Georgia State University
ScholarWorks @ Georgia State University

Educational Policy Studies Dissertations | Department of Educational Policy Studies

5-16-2008

Using Three Different Categorical Data Analysis Techniques to Detect Differential Item Functioning

Torie Amelia Stephens-Bonty

Follow this and additional works at: http://scholarworks.gsu.edu/eps_diss

Recommended Citation
Stephens-Bonty, Torie Amelia, "Using Three Different Categorical Data Analysis Techniques to Detect Differential Item Functioning" (2008). *Educational Policy Studies Dissertations*. Paper 24.

This Dissertation is brought to you for free and open access by the Department of Educational Policy Studies at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Educational Policy Studies Dissertations by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.

ACCEPTANCE

This dissertation, USING THREE DIFFERENT CATEGORICAL DATA ANALYSIS TECHNIQUES TO DETECT DIFFERENTIAL ITEM FUNCTIONING, by TORIE AMELIA STEPHENS-BONTY, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree Doctor of Philosophy in the College of Education, Georgia State University.

The Dissertation Advisory Committee and the student's Department Chair, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty. The Dean of the College of Education concurs.

_____
Carolyn F. Furlow, Ph.D.
Committee Chair

_____
T. Chris Oshima, Ph.D.
Committee Member

_____
Phillip Gagne, Ph.D.
Committee Member

_____
Chris Domaleski, Ph.D.
Committee Member

_____
Date

_____
Sheryl A. Gowen, Ph.D.
Chair, Department of Educational Policy Studies

_____
R. W. Kamphause, Ph.D.
Dean and Distinguished Research Professor
College of Education

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education's director of graduate studies and research, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without written permission.

_____

Torie Amelia Stephens-Bonty

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Torie Amelia Stephens-Bonty
2244 Hudson Drive
Lilburn, GA 30047

The director of this dissertation is:

Dr. Carolyn F. Furlow
Department of Educational Policy Studies
College of Education
Georgia State University
Atlanta, GA 30303

VITA

Torie A. Stephens-Bonty

ADDRESS: 2244 Hudson Drive
Lilburn, GA 30047

EDUCATION:

Ph.D.	2008	Georgia State University Educational Policy Studies
M.P.H.	1995	University of California, Los Angeles Public Health
Teaching Credential	1991	University of California, Santa Barbara Science Education
B.A.	1989	University of California, Santa Barbara Chemistry

PROFESSIONAL EXPERIENCE:

2007-Present	Science Teacher Temima High School, Atlanta, GA
2006-2007	Test Developer Atlanta City Public Schools, Atlanta, GA
2005-2006	Education Specialist University of Georgia, Athens, GA
2002-2003	Test Developer Atlanta City Public Schools, Atlanta, GA
2000-2004	Graduate Assistant Georgia State University, Atlanta, GA
1996-1999	Physics teacher Lakeside High School, Atlanta, GA
1994-1996	AIDS Research AIDS Program, Los Angeles, CA
1991-1994	Chemistry and Physics Teacher Venice High School, Venice CA

PRESENTATIONS AND PUBLICATIONS:

Stephens-Bonty, T. A. (2002). *An Evaluation of Georgia's HOPE Scholarship.* Paper presented at the 2002 Georgia Educational Research Association conference, Savannah, GA.

Neel, J., Monaco, M. K., Domaleski, C. S., & Stephens-Bonty, T. (2001, November). *Levenes's Test Revisited*. Paper presented at the annual meeting of the Georgia Educational Research Association conference, Atlanta, GA.

Stephens-Bonty, T. A. (2001). *An Investigation of the Value of Policy in Rural Schools*. Paper presented at the Dogwood Conference, Emory University, GA.

Stephens-Bonty, T. A. (2001). An Investigation of the Value of Policy in Rural Schools. *Emory University Dogwood Journal*, 2001.

ABSTRACT

USING THREE DIFFERENT CATEGORICAL DATA ANALYSIS TECHNIQUES TO DETECT DIFFERENTIAL ITEM FUNCTIONING
by
Torie A. Stephens-Bonty

Diversity in the population along with the diversity of testing usage has resulted in smaller identified groups of test takers. In addition, computer adaptive testing sometimes results in a relatively small number of items being used for a particular assessment. The need and use for statistical techniques that are able to effectively detect differential item functioning (DIF) when the population is small and or the assessment is short is necessary. Identification of empirically biased items is a crucial step in creating equitable and construct-valid assessments.

Parshall and Miller (1995) compared the conventional asymptotic Mantel-Haenszel (MH) with the exact test (ET) for the detection of DIF with small sample sizes. Several studies have since compared the performance of MH to logistic regression (LR) under a variety of conditions. Both Swaminathan and Rogers (1990), and Hildalgo and López-Pina (2004) demonstrated that MH and LR were comparable in their detection of items with DIF. This study followed by comparing the performance of the MH, the ET, and LR performance when both the sample size is small and test length is short.

The purpose of this Monte Carlo simulation study was to expand on the research done by Parshall and Miller (1995) by examining power and power with effect size measures for each of the three DIF detection procedures. The following variables were

manipulated in this study: focal group sample size, percent of items with DIF, and magnitude of DIF. For each condition, a small reference group size of 200 was utilized as well as a short, 10-item test. The results demonstrated that in general, LR was slightly more powerful in detecting items with DIF. In most conditions, however, power was well below the acceptable rate of 80%. As the size of the focal group and the magnitude of DIF increased, the three procedures were more likely to reach acceptable power. Also, all three procedures demonstrated the highest power for the most discriminating item. Collectively, the results from this research provide information in the area of small sample size and DIF detection.

USING THREE DIFFERENT CATEGORICAL DATA ANALYSIS TECHNIQUES TO DETECT DIFFERENTIAL ITEM FUNCTIONING

by
Torie A. Stephens-Bonty

Presented in Partial Fulfillment of Requirements for the
Degree of
Doctor of Philosophy
in
Educational Policy Studies
in
the Department of Educational Policy Studies
in
the College of Education
Georgia State University

Atlanta, Georgia
2008

Copyright by
Torie A. Stephens-Bonty
2008

ACKNOWLEDGMENTS

I dedicate this dissertation to my grandmother, Olga Stephens. Her love of reading and thirst for knowledge has been an inspiration to me.

It is impossible to accomplish a goal such as this without the help of many people. I would like to start by thanking Dr. Carolyn Furlow for her guidance and expertise throughout this process. I am not certain I could have accomplished this without her. She has been a teacher, coach, friend, and cheerleader at times when I needed it most. I would also like to thank the rest of my committee members; Dr. Chris Oshima for her ongoing support; Dr. Phil Gagné for his technical and mathematical expertise; and Dr. Chris Domaleski for his generosity with his time. Dr. John Wiggins, although not a member of my committee, provided me with well needed encouragement.

The support and prayers of my family has been important to me. I am thankful to my father, Wendell, who always believed in me and my mother, Elaine, for her wisdom and strength. To my sister, Sherie, who patiently listened and provided guidance whenever I called; my sister, Katarina, who gave up her spring break to watch my children so I could research and write; and my sister Erika for her prayers and support. Thanks you guys!

My husband, Tyson, has been my greatest source of encouragement. He is both the love of my life and my best friend. Having his hand to hold has made all the difference. Joshua and Jessica are the joy of my life and have kept me grounded throughout this process.

Thanks to all of my friends and family who have helped me in a myriad of ways.

TABLE OF CONTENTS

Page

List of Tables v
List of Equations vii
Abbreviations viii

Chapter
1 INTRODUCTION……1

2 REVIEW OF THE LITERATURE……..4
Methods of Identifying DIF…….6
DIF Detection Methods Based on Categorical Data Analysis……...10
Factors Influencing DIF Detection…….16
Comparison of DIF Detection Methods…….19
Practical Significance…….24
Purpose……...26

3 METHODS…….28
Data Generation……..31
DIF Detection…….33

4 RESULTS……...35
Data Analysis…….35
Item Discrimination and Item Difficulty……37
Comparison Based on 10% of Items Containing DIF……38
Comparison Based on 20% of Items Containing DIF……42
Comparison Based on 30% of Items Containing DIF……47
Practical Significance…….52

5 CONCLUSION……..61
Summary……61
Limitations…….66
Implications……67
Future Research……..68

References…….71

Appendixes……...……76

LIST OF TABLES

Tables

1 2 x 2 Contingency Table……..……..9

2 Tabular Representation of the Frequency Counts in the *j*-th 2 x 2 Contingency Table *j* = 1, …, *J*)……...12

3 Study Design…….……...……30

4 Generated Item Parameters...……...31

5 DIF Magnitude Values…….32

6 Mean Power Rates Across Conditions for Each Method by percent of Items Induced with DIF…….……..……...38

7 10% of the Items with DIF: Item One. Power and Effect Sizes by Focal Group Size and DIF Magnitude (Percent)…….……...……40

8 20% of the Items with DIF: Item One. Power and Effect Sizes by Focal Group Size and DIF Magnitude (Percent)…….…….....……44

9 20% of the Items with DIF: Item Two. Power and Effect Sizes by Focal Group Size and DIF Magnitude (Percent)…….……...…….…46

10 30% of the Items with DIF: Item One. Power and Effect Sizes by Focal Group Size and DIF Magnitude (Percent)…….…..……..……48

11 30% of the Items with DIF: Item Two. Power and Effect Sizes by Focal Group Size and DIF Magnitude (Percent)…….…..……..……49

12 30% of the Items with DIF: Item Three. Power and Effect Sizes by Focal Group Size and DIF Magnitude (Percent)…...…….51

13 10% of Items with DIF: Item One. Percent of Category A, B, and C Item By Focal Group Size and Magnitude of DIF…….…...……..53

14 20% of Items with DIF: Item One. Percent of Category A, B, and C Item By Focal Group Size and Magnitude of DIF…..……...……..54

15 20% of Items with DIF: Item Two. Percent of Category A, B, and C Item By Focal Group Size and Magnitude of DIF……..……...……..55

16 30% of Items with DIF: Item One. Percent of Category A, B, and C Item By Focal Group Size and Magnitude of DIF……..……...……..56

17 30% of Items with DIF: Item Two. Percent of Category A, B, and C Item By Focal Group Size and Magnitude of DIF…..……...……..57

18 30% of Items with DIF: Item Three. Percent of Category A, B, and C Item By Focal Group Size and Magnitude of DIF…..……...……..……..58

19 Percent of an Unattainable MH D-DIF Value per Run per Condition for MH…….……...……60

LIST OF EQUATIONS

Equation

1 Mantel-Haenszel Test Statistic……....12

2 Variance…….……..12

3 Hypergeometric Distribution……...……13

4 Hypergeometric Distribution in Factorial Form……..……...14

5 Logistic Regression Model……..……15

6 Logistic Regression Function……..…15

7 Delta Scale Transformation.....25

8 Logistic Regression Effect Size with Uniform DIF: Step One.....25

9 Logistic Regression Effect Size with Uniform DIF: Step Two.....25

10 Two Parameter IRT Model……..……32

ABBREVIATIONS

2PL	Two Parameter Logistic
AYP	Adequate Yearly Progress
CTT	Classical Test Theory
DIF	Differential Item Functioning
ES	Effect Size
ET	Exact Test
ETS	Educational Testing Service
FA	Factor Analysis
GMH	Generalized Mantel-Haenszel
GROUP	Group Membership
ICC	Item Characteristic Curve
IRT	Item Response Theory
LDFA	Logistic Discriminant Function Analysis
LR	Logistic Regression
MH	Mantel-Haenszel
NAEP	National Assessment of Educational Progress
NCLB	No Child Left Behind
SAS	Statistical Application Software
SES	Socioeconomic Status
SIBTEST	Simultaneous Item Bias Test

SMD	Standard Mean Difference
TOT	Total Score
UCLOLR	Unconstrained Cumulative Logits Ordinal Logistic Regression

CHAPTER 1

INTRODUCTION

Identification of biased items is a crucial step in creating equitable and construct-valid assessments. A biased item is by definition one that contains a systematic error which causes the results of tests to be invalid (Camilli & Shepard, 1994). The term *differential item functioning* (DIF) describes the empirical evidence used to support or refute bias. An item is said to exemplify DIF if individuals with equal ability but different group membership have a different probability of solving an item correctly (Swaminathan & Rogers, 1990).

