed no statistically significant differences for the within-subjects effects test for an interaction for method by sample size combination by population distribution ($F(df=12, 131,538) = .667, p < .785, \text{effect size} < 0.00\%$).

RM-ANOVA revealed no statistically significant differences for the within-subjects effects test for an interaction for method by ability distribution by population distribution ($F(df=8, 131,538) = 1.539, p < .138, \text{effect size} < 0.00\%$).

RM-ANOVA revealed no statistically significant differences for the within-subjects effects test for an interaction for method by sample size combination by ability distribution by population distribution ($F(df=24, 131,538) = 1.369, p < .107, \text{effect size} < 0.00\%$).

RM-ANOVA revealed statistically significant differences for the within-subjects effects test for an interaction for method by ability distribution by sample size combinations ($F(df=6, 131,538) = 23.974, p < .001, \text{effect size} = 0.11\%$). Table 26 lists descriptive statistics for estimated marginal means for LR DIF and MH DIF for the

interaction effect of method by sample size combination by ability distribution. Figure 10 through Figure 12 illustrate the interaction effects for method by sample size combinations by ability distribution. Figures are shown for each ability distribution level. Figure 10 illustrates the estimated marginal means for sample size combinations for ability distribution level *normal*. Figure 11 illustrates the estimated marginal means for sample size combinations for ability distribution level *moderate*. Figure 12 illustrates the estimated marginal means for sample size combinations for ability distribution level *severe*.

Table 26 Estimated Marginal Means for Method by Ability Distribution by Sample Size Combinations (p Values)

Sample Size	Ability	Method	Mean	Std. Error
1000/100	Normal	LR DIF	.320	.003
		MH DIF	.304	.003
	Moderate	LR DIF	.207	.003
		MH DIF	.310	.003
	Severe	LR DIF	.240	.003
		MH DIF	.314	.003
500/100	Normal	LR DIF	.329	.003
		MH DIF	.317	.003
	Moderate	LR DIF	.227	.003

Table 26 Continued

Sample Size	Ability	Method	Mean	Std. Error
		MH DIF	.322	.003
	Severe	LR DIF	.273	.003
		MH DIF	.334	.003
300/300	Normal	LR DIF	.256	.003
		MH DIF	.254	.003
	Moderate	LR DIF	.162	.003
		MH DIF	.239	.003
	Severe	LR DIF	.238	.003
		MH DIF	.272	.003
1000/300	Normal	LR DIF	.200	.003
		MH DIF	.200	.003
	Moderate	LR DIF	.123	.003
		MH DIF	.220	.003
	Severe	LR DIF	.196	.003
		MH DIF	.216	.003

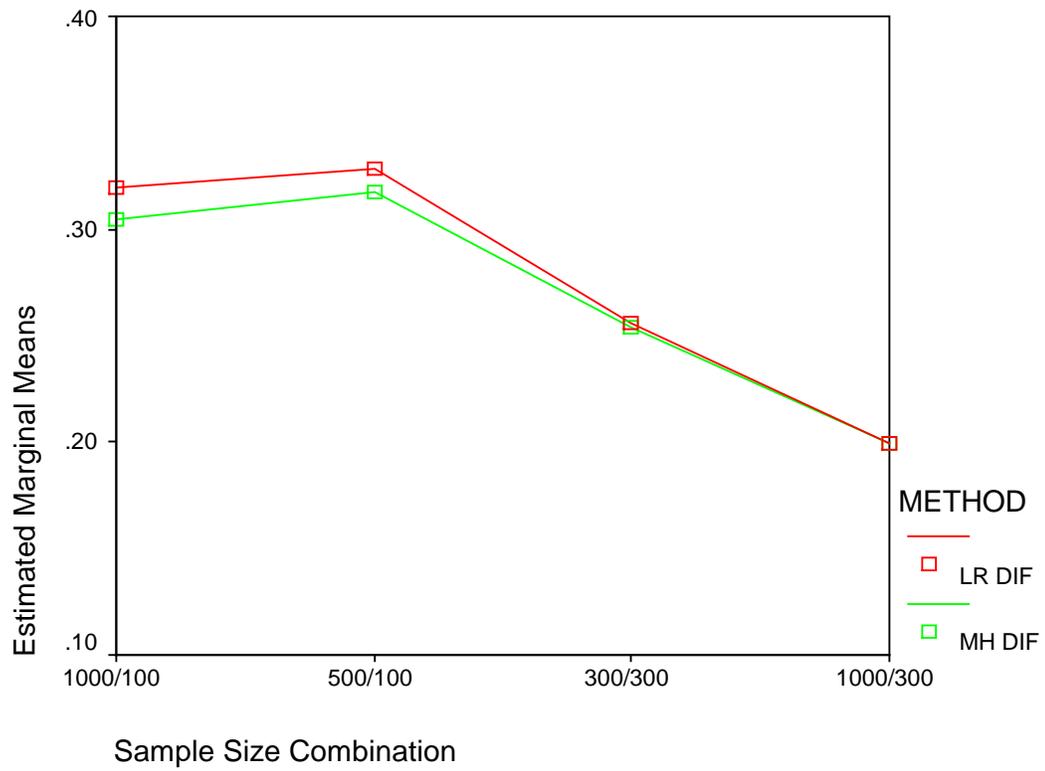


Figure 10. Estimated Marginal Means for Normal Ability Distributions (Method by Sample Size) (p Values)

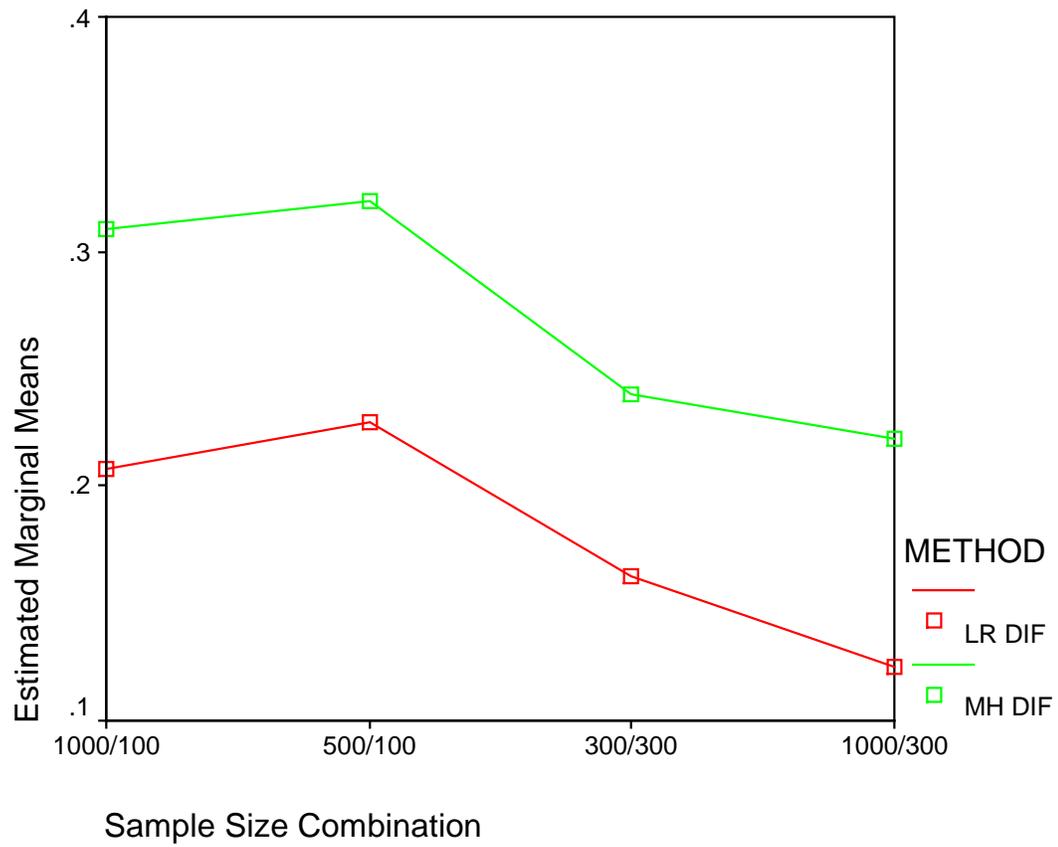


Figure 11. Estimated Marginal Means for Moderate Ability Distributions (Method by Sample Size) (p Values)

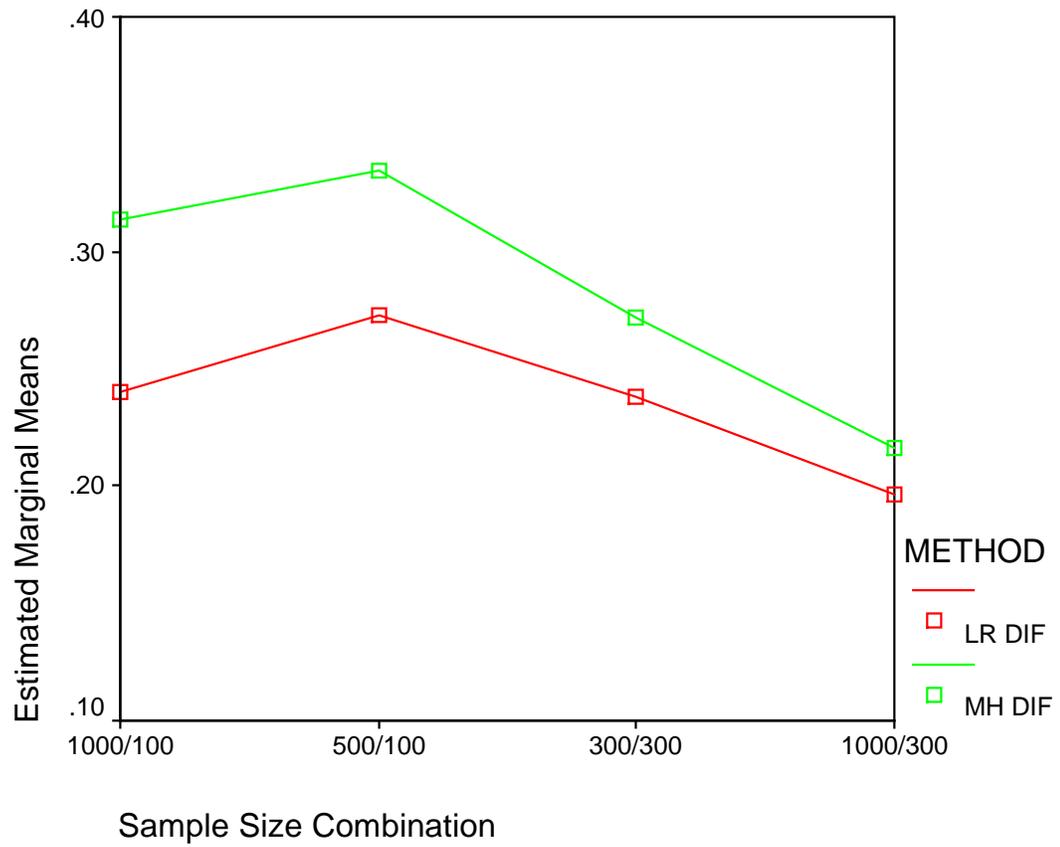


Figure 12. Estimated Marginal Means for Severe Ability Distributions (Method by Sample Size) (p Values)

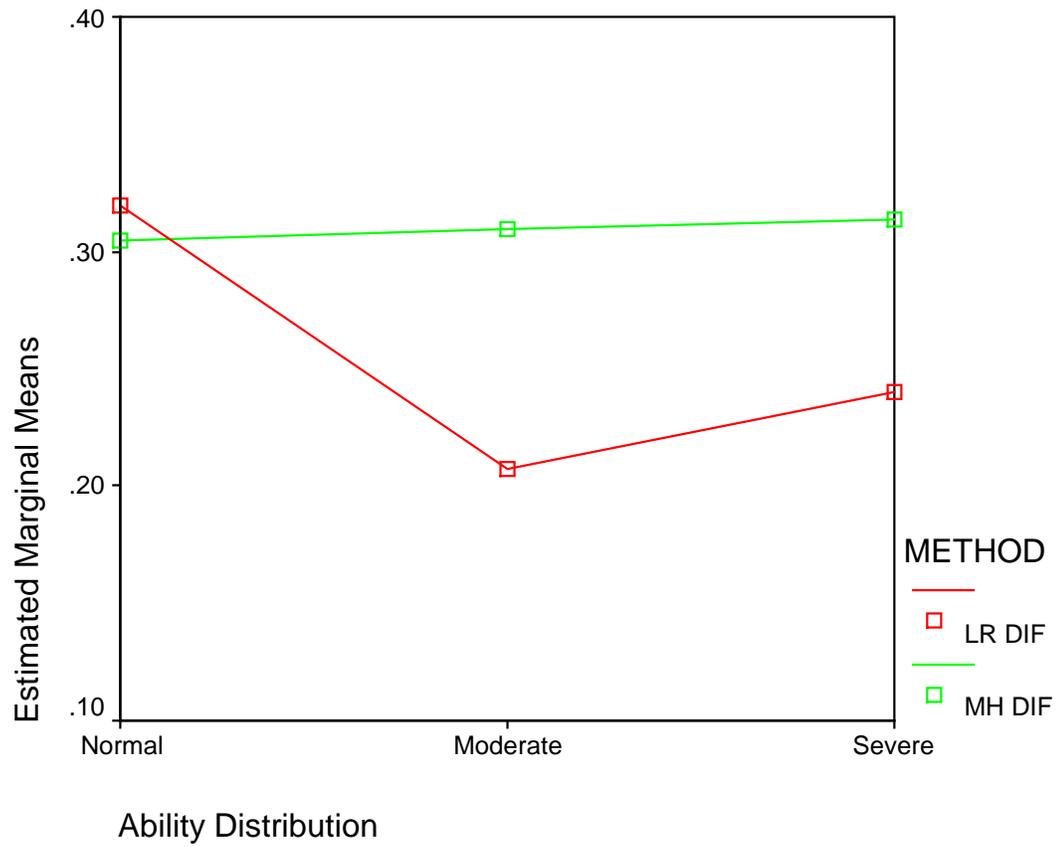


Figure 13. Estimated Marginal Means for Sample Size Combination 1000/100 (Method by Ability Distribution) (p Values)

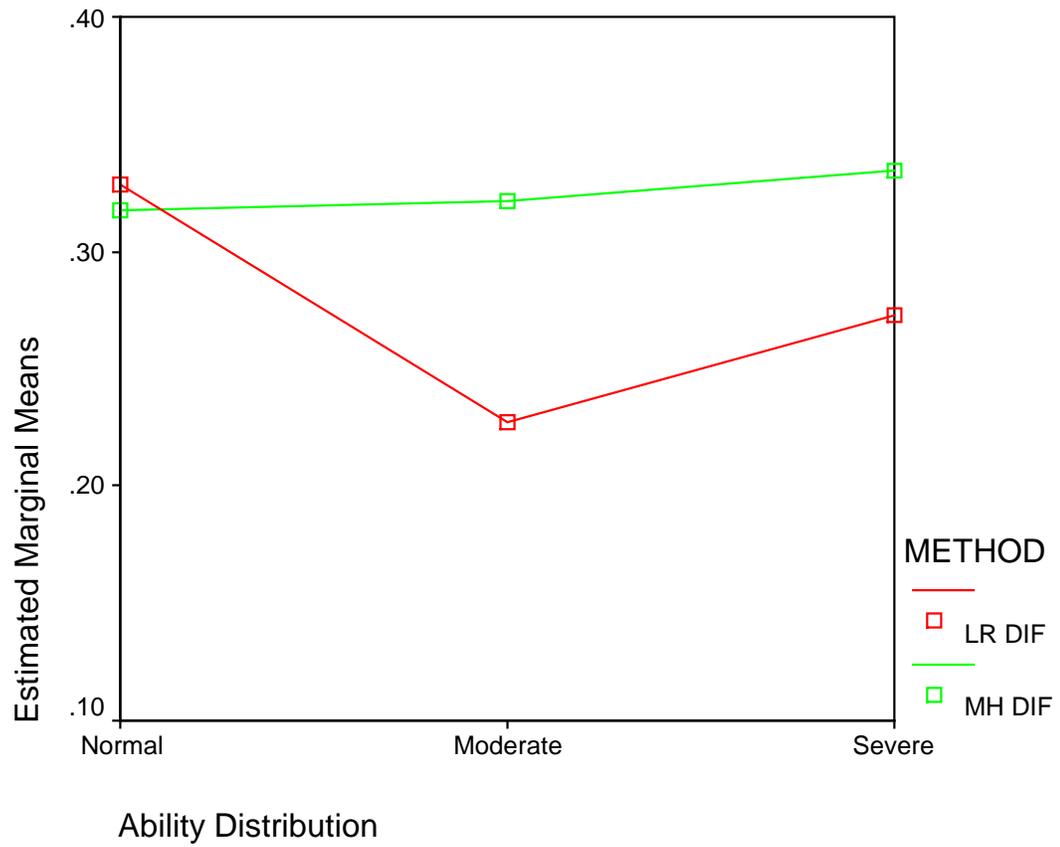


Figure 14. Estimated Marginal Means for Sample Size Combination 500/100 (Method by Ability Distribution) (p Values)

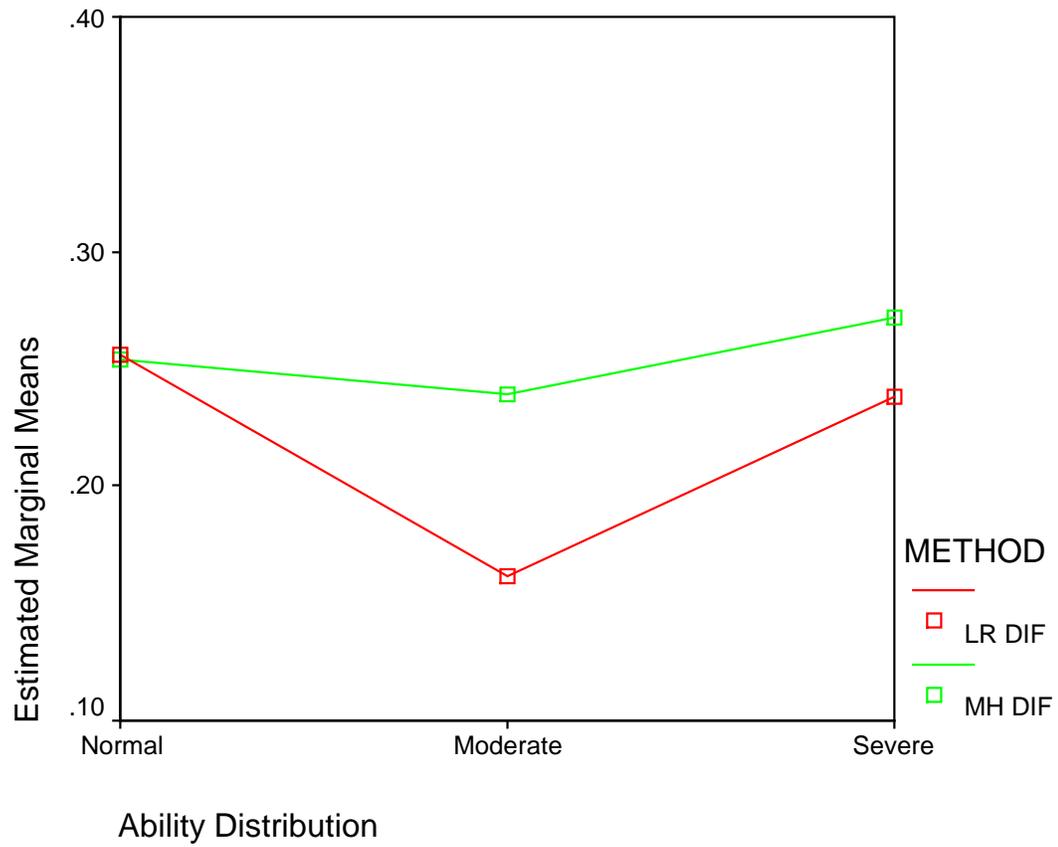


Figure 15. Estimated Marginal Means for Sample Size Combination 300/300 (Method by Ability Distribution) (p Values)

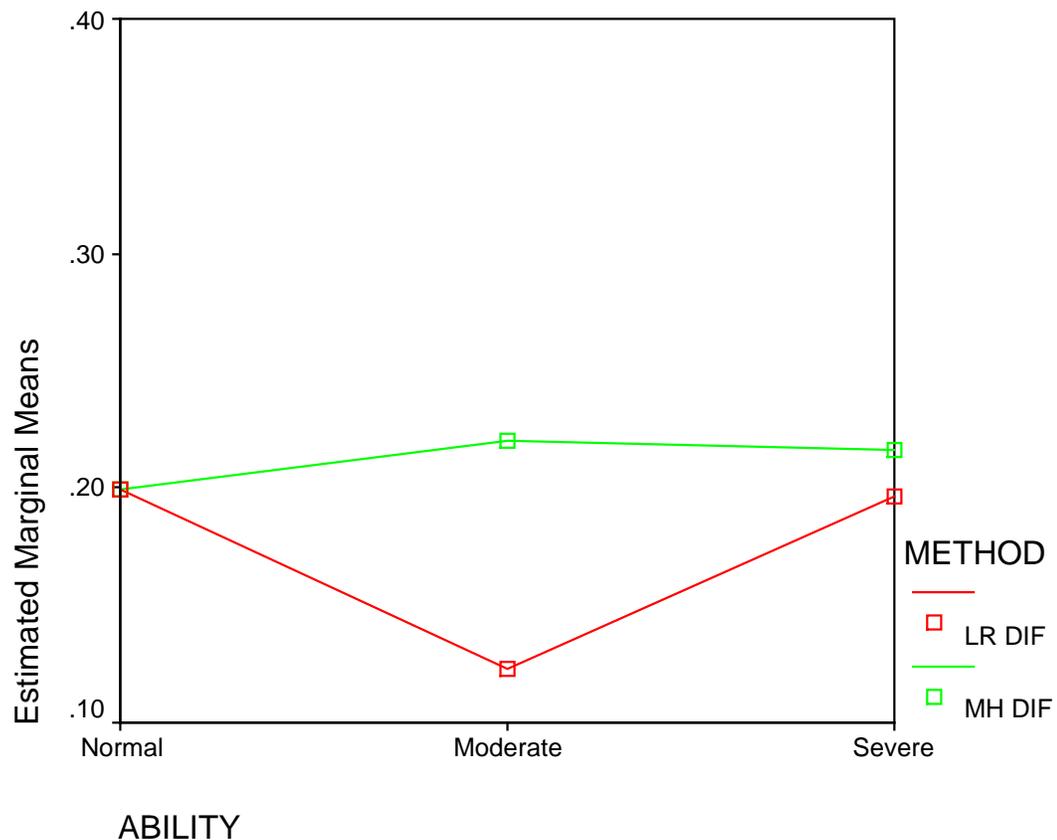


Figure 16. Estimated Marginal Means for Sample Size Combination 1000/300 (Method by Ability Distribution) (p Values)

Custom contrasts were run for the interaction effect of method by ability by sample. Contrasts were run for method by sample size combinations across ability distribution levels and are illustrated in Figures 10 through Figure 12. Contrasts were run for method by ability distribution across sample size combinations and are illustrated in Figures 13 through Figure 16.

The method by ability by sample linear contrast was statistically significant ($F(1, 131538) = 36.574, p < .001, \text{effect size} < .00\%$), the ability by sample quadratic contrast

was statistically significant ($F(1, 131538) = 4.118, p = .04$, effect size $< .00\%$), and ability by sample cubic contrast was statistically significant ($F(1, 131538) = 5.485, p = .02$, effect size $< .00\%$). The ability by sample quadratic contrast was the best fit due to its small F ratio and effect size. This is illustrated in Figure 13 through Figure 16 across ability distribution levels.

The method by sample by ability linear contrast was statistically significant ($F(4, 131538) = 607.618, p < .001$, effect size = $.45\%$) and the method by sample by ability quadratic contrast was statistically significant ($F(4, 131538) = 1161.059, p < .001$, effect size = $.86\%$). The method by sample by ability linear contrast was a better fit than the sample by ability quadratic contrast. Figure 10 through Figure 12 illustrate how the quadratic trend is slightly a better fit than the linear trend across sample size combinations. Notably, both contrasts had high F statistics and effect sizes compared to effect sizes throughout all analyses. Table 27 lists calculated statistics for all interaction tests.

Table 27 Contrast Trends for Method by Ability Distribution by Sample Size

Combinations (p Values)

Effect	Type I SOS	df	Mean Square	F	Sig.	Effect Size (%)
Method	129.333	1	129.333	1901.956	< .00	1.4
method * sample* alin	41.318	4	10.326	607.618	< .00	.45
method * sample * aquad	78.952	4	19.738	1161.059	< .00	.86
method * ability * slin	2.487	1	2.487	36.574	< .00	.03
method * ability * squad	.280	1	.280	4.118	.04	< .00
method * ability * scub	.373	1	.373	5.485	.02	< .00
Error (method)	8970.543	131538	.068			
Total	9227.018					

Note. df = degrees of freedom, F = F ratio, Sig. = significance, slin = sample linear contrast, squad = sample quadratic contrast, alin = ability linear contrast, aquad = ability quadratic contrast, acub = ability cubic contrast.

Research Question 7

Research question 7 addresses whether the detection of DIF in an item is more likely to occur with a weighted-least-squares (WLS) R^2 (effect size for LR DIF) or log odds ratio (effect size for MH DIF) with varying ability distributions. RM-ANOVA with effect sizes as the dependent variable and ability distribution as the independent variable was used to answer this research question. The independent variable's levels were *normal* (mM0/mF0/stM1/stF1), *moderate* (mM1/mF1/stM1/stF2), and *severe* (mM0/mF1/stM1/stF2) differences between focal and reference ability distributions. Within-subjects effect was the focus of the RM-ANOVA to reveal possible differences in methods across ability distributions. Descriptives for the RM-ANOVA including varying ability distributions as an independent variable are in appendix M.

RM-ANOVA revealed statistically significant differences in the within-subjects effects test across methods for ability distribution levels ($F(df=2, 131,538) = 163.441, p < .001, \text{effect size} = 0.10\%$). The overall effect sizes for MH DIF were different than LR DIF indicating MH DIF and LR DIF do not similarly detect DIF across varying ability distributions. Table 28 lists descriptive statistics for estimated marginal means for LR DIF and MH DIF for ability distribution levels. Figure 17 illustrates the linear and quadratic trends for method by ability distribution with estimated marginal means for ability distributions. Figure 13 illustrates the method by ability distribution effect for the estimated marginal means from Table 28.

Table 28 Estimated Marginal Means for Method by Ability Distribution (Effect Sizes)

Ability	Method	Mean	Std. Error
Normal	LR DIF	1.505	.003
	MH DIF	2.802	.005
Moderate	LR DIF	1.593	.003
	MH DIF	2.773	.005
Severe	LR DIF	1.531	.003
	MH DIF	2.728	.005

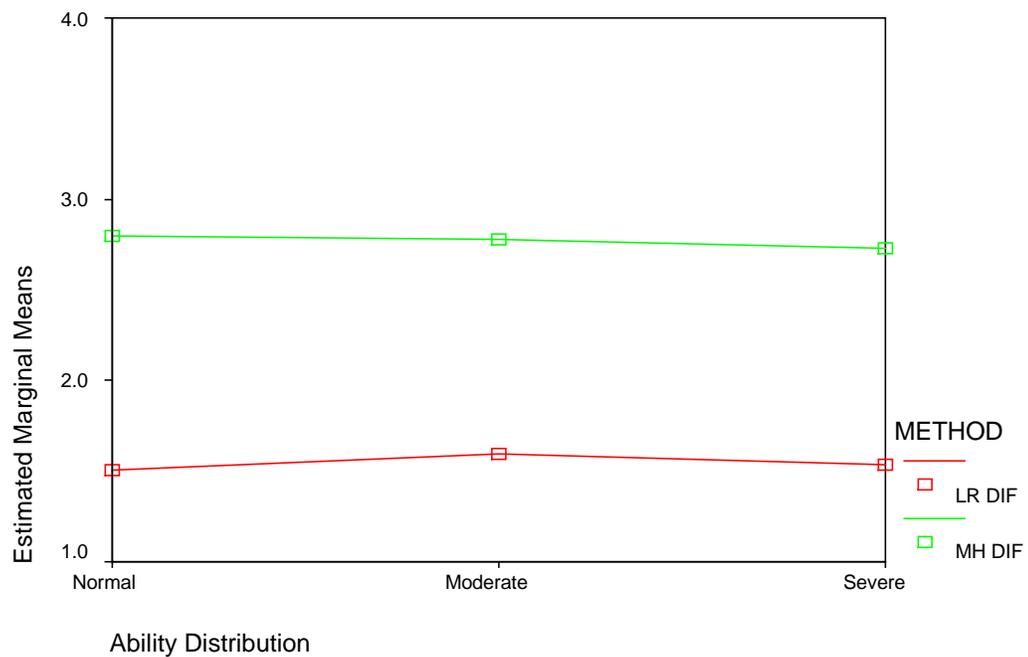


Figure 17. Estimated Marginal Means for MH DIF and LR DIF Ability Distributions (Effect Sizes)

Linear and quadratic contrasts were run for method by ability distribution. The interaction effect had a statistically significant linear trend ($F(df=1, 131538) = 274.549$, $p < .001$, effect size = 0.09%) and a statistically significant quadratic trend ($F(df=1, 131538) = 52.180$, $p < .001$, effect size = 0.02%). The quadratic trend seems to be the best fit for the method by ability distribution interaction effect due to the small F statistic and effect size. Table 29 lists the statistics calculated for the linear and quadratic trends for method by ability distribution. Figure 17 illustrates the trends for each method and for the interaction of method by ability distribution.

Table 29 Linear and Quadratic Trends for Method by Ability Distribution (Effect Sizes)

Effect	Type I SOS	df	Mean Square	F	Sig.	Effect Size (%)
method	98758.552	1	98758.552	186689.134	< .001	58
method * alin	111.644	1	111.644	211.241	< .001	.09
method * aquad	27.532	1	27.532	52.093	< .001	.02
Error (method)	69548.477	131592	.528			
Total	169031.033					

Note. alin = ability linear contrast, aquad = ability quadratic contrast.

Research Question 8

Research question 8 addresses whether the detection of DIF in an item is more likely to occur with a weighted-least-squares (WLS) R^2 (effect size for LR DIF) or log odds ratio (effect size for MH DIF) with varying population distributions. RM-ANOVA with effect sizes as the dependent variable and population distribution as the independent variable was used to answer this research question. The independent variable's levels were *normal* (mM0/mF0/stM1/stF1), *moderate* (mM1/mF1/stM1/stF2), and *severe* (mM0/mF1/stM1/stF2) differences between focal and reference ability distributions. Within-subjects effect was the focus of the RM-ANOVA to reveal possible differences in methods across population distributions. Descriptives for the RM-ANOVA including varying population distributions as an independent variable are in appendix M.

RM-ANOVA revealed statistically significant differences in the within-subjects effects test across methods for population distribution levels (F (df== 4, 131538) = .613, $p = .653$, effect size < 0.00%). The overall effect sizes for MH DIF were different than LR DIF indicating MH DIF and LR DIF do not similarly detect DIF across varying ability distributions.

There was not an overall within-subjects effect; therefore, simple custom contrasts were run on method by population distribution. The overall simple contrast was not statistically significant (F (df = 4, 131538) = 1.428, $p = .222$, effect size < 0.00%) for population distribution. Level one, *normally distributed*, was the reference for the custom contrasts and not statistically significant versus level 2 ($p = .460$), level 4 ($p = .083$), and level 5 ($p = .315$). The simple contrast for level 1 versus level 3 was

statistically significant ($p = .031$), with a difference of .014 between levels. Figure 18 illustrates the method by ability distribution effect with estimated marginal means for population distributions. Table 30 lists descriptive statistics for estimated marginal means for LR DIF and MH DIF for population distribution levels.

Table 30 Estimated Marginal Means for Method by Population Distribution (Effect Sizes)

Population Levels	Method	Mean	Std. Error
Normally Dist.	LR DIF	1.539	.007
	MH DIF	2.758	.014
Moderately Dist.	LR DIF	1.541	.007
	MH DIF	2.765	.014
Skew/Lepto.	LR DIF	1.547	.007
	MH DIF	2.777	.014
Skew/Extremely Lepto	LR DIF	1.544	.007
	MH DIF	2.774	.014
Platykurtic	LR DIF	1.543	.007
	MH DIF	2.766	.014

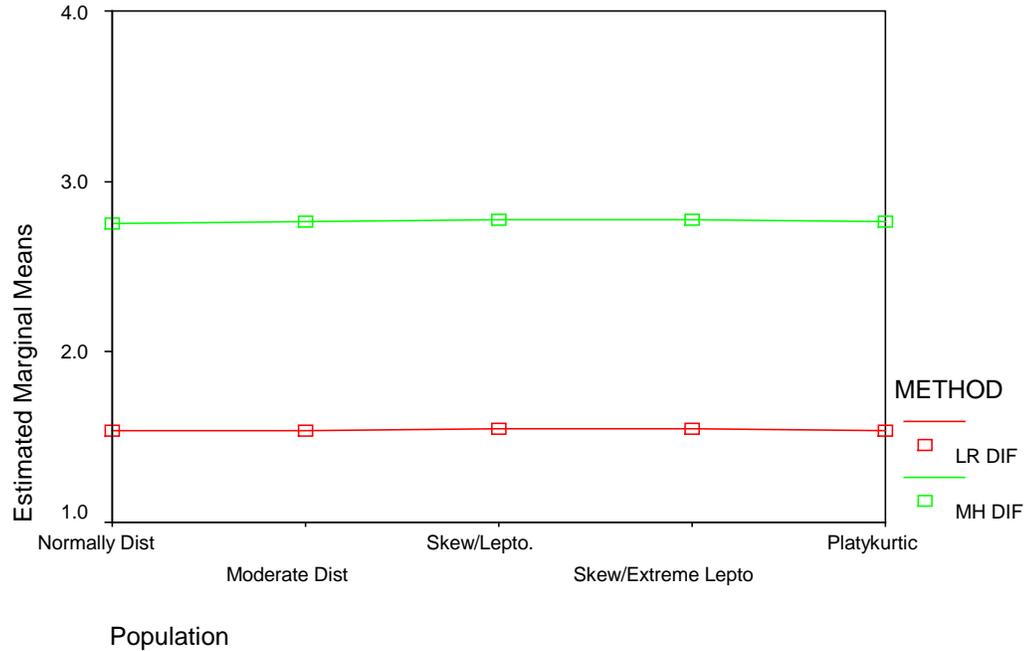


Figure 18. Estimated Marginal Means for MH DIF and LR DIF Population Distributions (Effect Sizes)

Research Question 9

Research question 9 addresses whether the detection of DIF in an item is more likely to occur with weighted-least-squares WLS R^2 (effect size for LR DIF) or log odds ratio (effect size for MH DIF) with varying sample size combinations. RM-ANOVA was run with effect sizes as the dependent variable and sample size combination as the independent variable. The four sample size combination levels were 1000_r/100_f, 500_r/100_f, 300_r/300_f, 1000_r/300_f.

Within-subjects effect was the focus of the RM-ANOVA to reveal possible differences in methods across sample size combination levels. Descriptives for the RM-ANOVA including the independent variable sample size combination are in appendix M.

RM-ANOVA revealed statistically significant differences in the within-subjects effects test across methods for sample size combinations ($F(df=3, 131,538) = 342.169, p < .001, \text{effect size} = 0.32\%$). The overall effect sizes for MH DIF were statistically significantly different than LR DIF indicating MH DIF and LR DIF do not detect DIF similarly across sample size combinations. Table 31 lists descriptive statistics for estimated marginal means for sample size combinations for LR DIF effect size and MH DIF effect size. Figure 19 illustrates the method by sample size combinations effect using the estimated marginal means from Table 31.

Table 31 Estimated Marginal Means for Method by Sample Size Combination (Effect Sizes)

Sample Size Levels (r/f)	Method	Mean	Std. Error
1000/100	LR DIF	1.391	.003
	MH DIF	2.751	.006
500/100	LR DIF	1.542	.003
	MH DIF	2.769	.006
300/300	LR DIF	1.670	.003
	MH DIF	2.775	.006
1000/300	LR DIF	1.569	.003
	MH DIF	2.777	.006

Note. r/f = reference/focal.

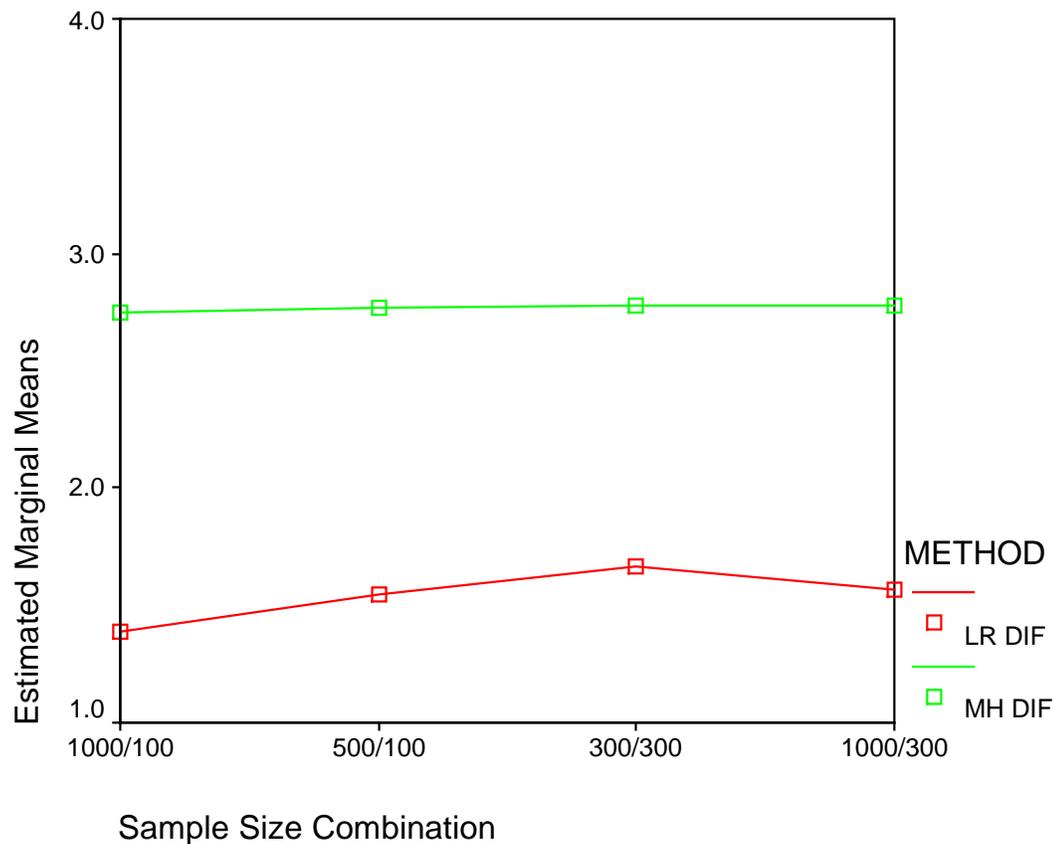


Figure 19. Estimated Marginal Means for MH DIF and LR DIF Sample Size Combination (Effect Sizes)

Linear, quadratic, and cubic contrasts were run for method by sample size combinations. The interaction effect had a statistically significant linear trend ($F(df=1, 131538) = 522.438, p < .001, \text{effect size} = 0.16\%$), a statistically significant quadratic trend ($F(df=1, 131538) = 434.195, p = .001, \text{effect size} = 0.14\%$), and a statistically significant cubic trend ($F(df=1, 131538) = 69.803, p = .001, \text{effect size} = 0.02\%$). The cubic trend seems to be the best fit for the method by ability distribution interaction effect due to the smallest F statistic and effect size. Table 32 lists the statistics

calculated for the linear and quadratic trends for method by ability distribution. Figure 19 illustrates the trends for each method and for the interaction of methods by sample size combinations.

Table 32 Linear and Quadratic Trends for Method by Sample Size Combination (Effect Sizes)

Effect	Type I SOS	df	Mean Square	F	Sig.	Effect Size (%)
Method	98758.552	1	98758.552	186689.134	< .001	58
Method * slin	275.847	1	275.847	522.438	< .001	0.16
Method * squad	229.255	1	229.255	434.195	< .001	0.14
Method * scub	36.856	1	36.856	69.803	< .001	0.02
Error (method)	69548.477	131592	.528			
Total	169031.033					

Note. df = degrees of freedom, F = F ratio, Sig. = significance, slin = sample linear contrast, squad = sample quadratic contrast, scub = sample cubic contrast.