Statistical techniques able to effectively calculate DIF are needed in the current testing market where diversity in the population, along with the diversity of testing usage, has resulted in varied testing conditions. States are now required to show evidence of DIF consideration in the test development process (Standards and Assessments Peer Review Guidance, 2004). The No Child Left Behind Act (NCLB, 2002) requires schools to demonstrate adequate yearly progress (AYP) for each identified student group (e.g., groups based on ethnicity, socioeconomic status, or disability) each of these groups might only contain a small number of students. Small numbers of participants per test is common in translation and adaptation tests as well. Also, computer adaptive testing sometimes results in a relatively small number of examinees answering a particular question or item. This has resulted in a re-examination of DIF detection methods when the testing population is small or when the focal group (e.g., minority group) is small in

size. As tests and technology change, methods of detecting DIF have had to adjust accordingly. The result is myriad detection methods continually becoming more effective in detecting DIF.

Identifying which DIF detection method (or methods) is most effective is vital in the current testing climate. The relationship between sample size and DIF detection performance should not be ignored. Each detection method has limitations regarding sample size. Identifying the confines of each method is essential. A comparison of methods would provide researchers with data useful in selecting a DIF detection method when there is a small sample size for the focal group.

This study compares the effectiveness of three DIF detection methods in their performance when sample sizes are small and the test length is short. The first method, the Mantel-Haenszel (MH), was chosen because of its widespread usage in the testing industry. The MH is based on an asymptotic approximation of an exact distribution. Because it is an approximation, its performance is expected to weaken as sample size decreases. This is in contrast to the second method, the exact test (ET), which allows users to calculate exact probabilities as opposed to relying on the asymptotic approximation. Because the ET relies on exact probabilities, it is expected to be more powerful in detecting DIF than asymptotic methods with small sample sizes (Agresti, 1996). The third DIF detection method, logistic regression (LR), was included in the study because other studies revealed it to be as effective in DIF detection as the MH. Conversely, small sample sizes may present a problem for LR. Typically, small sample sizes inhibit the estimation of the model parameters used in the regression equation (Agresti, 1996).

The primary purpose of this study was to extend the work done by Parshall and Miller (1995) in which they compared the performance of the exact test (Agresti, 1996) with the conventional asymptotic Mantel-Haenszel (Mantel & Haenszel, 1959) test for the detection of DIF with small sample sizes. Few published studies prior to Parshall and Miller's (1995) compared methods regarding their effectiveness in DIF detection with small sample sizes. In addition to the MH and the exact test, the performance of logistic regression was examined in this study along with measures of practical significance. Additional conditions were also examined including larger DIF magnitudes and smaller focal group sample sizes.

This study demonstrated how each of the three techniques performed in detecting DIF under small sample size conditions. Given the increased task of evaluating the effects of DIF on small sample subgroups, it is important to identify what circumstances restrict each method's effectiveness. With this in mind, comparing the exact test, MH, and logistic regression will identify strengths and weaknesses in DIF detection with small sample sizes. Effect size (ES) measures will assist in identifying the practical significance of the results.

CHAPTER 2

REVIEW OF LITERATURE

The review of the literature begins with an evaluation of relevant terminology and an overview of previous research on the effectiveness of statistical testing in detecting DIF with regard to sample size. The examination then proceeds to a brief discussion of DIF. Next, an overview of DIF detection methods is given. Research on the effectiveness of the MH, the exact test, and logistic regression in detecting DIF when sample sizes are small is then presented.

There are a number of important terms that are used when referencing the literature on DIF detection. This section provides a review of terms used throughout the study including: (1) group membership, (2) test, (3) ability, (4) DIF, and (5) bias.

(1) Group membership. In the field of DIF detection, group membership refers to the label given to a set of examinees. Often the label is assigned based on a certain demographic characteristic. For example, race, ethnicity, socioeconomic status (SES), gender, or native language can be used to define a group. Individuals belonging to the majority group (e.g., English speakers, Caucasians) are typically categorized as the reference group. The focal group is made up of individuals belonging to the identified minority (e.g., non-English speakers, Hispanics). DIF detection is a comparison of the relative performance of the focal group to the reference group.

(2) Test. A test is defined in this study as an instrument which measures an attribute that is not clearly observable. A test can be further explained as a collection of items believed to be a representative sample of the behavior or trait that is being measured. Thorndike (1997) states, "we never measure a thing or a person. We always measure a quality or an attribute of the thing or the person" (p.9). This not only gives us a working definition of test but it also reminds us that what is being measured by the test is used to reveal something about the examinee.

(3) Ability. Tests are typically used to measure the ability (e.g., math or reading) of the test taker. Ability refers to the proficiency of a person in a specific area. Rasch (1993) said, "A person having a greater ability than another should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another one means that for any person the probability of solving the second item correctly is the greater one" (p. 117).

(4) Differential Item Functioning. An item is said to contain DIF if individuals having the same ability but belonging to different groups perform differently on that item. Hambleton, Swaminathan, and Rogers (1991) define DIF as the *empirical* evidence obtained in investigations of bias. The presence of DIF is seen as evidence against test fairness. Oftentimes this occurs when the focal group's performance is poorer than that of the reference group's on an item after the two groups have been matched by ability.

An illustrative example of DIF could be a situation where a student's reading comprehension is being assessed. The girls (focal group) taking the test are,

however, impeded by items referencing boxing. The poor performance of girls observed on the boxing items (comparable to boys) may not be due to their reading comprehension ability but instead due to their lack of knowledge about the sport.

(5) Bias. When a test performs as it was designed, examinees with the same ability have similar if not identical total scores. When this does not occur, it is possible that the test or a specific item (or items) on the test may be biased. To be identified as biased, an item must measure some construct other than the construct it was intended to measure. The first step in detecting bias is to identify if DIF exist for the item in question. Secondly, when there is empirical evidence, DIF, the cause of the difference in performance on the item between the two groups must be identified. When the cause of the DIF is not relevant to the construct being assessed, the item is identified as bias. However, if the cause of DIF is related to the construct being assessed, the item is not bias. The term bias is more specific than the term DIF. An item identified as biased can alter the test's meaning by assessing a construct that differs from the one the test was intended to measure.

Methods of Identifying DIF

Since tests are designed to estimate an individual's true ability despite group membership, DIF detection is an important topic for educational researchers and has thus been extensively studied. The history of the development of methods to detect DIF partly overlaps with the history of item response theory (IRT). It may be insufficient to discuss DIF without mentioning IRT-based methods of detecting DIF. DIF identification methods can be divided into two general categories: (1) IRT-based methods and (2) non-

IRT methods. The first group of methods is comprised of the comparison of item parameters and the area between Item Characteristic Curves (ICC; Hambleton, Swaminathan, & Rogers, 1991). Included in the second group are classical test theory (CTT) methods, factor analysis (FA) methods, and categorical-data-analysis-based methods (Gómez-Benito & Navas-Ara, 2000).

In categorizing the methods to detect DIF, it is important to note that the model chosen to analyze the data does not necessarily dictate the model used to identify DIF; this is the case with IRT models. Large testing companies sometimes use IRT-based models to analyze their data and non-IRT methods for DIF detection. This is not unusual, given the weakness of DIF identification when IRT-based DIF detection models are used with smaller sample sizes (Crane et al., 2004). The type of estimation used with IRT methods requires a large number of examinees (at least a size of 250 in each group) and a large ability range to be effective (Embretson & Riese, 2000). The use of IRT in identifying DIF typically requires that the item parameters are estimated separately for the reference group and the focal group (Hambleton et. al, 1991). A major advantage of non-IRT models over IRT models is that they are typically non-parametric and do not require the assumptions needed with IRT-based methods and thus they are frequently able to detect DIF with smaller sample sizes. As a result, non-IRT methods are often utilized for identifying DIF. A large sample size typically refers to reference and focal group sizes that are approximately at least 250 to 500 individuals in each (Embretson & Riese, 2000). So while IRT-based methods are popular in the area of DIF detection, because the focus of this study is on small sample sizes, their use is not feasible.

Although the choice of DIF detection method is important, treatment and selection of the matching criteria are also important. The matching criterion is the variable used to equate members of both the reference and the focal groups to determine if there is a difference in performance on a particular item. Members of each group are paired by a common trait level. The trait is a proxy for the individual's performance in the area being assessed. Typically this is the individual's ability in the investigated area or total score on the test. Matching individuals on a common trait allows researchers to predict the outcome of one individual based on the performance of another. In essence matching permits DIF analysis to occur by enabling a relative comparison of the focal group to the reference group.

Types of matching include thin and thick matching. Thin matching is the term used to describe association based on the total score. It requires each level of the matching criteria to be assigned a value (or weight) based on the frequency of the contingency table (Donoghue & Allen, 1993). For each of the matched ability levels, there is a contingency table of item responses. With dichotomous item responses, the 2 x 2 contingency table is set up where group membership (focal or reference group) and item response (right or wrong) represent the two variables used to categorize item responses. Table 1 presents a 2 x 2 contingency table for each matched level. When the 2 x 2 table has a frequency of zero in a row or column, the matching criteria is assigned a weight of zero. Because thin matching assigns each ability level a value based on the contingency table, those tables with zero frequencies are not used in the analysis and data are lost.

Thick matching is the result of combining total score levels. The process is not limited by the cell count of each level of the matching criteria. Instead levels are pooled to eliminate 2 x 2 tables with zero frequencies. Donoghue and Allen (1993) identified three methods of thick matching that yield results superior to thin matching including: combining extreme levels until each cell of the 2 x 2 table has a minimum of 1 observation, combining levels until each cell of the 2 x 2 table has at least 1 observation per every level of the matching variable, and combining every two levels of the matching variable. Donoghue and Allen's simulation study concluded that for short tests, 10 items or less, thick matching improved DIF detection when compared to thin matching. They concluded two methods of thick matching were superior to thin matching for short tests. One method required pooling the number of examinees to result in an equal number per matching level. The second was similar; it required pooling the members of the focal group until an equal number existed per matching level. With shorter tests (10 items or less), thin matching tended to result in inflated Type I error rates.

Table 1

2 x 2 Contingency Table

	Item Correct	Item Incorrect
Reference Group	*y*	*X*
Focal Group	*y'*	*x'*

Some researchers recommend a two-step purification process to increase detection rates for DIF items (Gierl et al., 2000). Purification is a process used to eliminate items containing DIF from the matching criteria. Miller and Oshima (1992)

found that detection of DIF items was not improved by purification when the proportion of DIF items was small, 10% or less. As the proportion of DIF items exceeded 10%, however, purification resulted in an increase of power and a reduction in Type I error. These results have been supported by the previous work of Holland and Thayer (1986) and more recently by Fildago, Mellenbergh, and Muñiz (2000). The two-step procedure used by Educational Testing Services (ETS) allows researchers to refine the matching criteria based on the effect size of DIF regardless of statistical significance (Gierl, Jodoin, & Ackerman, 2000). The first step involves including all items in the total score (the matching criterion). Secondly, DIF detection takes place for all items and those items that are identified as containing DIF have their item responses removed from the total score for matching. This final step results in a "pure" matching criterion. DIF detection is then done for each of the items previously flagged as having DIF. The matching criteria used for each test includes all of the non-DIF items plus the one DIF item under investigation (Donoghue, Holland, & Thayer, 1993; Holland & Thayer, 1986).

DIF Detection Methods Based on Categorical Data Analysis

The focus of this study is to examine the performance of three methods (all categorical data analysis based) for DIF detection under small sample size conditions: the Mantel-Haenszel test, Fisher's exact test, and Logistic Regression. An individual description of each of these methods follows.

Mantel-Haenszel. A common non-IRT method used today to identify DIF is the Mantel-Haenszel (MH; Mantel & Haenszel, 1959). Mantel and Haenszel (1959) proposed a simple estimator for the common odds ratio in a series of 2 x 2 tables. The MH compares the odds ratios of the focal and the reference groups. Here, the odds ratio is identified as the ratio of two odds (one from each of the groups), where each of the odds

is a fraction defined as the probability of success over the probability of failure. The common odds ratio is represented by αMH. This value can be transformed to any scale enabling it to be used in a plethora of ways. The MH is asymptotically distributed with degrees of freedom equal to 1 (Agresti, 2002).

The MH procedure requires that the reference and focal groups are matched on total test score using every *J* test score (thin matching) or some thick matching criteria (where total scores are lumped into *J* number of groups). The matching results in *J* separate 2 x 2 tables (Gómez-Benito & Navas-Ara, 2000). In the context of DIF, it is a comparison of performance where the two groups are matched on total score or score categories. MH tests the null hypothesis that the odds of a correct response are the same in both groups. In other words, the odds of a correct response for the focal and reference group when matched on ability and no DIF is present should be the same. The resulting odds ratio would be one and the null hypothesis would not be rejected. When there is a significant difference in the probability of getting an item wrong across the matched groups, then DIF is present and the null hypothesis is rejected (Fischer, 1995). The hypothesis test for the Mantel-Haenszel has a chi-square distribution with $df = 1$. Once the odds of a correct response are calculated for both groups, the values are compared.

Table 2

Tabular Representation of the Frequency Counts in the j-th 2 x 2 Contingency Table j = 1, ..., J)

	Item Correct	Item Incorrect	Row Total
Reference Group	y_j	x_j	m_j
Focal Group	y_j'	x_j'	m_j'
Column Total	n_j	n_j'	N_j

The MH test statistic and variance have the following form:

$$X^2_{\mathrm{MH}} = \frac{\left[\sum_{j=1}^{J} \frac{\left(y_j x_j' - y_j' x_j\right)}{N_j}\right]^2}{\sum_{j=1}^{J} Var} \tag{1}$$

$$Var = \sum_{j=1}^{J} \frac{m_j m_j' n_j n_j'}{\left(N_j - 1\right) N_j^2} \tag{2}$$

where y_j, x_j, y_j', and x_j' are the frequency counts of the (1, 1), (1, 2), (2,1), and (2, 2) elements, respectively, in each of the j-th 2 x 2 contingency table and N_j is the total count of all cells for each J category (see Table 2). The variance, *Var,* is the product of row and column totals, divided by the product of the total count of all cells squared and the total cell count minus one, for each J category (see Table 2). Equation 1 provides a

test for association between the two binary variables of interest. A sample calculation using X^2_{MH} is demonstrated in the Appendix.

Exact Test. Fisher developed the exact test (ET) in 1934 (Agresti, 1996) to test the hypothesis of the conditional independence between the reference and focal group using the 2 x 2 x *J* contingency table (see Table 2). The tested hypothesis states that the odds ratio will equal one when there is no difference between the two groups. Because its probabilities are not based on approximating values, the results are believed to be superior to tests that use approximations like MH (Agresti, 1996). When sample sizes are small (100 or less in each group), the accuracy of approximations decreases, giving the exact test a possible advantage (Hambleton et. al., 1993).

Fisher's exact test probabilities are based on a hypergeometric distribution. A hypergeometric distribution, is one in which the number of successes in a sequence of selections from a finite population without replacement can be described by a discrete probability distribution. In a hypergeometric distribution the cell count probabilities of all cells are determined by the count of the y_j cell, element (1, 1) (see Table 2). The probability of y_j is defined by the following equation (Agresti, 1996),

$$P\left(y_j\right) = \frac{\binom{m_j}{y_j}\binom{m_j'}{n_j - y_j}}{\binom{N_j}{n_j}} \tag{3}$$

$$P(y_j) = \frac{\left[\dfrac{m_j!}{y_j!(m_j - y_j)!}\right]\left[\dfrac{m_j^{'}!}{(n_j - y_j)!(m_j^{'} - n_j - y_j)!}\right]}{\dfrac{N_j!}{n_j!(N_j - n_j)!}} \tag{4}$$

where y_j corresponds to the probability of a correct response by the focal group, m_j is the number of correct responses for both groups (row 1 total), $m_j^{'}$ is the number of incorrect responses by both groups (row 2 total), n_j represents the total items taken by the focal group (column 1 total), and N_j is the total number of items (see Table 1). Equation 3 presents the probability of a specific value, y_j, when the odds ratio is equal to one. It may be rewritten as a factorial (Equation 4). A sample calculation using the hypergeometric distribution is demonstrated in the Appendix.

An item is identified as having DIF when the odds ratio calculated by the hypergeometric distribution (Equation 3) exceeds one. When there is no DIF, the odds ratio is equal to one. This means that the probability of a correct response is independent of group membership. It then follows that evidence against the null hypothesis of independence is strengthened as the probability of the odds ratio strays from one. When the experimental value of the odds ratio and one are significantly different, the null hypothesis is rejected.

Logistic Regression. A third method for DIF detection is logistic regression. One advantage of the logistic regression model is that it can be used in identifying uniform and non-uniform DIF (Swaminathan & Rogers, 1990). Zumbo (1999) demonstrated how to classify a DIF item as uniform or non-uniform. If the probability of a correct response

is the same across all ability levels and the item has been flagged as having DIF, it is classified as having uniform DIF. If, however, there is an interaction between group membership and ability level, then it is called non-uniform DIF.

The following equation is used when testing simultaneously for uniform and non-uniform DIF for an item:

$$Y = b_0 + b_1 TOT + b_2 GROUP + b_3 TOT * GROUP \tag{5}$$

where *Y* is the function of the linear combination of the predictor variables, *TOT* represents the total score or created total score using thick matching for each individual, *GROUP* refers to group membership (reference or focal), and *TOT*GROUP* is the interaction between group and total score. *Y* can also be described as the dependent variable which is equal to the natural log of the probability of a correct response, *p*, divided by the probability of an incorrect response, 1 - *p* (Equation 6). DIF detection with

$$Y = \ln\left[\frac{p}{(1-p)}\right] \tag{6}$$

regard to LR is the result of evaluating the contribution of each model term (*TOT*, *GROUP*, and *TOT*GROUP*) successively to test for improvement of fit. Uniform DIF is present when *GROUP* is statistically significant and *TOT*GROUP* is not and non-uniform DIF is present when the interaction term is statistically significant regardless of the significance of the *GROUP* term (Hidalgo & Lopez-Pina, 2004). This process allows for the identification of uniform and non-uniform DIF simultaneously.

Factors Influencing DIF Detection

Item discrimination. Differences in an item's ability to discriminate have been shown to impact the power to detect DIF. Chang, Mazzeo, and Roussos (1996) conducted a simulation study to examine the impact of variation in the discriminating parameter when comparing the performance of MH, SIBTEST (Simultaneous Item Bias Test), and SMD (standard mean difference) in detecting items embedded with DIF using the 3 parameter logistic model. The ability of the focal group was sampled from N(-1, 1) while that of the reference group was sampled from N(0, 1). The discrimination parameter value ranged from .15 to 2.0 and included 11 different values. Both the reference and focal group included three different sample size values (500, 1000, and 2000). This study found that as sample size increased, power increased for all three procedures. For items that were more discriminating, the increase was more pronounced. Specifically, power increased from .131 to 1.00 for MH, from .138 to 1.00 for the SIBTEST, and from .136 to .985 for SMD as the item discriminating parameter increased. Their simulation presented a relationship between DIF detection method, discriminating parameter, and sample size.

Kristjansson, Aylesworth, McDowell, and Zumbo (2005) used a simulation study to compare four methods in their DIF identification rates for items with polytomous responses. Their study supports the findings of Chang, Mazzeo, and Roussos (1996). The four methods used in the investigation were the Mantel, generalized Mantel-Haenszel (GMH), logistic discriminant function analysis (LDFA), and unconstrained cumulative logits ordinal logistic regression (UCLOLR). A primary goal of the study was to identify the influence the item discrimination parameter had on power and Type I error rates when evaluating uniform DIF. Three items were classified by their corresponding item discrimination value using the following criteria: the item with a value of 0.8 was

categorized as low, the item with a value equal to 1.2 was categorized as moderate, and the item with a discrimination value of 1.6 was labeled high. Their results identified a relationship between power and item discrimination when evaluating uniform DIF. As item discrimination increased from low to high, power increased as well. The Mantel's power went from 98.3% to 100%, GMH increased from 95.1% to 99.9%, LDFA increased from 97.0% to 100%, and UCLOLR went from 90.0% to 99.9%.

Sample size, DIF magnitude, and item difficulty. Using a simulation study, Mazor, Clauser, and Hambleton (1992) identified three characteristics which were associated with MH's power to detect DIF: (1) sample size, (2) DIF magnitude, and (3) item difficulty. They used a three parameter logistic model to generate a pool of three datasets of 2000 examinees per replication. The first two datasets represented the reference and the focal group (focal group 1). Both groups had a mean ability distribution of zero. The third set, representing focal group 2, had a mean ability distribution of -1.0. A test with 75 items was generated with 16 items containing DIF was used in the simulation. Four datasets were randomly generated from the sample of 2000. This resulted in the following five sample sizes: 100, 200, 500, 1000, and 2000. Four levels of DIF were added to the focal group, 0.25, 0.50, 1.00, and 1.50, making the item more difficult for the focal group. As the sample size decreased, MH's ability to correctly identify DIF items also decreased. When the focal group size was 2000, focal group 1 had a mean detection rate of 74% across all conditions while focal group 2 had a mean DIF detection rate of 64%. DIF detection rates decreased when sample size decreased. Also, MH had higher power when the ability of the two groups did not differ. This was consistent across all sample sizes and conditions except one. MH was more likely to correctly identify DIF items

when there was an ability difference between the two groups and the item was classified as easy.

The impact of sample size and item characteristics on MH's ability to correctly identify DIF items was further studied by Roussos and Stouts (1996). Their study compared the MH procedure to SIBTEST under small and large sample size conditions involving two simulation studies. The first study examined conditions with small to moderate sample sizes. The focal and reference groups had equal sizes of 100, 200, 500, and 1000. The mean differences in ability between the reference and focal groups were, 0.0, 0.5, and 1.0. When the size of each group was 1000 and the ability difference between them was 1.0, the mean Type I error rate for MH was 6% and decreased to 2.3% when there was no difference in ability and sample size was 100. The results of SIBTEST were similar under the same conditions.

The second study (an examination of Type I error) increased the number of non-DIF items and focused on moderate to large sample sizes: 500, 1000, and 2000 per group. Focal and reference group sizes were once again set equal to each other. The focus of this simulation was to identify the rate at which each procedure falsely characterized items as having moderate to high DIF. The DIF items were pre-identified as having moderate or high DIF. When the difference in ability between the two groups was 1.0 and sample size was 500, MH falsely categorized items as moderate or high at a rate of 9% compared to 7% when sample size increased to 3000. Sample size affected SIBTEST's power as well. These findings support the research comparing MH and SIBTEST in simulation research (Shealy & Stout, 1993; Holland & Thayer, 1988). It is insufficient to explore the impact

of small sample size on the effectiveness of DIF detection methods without recognizing the impact of the item parameters.

Comparison of DIF Detection Methods

Mantel-Haenszel versus the exact test. Until fairly recently, the exact test was rarely used in applied statistical research because it requires a large amount of computation. With recent advances in software capability, the exact test has become feasible for researchers to utilize. Its use, however, is still relatively uncommon among researchers in detecting DIF. To date, only two studies have compared the performance of the exact test with MH. Parshall and Miller (1995) examined the relative performance of the exact test against the MH test using dichotomous items in a simulation study, while Meyer, Huynh, and Seaman (2004) compared the performance of both methods when using polytomous items with a real dataset. No published research has yet compared the exact test and logistic regression.

Parshall and Miller's (1995) simulation was comprised of three studies. The first study consisted of a 25 item test with parameters generated according to the three-parameter IRT model, with only a single DIF item. The *a* and *c* parameters were generated from the log-normal distribution (0, 0.5) and a beta distribution, respectively. The *b* parameters were generated from the normal $N(0, 0.75)$. Three levels of DIF magnitude were added to the generating parameters for the focal group: 0.25, 0.50, and 0.75; thus making the item more difficult for the focal group. A sample size of 500 was used for the reference group and sample sizes of 25, 50, 100, and 200 were used for the focal group. Both groups employed a normal ability distribution. No substantial differences in power were discovered between the ET and MH in the detection of DIF.

DIF items were detected 88% of the time by the Exact Test and 90% of the time when MH was utilized (using alpha of 0.05) when the sample size was 100 or greater and the DIF magnitude was 0.75. When the alpha level was more conservative, 0.01, a sample size of 200 was required to have a rejection rate of 87% for both methods. The exact test tended to be slightly more conservative than the MH test with the Type I error rates, while the MH was closer to the nominal alpha level.