Research Question 10

Research question 10 addresses whether the detection of DIF in an item is more likely to occur with weighted-least-squares WLS R^2 (effect size for LR DIF) or log odds ratio (effect size for MH DIF) with varying ability distributions, population distributions, and sample size combinations. RM-ANOVA was run with effect sizes (WLS R^2 and log odds ratio) as the dependent variable and ability distribution, population distribution, and sample size combination as the independent variables.

Within-subjects effect was the focus of the RM-ANOVA to reveal possible differences in methods across ability distribution, population distribution, and sample size combinations. Descriptives for the RM-ANOVA including sample size combinations as an independent variable are in appendix M.

RM-ANOVA revealed no statistically significant differences for the within-subjects effects test for an interaction for method by sample size combination by population distribution ($F(df=12, 131,538) = .987, p < .458, \text{effect size} < .00\%$).

RM-ANOVA revealed no statistically significant differences for the within-subjects effects test for an interaction for method by ability distribution by population distribution ($F(df=8, 131,538) = .818, p < .587, \text{effect size} < .00\%$).

RM-ANOVA revealed no statistically significant differences for the within-subjects effects test for an interaction for method by sample size combination by ability distribution by population distribution ($F(df=24, 131,538) = .10311, p < .141, \text{effect size} < .00\%$).

RM-ANOVA revealed statistically significant differences for the within-subjects effects test for an interaction for method by ability distribution by sample size combinations ($F(df=6, 131,538) = 29.635, p < .001, \text{effect size} = 0.10\%$).

Table 33 lists descriptive statistics for estimated marginal means for LR DIF and MH DIF for the interaction effect of method by sample size combination by ability distribution by population distribution. Figure 20 through Figure 22 illustrate the interaction effects for method by sample size combinations by ability distribution. Figures are shown for each ability distribution level. Figure 20 illustrates the estimated marginal means for sample size combinations for ability distribution level *normal*. Figure 21 illustrates the estimated marginal means for sample size combinations for ability distribution level *moderate*. Figure 22 illustrates the estimated marginal means for sample size combinations for ability distribution level *severe*.

Table 33 Estimated Marginal Means for Method by Ability Distribution by Sample Size Combinations (Effect Sizes)

Sample Size	Ability	Method	Mean	Std. Error
1000/100	Normal	LR DIF	1.324	.005
		MH DIF	2.787	.011
	Moderate	LR DIF	1.478	.005
		MH DIF	2.744	.011
	Severe	LR DIF	1.370	.005
		MH DIF	2.723	.011

Table 33 Continued

Sample Size	Ability	Method	Mean	Std. Error
500/100	Normal	LR DIF	1.484	.005
		MH DIF	2.788	.011
	Moderate	LR DIF	1.630	.005
Sample Size	Ability	Method	Mean	Std. Error
		MH DIF	2.801	.011
	Severe	LR DIF	1.513	.005
		MH DIF	2.718	.011
300/300	Normal	LR DIF	1.678	.005
		MH DIF	2.801	.011
	Moderate	LR DIF	1.647	.005
		MH DIF	2.798	.011
	Severe	LR DIF	1.684	.005
		MH DIF	2.727	.011
1000/300	Normal	LR DIF	1.532	.005
		MH DIF	2.831	.011
	Moderate	LR DIF	1.616	.005
		MH DIF	2.756	.011
	Severe	LR DIF	1.559	.005
		MH DIF	2.743	.011

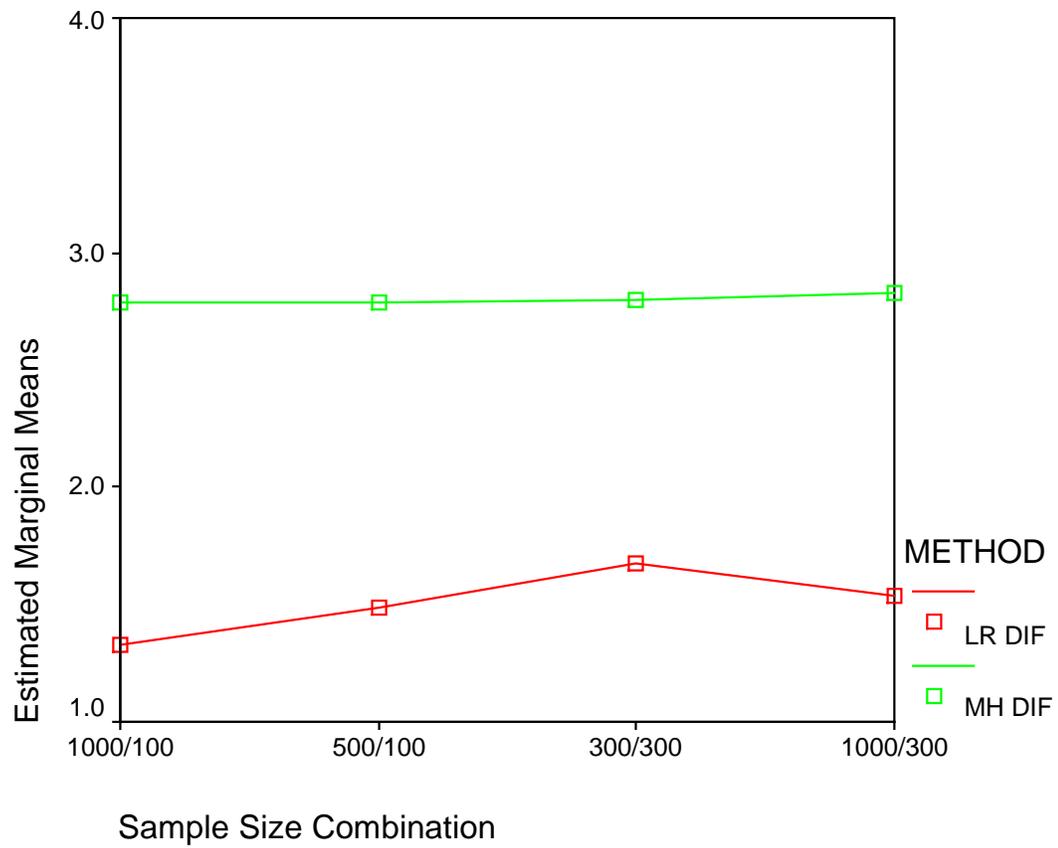


Figure 20. Estimated Marginal Means for Normal Ability Distributions (Method by Sample Size) (Effect Sizes)

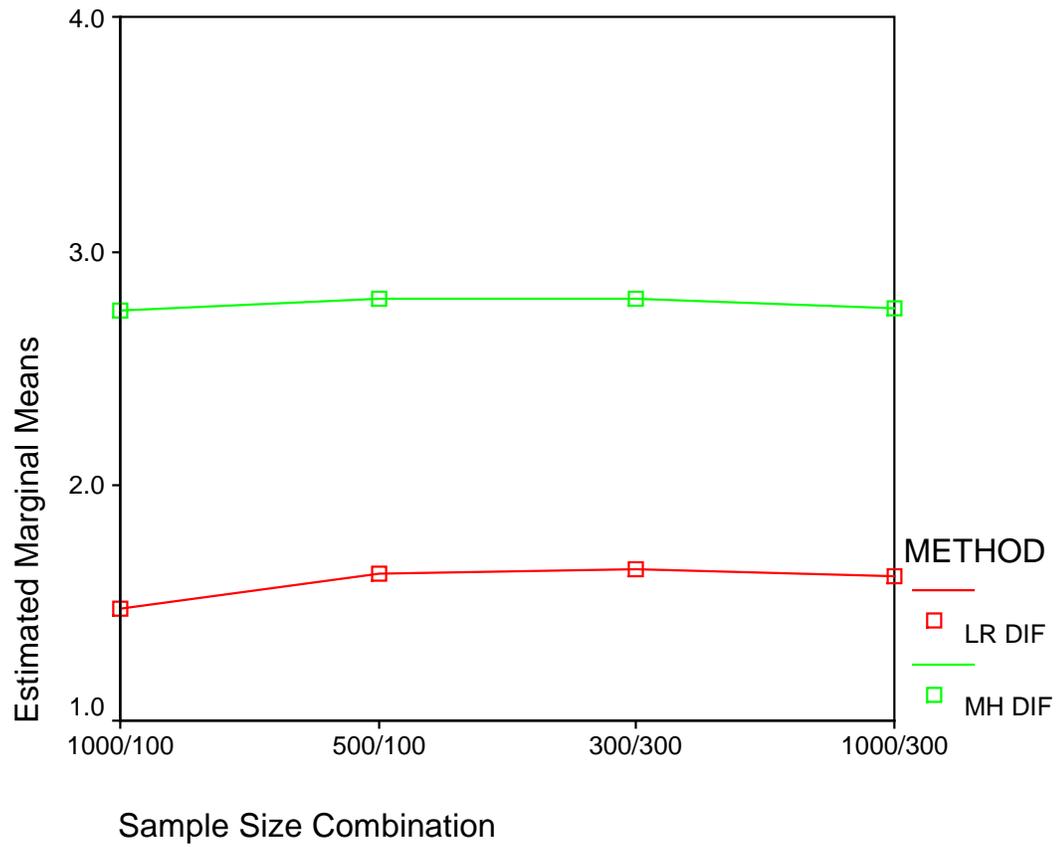


Figure 21. Estimated Marginal Means for Moderate Ability Distributions (Method by Sample Size) (Effect Sizes)

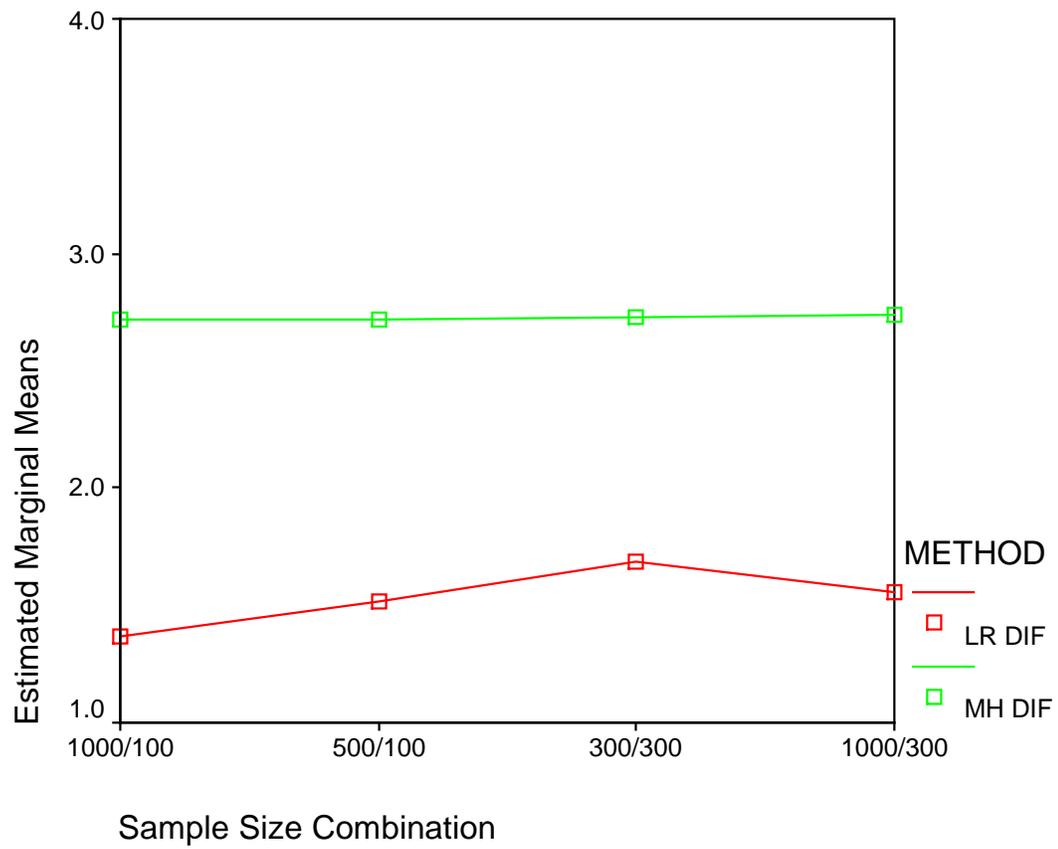


Figure 22. Estimated Marginal Means for Severe Ability Distributions (Method by Sample Size) (Effect Sizes)

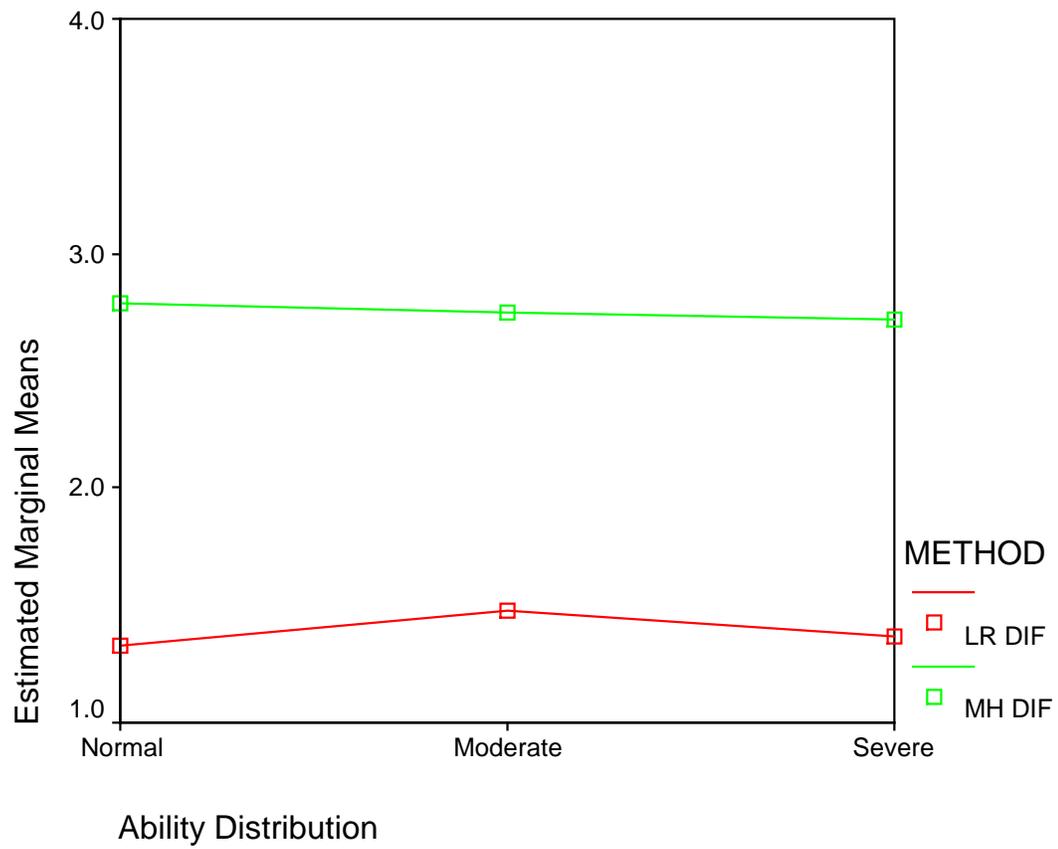


Figure 23. Estimated Marginal Means for Sample Size Combination 1000/100 (Method by Ability Distribution) (Effect Sizes)

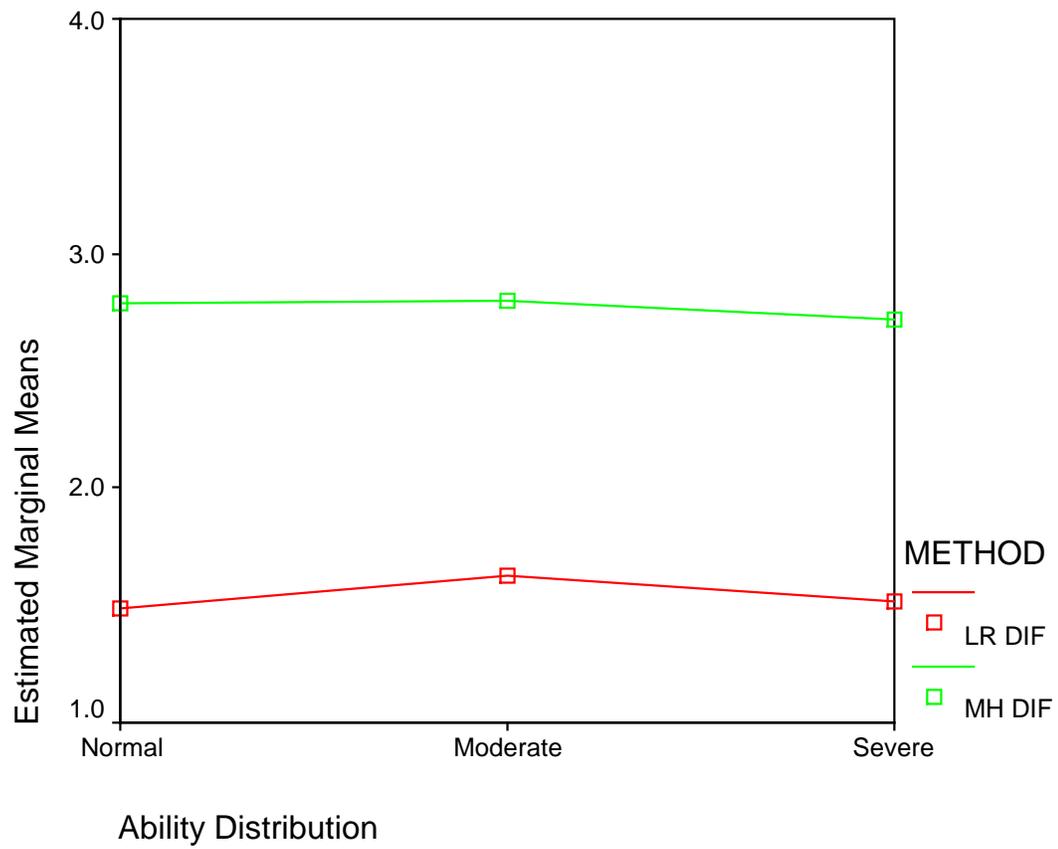


Figure 24. Estimated Marginal Means for Sample Size Combination 500/100 (Method by Ability Distribution) (Effect Sizes)

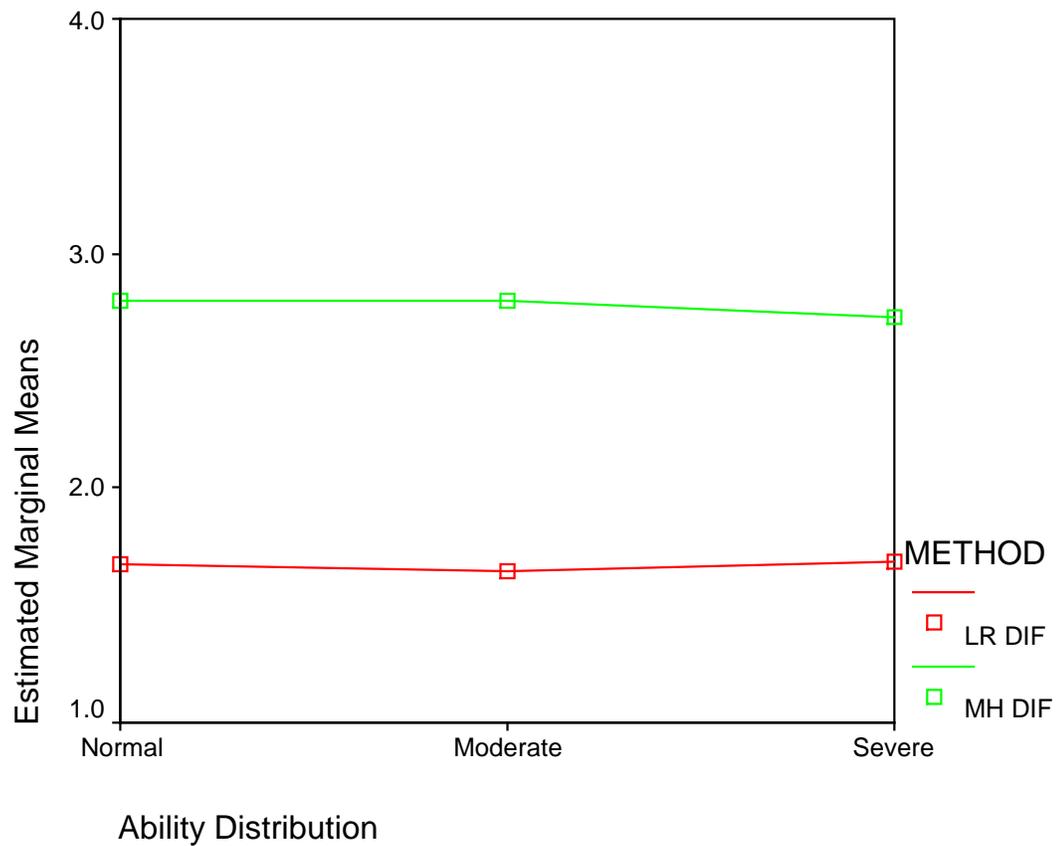


Figure 25. Estimated Marginal Means for Sample Size Combination 300/300 (Method by Ability Distribution) (Effect Sizes)