In the second simulation study from Parshall and Miller (1995), the focal and reference groups' abilities were generated from different distributions. The focal group's ability distribution was one standard deviation below the reference group's ability distribution, thus creating a difference in ability also known as impact. Beyond this change, the conditions were consistent with those of the first study. The presence of impact negatively affected power for both methods. A focal group size of 200 and DIF magnitude of 0.75 was required to produce a rejection rate of only 29% for the ET and 36% for MH when alpha was 0.05. The rejection rate decreased as the alpha level became more conservative, 0.01, under the same conditions listed. The exact test's rate decreased to 15% while MH's was 16%. Again, the differences in DIF detection between the exact test and MH procedures were not substantial, and when differences emerged they tended to slightly favor the MH method.

The third study used data generated from parameters derived from an administration of the 40-item ACT Assessment Mathematics test. The parameters were estimated separately for a sample of 2,000 White examinees and a sample of 2,000 African American examinees. The estimates were then placed on the same scale using a set of linear transformations. The scaled estimates served as the item parameters from

which the data were generated. As in the first study, both the reference and focal groups were generated under a normal distribution. The sample sizes were the same as those from the first study. However, in the third study, no DIF was added. Using alpha levels of both 0.01 and 0.05, the Type I error rates for both methods were found to be similar, with a slight conservative tendency for the exact method. Again, the main finding was that the exact test offered no particular advantage over the MH test when the focal group sample size was small.

Meyer et al. (2004) contributed to the comparison of the performance of MH with the exact test in an applied scenario. They investigated two areas not addressed by Parshall and Miller: the use of polytomous items and the use of effect sizes with real data. Their study included 375 participants. Typically with DIF detection, females are considered the focal group, but in this study they were in the majority (n= 299). So the researchers decided to make the men the focal group (n = 76). The researchers employed Donoghue and Allen's (1993) thick matching procedure. The survey was made up of 30 Likert scale items, of which 10% of the items were classified as containing DIF. Their study findings regarding statistical significance were similar to that of Parshall and Miller (1995). Both found the MH to detect DIF slightly more frequently than the exact test in identifying the target item. However, effect size calculations resulted in a similar number of items classified as containing large DIF based on NAEP classifications.

MH versus logistic regression. Hildalgo and López-Pina (2004) compared MH and logistic regression for the identification of DIF. In this simulation study, both the focal and reference groups had a sample size of 1000. The study used the two-parameter logistic (2PL) as the generating model. A total of 25 conditions were run using a 75 item

test. In the test, 16 items contained DIF. Each DIF item had a unique magnitude (resulting in a total of 16 x 25 = 400 different DIF items). The 400 items containing DIF were created using the following factors: five levels of the *b* parameter (-1.5, -1.0, 0, 1.0, and 1.5), four levels of the *a* parameter (0.25, 0.60, 0.90, and1.25), four levels of change in the *b* parameter (0, 0.30, 0.60, and 1.00), and five levels of change in the *a* parameter (0, 0.25, 0.50, 0.75, and 1.00). In each condition, MH and LR were used to detect DIF. Across all conditions with uniform DIF, MH correctly identified DIF items at a rate of 55%, while logistic regression had a detection rate of 53.33%. In the non-uniform DIF conditions, the strength of each procedure's findings was reversed. Logistic regression's non-uniform DIF detection rate of 68.75% across conditions was more powerful than MH's rate of 61.25%. For both procedures the detection rate increased as the items became more difficult and more discriminating. This was exemplified by large differences in the *a* and *b* parameters. When the change in the *a* parameters was 0.5 or greater and the change in the *b* parameters was 0.6 or greater, the overall detection rate was 99%. This is an increase from 35% for the remaining conditions.

Swaminathan and Rogers (1990) also compared MH with LR in the detection of DIF. In their simulation study the following factors were manipulated: sample size per group (250 and 500), test length (40, 60, and 80 items), type of DIF (uniform or non-uniform), and the magnitude of DIF for both uniform and non-uniform conditions (0.6 and 0.8). In each of the conditions, 20% of the items were identified as having DIF. There were equal numbers of uniform and non-uniform DIF items. When the sample size was 500, both methods effectively identified uniform DIF with 100% accuracy. This rate dropped to 75% for both procedures when the sample size decreased to 250 per group.

Non-uniform DIF detection was influenced by both sample size and test length. Logistic regression had a DIF detection rate of 50% when the test had 40 items regardless of sample size. The non-uniform DIF detection rate increased to 75% when the test length doubled and the per group sample size was 500. LR's performance improved as the sample size of the group increased and test length doubled. MH had a detection rate of 0% for non-uniform DIF across all conditions. Demonstrating its' weakness in detecting non-uniform DIF. While the two methods performed similarly with uniform DIF, LR had the distinct advantage of also being able to detect non-uniform DIF.

Summary of DIF detection methods. Each of the three methods discussed above possess strengths and weakness which encourage a comparative study of their effectiveness to detect DIF items when sample sizes are small, particularly with a small focal group size. MH is the most commonly used of the three (Meyer, Huynh, & Seaman, 2004) and has been shown to be effective in detecting uniform DIF (Swaminathan & Rogers, 1990). Because it is an asymptotic approach, however, small sample sizes may pose a problem. As the number of examinees in a particular score group decreases, the likelihood of an empty cell in a 2 x 2 table increases. Effective calculation of the odds ratio is hindered by having cell counts of 0. LR boasts several advantages over MH including the ability to include a variety of independent variables into the model equation to predict an examinee's performance (Kelderman & Macready, 1990) as well as the ability to detect non-uniform DIF. Its performance has not yet been evaluated for DIF detection with small sample sizes.

The exact test differs from the previous two in that it does not rely on approximations. As a result, it is not inhibited by small sample sizes to derive an accurate

calculation. Small sample sizes of the focal and reference group and short test length are not expected to negatively impact its power (Agresti, 1996). However, its strength, non-reliance on approximations, has in the past been its greatest weakness. ET uses software that is not as accessible as those for MH and LR. The exact test has been used in the past when the sample sizes were too small for the approximations used by other procedures (Agresti, 1996).

Practical Significance

Measures of effect size are central in identifying the efficiency of a technique in detecting DIF that is of practical significance. When sample sizes are small, detecting an effect using statistical significance is typically more difficult. Practical significance supplies information that enhances statistical significance values by providing a measure of meaningfulness. Effect size measurements yield categories based on the strength of the DIF detected in an item. Items flagged as having DIF are typically divided into three categories: negligible, moderate, and large DIF. Effect size measures used for dichotomous data are based on a logarithmic transformation of the odds ratio (Holland & Thayer, 1988). This is true for those procedures producing an odds ratio value. The result is a value that has a symmetrical scale: 0 indicates the absence of DIF, while a negative value signifies that the item favors the focal group, and a positive value indicates favoritism towards the reference group. The magnitude of the transformed odds ratio yields the strength of the DIF item, its practical significance. This is the criterion preferred by ETS and used by National Assessment of Educational Progress (NAEP) to categorize DIF items (Meyer et al., 2004).

The MH common odds ratio, αMH, may be transformed to a delta scale, MH D-DIF (Dorans & Holland, 1993). This process utilizes the following equation:

$$\text{MH D - DIF} = -2.35 \ln [\alpha_{MH}] \tag{7}$$

An item is categorized as favoring the focal group when the MH D - DIF value is positive and favoring the reference group when the value is negative. Using statistical significance and the magnitude of the MH D - DIF value, the Educational Testing Service classifies DIF into three categories: type A, negligible DIF, |MH D - DIF| < 1; type C, large DIF, |MH D - DIF| ≥ 1.5; type B, intermediate DIF, 1 ≤ |MH D - DIF| > 1.5. In order to be classified into categories B or C, the item must be statistically significant as well.

A similar procedure can be employed in assessing the magnitude of uniform DIF with logistic regression. Calculation of the effect size with uniform DIF is a two step process. Zumbo (1999) outlines this method utilizing R^2. The proportion of explained variation, R^2, is compared between two models. R_1^2 is initially calculated by entering total score, b_1TOT, (Equation 8). R_2^2 is then calculated with the addition of group membership, b_2GROUP, (Equation 9). The difference between the two R^2 values, ΔR^2, is the variance explained by group membership.

$$Y = b_o + b_1 TOT \tag{8}$$

$$Y = b_o + b_1TOT + b_2GROUP \tag{9}$$

The following guidelines were used by Zumbo (1999) to categorize DIF items: type A, negligible DIF, $\Delta R^2 < .13$; type C, large DIF, $\Delta R^2 > .26$; type B, intermediate DIF, $.13 \leq \Delta R^2 \leq .26$. Categories B and C require statistical significance as well.

Purpose

Parshall and Miller's (1995) study was significant as the first simulation comparing the performance of MH and the exact test in DIF detection with small focal group sample sizes. The purpose of this simulation study was to expand on Parshall and Miller's (1995) research in order to augment knowledge about strengths and limitations in DIF identification under small sample size conditions. Several of these areas of expansion include additional focal group sample sizes, smaller reference group sample size, increasing the magnitude and number of DIF items, including a comparison of MH and the ET with logistic regression, and using measures of practical significance along with statistical significance to detect DIF.

In Parshall and Miller's (1995) study, a reference group sample size of 500 was used with focal group sizes in the amounts of 25, 50, 100, and 200. As discussed previously, there are a number of scenarios where small sample sizes in the focal and reference group may occur in testing. It is important to understand how DIF detection methods work under very small sample sizes and where methods' identification of DIF breaks down. Thus in this study a variety of additional small sample sizes will be included for the focal group. In addition to examining other sample sizes, the interaction between small sample size and the amount of DIF may be significant. In Parshall and Miller's (1995) study only one DIF item was included whereas in actual testing scenarios it is likely that more DIF items would be present. Therefore this study included additional DIF conditions incorporating different magnitudes of DIF. In addition, previous research has found that logistic regression performed as effectively as MH in detecting DIF. This examination of its performance when the sample size is small was compared in this

study. Measures of practical significance in this area have also not been examined in simulation research. Therefore to determine the sensitivity of each procedure to small sample sizes, measures of practical significance along with statistical significance were examined.

CHAPTER 3

METHOD

A Monte Carlo simulation study was conducted to assess the comparative performance of three DIF detection methods: MH, the exact test, and logistic regression, under small sample size conditions using a 10-item test. The results of short tests are often used in decision making, despite possible validity concerns (Emons, Sijtsma, & Meijer, 2007). DIF detection rates were examined based on statistical significance alone as well as the use of statistical significance and a measure of practical significance. Several conditions were varied, including the focal group sample size, the DIF magnitude, and the percent of items with DIF. For each condition, the DIF detection rates of the MH, ET, and LR were compared.

Focal group sample size. The first design factor varied in this study was the focal group sample size. Embretson and Riese (2000) described small focal group size as having less than 250 individuals in the group. Because this study focused on the effectiveness of DIF detection with small sample sizes, the reference group size was set at 200. The focal group sizes consisted of 5, 10, 20, 40, 60, 80, and 100 examinees. The initial focal group size was large enough to ensure that all three methods could be calculated and small enough to be categorized as a small sample size (Parshall & Miller, 1995).

DIF magnitude. The second factor manipulated was the magnitude of the DIF. Parshall and Miller incorporated three levels of DIF magnitude in their simulation study:

0.25, 0.50, and 0.75. Both Hidalgo and López-Pina (2004) and Fidalgo, Ferreres, and Muñiz (2004) included 1.0 as the largest magnitude of DIF in their examination of small sample sizes. As a result, four levels of DIF were employed in this study: 0.25, 0.50, 0.75, and 1.00, to represent small-to-large DIF magnitudes.

Percentage of items with DIF. The third factor varied was the percent of items containing DIF. Three levels of DIF amount were used in this study: 10%, 20%, and 30%. Augmenting the number of DIF items allowed for an investigation of the relationship between small sample size and concentration of DIF items. The conditions of this study design are listed in Table 3.

Table 3

Study Design

DIF Detection Methods Compared
1. Mantel-Haenszel
2. Exact Test
3. Logistic Regression

Focal Group Sample Size
1. 5
2. 10
3. 20
4. 40
5. 60
6. 80
7. 100

DIF Magnitude
1. 0.25
2. 0.50
3. 0.75
4. 1.00

Percentage of Items with DIF
1. 10%
2. 20%
3. 30%

By expanding on Parshall and Miller's (1995) work, the following factors were examined in this study: seven levels of focal group sample size (5, 10, 20, 40, 60, 80, and 100), four levels of DIF magnitude (0.25, 0.50, 0.75, and 1.00), and three levels of percentage of DIF items included (10%, 20%, and 30%). These levels were fully crossed resulting in 84 conditions. The 2PL IRT model was used in this study incorporating both item discrimination (*a*) and item difficulty (*b*) parameters.