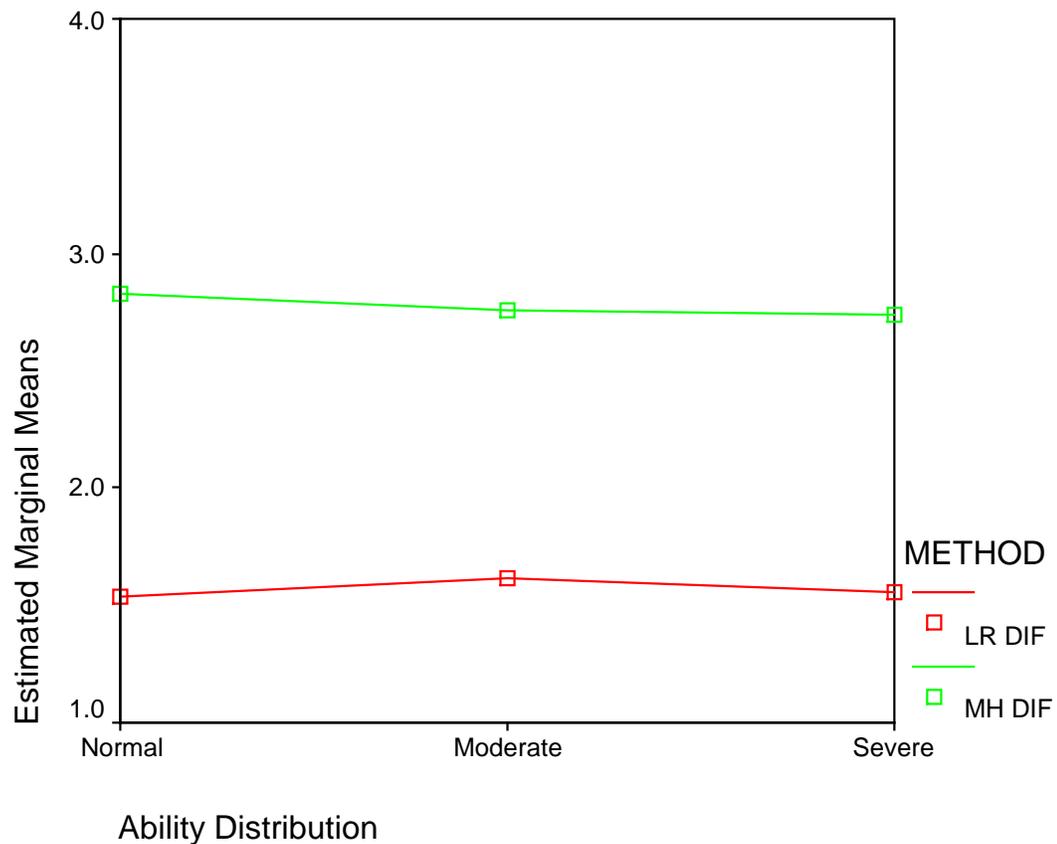


Figure 26. Estimated Marginal Means for Sample Size Combination 1000/300 (Method by Ability Distribution) (Effect Sizes)

Custom contrasts were run for the interaction effect of method by ability by sample. Contrasts were run for method by sample size combinations across ability distribution levels and are illustrated in Figures 20 through Figure 22. Contrasts were run for method by ability distribution across sample size combinations and are illustrated in Figures 23 through Figure 26. Table 34 lists calculated statistics for all interaction tests.

The method by sample by ability linear contrast was statistically significant ($F(4, 131538) = 215.064, p < .001, \text{effect size} = .07\%$) and the method by sample by ability quadratic contrast was statistically significant ($F(4, 131538) = 289.718, p < .001, \text{effect size} = .09\%$). The method by sample by ability linear contrast was a better fit than the sample by ability quadratic contrast. When observing the plots in Figure 20 through 22 there do not seem to be any trends across sample size combinations. Notably, all contrasts that were statistically significant might be due to the large sample size.

The method by ability by sample linear contrast was statistically significant ($F(1, 131538) = 522.236, p < .001, \text{effect size} = .16\%$), the ability by sample quadratic contrast was statistically significant ($F(1, 131538) = 434.063, p < .001, \text{effect size} = .13\%$), and ability by sample cubic contrast was statistically significant ($F(1, 131538) = 69.746, p < .001, \text{effect size} = .02\%$). The method by ability by sample cubic contrast was the best fit due to its small F ratio and effect size, and the method by ability by sample quadratic contrast was a better fit than the method by ability by sample linear contrast. When observing the plots in Figure 23 through 26 there do not seem to be any trends across sample size combinations. Notably, all contrasts that were statistically significant might be due to the large sample size.

Table 34 Contrast Trends for Method by Ability Distribution by Sample Size

Combination (Effect Sizes)

Effect	Type I SOS	df	Mean Square	F	Sig.	Effect Size (%)
Method	129.333	1	129.333	1901.95	< .00	1.4
Method * sample * alin	113.554	4	28.389	215.06	< .00	.07
Method * sample * aquad	152.971	4	38.243	289.72	< .00	.09
Method * ability * slin	275.788	1	275.788	522.33	< .00	.16
Method * ability * squad	229.185	1	229.185	434.06	< .00	.13
Method * ability * scub	36.826	1	36.826	69.75	< .00	.02
Error (method)	69427.025	131538	.528			
Total	169031.033					

Note. slin = sample linear contrast, squad = sample quadratic contrast, alin = ability linear contrast, aquad = ability quadratic contrast, acub = ability cubic contrast.

CHAPTER V

DISCUSSION AND CONCLUSIONS

This study focused on whether or not incorporating an effect size for LR DIF more accurately detects DIF while comparing the utility of an effect size index across MH DIF and LR DIF methods. A secondary focus of the present research was how various conditions, such as sample size, ability distributions, and population distributions, affect the detection of DIF with MH DIF and LR DIF methods.

Overall, the detection of DIF was more accurate when an effect size measure was included for LR DIF, but inclusion did not improve the detection of DIF for the MH DIF method. When detecting DIF across varying sample size combinations, ability distributions, and population distributions, both sample size and ability distributions affected the likelihood of detecting DIF in an item. Population distribution variation did not affect detection of DIF. Research questions one through ten are specifically discussed below.

Research Question 1

The LR DIF method is not as accurate of an analysis for detecting predetermined DIF as previously suggested. The p value is not a strong predictor of DIF and should not be trusted as the only determining factor for whether DIF is present in an item. Interestingly, the p value was a better predictor for Non-DIF items, while the p value was not a good predictor for DIF items. The effect size statistic is a more accurate predictor than the p value for Non-DIF and DIF items; however, the p value and the

effect size only predicted 45% DIF items correctly. If LR DIF is the selected DIF method, then an effect size should be included in the analysis and interpretation stages.

Research Question 2

The MH DIF method was more accurate than the LR DIF method for the p value and the effect size statistics. The p value and the effect size were both similarly accurate when detecting DIF and Non-DIF items. When the effect size, log odds ratio, was included in the MH DIF method the accuracy of the MH DIF method improved. The inclusion of an effect size improved the accuracy of the detection of DIF slightly; however both statistics would be beneficial when using the MH DIF method.

Research Question 3

Normal ability distributions result in an overall higher p value for LR DIF than MH DIF. As the ability distributions differ more between focal and reference groups the overall p value for LR DIF decreases and again increases at the severe level for ability distribution. The MH DIF p value remains higher and somewhat constant across levels of ability distribution. The LR DIF seems most sensitive to moderate levels in ability distribution differences, detecting DIF more often, while normal and severe levels in ability distributions result in higher p values, thus less likely to detect DIF. The MH DIF method seems to be fairly constant across ability distribution differences in focal and reference groups and detecting DIF less often overall than LR DIF.

Research Question 4

Overall, there is no evidence of a statistically significant increase or decrease in detection of DIF for LR DIF across population distribution levels or MH DIF across

population distribution levels. There were no statistically significant differences between levels. The p value dependent variable for LR DIF is more liberal and is overall lower than the dependent variable p value for MH DIF.

Research Question 5

Sample size and difference in focal and reference subgroup sample sizes seem to be relevant to the detection of DIF. Overall, as sample sizes increase and the dependent variable p value decreases for both MH and LR DIF. Specifically, for the level 1000/100 a lower dependent p value is found than for the level 500/100, which might be due to the larger overall sample size. This indicates overall sample size is more relevant than percentage of difference in focal and reference subgroup sample size. Comparatively, when looking at the decrease in dependent variable p value for 300/300 level versus the 500/100 level the lower percentage difference is more likely to detect DIF. The trends for MH DIF and LR DIF are similar and tend to decrease with the higher the overall sample size.

Research Question 6

Overall, ability distribution levels and sample size combinations seem to affect MH DIF and LR DIF methods' detection of DIF. When looking across sample size combinations there seems to be a linear trend. As the overall sample size increases the dependent variable p value decreases. As the percentage difference in focal and reference subgroups increases the dependent variable p value also increases; therefore, detecting DIF less often. Using large sample sizes and small subgroup differences

between focal and reference groups are the optimal parameters for a study for detection of DIF when DIF is present.

When looking across ability distribution levels there seems to be a quadratic trend across levels. Moderate ability distributions seem to detect DIF more often across sample size combinations than normal and severe ability distributions. Moderate ability distributions had lower dependent variable p values than normal and severe ability distributions. As for the ability distribution level interaction, less variability occurred across sample size combinations and MH DIF had overall higher p values than LR DIF, thus MH DIF was less likely to detect DIF. Notably, the moderate ability distribution level varied the most while the normal and severe levels remained somewhat the same.

Overall, MH DIF seemed less likely to detect DIF across ability distribution levels and sample size combinations. One exception for this was in the normal ability distribution level for the two smallest sample size combinations and largest percentage difference in focal/reference subgroups where MH DIF was slightly lower than the LR DIF.

Research Question 7

Overall, LR DIF seems to be more conservative when detecting DIF compared to MH DIF across varying ability distributions. The linear and quadratic trends were statistically significant, which might be due to the large sample size. There is at best a slight bend in the trend in the dependent variable method. The more severe the ability distribution, the lower the MH DIF effect size, and the moderate level of ability distribution for LR DIF seemed to have the highest effect size. Both normal and severe

levels for LR DIF were more conservative detecting DIF, with the normal ability distribution level having a detection of DIF less often. MH DIF does not seem to be as affected by varying ability distributions as LR DIF. A note of caution is required when interpreting results that used effect sizes for MH DIF and LR DIF since there is the restriction of range for the dependent variable.

Research Question 8

Overall, effect sizes do not show evidence of increase or decrease association with the detection of DIF for LR DIF across population distribution levels or for MH DIF across population distribution levels. The effect size WLS R^2 for LR DIF is more conservative and has an overall lower magnitude than log odds ratio for MH DIF, thus a more conservative effect size for the parameters of the present study. The detection of DIF does not seem to be affected by varying shapes of population distributions when MH DIF or LR DIF are used to detect DIF. A note of caution when interpreting results that used effect sizes for MH DIF and LR DIF is the restriction of range for the dependent variable.

Research Question 9

Sample size combinations do not affect the detection of DIF for MH DIF, but do seem to affect the detection of DIF for LR DIF. MH DIF is not conservative with the log odds ratios across sample size combinations all nearing moderate effect sizes (category 3). LR DIF is more conservative with the WLS R^2 , falling within negligible effect sizes in category 1. MH DIF effect size seemed to decrease as the ability distribution got more severe. LR DIF effect size was highest for the moderate level

while the normal and severe levels were slightly lower than the moderate level. Overall, LR DIF has a more conservative effect size than MH DIF. A note of caution when interpreting results that used effect sizes for MH DIF and LR DIF is the restriction of range for the dependent variable.

Research Question 10

Overall, ability distribution levels and sample size combinations do not seem to affect MH DIF and LR DIF methods detection of DIF. When looking across sample size combinations there seems to be a very slight change in LR DIF across sample size combinations where the larger the sample size and smaller the percentage difference in focal/reference subgroups the higher the effect size; however, MH DIF effect size remains constant and much larger across sample size combinations. The contrasts tests showed more of a cubic trend than quadratic or linear, but this is most likely due to the large sample size. When looking across ability distribution levels there seems to be a slight increase in the LR DIF method effect size for the moderate ability distribution level, while for the normal and severe ability distribution, effect size levels remain lower across sample size combinations. This could be a quadratic trend, but the contrasts tests showed a better fit for the linear trend. This linear trend is supported slightly for the MH DIF effect sizes.

Overall, MH DIF was more likely to detect DIF across ability distribution levels and sample size combinations by nearly one full category. A note of caution when interpreting the effect sizes for MH DIF and LR DIF is the restriction of range for the dependent variable.

Conclusions

The inclusion of an effect size improves the accuracy of both the LR DIF and MH DIF methods. While the p value was not very stable for the MH DIF method, the p value did seem more accurate for the LR DIF method. Based on the present study, there is strong evidence to include an effect size in either method when determining if DIF is present in an item.

When using MH DIF or LR DIF methods to determine DIF the researcher needs to take caution if ability distributions or sample size combinations vary. Ability distributions seem to be more affected when distribution differences are moderate and when using the LR DIF method. Varying sample size combinations were not as sensitive as the ability distribution differences, but did increase the p value and decrease the effect size somewhat when sample size combinations were more discrepant and overall sample sizes were small. A minimum sample size 300-500 is recommended for focal and reference groups. The MH DIF method yielded higher effect sizes across the majority of ability distribution levels and sample size combinations. The only comparison that MH DIF was lower than LR DIF was for the p values in the normal ability distribution level for all sample size combinations. Population distributions did not seem to affect MH DIF or LR DIF methods for p values or effect sizes.

The present study had items with predetermined DIF; therefore, the focus was to find if the MH DIF and LR DIF methods could detect DIF, not determine if DIF was present. In actual studies not involving a simulation, more caution should be taken when determining if DIF is included in an item. MH DIF and LR DIF are analyses to give the

researcher or psychometrician empirical knowledge to assist in their decisions about items.

Strengths and Limitations

This study encompassed 60 conditions to control for as many parameter differences as possible in order to generalize to realistic data, which created a strong research study; however, there are also some limitations.

The comprehensiveness of the study is a strength and profitable to the present and future research of DIF analyses. The current study gives many researchers a place to begin researching with prior knowledge of LR and MH DIF analyses. For example, population distributions do not seem to affect varying types of datasets. This simulation study shows that researchers do not have to worry about varying population distributions because they do not seem to affect the statistics for LR and MH DIF analyses. The present study was a systematic study that offers a baseline for many types of studies in LR and MH DIF to be researched.

One of the main limitations is that while the data was taken from real data (ACT examinee responses) there were not any real DIF items in the original.

The DIF items in this study were chosen as DIF items based on Raju's formula to measure the area between ICCs (.40 or higher) and similar studies that used DIF items from real data (Hidalgo and Lopez-Pina, 2004). One limitation is that some items were manipulated to contain DIF due to the low amount of items containing DIF from the ACT data. Secondly, these items might not be deemed DIF items for studies different from the present study.

Another limitation of the present study is that exclusion of analyzing non-uniform DIF items. Recent literature has made it more likely and possible to determine if items contain non-uniform DIF (Hidalgo and Lopez-Pina, 2004; Jodoin & Gierl, 2001). The present study only focused on uniform DIF for a more parsimonious study due to the 75 conditions and the comparison between two methods (MH DIF and LR DIF). Further research could include the analysis of non-uniform DIF items and the interactions with the detection of DIF.

Last, the limitation of computer technology. The simulation study for this study was for 200 replications. When choosing to run a simulation, the computer technology that is available can be a limitation to how large you can make your simulation. Depending on the size of the replications in the simulation, computers available to graduate students might not be powerful enough for such a study. Knowledge of computer technology is important for a study of this magnitude.

REFERENCES

- Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*, 113-127.
- Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.
- Angoff, W.H. (1972, September). *A technique for the investigation of cultural differences*. Paper presented at the annual meeting of the American Psychological Association, Honolulu. (ERIC Document Reproduction Service No. ED 069686)
- Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96-116). Baltimore: Johns Hopkins University Press.
- Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Thayer (Eds.), *Differential item functioning* (pp. 397-418) Hillsdale, NJ: Lawrence Erlbaum Associates.
- Angoff, W. H., & Sharon, A. T. (1974). The evaluation of differences in test performance of two or more groups. *Educational and Psychological Measurement, 34*, 807-816.
- Bertrand, R., & Boiteau, N. (2003). *Comparing the stability of IRT-based and non IRT-based DIF methods in different cultural contexts using TIMSS data*. (ERIC Document Reproduction Service No. ED 476924)

- Camille, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P.W. Holland & H. Thayer (Eds.), *Differential item functioning* (pp. 397-418). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cardall, C., & Coffman, W. E. (1964). A method for comparing the performance of different groups on the same items of a test. *Research and Development Reports*, 9, 64-65.
- Cleary, T. A. & Hilton, T. J. (1968). An investigation of item bias. *Educational and Psychological Measurement*, 5, 115-124.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 1304-1312.
- Cohen, J. (1992). The power primer. *Psychological Bulletin*, 112, 155-159.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.
- Cole, N.S. (1978). *Approaches to examining bias in achievement test items*. Paper presented at the national meeting of the American Personnel and Guidance Association, Washington, DC.
- Cole, N.S. (1993). History and development of DIF. In P. W. Holland & H. Thayer (Eds.), *Differential item functioning* (pp. 25-30), Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cook Johnson, C. (1993, November). *The effects of single and compound violations of data set assumptions when using the one-way, fixed effects analysis of variance and the one concomitant analysis*. Paper presented at the annual meeting of the

- Mid-South Educational Research Association, New Orleans. (ERIC Document
Reproduction Service No. ED 365730)
- Courville, T. G. (2004). *An empirical comparison of item response theory and classical
test theory item/person statistics*. Doctoral dissertation, Texas A&M University,
2004.
- Cromwell, S. (2001, February). *An introductory summary of various effect size
choices*. Paper presented at the annual meeting of the Southwest Educational
Research Association, San Antonio, TX. (ERIC Document Reproduction
Service No. ED 449212)
- Dorans, N.J. (1989). Two new approaches to assessing differential item functioning:
Standardization and the Mantel-Haenszel method. *Applied Measurement in
Education*, 2, 217-233.
- 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* (Research Rep. No. 83-9).
Princeton, NJ: Educational Testing Service.
- Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel
and Standardization. In P. W. Holland & H. Thayer (Eds.), *Differential item
functioning* (pp. 35-66), Hillsdale, NJ: Lawrence Erlbaum Associates.
- Dorans, N.J., & Potenza, M.T. (1994). *Equity assessment for polytomously scored
items: a taxonomy of procedures for assessing differential item functioning*
(Research Report RR-94-49). Princeton, NJ: Educational Testing

Service.

- Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*.
Multivariate Applications Books Series. Mahwah, NJ: Lawrence Erlbaum
Associates.
- Fan, X. (1998). Item Response Theory and classical test theory: an empirical
comparison of their item/person statistics. *Educational and Psychological
Measurement, 58*, 357-381.
- Fidler, F. (2002). The fifth edition of the *APA Publication Manual*: why its statistics
recommendations are so controversial. *Educational and Psychological
Measurement, 62*, 749-770.
- Finch, S., Cumming, G., & Thomason, N. (2001). Reporting statistical inference in the
Journal of Applied Psychology: Little evidence of reform. *Educational and
Psychological Measurement, 61*, 181-210.
- Fleishman, A.K. (1978). A method for simulating non-normal distributions.
Psychometrika, 43, 521-532.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *MMSS fundamentals of
Item Response Theory*. Newbury Park, CA: Sage Publications.
- Harlow, L. (1997). Significance testing introduction and overview. In L. Harlow, S.
Mulaik, & J. Steiger (Eds), *What if there were no significance tests?* (pp. 1-17).
Mahwah, NJ: Lawrence Erlbaum.
- Hidalgo, M. D., & Lopez-Pina, J.A. (2004). Differential item functioning detection and
effect size: a comparison between logistic regression and Mantel-Haenszel

- procedures. *Educational and Psychological Measurement*, 64(6), 903-915.
- Holland, P.W. (1985). *On the study of differential item performance without IRT*. Proceedings of the 27th Annual Conference of the Military Testing Association (Vol. 1, pp. 282-287). San Diego.
- Holland P.W., & Thayer, H. (1986, April). *Differential item performance and the Mantel-Haenszel Procedure*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Holland, P.W., & Thayer, H. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp.129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Huberty, C. (2002). A history of effect size indices. *Educational and Psychological Measurement*, 62, 227-240.
- Jenson, A. R. (1980). *Bias in mental testing*. New York: The Free Press.
- Jodoin, M. G., & Gierl, M.J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, 14, 329-349.
- Kirk, R. (1996). Practicial significance: a concept whose time has come. *Educational & Psychological Measurement*, 56, 746-759.
- Kirk, R. (2001). Promoting good statistical practices: some suggestions. *Educational and Psychological Measurement*, 61, 213-218.