Data Generation

The S-plus code by Douglas (personal communication, 2003) was adapted to generate *a* and *b* parameters for each of the 10 items on the simulated test. The distributions used by Parshall and Miller (1995) to generate item parameters were replicated in this study. The *a* parameters were generated by using a log-normal distribution (0, 0.5). A normal distribution (0, 0.75) was used to generate the *b* parameters. Four levels of uniform DIF (0.25, 0.50, 0.75, and 1.00) were added to the focal group's *b* parameter for each designated DIF item. The remaining items in the two groups had the same *b* parameter values. In conditions where 10% of the items contained DIF, DIF was added to the first item in the test. With 20% of the items containing DIF, DIF was added to the first two items and to the first three items when 30% of the items contained DIF. Table 4 includes the generated *a* and *b* parameters for all ten items. Table 5 lists the changes in the *b* parameter for the three items induced with DIF.

Table 4

Generated Item Parameters

Item	*a* Parameter	*b* Parameter
1	0.6115882	0.6520854
2	0.8894247	0.2586809
3	1.8544249	-0.2287187
4	0.3323106	0.1594453
5	0.5104524	0.2929465
6	1.7806141	0.4019744
7	0.3680925	-0.2654112
8	1.4723857	0.7996554
9	0.5198635	-0.3064966
10	0.3080919	-0.8650780

Table 5

DIF Magnitude Values

	30% of DIF Items		
	20% of DIF Items		
DIF Magnitude	10% of DIF Items		
	Item 1	Item 2	Item 3
.25	.9020854	.5086809	.0212813
.50	1.1520854	.7586809	.2712813
.75	1.4020854	1.0086809	.5212813
1.00	1.6520854	1.2586809	.7712813

IRTGEN (Whittaker, Fitzpatrick, Dodd, & Williams, 2003), a SAS/IML program, was used to generate the theta values using a normal distribution (0,1) and item responses for each simulee in the focal and reference groups based on the previously generated parameters. This process was first accomplished by randomly assigning each examinee a known theta value (θ) from a normal distribution. Secondly, the item parameters and theta values were combined to create a response score based on the 2PL IRT model using the following equation

$$P_i(\theta) = \frac{e^{Da_i(\Theta - b_i)}}{1 + e^{Da_i(\Theta - b_i)}} \qquad i = 1, 2, \ldots, n \tag{10}$$

where a_i is the discrimination parameter and b_i is the difficulty parameter for item i. D represents a scaling constant which is equal to 1.7. This process was repeated for 200 replications in each condition.

DIF Detection

StatXact (Mehta & Patel, 2001) is a statistical software program distributed by Cytel that permits users to formulate exact inferences based on algorithms as opposed to large scale assumptions. StatXact PROCs enables SAS users to access StatXact while working in SAS. This study used StatXact PROCs to examine the exact test.

For each replication, the exact test, MH, and logistic regression were used to detect DIF in each DIF item. Results were examined in terms of power, calculated as the proportion of times a DIF detection method correctly identified an item as displaying DIF using statistical significance, and then using statistical significance along with measures of practical significance. The logistic regression model used had two predictors and therefore only uniform DIF was calculated. The p-value associated with uniform DIF was used to compute statistical significance. A description of their implementation follows. The α level used in this study was 0.05. Power was calculated for each item embedded with DIF.

Power based on use of statistical significance along with practical significance was calculated using two steps. First effect sizes were calculated for each method of DIF detection. Both MH and the exact test used the criteria employed by ETS. The MH D-DIF value calculated using Equation 7. Second, the calculated value was categorized as follows: type A, negligible DIF, $|\text{MH D - DIF}| < 1$; type C, large DIF, $|\text{MH D - DIF}| \geq 1.5$; type B, intermediate DIF, $1 \leq |\text{MH D - DIF}| < 1.5$. Effect sizes for logistic regression

were determined by employing the method recommended by Thomas and Zumbo (1998), which requires the calculation of R^2. Once calculated, the items were classified as follows: type A, negligible DIF, $\Delta R^2 < .13$; type C, large DIF, $\Delta R^2 > .26$; type B, intermediate DIF, $.13 \leq \Delta R^2 \leq .26$. For both methods of effect size calculation, items categorized as type B or type C had to display statistical significance. For each condition and each method, items falling into categories A, B or C were tallied and divided by the number of replications to determine the power rates when statistical and practical significance were used together. Category A also included items that did not have statistical significance.

CHAPTER 4

RESULTS

A total of 84 unique conditions were used in this simulation study to make comparisons between each of the three DIF detection methods. In each condition, 200 replications were used, resulting in a total of 16,800 simulated data sets. Table 3 presents the study conditions, Table 4 lists the parameters of the test items, and Table 5 lists the DIF magnitude values used in the present study. These three tables collectively aid in understanding the factors examined in this study.

Data Analysis

Tables 7–12 present power in two different ways for each DIF item and study condition. First, power was assessed solely by statistical significance and second by statistical significance along with the presence of an intermediate or large effect size. In each table, power using statistical significance is compared to power using statistical significance along with effect sizes for each DIF detection method. To better understand the differences in magnitude of effect size among the three procedures, Tables 13–18 present the percent of items across replications that were in each of the three effect size categories for each study condition.

A proportion of MH items that demonstrated statistical significance resulted in an unattainable MH-DIF. Items with statistical significance which had an unattainable MH D-DIF value were not categorized. These items belonged to conditions that included the smallest sample sizes 5, 10 and less often 20. Complications arose in the calculation of

the effect sizes for the MH procedure when the focal group sizes were extremely small, 20 and below. The MH D-DIF value used to calculate effect size, see Equation 7, is based on a calculated α_{MH} value. When the focal group sample sizes were below 20, the likelihood that one or more cells in the 2x2 table may be empty increased. When this occurred, and a column and or row sum was zero the computation of the necessary α_{MH} value was inhibited. Although a *p*-value was assigned and was always 0.000, an effect size measure could not be determined because of computation limitations. Table 19 presents the proportion of statistically significant items that had unattainable effect sizes that could not be calculated.

This did not occur for the exact test or logistic regression. Of the simulated conditions with 10% of the items embedded with DIF, an average of 88% of the DIF item across conditions had an unattainable MH-DIF value when the focal group size was five across all DIF magnitude variations. This value decreased to 22% when the focal group size increased to 10. The largest focal group size to exhibit an unattainable MH-DIF value was 20, with a mean of 2%.

As the number of items embedded with DIF increased, the trend was replicated. The simulated condition with two items induced with DIF presented similar results for items one and two. When the focal group size was five, item one had a mean rate of 80% while item two's rate was 84%. Both items demonstrated a substantial decrease as sample size increased to 10. Under these conditions item one had a rate of 15% and item two had a rate of 17%. Item one was the only item unable to calculate a MH-DIF value when the sample size increased to 20. This occurred 4% of the time when the DIF magnitude was 0.50. As the number of DIF items increased from 20% to 30%, item one and two were

similar while item three followed the trend but had lower rates of MH-DIF values. For the focal group size of 5, items one and two had similar values, 75% and 74%, respectively. This was in contrast to 57% for item three. When the focal group size increased to 10, all items demonstrated a decrease in their inability to calculate the MH-DIF value. Item one's rate was 35%, item two's was 39% and item three's was 12%. Only item two and three were unable to calculate the value when the focal group sample size increased to 20. This occurred when the DIF magnitude was 1.00 at a rate of 2% for both items.

Item Discrimination and Item Difficulty

The item discrimination parameter, *a*, impacted the detection power for each of the three methods studied (see Table 6). The three items embedded with DIF had a different item discrimination value. Item one's value was 0.612, item two's was 0.889, and item three's value was 1.854 (see Table 4). Recall from Chapter 2, Kristjansson, Aylesworth, McDowell, and Zumbo (2005) defined an item discrimination parameter as low if its value was 0.8 or less, high if its value was at least 1.6, those values in between were labeled moderate. The value associated with the first item is defined as a low discrimination, item two's value is defined as moderate while item three's value is defined as high (Kristjansson, et al., 2005). In this study a relationship existed between power, item discrimination and sample size. The impact of a highly discriminating item (item 3) was greater when sample sizes were above 20. For sample sizes less than or equal to 20, item discrimination was not as influential (since power was always very low). This pattern was evident regardless of DIF detection method. The largest average increase in power was visible between item one (a = 0.612) and item three (a = 1.854)

when the focal group sizes were above 20. Under these conditions MH had a rate of 24.1% for item one compared to 63.9% for item three, the ET's rate was 22.1% compared to 61.9%, and LR's rate was 22.8% and 63.3%, respectively.

Table 6

Mean Power Rates Across Conditions for Each Method by Percent of Items Induced with DIF

Condition	Item 1	Item 2	Item 3
10% of DIF Items			
MH	17.6		
ET	15.3		
LR	19.6		
20% of DIF Items			
MH	16.6	27.6	
ET	14.3	24.9	
LR	18.8	29.1	
30% of DIF Items			
MH	17.5	26.8	45.4
ET	15.4	23.9	42.6
LR	19.7	28.3	45.1

Comparison Based on 10% of Items Containing DIF

Table 7 contains the results for the simulated conditions with 10% of the items embedded with DIF. As expected, an increase in the magnitude of DIF resulted in amplified power for all three procedures. To a lesser extent, increasing sample size increased power within each designated magnitude of DIF. As the magnitude of the DIF item increased, the impact of sample size on power also increased. The positive relationship between DIF magnitude and power, and sample size and power was exhibited by all three procedures, although the strength of the relationship varied. On

average across all conditions, LR had the highest DIF detection rate of 19.6%, MH was next with a rate of 17.6%, and ET was last with a rate of 15.3%. For all three methods overall, however, power was poor. Under the most favorable condition, focal group size of 100 and DIF magnitude of 1.00, power was still below the acceptable rate of 80% for all three procedures.

Table 7

10% of Items with DIF: Item One

Power and Power with Effect Sizes by Focal Group Size and DIF Magnitude (Percent)

$b_R - b_F$	Focal	MH		Exact		Logistic	
	N	Power	Power with ES	Power	Power with ES	Power	Power with ES
.25	5	2	1	2	2	4	4
	10	6	6	2	2	9	8
	20	6	6	6	6	8	8
	40	7	7	5	5	6	3
	60	6	6	5	5	7	7
	80	5	5	5	5	6	5
	100	6	6	5	5	5	4
.50	5	2	0	2	2	7	6
	10	9	8	7	7	9	7
	20	8	8	6	6	9	6
	40	14	14	9	9	16	14
	60	18	18	16	16	18	15
	80	18	18	16	16	21	19
	100	22	22	20	20	24	20
.75	5	7	1	3	3	11	10
	10	10	7	7	7	13	9
	20	12	12	9	9	13	10
	40	22	22	21	21	26	21
	60	28	28	27	27	30	25
	80	30	30	25	25	34	30
	100	34	34	32	32	32	27
1.00	5	6	1	4	4	19	15
	10	15	10	10	10	18	17
	20	21	20	18	18	22	17
	40	33	33	27	27	33	31
	60	38	38	37	37	39	33
	80	54	54	49	49	52	42
	100	55	55	53	53	57	45

Note. ES refers to effect size

When the focal group sample size was five, LR displayed the highest average power with a rate of 10.3%, MH was next with a rate of 4.3% and ET was last with a rate of 2.8% across all DIF magnitudes. Regardless of the magnitude of DIF, logistic regression's mean detection rate was 2.4 times as powerful as MH's and 3.7 times as powerful as the exact test's when the sample size was five. Increasing the sample size resulted in a general increase of power and robustness for all three techniques as expected. For the most extreme sample size conditions (focal group size less than or equal to 20) logistic regression exhibited the highest power followed by MH. The average power of logistic regression under these conditions was 11.8%, MH's value was 8.7% and the exact test had an average of 6.3%. Although sample size had an impact, the greatest differences were observed as a result of variation in DIF magnitude.