- Linn, R. L. & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement*, 18, 109-118.
- Lord, F.M. (1952). A theory of test scores. *Psychometric Monographs* (Whole No. 7), Richmond, VA: William Byrd Press.
- Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- MacDonald, P., & Paunonen, S.V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, 62, 921-943.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.
- Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement*, 32, 131-144.
- McKinley, R., & Mills, C. (1989). Item response theory: advances in achievement and attitude measurement. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 1, pp. 71-135). Greenwich, CT: JAI Press.

- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156-166.
- Millsap, R.E., & Everson, H.T. (1993). Methodology review: statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297-334.
- Monahan, P. (2000, April). *The effect of unequal variances in the ability distributions on they type I error rate of the Mantel-Haenszel chi-square test for detecting DIF*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans (ERIC Document Reproduction Service No. ED 442839)
- Paek, I. (2002). *Investigations of differential item functioning: comparisons among approaches, and extension to a multidimensional context*. Doctoral Dissertation, University of California, Berkeley, 2002.
- Pearson, K. (1895). Contributions to the mathematical theory of evolution: II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society, Ser. A*, *186*, 343-414.
- Penny, J., & Johnson, R. L. (1999). How group differences in matching criterion distribution and IRT item difficulty can influence the magnitude of the Mantel-Haenszel chi-square DIF index. *Journal of Experimental Education*, *67*(4), 343-366.
- Pommerich, M., Spray, J. A., & Parshal, C. G. (1994, April). *The performance of the Mantel-Haenszel DIF statistic when comparison group distributions are incongruent*. Paper presented at the Annual Meeting of the National Council on

Measurement in Education, New Orleans. (ERIC Document Reproduction Service No. ED372097)

Potenza, M.T., & Dorans, N.J. (1995). DIF assessment for polytomously scored items:

A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.

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.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmark's Paedagogiske Institut. (Republished in 1980 by the University of Chicago Press, Chicago).

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*, 45-58.

Rosnow, R.L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods, 1*, 331-340.

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

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). A Monte Carlo comparison of

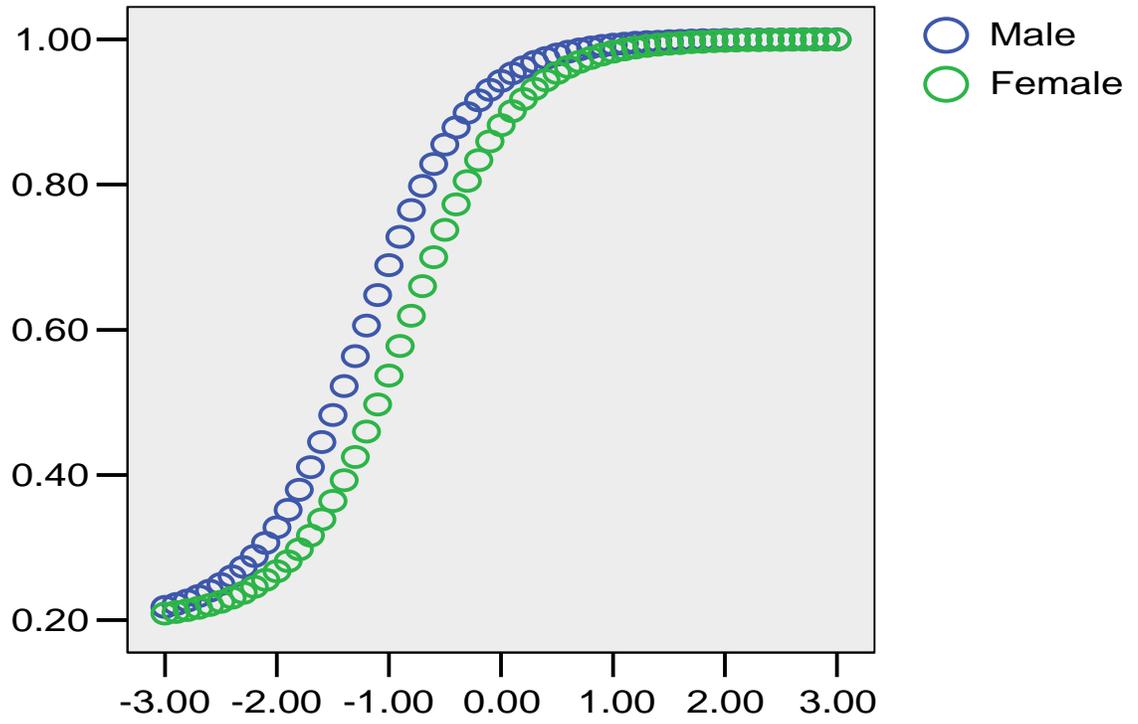
- seven biased item detection techniques. *Journal of Educational Measurement*, 17, 1-10.
- Scheuneman, J.D. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, 16, 143-152.
- Schiell, J.L., & King, J. E. (1999). *Accuracy of course placement validity statistics under various soft truncation conditions*. ACT Research Report Series 99-2.
- Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1, 115-129.
- Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Thayer (Eds.), *Differential item functioning* (pp. 281-315), Hillsdale, NJ: Lawrence Erlbaum Associates.
- Schnipke, D. L., Roussos, L. A., & Pashley, P. J. (2000). *A comparison of Mantel-Haenszel differential item functioning parameters*. Newtown, PA: Law School Admission Council.
- Scientific Software International. (2003). *BILOG-MG 3[®]*. St. Paul, MN: Assessment Systems Corporation.
- Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, 6, 317-375.
- Srivastava, M. S. (2002). *Methods of multivariate statistics*. New York: Wiley.

- Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370.
- Sweeney, K. P. (1996). *A Monte Carlo investigation of the likelihood-ratio procedure in the detection of differential item functioning*. Unpublished doctoral dissertation, Fordham University, New York, NY.
- Tabachnick, B.G., & Fidell, L.S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.
- Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Thayer (Eds.), *Differential Item Functioning* (pp. 67-114) Hillsdale, NJ: Lawrence Erlbaum Associates.
- Thomas, D. R., & Zumbo, B. D. (1998). *Variable importance in logistic regression based on partitioning an R-squared measure*. Paper presented at the Psychometric Society Meetings, Urbana, IL.
- Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: three suggested reforms. *Educational Researcher*, 25(3), 26-30.
- Thompson, B. (1998). Review of *What if there were no significance tests?* *Educational and Psychological Measurement*, 58, 332-344.

- Thompson, B. (1999). Statistical significance tests, effect size reporting, and the vain pursuit of pseudo-objectivity. *Theory & Psychology, 9*(2), 191-196.
- Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80*, 64-71.
- Thompson, B., & Snyder, P.A. (1998). Statistical significance and reliability analyses in recent JCD research articles. *Journal of Counseling and Development, 76*, 436-441.
- Thompson, B., & Keiffer, K. (2000). Interpreting statistical significance test results: A proposed new "what if" method. *Research in the Schools, 7*, 3-10.
- Traub, J. M. (1983). A priori considerations in choosing an item response model. *Journal of Educational Measurement, 13*, 28-34.
- Thurstone, L.L. (1925). A method of scaling educational and psychological tests. *Journal of Educational Psychology, 16*, 263-278.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Thayer (Eds.), *Differential item functioning* (pp. 337-347), Hillsdale, NJ: Lawrence Erlbaum Associates.
- Zumbo, B.D., & Thomas, D.R. (1996, October). *A measure of DIF effect size using logistic regression procedures*. Paper presented at the National Board of Medical Examiners, Philadelphia.

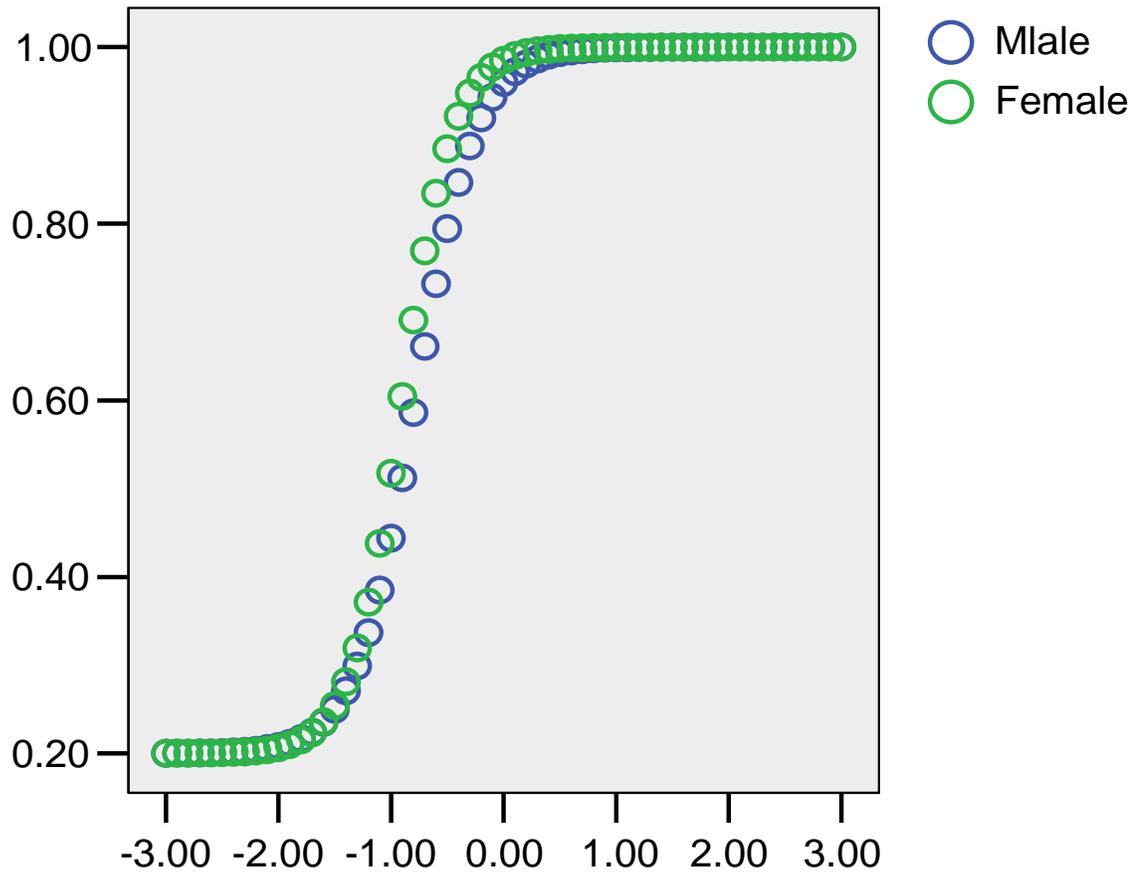
APPENDIX A

Non-DIF Item 3 ICC Curve



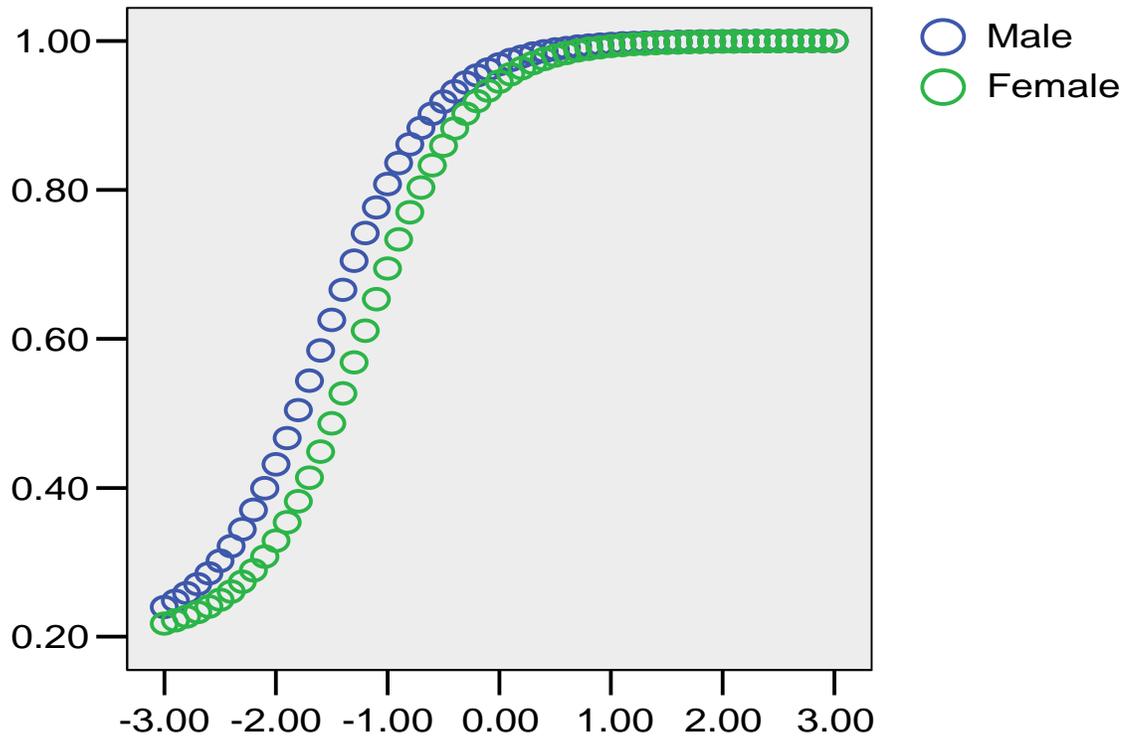
APPENDIX B

Non-DIF Item 5 ICC Curve



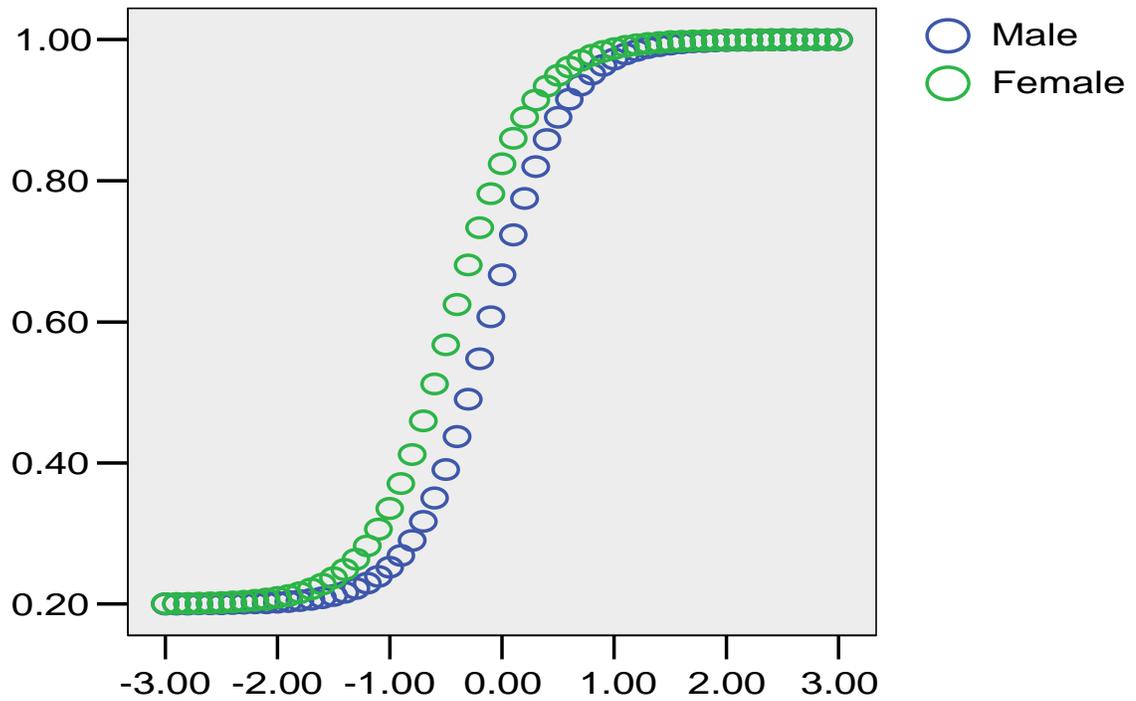
APPENDIX C

Non-DIF Item 6 ICC Curve



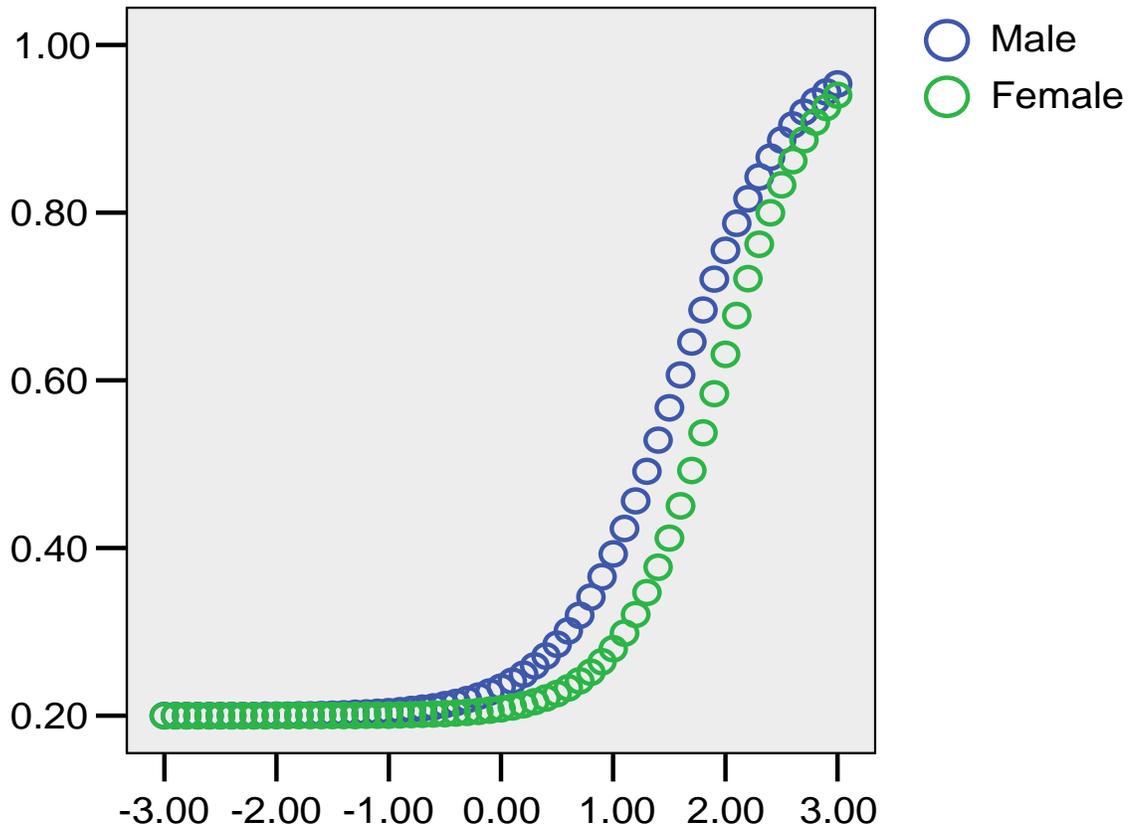
APPENDIX D

Non-DIF Item 32 ICC Curve



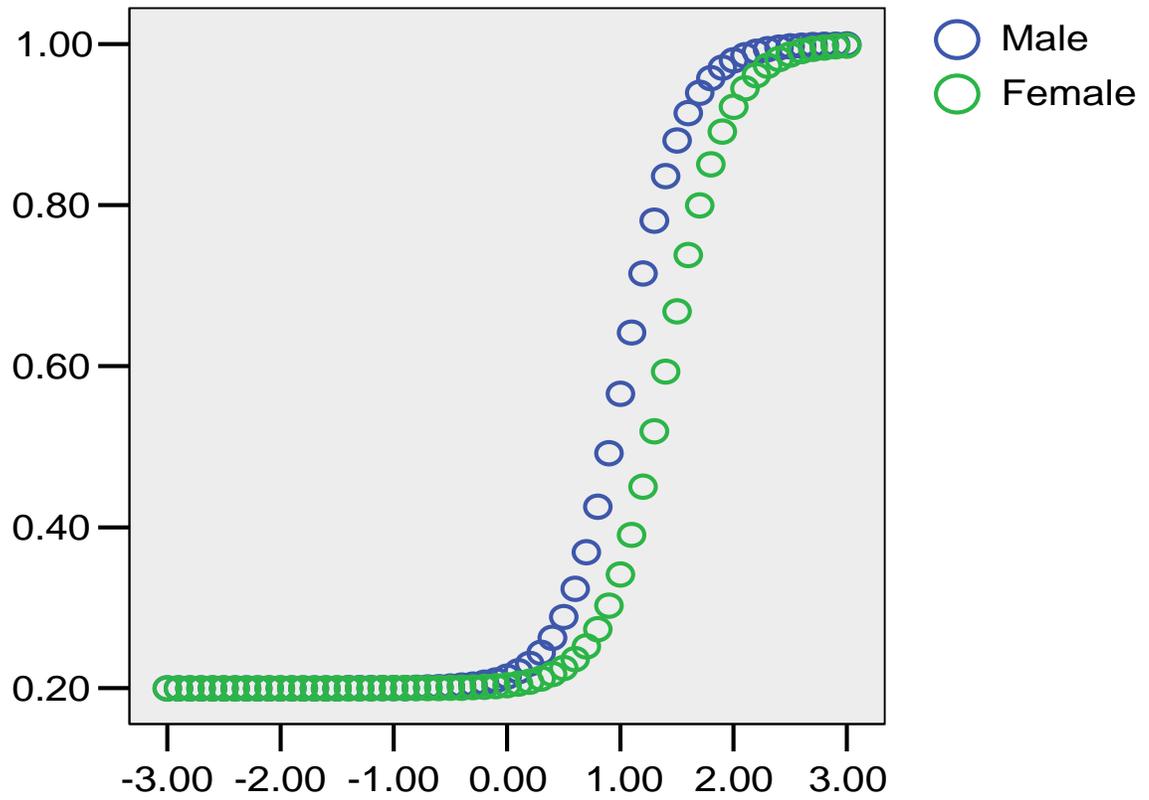
APPENDIX E

Non-DIF Item 55 ICC Curve



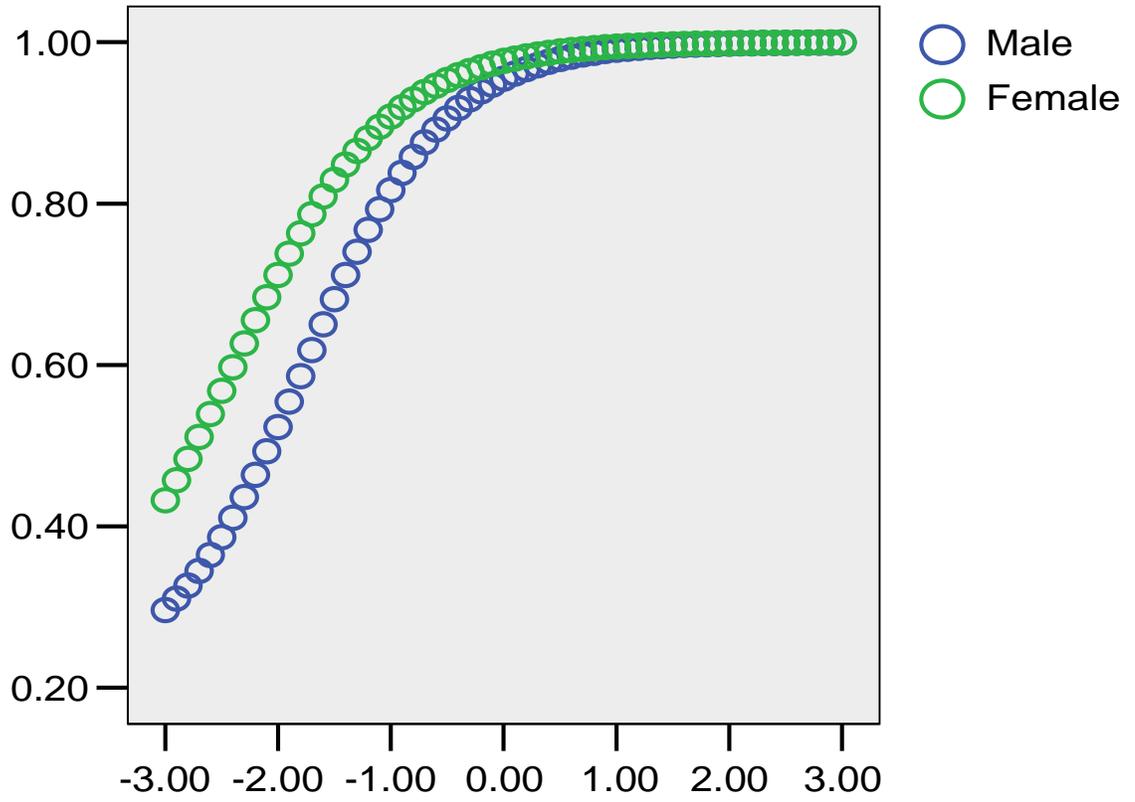
APPENDIX F

Non-DIF Item 58 ICC Curve



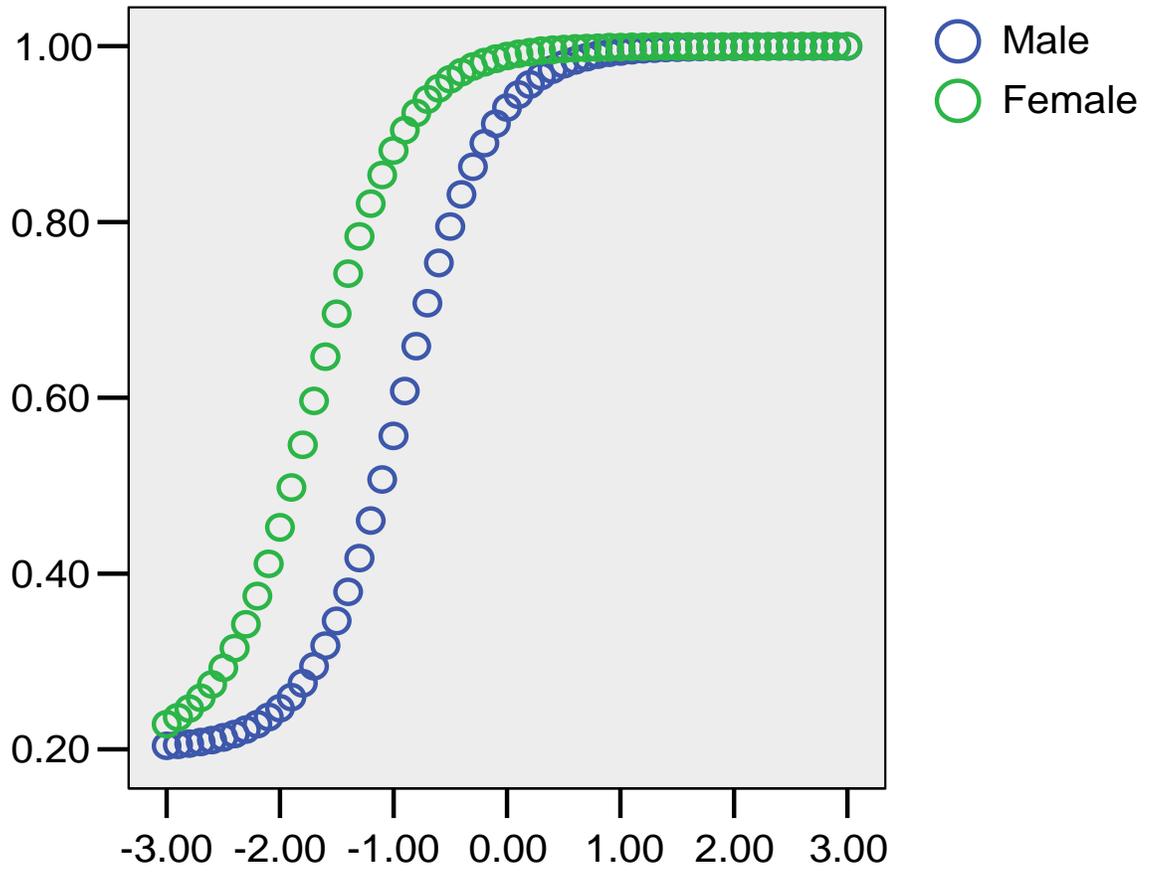
APPENDIX G

DIF Item 1 ICC Curve



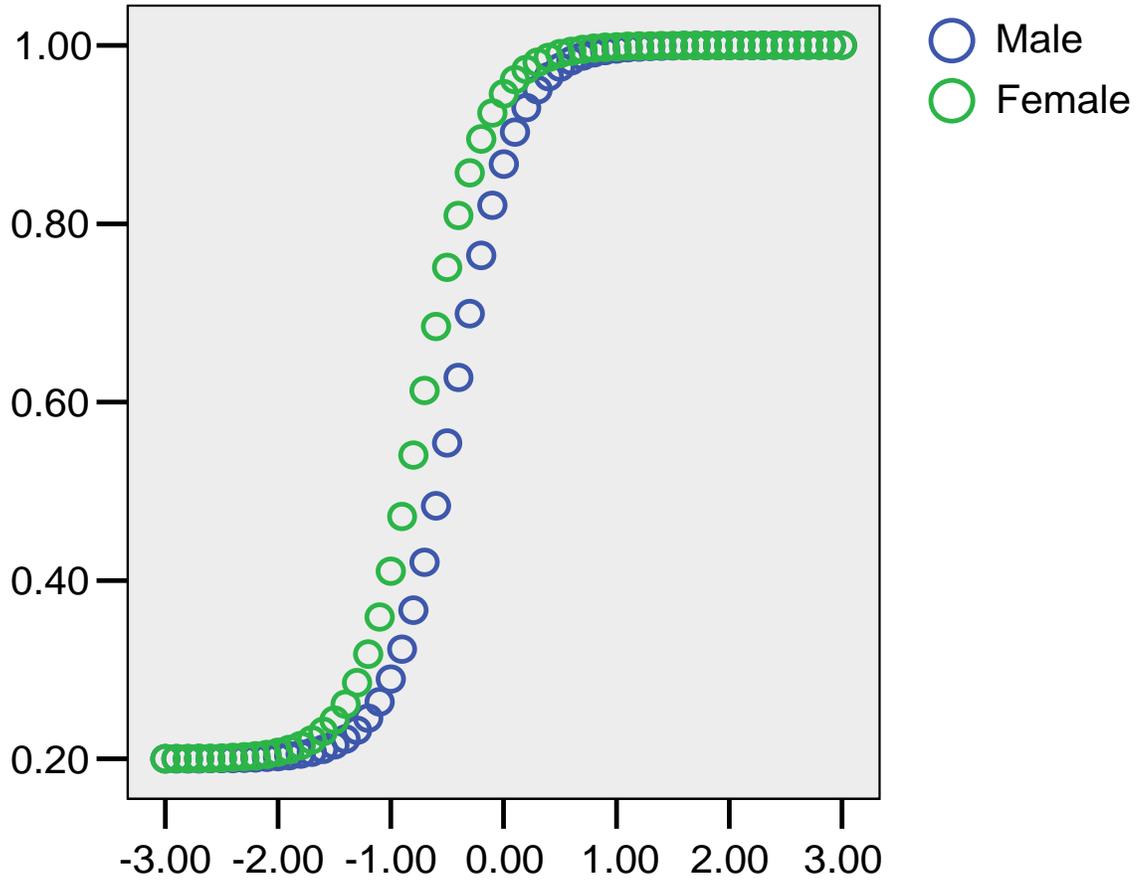
APPENDIX H

DIF Item 8 ICC Curve



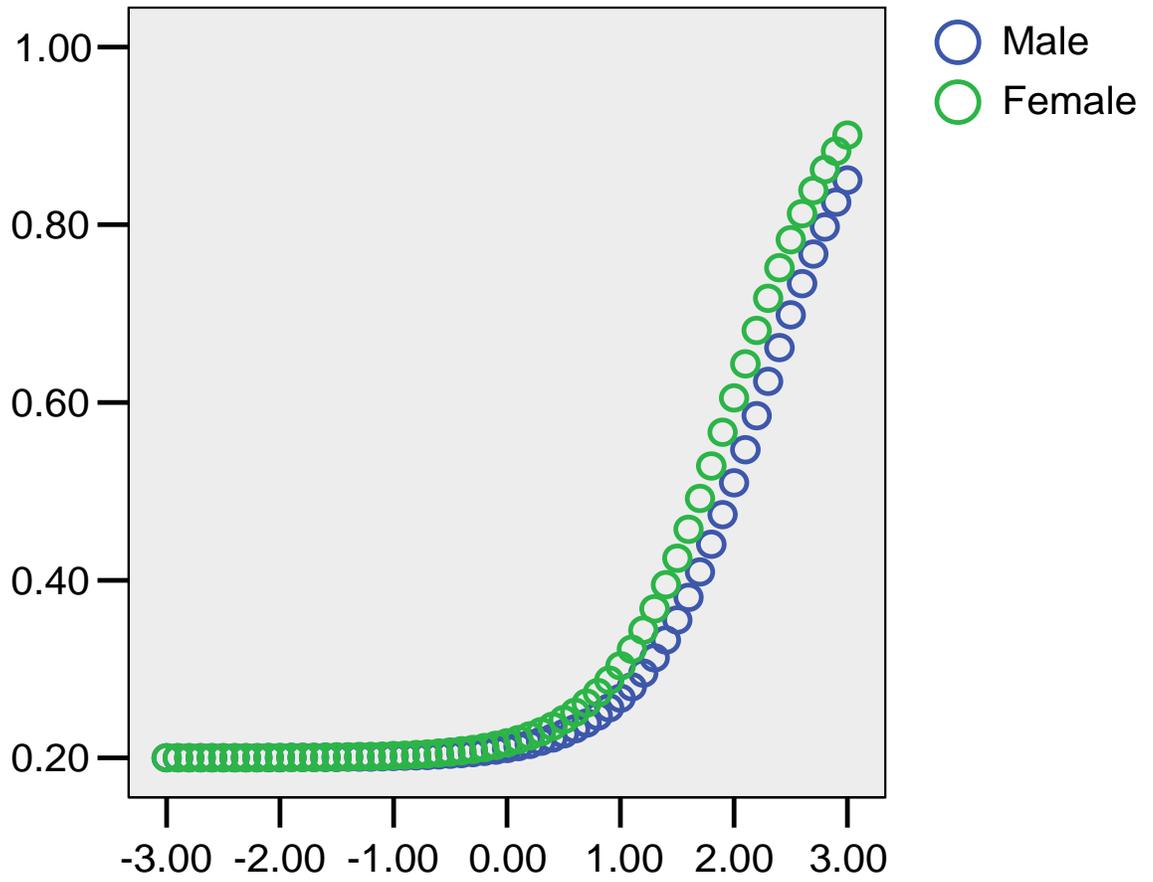
APPENDIX I

DIF Item 22 ICC Curve

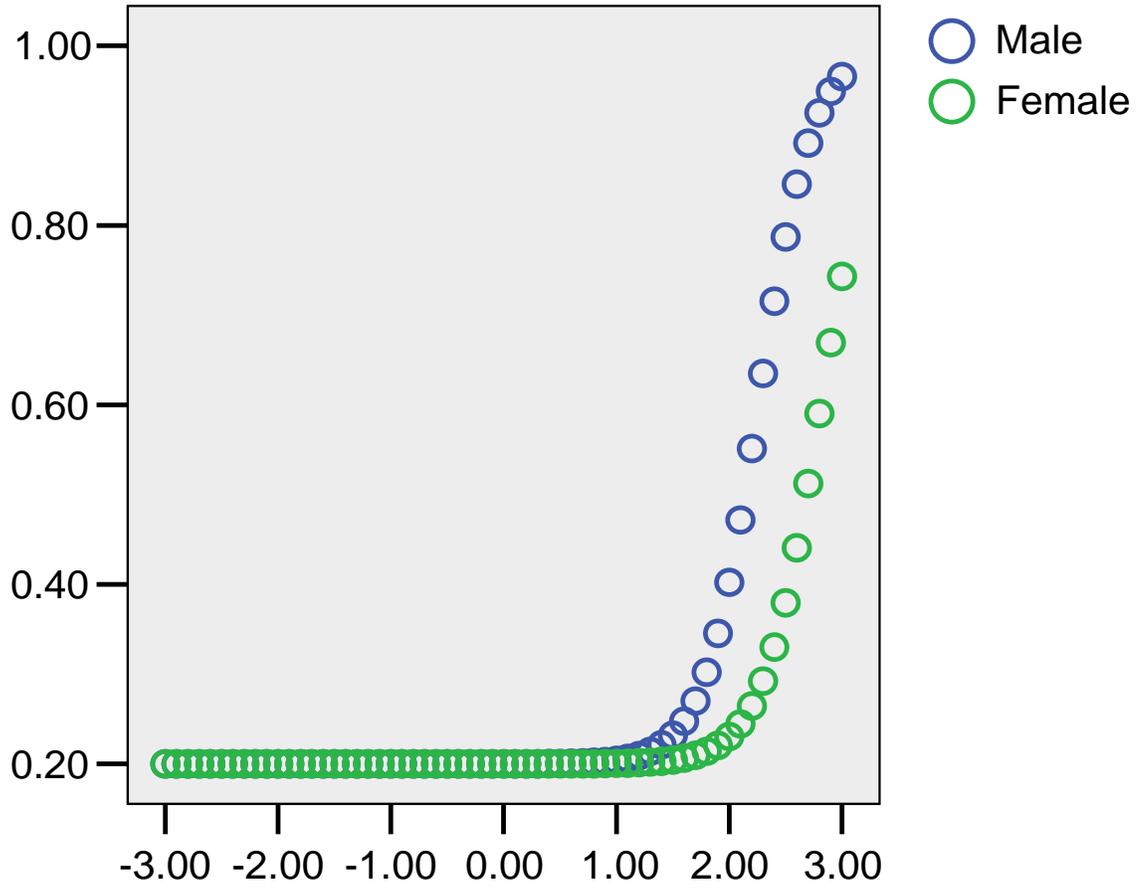


APPENDIX J

DIF Item 39 ICC Curve



APPENDIX K
DIF Item 60 ICC Curve



APPENDIX L

Table L1 Repeated Measures for p values Descriptive Statistics

	sample	ability	pop	Mean	Std. Deviation	N
p LR	1000/100	Normal	Normally Dist	.3232080	.29689911	2191
			Moderately Dist	.3172678	.29434781	2190
			Skew/Lepto.	.3264975	.30243670	2192
			Skew/Extremely Lepto	.3201505	.30006035	2194
			Platykurtic	.3125868	.28938341	2189
			Total	.3199444	.29664786	10956
		Moderate	Normally Dist	.2032242	.28100606	2192
			Moderately Dist	.2080706	.28425361	2189
			Skew/Lepto.	.2114970	.28318231	2192
			Skew/Extremely Lepto	.2075134	.28121592	2194
			Platykurtic	.2059483	.27548829	2192
			Total	.2072505	.28100643	10959
		Severe	Normally Dist	.2339055	.29138276	2195
			Moderately Dist	.2421713	.29800647	2192
			Skew/Lepto.	.2313887	.29246791	2193
			Skew/Extremely Lepto	.2488329	.30201945	2190
			Platykurtic	.2425086	.29733766	2196
			Total	.2397584	.29627947	10966
	Total	Normally Dist	.2534264	.29422668	6578	
		Moderately Dist	.2558397	.29575725	6571	
		Skew/Lepto.	.2564572	.29702701	6577	
		Skew/Extremely Lepto	.2588383	.29818640	6578	
		Platykurtic	.2536475	.29088853	6577	
		Total	.2556418	.29521676	32881	
	500/100	Normal	Normally Dist	.3358360	.30470987	2190
			Moderately Dist	.3304935	.29604188	2188
			Skew/Lepto.	.3201655	.29345808	2190
			Skew/Extremely Lepto	.3202400	.29926873	2190
			Platykurtic	.3368430	.30203143	2191
			Total	.3287160	.29916392	10949
		Moderate	Normally Dist	.2162400	.28353362	2193
			Moderately Dist	.2235848	.28393428	2195
			Skew/Lepto.	.2443089	.29787335	2192
			Skew/Extremely Lepto	.2337981	.29184314	2190
			Platykurtic	.2163857	.28023392	2199
			Total	.2268537	.28770039	10969
Severe		Normally Dist	.2714637	.29638900	2193	
		Moderately Dist	.2699389	.29736745	2193	
		Skew/Lepto.	.2707639	.29967486	2192	