The average power of LR increased by 12% as the DIF magnitude changed from 0.50 to 0.75, and the sample size remained at five. When DIF was weak (0.25), MH exhibited power of 5.4% while ET's mean detection rate was 4.3%. As DIF magnitude increased, however, the difference between MH and the exact test increased. The mean detection rate increased to 20.4% for MH and 17.7% for the exact test when the DIF magnitude was increased to 0.75. Logistic regression's mean detection rate across all sample size conditions increased to 23.9% when ($b_R - b_F$ = -0.75) from 6.4% when ($b_R - b_F$ = -0.25). In conditions with the highest magnitude of DIF ($b_R - b_F$ = -1.00) and focal group size of 100 the three procedures performed similarly with values ranging from 53% to 57%. While LR demonstrated the highest DIF detection rates overall, MH and the exact test performed similar to one another. Overall, sensitivity of the logistic regression procedure to changes in the magnitude of DIF and sample size resulted in a slightly better

performance when compared to MH and the exact test, although power in most conditions was still very low.

When effect sizes were taken into account, regardless of DIF magnitude, MH's performance resulted in slightly greater practical significance than both the exact test and logistic regression when the focal group size exceeded 20. The average power when including only with moderate to large effect sizes when sample sizes were above 20, across all magnitudes of DIF, was 24.4% for MH followed by 22.0% for the ET. The average power when including effect sizes for logistic regression was 21.3%, under the same conditions power without effect sizes was 25.4%. When the focal group size was 20 or below the average power for each of the methods decreased across all conditions of DIF magnitude. The average went from 24.4% to 8.7% for MH, 22.0% to 6.3% for ET, and 25.4% to 11.8% for LR. However, when the focal group size was 20 or below, MH experienced the largest decline in practical significance. LR had the highest power in identifying practically significant items when sample sizes were extremely small, although this power was still very small.

Comparison Based on 20% of Items Containing DIF

Item one. When 20% of the items had DIF, the same patterns were seen for DIF detection in item one as in the conditions where item one was the only DIF item (10% of the items had DIF, see Table 8). As the DIF magnitude increased from weak ($b_R - b_F =$ -0.25) to strong ($b_R - b_F =$ -1.00) logistic regression's average power across all of the focal group's sample sizes increased from 6.9% to 33.9%, for item one. Under the same conditions, MH's rate of DIF detection increased from 5.9% to 30.4% and the exact test's rate increased from 4.9% to 27.7%. Logistic regression exhibited the greatest gain in

power as the DIF magnitude increased from weak to strong, still well below 80%. As expected, all three procedures performed best when the magnitude of DIF was strongest. The overall average power for MH was 16.6%, 14.3% for the ET, and 18.8% for LR. These results demonstrate a slight decrease from the conditions where item one was the only DIF item (10% of the items had DIF). The overall average power for item one decreased by 1.0% for both MH and the ET, and 0.8% for LR as the number of items with DIF increased from 10% to 20%.

Table 8

20% of Items with DIF: Item One

Power and Power with Effect Sizes by Focal Group Size and DIF Magnitude (Percent)

$b_R - b_F$	Focal	MH		Exact		Logistic	
	N	Power	Power with ES	Power	Power with ES	Power	Power with ES
.25	5	5	1	4	4	10	9
	10	6	4	4	4	7	6
	20	5	5	3	3	5	5
	40	6	6	6	6	7	6
	60	4	4	4	4	6	5
	80	8	8	7	7	8	6
	100	7	7	6	6	5	5
.50	5	3	1	1	1	10	10
	10	5	4	4	4	7	6
	20	13	13	6	6	13	11
	40	8	8	7	7	6	6
	60	13	13	11	11	18	14
	80	16	16	14	14	18	18
	100	18	18	17	17	22	18
.75	5	6	2	2	2	13	10
	10	7	7	3	3	10	9
	20	9	9	7	7	9	6
	40	18	18	17	17	18	18
	60	22	22	20	20	24	23
	80	3	30	26	26	30	29
	100	44	44	38	38	44	37
1.00	5	3	1	2	2	14	13
	10	9	8	5	5	10	9
	20	21	21	17	17	24	18
	40	34	34	30	30	37	34
	60	43	43	41	41	46	40
	80	52	52	49	49	54	47
	100	51	51	50	50	52	48

Note. ES refers to effect size

Item two. Logistic regression slightly exceeded both MH and the exact test in detecting DIF for item two (see Table 9). Under the weakest DIF magnitude conditions the mean detection rate for logistic regression across focal group sizes was 10.6% compared to 8.9% for MH, and 7.7% for the ET. As the DIF magnitude increased in strength the performance of all three procedures increased as well with logistic regression slightly outperforming the other two. When the magnitude of DIF was strongest, logistic regression had an average power of 51.1%. Under the same conditions MH's average rate was 49.0% and the exact test's was 45.6%. The strength of each procedure's performance corresponded to the strength of the DIF magnitude. When focal group sample sizes were above 20 the average power across DIF magnitudes for LR was 39.9%. This decreased to an average of 14.8% for sample sizes of 20 and below. For these same conditions, MH went from 38.8% to 12.6% and the ET went from 36.3 % to 9.6%. The mean power rate across all conditions for item two was 29.1% for LR, 27.6% for MH, and 24.9% for the ET. The rates for item two were noticeably higher than the mean power rates for item one but not near the acceptable rate of 80%. All three procedures had satisfactory power measures for item two with a focal group size of 100 and DIF magnitude of 1.00.

Table 9

20% of Items with DIF: Item Two
Power and Power with Effect Sizes by Focal Group Size and DIF Magnitude (Percent)

$b_R - b_F$	Focal	MH		Exact		Logistic	
	N	Power	Power with ES	Power	Power with ES	Power	Power with ES
.25	5	4	1	3	3	10	10
	10	4	3	3	3	7	7
	20	7	7	6	6	7	7
	40	11	11	11	11	12	12
	60	12	12	10	10	10	10
	80	8	8	7	7	10	10
	100	16	16	14	14	18	18
.50	5	8	2	5	5	9	9
	10	10	10	5	5	8	8
	20	11	11	11	11	12	12
	40	21	21	19	19	21	21
	60	21	21	19	19	22	22
	80	33	33	27	27	31	31
	100	31	31	27	27	34	34
.75	5	7	1	4	4	10	10
	10	12	10	9	9	13	13
	20	19	19	16	16	18	18
	40	29	29	27	27	32	31
	60	39	39	33	33	44	44
	80	60	60	55	55	60	59
	100	66	66	66	66	69	69
1.00	5	14	2	9	9	27	27
	10	23	18	16	16	24	24
	20	32	32	28	28	32	32
	40	50	50	47	47	51	50
	60	67	67	66	66	67	67
	80	72	72	70	70	71	70
	100	85	85	83	83	86	86

Note. ES refers to effect size

An analysis of power with effect size identified logistic regression as slightly outperforming MH and the exact test across all conditions for both items one and two. Logistic regression's effect size and power calculations had almost identical values for item one and two; this was also the case with the exact test. The MH procedure illustrated the greatest difference in power and power with effect size. As conditions became less extreme, the difference between the procedures decreased. Power when effect sizes were included identified logistic regression as having slightly higher rates, this advantage increased as sample sizes became increasingly large.

Comparison Based on 30% of Items Containing DIF

Items one and two. As the percentage of items induced with DIF increased from 10% to 20% and 20% to 30%, the average power for all three methods remained relatively constant. When 30% of the items were embedded with DIF, item one demonstrated a mean power rate of 17.5% for MH, 15.4% for the ET, and 19.7% for LR while item two's values were 26.8% for MH, 23.9% for the ET, and 28.3% for LR (see Tables 10 and 11). These averages are similar to the values calculated when 10% of the items had DIF and slightly higher than the values calculated when 20% of the items had DIF. There was a slight increase for all three methods as the items with DIF increased from 20% to 30%. Both MH and LR's mean power increased by 0.9%, while the ET's mean power increased by 1.1%. The percent of items with DIF did not impact overall power rates for items one and two when compared with conditions with 10% or 20% of items with DIF. Similar to these conditions, both items displayed a power rate closer to the acceptable rate of 80% as the percent of DIF items increased. This was observed when the DIF magnitude was1.00 and focal group size was 100 for item one and DIF magnitude was 1.00 and focal group size was greater than 20 for item two.

Table 10

30% of Items with DIF: Item One
Power and Power with Effect Sizes by Focal Group Size and DIF Magnitude (Percent)

$b_R - b_F$	Focal	MH		Exact		Logistic	
	N	Power	Power with ES	Power	Power with ES	Power	Power with ES
.25	5	3	2	3	3	11	10
	10	8	4	5	5	9	8
	20	5	5	4	4	5	4
	40	7	7	6	6	7	7
	60	7	7	7	7	9	8
	80	6	6	5	5	6	6
	100	6	6	6	6	9	8
.50	5	6	2	3	3	12	12
	10	11	7	7	7	10	9
	20	8	8	8	8	9	8
	40	10	10	10	10	12	12
	60	14	14	12	12	15	14
	80	20	20	17	17	18	16
	100	22	22	20	20	23	22
.75	5	4	1	3	3	12	12
	10	11	9	7	7	14	13
	20	10	10	7	7	11	11
	40	20	20	16	16	21	19
	60	25	25	25	25	26	25
	80	35	35	32	32	34	31
	100	34	34	30	30	35	33
1.00	5	6	2	4	4	21	19
	10	10	7	8	8	12	11
	20	24	24	20	20	28	26
	40	32	32	29	29	33	29
	60	41	41	37	37	42	39
	80	45	45	41	41	46	39
	100	61	61	60	60	61	57

Note. ES refers to effect size

Table 11

30% of Items with DIF: Item Two
Power and Power with Effect Sizes by Focal Group Size and DIF Magnitude (Percent)

$b_R - b_F$	Focal	MH		Exact		Logistic	
	N	Power	Power with ES	Power	Power with ES	Power	Power with ES
.25	5	6	2	4	4	10	10
	10	3	2	2	2	3	3
	20	7	7	5	5	6	6
	40	11	11	8	8	12	12
	60	11	11	9	9	12	11
	80	11	11	9	9	12	12
	100	14	14	12	12	15	15
.50	5	7	2	5	5	9	9
	10	9	6	7	7	11	11
	20	8	8	7	7	12	11
	40	16	16	14	14	15	15
	60	22	22	20	20	22	21
	80	26	26	24	24	29	29
	100	28	28	26	26	27	27
.75	5	4	0	2	2	12	12
	10	12	10	10	10	16	16
	20	21	21	16	16	21	21
	40	38	38	34	34	41	41
	60	44	44	41	41	44	44
	80	50	50	45	45	50	50
	100	61	61	58	58	63	63
1.00	5	10	0	6	6	17	16
	10	13	9	10	10	15	15
	20	32	32	26	26	32	32
	40	60	60	55	55	57	57
	60	70	70	65	65	72	72
	80	72	72	67	67	74	74
	100	83	83	81	81	83	83

Note. ES refers to effect size

Item three. The impact of DIF magnitude on power was amplified for item three for all three procedures (see Table 12). As the DIF magnitude increased from weak ($b_R - b_F$ = -0.25) to strong ($b_R - b_F$ = -1.00) logistic regression's average power across all of the focal group's sample sizes increased from 11.9% to 75.1%. Under the same conditions MH's rate of DIF detection increased from 11.4% to 74.6% and the exact test's rate increased from 10.3% to 71.1%. As the strength of the DIF magnitude increased the performance of each method also increased. The impact of sample size on item three was consistent with the results of items one and two; as the focal group size increased the three procedures increased in performance. Each of the three methods had acceptable detection rates when the DIF magnitude was 0.75 and focal group sample size was above 40. When the DIF magnitude increased to 1.00 and focal group sample size was greater than 20 all three procedures had a detection rate of 100%.