Table L1 continued

	sample	ability	pop	Mean	Std. Deviation	N		
300/300	Total		Skew/Extremely Lepto	.2748981	.30268782	2191		
			Platykurtic	.2775471	.30264040	2192		
			Total	.2729217	.29972141	10961		
			Normally Dist	.2744853	.29897783	6576		
			Moderately Dist	.2746144	.29571757	6576		
			Skew/Lepto.	.2784001	.29862866	6574		
			Skew/Extremely Lepto	.2763118	.30000745	6571		
			Platykurtic	.2768518	.29916364	6582		
			Total	.2761326	.29848793	32879		
			Normally Dist	.2540557	.28626229	2196		
			Moderately Dist	.2563716	.29059650	2192		
			Skew/Lepto.	.2570282	.29305261	2194		
	Normal	Moderate		Skew/Extremely Lepto	.2508354	.28483453	2194	
				Platykurtic	.2616736	.29801361	2192	
				Total	.2559914	.29055762	10968	
				Normally Dist	.1602221	.25004765	2191	
				Moderately Dist	.1583251	.24777810	2191	
				Skew/Lepto.	.1695220	.26413597	2191	
		Severe			Skew/Extremely Lepto	.1576798	.24978586	2189
					Platykurtic	.1632725	.25740615	2196
					Total	.1618057	.25389689	10958
					Normally Dist	.2438325	.28807531	2197
					Moderately Dist	.2284593	.27257142	2191
					Skew/Lepto.	.2320692	.27692020	2199
Total	Moderate		Skew/Extremely Lepto	.2337346	.27852253	2196		
			Platykurtic	.2498289	.28739050	2195		
			Total	.2375870	.28082620	10978		
			Normally Dist	.2194187	.27851874	6584		
			Moderately Dist	.2143917	.27396961	6574		
			Skew/Lepto.	.2195721	.28067245	6584		
	Severe			Skew/Extremely Lepto	.2141321	.27445394	6579	
				Platykurtic	.2248989	.28480772	6583	
				Total	.2184844	.27852691	32904	
				Normally Dist	.1883722	.25940598	2192	
				Moderately Dist	.1915250	.26216891	2196	
				Skew/Lepto.	.2191183	.27045804	2194	
Normal			Skew/Extremely Lepto	.2003502	.26727179	2197		
			Platykurtic	.1982949	.27143438	2192		
			Total	.1995331	.26635583	10971		
			Normally Dist	.1238915	.22621062	2196		
			Moderately Dist	.1264924	.23396795	2197		

Table L1 continued

	sample	ability	pop	Mean	Std. Deviation	N
			Skew/Lepto.	.1184398	.21811718	2197
			Skew/Extremely Lepto	.1269064	.23354179	2198
			Platykurtic	.1190502	.22343261	2196
			Total	.1229567	.22712298	10984
		Severe	Normally Dist	.1971288	.28986625	2195
			Moderately Dist	.1952622	.28777868	2198
			Skew/Lepto.	.1962220	.28925130	2194
			Skew/Extremely Lepto	.2016681	.29247274	2196
			Platykurtic	.1921132	.28656519	2196
			Total	.1964786	.28915736	10979
		Total	Normally Dist	.1697820	.26180004	6583
			Moderately Dist	.1710938	.26408665	6591
			Skew/Lepto.	.1778996	.26449454	6585
			Skew/Extremely Lepto	.1762969	.26776919	6591
			Platykurtic	.1698021	.26428104	6584
			Total	.1729754	.26450014	32934
	Total	Normal	Normally Dist	.2753390	.29328854	8769
			Moderately Dist	.2738412	.29133091	8766
			Skew/Lepto.	.2806739	.29346925	8770
			Skew/Extremely Lepto	.2728476	.29251111	8775
			Platykurtic	.2773307	.29521869	8764
			Total	.2760062	.29316609	43844
		Moderate	Normally Dist	.1758771	.26375307	8772
			Moderately Dist	.1790959	.26620096	8772
			Skew/Lepto.	.1859053	.27157182	8772
			Skew/Extremely Lepto	.1814393	.26833703	8771
			Platykurtic	.1761643	.26289129	8783
			Total	.1796955	.26658227	43870
		Severe	Normally Dist	.2365763	.29260551	8780
			Moderately Dist	.2339362	.29031436	8774
			Skew/Lepto.	.2326021	.29083338	8778
			Skew/Extremely Lepto	.2397572	.29521725	8773
			Platykurtic	.2404815	.29512398	8779
			Total	.2366706	.29282925	43884
		Total	Normally Dist	.2292611	.28647946	26321
			Moderately Dist	.2289479	.28549575	26312
			Skew/Lepto.	.2330567	.28805776	26320
			Skew/Extremely Lepto	.2313550	.28809087	26319
			Platykurtic	.2312909	.28785008	26326
			Total	.2307824	.28719639	131598
p MH	1000/100	Normal	Normally Dist	.3064717	.30123688	2191

Table L1 continued

	samplen	ability	pop	Mean	Std. Deviation	N
			Moderately Dist	.3042163	.30307826	2190
			Skew/Lepto.	.3112017	.30544466	2192
			Skew/Extremely Lepto	.3046938	.30390167	2194
			Platykurtic	.2944899	.29330742	2189
			Total	.3042172	.30142060	10956
		Moderate	Normally Dist	.3027072	.29928341	2192
			Moderately Dist	.3140314	.29976724	2189
			Skew/Lepto.	.3126838	.30034731	2192
			Skew/Extremely Lepto	.3094056	.30108387	2194
			Platykurtic	.3104284	.30283843	2192
			Total	.3098501	.30063769	10959
		Severe	Normally Dist	.3149738	.29726162	2195
			Moderately Dist	.3050284	.30183826	2192
			Skew/Lepto.	.3165872	.29733437	2193
			Skew/Extremely Lepto	.3182036	.29919377	2190
			Platykurtic	.3132314	.30421456	2196
			Total	.3136046	.29996184	10966
		Total	Normally Dist	.3080543	.29926234	6578
			Moderately Dist	.3077569	.30155160	6571
			Skew/Lepto.	.3134914	.30102295	6577
			Skew/Extremely Lepto	.3107631	.30140681	6578
			Platykurtic	.3060595	.30023268	6577
			Total	.3092253	.30068935	32881
	500/100	Normal	Normally Dist	.3217876	.30007762	2190
			Moderately Dist	.3163256	.29867186	2188
			Skew/Lepto.	.3108028	.30027055	2190
			Skew/Extremely Lepto	.3118865	.30361244	2190
			Platykurtic	.3255493	.30529106	2191
			Total	.3172713	.30159383	10949
		Moderate	Normally Dist	.3230170	.30629429	2193
			Moderately Dist	.3164260	.30041123	2195
			Skew/Lepto.	.3204467	.29970118	2192
			Skew/Extremely Lepto	.3397657	.30253755	2190
			Platykurtic	.3079612	.29878868	2199
			Total	.3215101	.30168270	10969
		Severe	Normally Dist	.3228770	.29977004	2193
			Moderately Dist	.3483279	.31056978	2193
			Skew/Lepto.	.3373588	.30116704	2192
			Skew/Extremely Lepto	.3301273	.30319229	2191
			Platykurtic	.3330084	.30337107	2192
			Total	.3343405	.30369901	10961

Table L1 continued

	samplen	ability	pop	Mean	Std. Deviation	N
		Total	Normally Dist	.3225609	.30201773	6576
			Moderately Dist	.3270314	.30359352	6576
			Skew/Lepto.	.3228731	.30053504	6574
			Skew/Extremely Lepto	.3272602	.30328871	6571
			Platykurtic	.3221573	.30262803	6582
		Total		.3243759	.30240472	32879
	300/300	Normal	Normally Dist	.2564343	.29898876	2196
			Moderately Dist	.2554304	.29557326	2192
			Skew/Lepto.	.2461547	.29132105	2194
			Skew/Extremely Lepto	.2546067	.29304808	2194
			Platykurtic	.2553027	.29555660	2192
		Total		.2535856	.29487974	10968
		Moderate	Normally Dist	.2369647	.29238989	2191
			Moderately Dist	.2342836	.28913482	2191
			Skew/Lepto.	.2398501	.29110337	2191
			Skew/Extremely Lepto	.2332648	.28937596	2189
			Platykurtic	.2491674	.30137584	2196
		Total		.2387119	.29271764	10958
		Severe	Normally Dist	.2839376	.30141629	2197
			Moderately Dist	.2749996	.29662059	2191
			Skew/Lepto.	.2778562	.29949122	2199
			Skew/Extremely Lepto	.2614565	.30000355	2196
			Platykurtic	.2619360	.29041677	2195
		Total		.2720394	.29769697	10978
		Total	Normally Dist	.2591328	.29820532	6584
			Moderately Dist	.2549046	.29422059	6574
			Skew/Lepto.	.2546447	.29442917	6584
			Skew/Extremely Lepto	.2497921	.29438100	6579
			Platykurtic	.2554678	.29581896	6583
		Total		.2547891	.29541243	32904
	1000/300	Normal	Normally Dist	.1979128	.27484732	2192
			Moderately Dist	.1902225	.26829301	2196
			Skew/Lepto.	.1998810	.26894977	2194
			Skew/Extremely Lepto	.2095934	.28361553	2197
			Platykurtic	.2013148	.27748892	2192
		Total		.1997859	.27471889	10971
		Moderate	Normally Dist	.2263342	.29904234	2196
			Moderately Dist	.2172151	.29222840	2197
			Skew/Lepto.	.2202757	.29292248	2197
			Skew/Extremely Lepto	.2246103	.29891118	2198
			Platykurtic	.2135532	.28652499	2196

Table L1 continued

	sample	ability	pop	Mean	Std. Deviation	N
			Total	.2203982	.29394773	10984
		Severe	Normally Dist	.2138932	.28424992	2195
			Moderately Dist	.2142435	.28920730	2198
			Skew/Lepto.	.2120549	.28673037	2194
			Skew/Extremely Lepto	.2192269	.28889642	2196
			Platykurtic	.2225567	.28792514	2196
			Total	.2163957	.28738217	10979
		Total	Normally Dist	.2127222	.28641974	6583
			Moderately Dist	.2072306	.28366042	6591
			Skew/Lepto.	.2107415	.28313557	6585
			Skew/Extremely Lepto	.2178110	.29056751	6591
			Platykurtic	.2124817	.28411149	6584
			Total	.2121975	.28559581	32934
	Total	Normal	Normally Dist	.2706294	.29790764	8769
			Moderately Dist	.2664825	.29584876	8766
			Skew/Lepto.	.2669800	.29552555	8770
			Skew/Extremely Lepto	.2701554	.29898199	8775
			Platykurtic	.2691491	.29668311	8764
			Total	.2686795	.29698371	43844
		Moderate	Normally Dist	.2722447	.30209529	8772
			Moderately Dist	.2704636	.29880624	8772
			Skew/Lepto.	.2732877	.29923189	8772
			Skew/Extremely Lepto	.2767339	.30199327	8771
			Platykurtic	.2702721	.30019994	8783
			Total	.2725997	.30046384	43870
		Severe	Normally Dist	.2839115	.29880717	8780
			Moderately Dist	.2856094	.30355205	8774
			Skew/Lepto.	.2859445	.30000160	8778
			Skew/Extremely Lepto	.2822017	.30115040	8773
			Platykurtic	.2826626	.29967025	8779
			Total	.2840660	.30063042	43884
		Total	Normally Dist	.2755983	.29965579	26321
			Moderately Dist	.2741878	.29952229	26312
			Skew/Lepto.	.2754071	.29835294	26320
			Skew/Extremely Lepto	.2763632	.30073991	26319
			Platykurtic	.2740301	.29890792	26326
			Total	.2751173	.29943356	131598

APPENDIX M

Table M1 Repeated Measures for Effect Size Descriptive Statistics

	sample	ability	pop	Mean	Std. Deviation	N
rs2cat	1000/100	Normal	Normally Dist	1.3131	.46386	2191
			Moderately Dist	1.3297	.47214	2190
			Skew/Lepto.	1.3307	.47156	2192
			Skew/Extremely Lepto	1.3314	.47178	2194
			Platykurtic	1.3175	.46561	2189
			Total	1.3245	.46898	10956
		Moderate	Normally Dist	1.4895	.53698	2192
			Moderately Dist	1.4619	.53406	2189
			Skew/Lepto.	1.4772	.53061	2192
			Skew/Extremely Lepto	1.4813	.53587	2194
			Platykurtic	1.4790	.52550	2192
			Total	1.4778	.53260	10959
		Severe	Normally Dist	1.3786	.48796	2195
			Moderately Dist	1.3850	.48859	2192
			Skew/Lepto.	1.3611	.48139	2193
	Skew/Extremely Lepto		1.3502	.47811	2190	
	Platykurtic		1.3739	.48675	2196	
	Total		1.3698	.48465	10966	
	Total	Normally Dist	1.3937	.50242	6578	
		Moderately Dist	1.3922	.50180	6571	
		Skew/Lepto.	1.3897	.49912	6577	
		Skew/Extremely Lepto	1.3877	.50049	6578	
		Platykurtic	1.3901	.49770	6577	
		Total	1.3907	.50028	32881	
	500/100	Normal	Normally Dist	1.4817	.53253	2190
			Moderately Dist	1.4744	.53140	2188
			Skew/Lepto.	1.4959	.52938	2190
			Skew/Extremely Lepto	1.4963	.53880	2190
			Platykurtic	1.4696	.54969	2191
			Total	1.4836	.53643	10949
Moderate		Normally Dist	1.6133	.58263	2193	
		Moderately Dist	1.6150	.57829	2195	
		Skew/Lepto.	1.6638	.57453	2192	
		Skew/Extremely Lepto	1.6073	.55992	2190	
		Platykurtic	1.6498	.59356	2199	
		Total	1.6299	.57824	10969	
Severe		Normally Dist	1.5166	.53935	2193	
		Moderately Dist	1.5093	.52669	2193	
		Skew/Lepto.	1.5155	.52915	2192	

Table M1 continued

	sample	ability	pop	Mean	Std. Deviation	N
			Skew/Extremely Lepto	1.5244	.53055	2191
			Platykurtic	1.4986	.53023	2192
		Total	Total	1.5129	.53118	10961
			Normally Dist	1.5373	.55467	6576
			Moderately Dist	1.5330	.54916	6576
			Skew/Lepto.	1.5584	.54983	6574
			Skew/Extremely Lepto	1.5427	.54518	6571
			Platykurtic	1.5395	.56398	6582
			Total	1.5422	.55264	32879
300/300		Normal	Normally Dist	1.6762	.65119	2196
			Moderately Dist	1.6816	.63149	2192
			Skew/Lepto.	1.6737	.64145	2194
			Skew/Extremely Lepto	1.6805	.62805	2194
			Platykurtic	1.6775	.65603	2192
			Total	1.6779	.64163	10968
		Moderate	Normally Dist	1.6426	.55771	2191
			Moderately Dist	1.6490	.55604	2191
			Skew/Lepto.	1.6495	.56001	2191
			Skew/Extremely Lepto	1.6487	.55783	2189
			Platykurtic	1.6471	.57570	2196
			Total	1.6474	.56141	10958
		Severe	Normally Dist	1.6800	.60735	2197
			Moderately Dist	1.6878	.59850	2191
			Skew/Lepto.	1.6826	.59816	2199
			Skew/Extremely Lepto	1.6981	.61745	2196
			Platykurtic	1.6738	.63132	2195
			Total	1.6845	.61063	10978
		Total	Normally Dist	1.6663	.60680	6584
			Moderately Dist	1.6728	.59630	6574
			Skew/Lepto.	1.6686	.60088	6584
			Skew/Extremely Lepto	1.6758	.60220	6579
			Platykurtic	1.6661	.62196	6583
			Total	1.6699	.60567	32904
1000/300		Normal	Normally Dist	1.5274	.53039	2192
			Moderately Dist	1.5305	.52585	2196
			Skew/Lepto.	1.5365	.52981	2194
			Skew/Extremely Lepto	1.5253	.54147	2197
			Platykurtic	1.5429	.53194	2192
			Total	1.5325	.53186	10971
		Moderate	Normally Dist	1.6088	.51889	2196
			Moderately Dist	1.6113	.51305	2197

Table M1 continued

	sample	ability	pop	Mean	Std. Deviation	N
			Skew/Lepto.	1.6213	.51962	2197
			Skew/Extremely Lepto	1.6201	.51460	2198
			Platykurtic	1.6184	.52553	2196
			Total	1.6160	.51828	10984
		Severe	Normally Dist	1.5408	.52603	2195
			Moderately Dist	1.5605	.51268	2198
			Skew/Lepto.	1.5533	.51880	2194
			Skew/Extremely Lepto	1.5697	.51419	2196
			Platykurtic	1.5697	.51772	2196
			Total	1.5588	.51792	10979
		Total	Normally Dist	1.5590	.52625	6583
			Moderately Dist	1.5674	.51822	6591
			Skew/Lepto.	1.5704	.52397	6585
			Skew/Extremely Lepto	1.5717	.52493	6591
			Platykurtic	1.5770	.52594	6584
			Total	1.5691	.52387	32934
Total		Normal	Normally Dist	1.4997	.56372	8769
			Moderately Dist	1.5041	.55762	8766
			Skew/Lepto.	1.5092	.55998	8770
			Skew/Extremely Lepto	1.5084	.56159	8775
			Platykurtic	1.5019	.56999	8764
			Total	1.5047	.56258	43844
		Moderate	Normally Dist	1.5886	.55258	8772
			Moderately Dist	1.5844	.55055	8772
			Skew/Lepto.	1.6029	.55153	8772
			Skew/Extremely Lepto	1.5893	.54603	8771
			Platykurtic	1.5987	.56023	8783
			Total	1.5928	.55222	43870
		Severe	Normally Dist	1.5290	.55227	8780
			Moderately Dist	1.5357	.54401	8774
			Skew/Lepto.	1.5283	.54568	8778
			Skew/Extremely Lepto	1.5357	.55171	8773
			Platykurtic	1.5290	.55495	8779
			Total	1.5315	.54973	43884
		Total	Normally Dist	1.5391	.55742	26321
			Moderately Dist	1.5414	.55172	26312
			Skew/Lepto.	1.5468	.55388	26320
			Skew/Extremely Lepto	1.5445	.55415	26319
			Platykurtic	1.5432	.56321	26326
			Total	1.5430	.55609	131598

Table M1 continued

	sample	ability	pop	Mean	Std. Deviation	N
odd2c at	1000/100	Normal	Normally Dist	2.7882	1.06367	2191
			Moderately Dist	2.7639	1.07951	2190
			Skew/Lepto.	2.8171	1.05824	2192
			Skew/Extremely Lepto	2.7835	1.08342	2194
			Platykurtic	2.7803	1.08195	2189
			Total	2.7866	1.07335	10956
		Moderate	Normally Dist	2.7464	1.20341	2192
			Moderately Dist	2.7268	1.19196	2189
			Skew/Lepto.	2.7436	1.19139	2192
			Skew/Extremely Lepto	2.7548	1.19241	2194
			Platykurtic	2.7500	1.20189	2192
			Total	2.7443	1.19604	10959
		Severe	Normally Dist	2.7276	1.13995	2195
			Moderately Dist	2.7541	1.16945	2192
			Skew/Lepto.	2.7155	1.15466	2193
	Skew/Extremely Lepto		2.7009	1.15834	2190	
	Platykurtic		2.7149	1.16078	2196	
	Total		2.7226	1.15660	10966	
	Total	Normally Dist	2.7540	1.13723	6578	
		Moderately Dist	2.7483	1.14794	6571	
		Skew/Lepto.	2.7587	1.13679	6577	
		Skew/Extremely Lepto	2.7464	1.14596	6578	
		Platykurtic	2.7484	1.14946	6577	
		Total	2.7512	1.14343	32881	
	500/100	Normal	Normally Dist	2.7689	1.07700	2190
			Moderately Dist	2.7888	1.07736	2188
			Skew/Lepto.	2.7740	1.09426	2190
			Skew/Extremely Lepto	2.7963	1.07998	2190
			Platykurtic	2.8138	1.05986	2191
			Total	2.7884	1.07767	10949
Moderate		Normally Dist	2.7241	1.22178	2193	
		Moderately Dist	2.7499	1.21890	2195	
		Skew/Lepto.	2.9193	1.17651	2192	
		Skew/Extremely Lepto	2.8237	1.18039	2190	
		Platykurtic	2.7863	1.21739	2199	
		Total	2.8006	1.20489	10969	
Severe		Normally Dist	2.7031	1.16827	2193	
		Moderately Dist	2.7150	1.14203	2193	
		Skew/Lepto.	2.7391	1.14983	2192	
	Skew/Extremely Lepto	2.7289	1.16002	2191		

Table M1 continued

	samplen	ability	pop	Mean	Std. Deviation	N
			Platykurtic	2.7039	1.16288	2192
		Total	Total	2.7180	1.15652	10961
		Total	Normally Dist	2.7321	1.15741	6576
		Total	Moderately Dist	2.7512	1.14785	6576
		Total	Skew/Lepto.	2.8108	1.14322	6574
		Total	Skew/Extremely Lepto	2.7830	1.14148	6571
		Total	Platykurtic	2.7680	1.14943	6582
		Total	Total	2.7690	1.14814	32879
	300/300	Normal	Normally Dist	2.7910	1.04371	2196
		Normal	Moderately Dist	2.8221	1.04325	2192
		Normal	Skew/Lepto.	2.7967	1.05424	2194
		Normal	Skew/Extremely Lepto	2.7967	1.03635	2194
		Normal	Platykurtic	2.7965	1.04208	2192
		Normal	Total	2.8006	1.04381	10968
		Moderate	Normally Dist	2.7878	1.23353	2191
		Moderate	Moderately Dist	2.7859	1.23691	2191
		Moderate	Skew/Lepto.	2.7891	1.22950	2191
		Moderate	Skew/Extremely Lepto	2.8273	1.23639	2189
		Moderate	Platykurtic	2.8010	1.22577	2196
		Moderate	Total	2.7982	1.23230	10958
		Severe	Normally Dist	2.7497	1.07681	2197
		Severe	Moderately Dist	2.7188	1.07797	2191
		Severe	Skew/Lepto.	2.7199	1.09125	2199
		Severe	Skew/Extremely Lepto	2.7322	1.11738	2196
		Severe	Platykurtic	2.7157	1.10782	2195
		Severe	Total	2.7273	1.09424	10978
		Total	Normally Dist	2.7761	1.12097	6584
		Total	Moderately Dist	2.7756	1.12318	6574
		Total	Skew/Lepto.	2.7685	1.12781	6584
		Total	Skew/Extremely Lepto	2.7854	1.13346	6579
		Total	Platykurtic	2.7711	1.12835	6583
		Total	Total	2.7753	1.12671	32904
	1000/300	Normal	Normally Dist	2.8225	1.01382	2192
		Normal	Moderately Dist	2.8342	1.02044	2196
		Normal	Skew/Lepto.	2.8400	1.02652	2194
		Normal	Skew/Extremely Lepto	2.8307	1.01649	2197
		Normal	Platykurtic	2.8298	1.01597	2192
		Normal	Total	2.8315	1.01849	10971
		Moderate	Normally Dist	2.7605	1.16752	2196
		Moderate	Moderately Dist	2.7656	1.18742	2197
		Moderate	Skew/Lepto.	2.7310	1.18814	2197

Table M1 continued

	sample	ability	pop	Mean	Std. Deviation	N
			Skew/Extremely Lepto	2.7639	1.18257	2198
			Platykurtic	2.7609	1.18984	2196
			Total	2.7564	1.18298	10984
		Severe	Normally Dist	2.7239	1.11358	2195
			Moderately Dist	2.7530	1.11938	2198
			Skew/Lepto.	2.7425	1.12621	2194
			Skew/Extremely Lepto	2.7523	1.09653	2196
			Platykurtic	2.7436	1.09475	2196
			Total	2.7431	1.11001	10979
		Total	Normally Dist	2.7690	1.10078	6583
			Moderately Dist	2.7843	1.11162	6591
			Skew/Lepto.	2.7711	1.11655	6585
			Skew/Extremely Lepto	2.7823	1.10101	6591
			Platykurtic	2.7781	1.10299	6584
			Total	2.7769	1.10656	32934
Total		Normal	Normally Dist	2.7927	1.04980	8769
			Moderately Dist	2.8023	1.05558	8766
			Skew/Lepto.	2.8070	1.05867	8770
			Skew/Extremely Lepto	2.8018	1.05439	8775
			Platykurtic	2.8051	1.05021	8764
			Total	2.8018	1.05370	43844
		Moderate	Normally Dist	2.7547	1.20681	8772
			Moderately Dist	2.7571	1.20895	8772
			Skew/Lepto.	2.7957	1.19865	8772
			Skew/Extremely Lepto	2.7924	1.19839	8771
			Platykurtic	2.7746	1.20877	8783
			Total	2.7749	1.20439	43870
		Severe	Normally Dist	2.7261	1.12507	8780
			Moderately Dist	2.7352	1.12766	8774
			Skew/Lepto.	2.7292	1.13060	8778
			Skew/Extremely Lepto	2.7286	1.13332	8773
			Platykurtic	2.7196	1.13186	8779
			Total	2.7277	1.12967	43884
		Total	Normally Dist	2.7578	1.12934	26321
			Moderately Dist	2.7649	1.13278	26312
			Skew/Lepto.	2.7773	1.13124	26320
			Skew/Extremely Lepto	2.7743	1.13065	26319
			Platykurtic	2.7664	1.13270	26326
			Total	2.7681	1.13135	131598

APPENDIX N

Figure N1 Flow Chart of the Methods Section

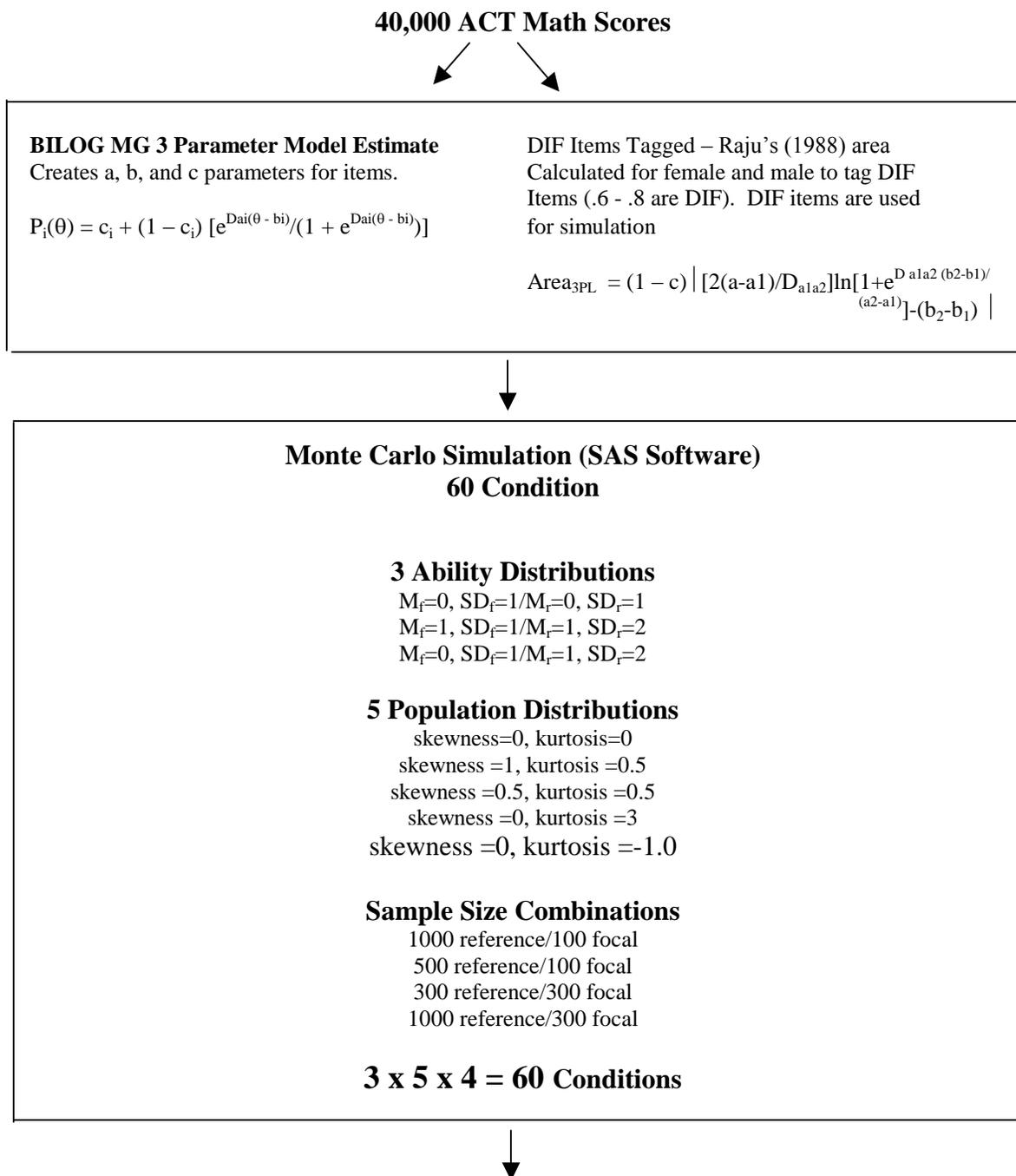
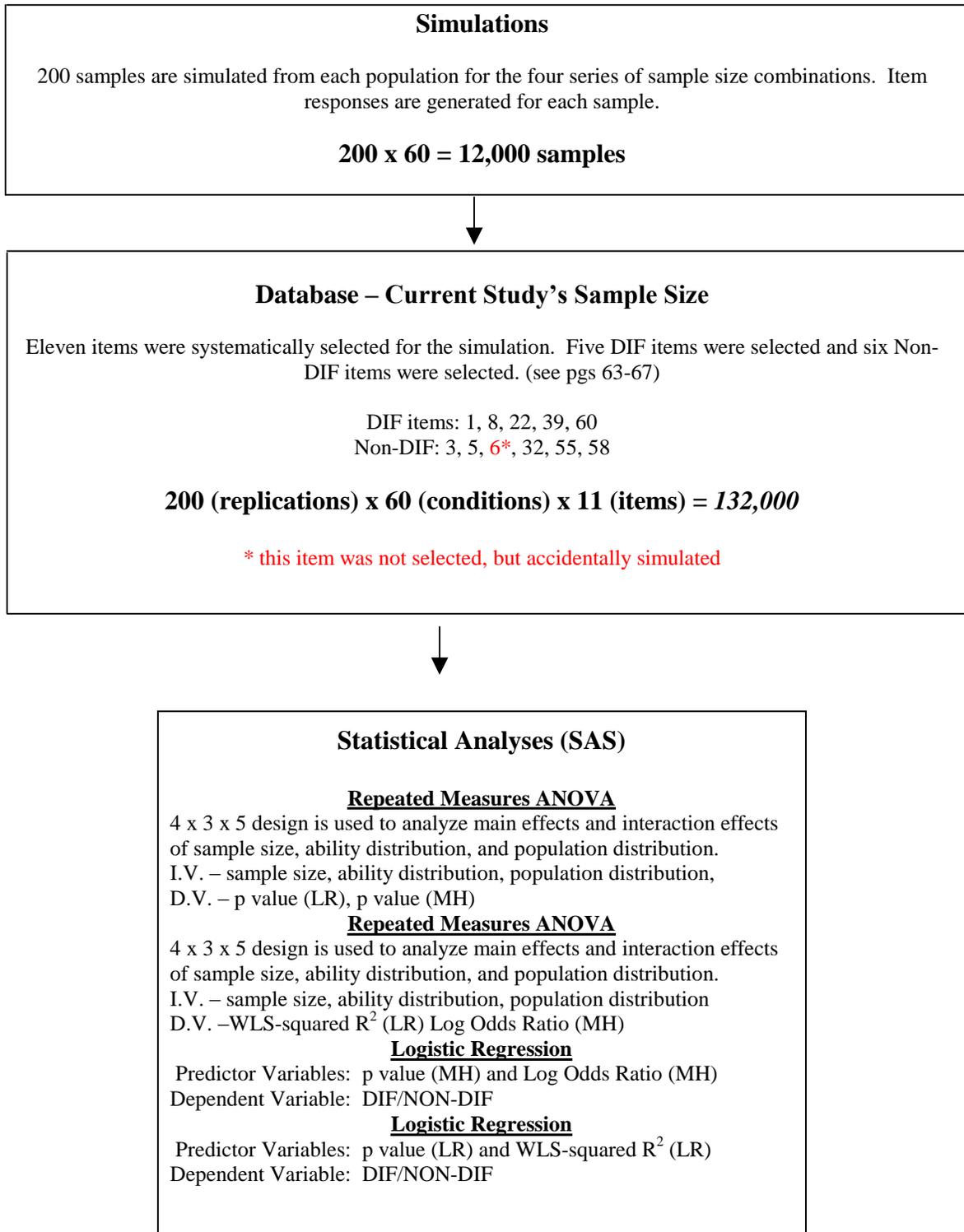


Figure N1 continued



APPENDIX O

Syntax for Simulation

```

proc printto log='c:\logfile.tmp';
%macro item;
proc datasets; delete irt1 irtm irtf r1m r1f irtt
    chi chi1 r r1 chimh chimh1 oddr oddr1
    tot1 tot2 tot3 tot4 tot5;
%DO DIS=1 %TO 3; /*DO LOOP FOR ABILITY DISTRIBUTION CONDITION*/
%IF &DIS=1 %THEN %DO; %LET MEANM=0.0;%LET STDm=1.0;%LET
MEANF=0.0;%LET STDF=1.0;%END;
%IF &DIS=2 %THEN %DO; %LET MEANM=1.0;%LET STDm=1.0;%LET
MEANF=1.0;%LET STDF=2.0;%END;
%IF &DIS=3 %THEN %DO; %LET MEANM=0.0;%LET STDm=1.0;%LET
MEANF=1.0;%LET STDF=2.0;%END;
%DO POP=1 %TO 5;
%IF &POP=1 %THEN %DO; %LET SKEW=0.0;%LET KURT=0.0;%LET A=0;%LET
B=1;%LET C=0;%LET D=0;%END;
%IF &POP=2 %THEN %DO; %LET SKEW=-1.0;%LET KURT=0.5;
%LET A=0.25852489125964;%LET B=1.11465523356736;
%LET C=-0.25852489125964;%LET D=-0.06601339414569;
%END;
%IF &POP=3 %THEN %DO; %LET SKEW=0.0;%LET KURT=3.0;
%LET A=0.0; %LET B=0.78235622045349;
%LET C=0.0; %LET D=0.06790455640586;
%END;
%IF &POP=4 %THEN %DO; %LET SKEW=0.0;%LET KURT=-1.0;
%LET A=0.0; %LET B=1.22100956933052;
%LET C=0.0; %LET D=-0.08015837236135;
%END;
%IF &POP=5 %THEN %DO; %LET SKEW=0.5;%LET KURT=0.5;
%LET A=-0.08045036185716; %LET B=0.97343106918044;
%LET C=0.08045036185716; %LET D=0.00664738328997;
%END;
%DO SMPLN=1 %TO 5; /*DO LOOP FOR SAMPLE SIZE CONDITIONS*/
%IF &SMPLN=1 %THEN %DO; %LET SMPLNM=1000; %LET
SMPLNF=100;%END;
%IF &SMPLN=2 %THEN %DO; %LET SMPLNM=500; %LET
SMPLNF=100;%END;
%IF &SMPLN=3 %THEN %DO; %LET SMPLNM=300; %LET
SMPLNF=300;%END;
%IF &SMPLN=4 %THEN %DO; %LET SMPLNM=1000; %LET
SMPLNF=300;%END;

```

```

%IF &SMPLN=5 %THEN %DO; %LET SMPLNM=1000; %LET
SMPLNF=1000;%END;
%do rep=1 %to 5000;
%do g=1 %to 2;
%if &g=1 %then %do;
data d1; set d1;
%do i=1 %to &smplnm;
proc iml;
start irtscore(theta,rrv,popa,popb,popc,score);
factnorm=probnorm(popb+(popa*theta));
pi=(popc+((1-popc)#factnorm))`;
score=pi>rrv;
finish;
use D1;
read all var{A} into popa;
read all var{b} into popb;
read all var{c} into popc;
nitems=nrow(popa);
theta=rannor(-123);
theta=&A + &B*theta + &C*theta**2 + &D*theta**3;
theta=&meanm + &stdm * theta;
rrv=j(1,nitems,0);
do k=1 to nitems;
rrv[1,k]=ranuni(-245);
end;
run irtscore(theta,rrv,popa,popb,popc,score);
score=score`;
gender=&g;
matrix=gender//theta//score;
matrix=matrix`;
create r1m from matrix[colname={gender ability t1 t2 t3 t4 t5 t6 t7 t8 t9 t10}];
append from matrix;
proc append out=irtm;
%end;
%end;
%if &g=2 %then %do;
data d1; set d2;
%do j=1 %to &smplnf;
proc iml;
start irtscore(theta,rrv,popa,popb,popc,score);
factnorm=probnorm(popb+(popa*theta));
pi=(popc+((1-popc)#factnorm))`;

```

```

score=pi>rrv;
finish;
use D1;
read all var{A} into popa;
read all var{b} into popb;
read all var{c} into popc;
nitems=nrow(popa);
theta=rannor(-123);
theta=&A + &B*theta + &C*theta**2 + &D*theta**3;
theta=&meanf + &stdf * theta;
rrv=j(1,nitems,0);
do k=1 to nitems;
rrv[1,k]=ranuni(-245);
end;
run irtscore(theta,rrv,popa,popb,popc,score);
score=score`;
gender=&g;
matrix=gender//theta//score;
matrix=matrix`;
create r1f from matrix[colname={gender ability t1 t2 t3 t4 t5 t6 t7 t8 t9 t10}];
append from matrix;
proc append out=irtf;
%end;
%end;
data irtt; set irtm irtf;
%end;
proc logistic data=irtt descending;
class gender;
model t1=ability gender /rsquare;
ods trace on;
ods output RSquare=r(keep=cValue2 rename=(cValue2=rsquare)) TypeIII=chi;
data chi1; set chi;
if Variable='GENDER';
run;
data tot1;
merge chi1 r;
data tot2;set tot1;
size=&smpln;
rep=&rep;
mstd=&DIS;
shape=&POP;
run;

```

```
data irt1; set irtt;
total=sum(of t1-t3);
proc freq data=irt1;
  tables total * t1 *gender /cmh noprint;
ods trace on;
ods output CMH=chimh (rename=(Prob=mhp Value=mhchisq))
CommonRelRisks=oddr(rename=(Value=oddratio));
data chimh1; set chimh;
keep mhp mhchisq; if AltHypothesis='Nonzero Correlation';
data oddr1; set oddr;
keep oddratio LowerCL UpperCL; if StudyType='Case-Control';
data tot3; merge chimh1 oddr1;
data tot4; set tot3;
size=&smpln;
rep=&rep;
MSTD=&DIS;
data tot5; merge tot2 tot4;
by size rep mstd;
proc append out=tot6;
run;
%end;
%end;
%end;
%end;
%mend item;
%item;
run;
quit;
```

VITA

Susan Cromwell Duncan
 6025 Hidden Way Lane
 Trussville, AL 35173
 isajaduncan@yahoo.com

- 1990-1994 **Sam Houston State University**, Huntsville, TX.
 Bachelor of Science in Psychology
- 1996-1998 **Abilene Christian University**, Abilene, TX.
 Master of Science in Clinical Psychology
- 2000-2006 **Texas A&M University**, College Station, TX.
 Doctor of Philosophy in Educational Psychology
 Specialization: Research, Measurement, and Statistics

Research Experience

- 1997-1998 *Negative and Positive Feedback as a Factor of Memory Self-Efficacy and Memory Performance*, Thesis for completion of Master of Science in Clinical Psychology.

Presentations and Conferences

- 2001 Cromwell, S. *An Introductory Summary of Various Effect Size Choices*, Southwest Educational Research Association, New Orleans, February
- 2002 Cromwell, S. *A Primer on Ways to explore Item Bias*, Southwest Educational Research Association, Austin, February
- Cromwell, S. *Applying the Bootstrap to Multivariate Analyses through Various Software*, American Educational Research Association, New Orleans, March.
- 2003 Cromwell, S. *Effect Sizes: Reporting Them Didn't Come out of Thin Air*, Southwest Educational Research Association, San Antonio, February
- Cromwell, S., & McNamara, J. *Demystifying Types I, II, and III Sums of Squares*, Southwest Educational Research Association, San Antonio, February
- 2004 Cromwell, S. *Item Response Theory and DIF Analyses: Software Package*, Southwest Educational Research Association, Dallas, February