Table 12

30% of Items with DIF: Item Three
Power and Power with Effect Sizes by Focal Group Size and DIF Magnitude (Percent)

$b_R - b_F$	Focal	MH		Exact		Logistic	
	N	Power	Power with ES	Power	Power with ES	Power	Power with ES
.25	5	3	1	2	2	5	5
	10	4	4	3	3	3	3
	20	7	7	6	6	7	7
	40	13	13	12	12	15	15
	60	12	12	12	12	12	12
	80	20	20	18	18	20	20
	100	21	21	19	19	21	21
.50	5	8	2	6	6	9	9
	10	10	9	8	8	10	10
	20	19	19	16	16	19	19
	40	34	34	27	27	29	29
	60	51	51	47	47	48	48
	80	59	59	54	54	54	54
	100	75	75	73	73	76	76
.75	5	10	5	6	6	13	13
	10	22	21	17	17	17	17
	20	40	40	34	34	39	39
	40	71	71	67	67	71	71
	60	89	89	86	86	87	87
	80	91	91	91	91	91	91
	100	91	91	91	91	92	92
1.00	5	22	15	11	11	23	23
	10	37	31	32	32	38	38
	20	67	66	61	61	69	69
	40	96	96	94	94	96	96
	60	100	100	100	100	100	100
	80	100	100	100	100	100	100
	100	100	100	100	100	100	100

Note. ES refers to effect size

Along with changes in DIF magnitude and focal group sample size, each method demonstrated sensitivity to parameter changes. The magnitude difference between the *b* parameters of items one and two was approximately equal to the magnitude difference between items two and three. Item three was the least difficult of the DIF items, the *b*-value was -0.229 (see Table 4). For item three, both statistical and practical significance measures had a mean of 45.1% for logistic regression across all conditions.

Similar to the results of conditions with 10% of items with DIF and 20% of items with DIF logistic regression demonstrated slightly higher power than MH and the exact test in identifying DIF items when 30% of items contained DIF. The exact test displayed a slightly poorer performance than the other two procedures for item three, with average power of 42.6% compared to approximately 45% for both MH and logistic regression. When effect size measures were taken into account, MH had a slightly superior performance to logistic regression only when sample sizes were above 20. As sample sizes decreased to 20 and below, LR consistently demonstrated slightly greater practical significance relative to MH and ET but still not high power.

Practical Significance

The items flagged as having statistically significant DIF were then categorized as having either negligible (category A), intermediate (category B) or large (category C) DIF. The proportions of these statistically significant DIF items that fell in the three categories are presented in Tables 13 through 18 for each of the three DIF detection procedures. Recall category A items, negligible DIF, included those items that did not have statistical significance. For some of the conditions, the percentages in each category (A, B, and C) do not add up to 100%. This is because unattainable MH-DIF values could

not be categorized and therefore were not included in Tables 13 through 18. These values are listed in Table 19.

Table 13

10% of Items with DIF: Item One

Percent of Category A, B, and C Item by Focal Group Size and Magnitude of DIF

$b_R - b_F$	Focal	MH			Exact			Logistic		
	N	A	B	C	A	B	C	A	B	C
.25	5	98.5	0.0	0.5	98.5	0.0	1.5	96.0	3.5	0.5
	10	94.0	0.0	5.5	94.0	0.0	6.0	92.0	8.0	0.0
	20	94.0	0.0	6.0	94.0	0.0	6.0	92.5	7.5	0.0
	40	93.5	0.0	6.5	93.5	0.0	6.5	97.0	3.0	0.0
	60	94.0	0.0	6.0	94.0	0.0	6.0	93.5	6.5	0.0
	80	95.5	0.5	4.0	95.5	0.5	4.0	90.5	4.5	0.0
	100	94.5	3.0	2.5	94.5	3.0	2.5	96.0	5.0	0.0
.50	5	98.0	0.0	0.0	98.0	0.0	2.0	94.5	5.5	0.0
	10	91.0	0.0	7.5	91.0	0.0	9.0	93.0	7.0	0.0
	20	92.5	0.0	7.5	92.5	0.0	7.5	94.0	6.0	0.0
	40	86.0	0.0	14.0	86.0	0.0	14.0	86.5	13.5	0.0
	60	82.0	0.0	18.0	82.0	0.0	18.0	85.0	14.0	0.0
	80	82.0	1.5	16.5	82.0	1.5	16.5	81.5	18.5	0.0
	100	78.5	6.0	15.5	78.5	6.0	15.5	80.0	20.0	0.0
.75	5	93.0	0.0	0.5	93.0	0.0	7.0	85.0	15.0	0.0
	10	90.5	0.0	6.5	90.5	0.0	9.5	91.0	9.0	0.0
	20	88.0	0.0	12.0	88.0	0.0	12.0	90.5	8.5	1.0
	40	78.5	0.0	21.5	78.5	0.0	21.5	79.0	21.0	0.0
	60	72.0	0.0	28.0	72.0	0.0	28.0	75.0	24.5	0.5
	80	70.0	1.5	29.0	70.0	1.5	29.0	70.0	30.0	0.0
	100	66.0	4.0	30.0	66.0	4.0	30.0	73.0	27.0	0.0
1.00	5	94.5	0.0	0.5	94.5	0.0	5.5	90.5	9.5	0.0
	10	85.5	0.0	9.5	85.5	0.0	14.5	83.0	17.0	0.0
	20	79.5	0.0	19.5	79.5	0.0	20.5	83.0	16.0	1.0
	40	67.5	0.0	32.5	67.5	0.0	32.5	70.0	30.0	0.5
	60	62.0	0.0	38.0	62.0	0.0	38.0	67.5	32.5	0.0
	80	56.0	1.0	53.0	56.0	1.0	53.0	58.5	41.5	0.0
	100	45.5	4.0	50.5	45.5	4.0	50.5	55.5	44.5	0.0

Note. A, B, and C may not total 100% since unattainable values could not be categorized.

Table 14

20% of Items with DIF: Item One

Percent of Category A, B, and C Item by Focal Group Size and Magnitude of DIF

$b_R - b_F$	Focal	MH			Exact			Logistic		
	N	A	B	C	A	B	C	A	B	C
.25	5	95.5	0.0	2.0	95.5	0.0	4.5	91.0	9.0	0.0
	10	94.5	0.0	1.0	94.5	0.0	5.5	94.5	5.0	0.5
	20	95.5	0.0	4.5	95.5	0.0	4.5	95.0	5.0	0.0
	40	94.0	0.0	6.0	94.0	0.0	6.0	94.0	5.5	0.5
	60	96.0	0.0	4.0	96.0	0.0	4.0	95.5	4.5	0.0
	80	92.5	0.0	7.5	92.5	0.0	7.5	94.0	6.0	0.0
	100	93.0	2.0	5.0	93.0	2.0	5.0	95.0	5.0	0.0
.50	5	97.0	0.0	0.5	97.0	0.0	3.0	90.5	9.5	0.0
	10	95.0	0.0	4.0	95.0	0.0	5.0	94.5	0.5	0.5
	20	87.0	0.0	12.5	87.0	0.0	13.0	89.5	10.5	0.0
	40	92.0	0.0	8.0	92.0	0.0	8.0	94.5	5.0	0.5
	60	87.5	0.0	12.5	87.5	0.0	12.5	86.0	14.0	0.0
	80	84.5	0.5	15.5	84.5	0.5	15.5	82.5	17.5	0.0
	100	82.0	2.5	16.0	82.0	2.5	16.0	82.5	17.5	0.0
.75	5	94.0	0.0	1.5	94.0	0.0	6.0	90.0	9.5	0.5
	10	93.5	0.0	6.5	93.5	0.0	6.5	91.5	8.5	0.0
	20	91.5	0.0	8.5	91.5	0.0	8.5	94.0	5.5	0.5
	40	82.0	0.0	18.0	82.0	0.0	18.0	82.5	16.5	1.0
	60	78.0	0.0	22.0	78.0	0.0	22.0	77.5	22.5	0.0
	80	70.0	1.5	27.5	70.0	1.5	27.5	71.5	27.0	1.5
	100	56.5	7.5	36.0	56.5	7.5	36.0	63.0	36.0	1.0
1.00	5	97.0	00	0.5	97.0	0.0	3.0	87.5	12.5	0.0
	10	91.0	0.0	8.0	91.0	0.0	9.0	91.5	7.5	1.0
	20	79.0	0.0	21.0	79.0	0.0	21.0	82.0	15.5	2.5
	40	66.0	0.0	34.0	66.0	0.0	34.0	61.5	32.5	1.0
	60	57.0	0.0	43.0	57.0	0.0	43.0	60.0	40.0	0.0
	80	48.0	1.5	50.5	48.0	1.5	50.5	53.5	46.5	0.0
	100	49.0	3.0	48.0	49.0	3.0	48.0	52.0	47.0	1.0

Note. A, B, and C may not total 100% since unattainable values could not be categorized.

Table 15

20% of Items with DIF: Item Two

Percent of Category A, B, and C Item by Focal Group Size and Magnitude of DIF

$b_R - b_F$	Focal	MH			Exact			Logistic		
	N	A	B	C	A	B	C	A	B	C
.25	5	96.5	0.0	0.5	96.5	0.0	3.5	90.0	7.5	2.5
	10	96.0	0.0	3.0	96.0	0.0	4.0	93.5	5.5	1.0
	20	93.0	0.0	7.0	93.0	0.0	7.0	92.5	6.5	2.0
	40	89.0	0.0	11.0	89.0	0.0	11.0	88.0	9.5	2.5
	60	88.5	0.0	11.5	88.5	0.0	11.5	92.0	7.5	0.5
	80	82.5	0.5	7.0	82.5	0.5	7.0	90.5	8.5	1.0
	100	84.5	2.0	13.5	84.5	2.0	13.5	82.0	15.5	2.5
.50	5	92.0	0.0	2.0	92.0	0.0	8.0	91.5	6.0	2.5
	10	90.5	0.0	9.5	90.5	0.0	9.5	92.0	7.0	1.0
	20	89.0	0.0	11.0	89.0	0.0	11.0	88.0	10.0	2.0
	40	79.5	0.0	20.5	79.5	0.0	20.5	79.5	16.0	4.5
	60	79.0	0.0	21.0	79.0	0.0	21.0	78.0	20.5	1.5
	80	67.5	2.0	30.5	67.5	2.0	30.5	69.5	26.0	4.5
	100	69.0	6.0	25.0	69.0	6.0	25.0	66.5	31.5	2.0
.75	5	93.0	0.0	1.5	93.0	0.0	6.0	90.5	8.0	1.5
	10	88.0	0.0	6.5	88.0	0.0	6.5	87.5	9.0	3.5
	20	81.5	0.0	8.5	81.5	0.0	8.5	82.0	14.5	3.5
	40	71.0	0.0	18.0	71.0	0.0	18.0	69.0	23.0	8.0
	60	61.0	0.0	22.0	61.0	0.0	22.0	56.0	37.5	6.5
	80	40.0	2.0	58.0	40.0	2.0	58.0	41.0	49.5	9.5
	100	36.0	7.5	58.5	36.0	7.5	58.5	31.0	56.5	12.5
1.00	5	86.5	0.0	0.5	86.5	0.0	3.0	73.5	18.0	8.5
	10	77.5	0.0	8.0	77.5	0.0	9.0	25.5	17.5	7.0
	20	68.5	0.0	21.0	68.5	0.0	21.0	68.0	24.0	8.0
	40	50.5	0.0	34.0	50.5	0.0	34.0	50.0	37.0	13.0
	60	33.0	0.0	43.0	33.0	0.0	43.0	35.5	58.5	8.0
	80	28.5	1.5	50.5	28.5	1.5	50.5	30.0	60.5	9.5
	100	15.0	3.0	48.0	15.0	3.0	48.0	14.5	72.0	13.5

Note. A, B, and C may not total 100% since unattainable values could not be categorized.

Table 16

30% of Items with DIF: Item One

Percent of Category A, B, and C Item by Focal Group Size and Magnitude of DIF

$b_R - b_F$	Focal	MH			Exact			Logistic		
	N	A	B	C	A	B	C	A	B	C
.25	5	97.5	0.0	2.5	97.5	0.0	3.5	90.0	9.5	0.5
	10	92.5	0.0	7.5	92.5	0.0	11	92.0	7.0	1.0
	20	95.0	0.0	5.0	95.0	0.0	5.0	96.5	3.5	0.0
	40	93.0	0.0	7.0	93.0	0.0	5.0	93.5	6.5	0.0
	60	93.5	0.0	6.5	93.5	0.0	6.5	92.0	8.0	0.0
	80	94.5	0.5	5.0	94.5	0.5	5.0	94.5	5.0	0.5
	100	94.0	1.0	5.0	94.0	1.0	5.0	92.0	7.5	0.5
.50	5	94.5	0.0	5.5	94.5	0.0	10.0	88.5	10.0	1.5
	10	89.5	0.0	10.5	89.5	0.0	14.0	91.0	7.5	1.5
	20	92.0	0.0	8.0	92.0	0.0	8.0	92.0	7.0	1.0
	40	90.0	0.0	10.0	90.0	0.0	10.0	88.5	10.0	1.5
	60	86.0	0.0	14.0	86.0	0.0	14.0	86.0	14.0	0.0
	80	80.0	1.0	19.0	80.0	1.0	19.0	84.5	15.5	0.0
	100	78.0	5.0	17.0	78.0	5.0	17.0	78.5	21.0	0.5
.75	5	96.5	0.0	3.5	96.5	0.0	6.5	88.0	12.0	0.0
	10	89.0	0.0	11.0	89.0	0.0	13.5	87.5	12.0	0.5
	20	90.5	0.0	9.5	90.5	0.0	9.5	89.0	10.5	0.5
	40	80.5	0.0	19.5	80.5	0.0	19.5	81.0	19.0	0.0
	60	75.0	0.5	24.5	75.0	0.5	24.5	75.0	24.5	0.5
	80	65.5	1.5	33.0	65.5	1.5	33.0	69.5	30.5	0.0
	100	66.5	3.5	30.0	66.5	3.5	30.0	67.0	32.0	1.0
1.00	5	94.5	0.0	5.5	94.5	0.0	10.5	81.0	17.0	2.0
	10	90.0	0.0	10.0	90.0	0.0	13.5	89.0	10.0	1.0
	20	76.5	0.0	23.5	76.5	0.0	23.5	74.5	24.0	1.5
	40	68.5	0.0	31.5	68.5	0.0	31.5	71.0	27.5	1.5
	60	59.5	0.0	40.5	59.5	0.0	40.5	61.0	37.5	1.5
	80	54.5	2.0	42.5	54.5	2.0	42.5	61.5	38.0	0.5
	100	39.0	9.0	56.5	39.0	9.0	56.5	43.0	54.0	3.0

Note. A, B, and C may not total 100% since unattainable values could not be categorized.

Table 17

30% of Items with DIF: Item Two

Percent of Category A, B, and C Item by Focal Group Size and Magnitude of DIF

$b_R - b_F$	Focal	MH			Exact			Logistic		
	N	A	B	C	A	B	C	A	B	C
.25	5	94.5	0.0	5.5	94.5	0.0	9.5	90.0	9.0	1.0
	10	97.5	0.0	2.5	97.5	0.0	3.0	97.0	2.5	0.5
	20	93.0	0.0	7.0	93.0	0.0	7.0	94.0	5.5	0.5
	40	89.0	0.0	11.0	89.0	0.0	11.0	89.0	10.5	1.0
	60	89.5	0.0	10.5	89.5	0.0	10.5	89.0	9.0	2.0
	80	89.0	0.0	11.0	89.0	0.0	11.0	88.5	7.0	4.5
	100	86.0	1.5	12.5	86.0	1.5	12.5	85.0	13.0	2.0
.50	5	93.5	0.0	6.5	93.5	0.0	11.5	91.5	7.0	1.5
	10	91.5	0.0	8.5	91.5	0.0	11.5	89.0	8.5	2.5
	20	92.0	0.0	8.0	92.0	0.0	8.0	89.0	8.5	2.5
	40	84.5	0.0	15.5	84.5	0.0	15.5	85.5	11.5	3.0
	60	78.5	0.0	21.5	78.5	0.0	21.5	79.0	14.5	6.5
	80	74.5	1.0	24.5	74.5	1.0	24.5	71.0	24.5	4.5
	100	72.5	0.0	27.5	72.5	0.0	27.5	73.5	22.0	4.5
.75	5	96.0	0.0	4.0	96.0	0.0	8.0	88.0	9.0	3.0
	10	88.0	0.0	12.0	88.0	0.0	14.5	84.0	13.0	3.0
	20	79.5	0.0	20.5	79.5	0.0	20.5	79.0	14.5	6.5
	40	62.5	0.0	37.5	62.5	0.0	37.5	59.5	32.0	8.5
	60	56.0	0.0	44.0	56.0	0.0	44.0	56.5	34.0	9.5
	80	50.5	0.0	49.5	50.5	0.0	49.5	50.0	42.5	7.5
	100	39.0	3.5	57.5	39.0	3.5	57.5	37.0	51.0	12.0
1.00	5	90.5	0.0	9.5	90.5	0.0	19.0	84.0	12.5	3.5
	10	87.5	0.0	12.5	87.5	0.0	16.0	85.5	11.5	3.0
	20	68.0	0.0	32.0	68.0	0.0	32.5	68.5	23.0	8.5
	40	40.5	0.0	59.5	40.5	0.0	59.5	43.5	44.5	12.0
	60	30.5	0.0	69.5	30.5	0.0	69.5	28.0	59.0	13.0
	80	28.0	1.0	80.5	28.0	1.0	80.5	26.5	57.5	16.0
	100	17.5	2.0	71.0	17.5	2.0	71.0	17.0	70.0	13.0

Note. A, B, and C may not total 100% since unattainable values could not be categorized.

Table 18

30% of Items with DIF: Item Three

Percent of Category A, B, and C Item by Focal Group Size and Magnitude of DIF

$b_R - b_F$	Focal	MH			Exact			Logistic		
	N	A	B	C	A	B	C	A	B	C
.25	5	97.0	0.0	3.0	97.0	0.0	4.5	95.5	1.5	3.0
	10	96.0	0.0	4.0	96.0	0.0	5.0	97.0	0.0	3.0
	20	93.5	0.0	6.5	93.5	0.0	6.5	93.5	1.0	5.5
	40	87.0	0.0	13.0	87.0	0.0	13.0	85.0	2.0	13.0
	60	88.5	0.0	11.5	88.5	0.0	11.5	88.5	1.0	10.5
	80	80.5	0.0	19.5	80.5	0.0	19.5	80.0	1.0	19.0
	100	79.0	2.5	18.5	79.0	2.5	18.5	79.5	3.0	17.5
.50	5	92.5	0.0	7.5	92.5	0.0	13.0	91.5	1.5	7.0
	10	90.0	0.0	10.0	90.0	0.0	11.0	90.0	0.5	9.5
	20	81.5	0.0	18.5	81.5	0.0	18.5	81.5	1.5	17.0
	40	66.0	0.0	34.0	66.0	0.0	34.0	71.0	2.0	27.0
	60	49.5	0.0	50.5	49.5	0.0	50.5	52.0	3.5	44.5
	80	41.5	0.0	58.5	41.5	0.0	58.5	46.0	3.0	51.0
	100	25.5	0.0	74.5	25.5	0.0	74.5	24.5	2.0	73.5
.75	5	90.0	0.0	10.0	90.0	0.0	15.5	87.5	2.5	10.0
	10	78.5	0.0	21.5	78.5	0.0	22.5	83.0	1.5	15.5
	20	60.0	0.0	40.0	60.0	0.0	40.0	61.5	4.5	34.0
	40	39.5	0.0	70.5	39.5	0.0	70.5	29.0	5.0	66.0
	60	11.5	0.0	88.5	11.5	0.0	88.5	13.0	5.0	82.0
	80	9.5	0.0	90.5	9.5	0.0	90.5	9.0	5.0	86.0
	100	9.0	0.0	91.0	9.0	0.0	91.0	8.0	0.0	92.0
1.00	5	81.5	0.0	21.5	81.5	0.0	28.5	77.5	1.5	21.0
	10	63.0	0.0	37.0	63.0	0.0	43.5	62.0	2.5	35.5
	20	33.5	0.0	66.5	33.5	0.0	67.5	31.5	3.0	65.5
	40	4.0	0.0	96.0	4.0	0.0	96.0	4.5	4.5	91.0
	60	0.0	0.0	100.0	0.0	0.0	100.0	0.0	2.5	97.5
	80	0.0	0.0	100.0	0.0.	0.0	100.0	0.0	2.0	98.0
	100	0.0	0.0	100.0	0.0	0.0	100.0	0.0	1.5	98.5

Note. A, B, and C may not total 100% since unattainable values could not be categorized.

The exact test identified statistically significant items as displaying category C, strong DIF, more frequently than MH or LR. The exact test displayed an advantage over MH and LR in identifying items as having statistical and practical significance when the sizes of the focal group were small across all conditions. When the focal group sample

sizes increased to include 60, MH performance equaled that of the exact test. LR did not exhibit this trend. As the conditions changed from 10% to 30% of the items having DIF, more items were classified as C for all three of the procedures when item three was examined. The greatest change was observed with logistic regression.

Logistic regression demonstrated the lowest frequency in categorizing items with statistical and practical significance as category C. It was the most robust procedure; however, LR was more likely to identify items that had statistical significance but lacked practical significance. This frequency difference in identifying statistical and practical significance highlighted a weakness in the LR method. However, as the percentage of items containing DIF increased so did logistic regression's detection of items with statistical and practical significance across all conditions.

Table 19 presents the proportion of statistically significant items that had unattainable effect sizes that could not be calculated.

Table 19

Percent of an Unattainable MH D-DIF Value per Run per Condition for MH

Condition		Item 1	Item 2	Item 3
10%	DIF Magnitude .25			
	Focal Group 5	67		
	Focal Group 10	8		
	DIF Magnitude .50			
	Focal Group 5	100		
	Focal Group 10	17		
	DIF Magnitude .75			
	Focal Group 5	93		
	Focal Group 10	32		
	DIF Magnitude 1.00			
	Focal Group 5	91		
	Focal Group 10	31		
	Focal Group 20	2		
20%	DIF Magnitude .25			
	Focal Group 5	78	86	
	Focal Group 10	27	25	
	DIF Magnitude .50			
	Focal Group 5	83	75	
	Focal Group 10	20	-	
	Focal Group 20	4	-	
	DIF Magnitude .75			
	Focal Group 5	75	86	
	Focal Group 10	-	21	
	DIF Magnitude 1.00			
	Focal Group 5	83	89	
	Focal Group 10	11	22	
30%	DIF Magnitude .25			
	Focal Group 5	40	20	67
	Focal Group 10	47	73	13
	DIF Magnitude .50			
	Focal Group 5	82	77	73
	Focal Group 10	33	35	10
	DIF Magnitude .75			
	Focal Group 5	86	100	55
	Focal Group 10	23	21	5
	DIF Magnitude 1.00			
	Focal Group 5	91	100	33
	Focal Group 10	35	28	18
	Focal Group 20	-	2	2

CHAPTER 5

CONCLUSION

This chapter presents a summary of the findings of this study. A comparison is first made between study findings and current research. Secondly, a comparison of the three methods investigated is presented. Strengths and limitations of this study are then discussed. Lastly, implications of this study are presented followed by a discussion on future research directions.

Summary

Identifying items containing DIF is a crucial step in providing valid assessments. This study focused on expanding the conditions Parshall and Miller (1995) examined in order to add to the research on DIF identification under extremely small sample sizes. To accomplish this goal, power and effect size measures of three categorical procedures, the exact test, MH, and logistic regression, were compared on how well they identified items embedded with DIF. Prior to this study, a comparison of all three methods had not been completed. The results revealed a number of characteristics which had an effect on DIF detection.

Results and comparisons with previous research. All three methods demonstrated acceptable mean power rates for the same conditions and items. The percentage of items with DIF did not impact the rate of acceptability for the three procedures. Instead, the item discrimination parameter, *a*, demonstrated a visible impact. None of the procedures were effective in obtaining an acceptable mean power rate for item one, regardless of the

study conditions. Item one had the lowest item discrimination value (0.612) of the three items with DIF. This was not the case for items two and three. Recall that the item discrimination values for all items are listed in Table 4. The most favorable condition for item two, focal group size of 100 and a DIF magnitude of 1.00 resulted in power rates above 80% for all three procedures. The number of conditions resulting in acceptable power rates increased from item two to item three. For item three, all three procedures demonstrated acceptable power rates when the focal group size exceeded 40 and $b_R - b_F$ = -0.75 and for focal group sizes above 20 when $b_R - b_F$ = -1.00. As expected, higher group sizes resulted in higher power for DIF detection.

The exact test did not offer a clear advantage over MH, despite the fact that MH is an approximation of the ET. This finding was consistent with that of Parshall and Miller (1995) despite the smaller focal and reference group sample sizes employed in this study. MH had slightly higher power compared with ET, even under the most extreme sample size conditions. Although MH was more powerful, it was unable to demonstrate an acceptable rate of power, 80% or above, in most conditions. Unlike Parshall and Miller (1995), Meyer et al.'s (2004) applied study included effect size measures in their comparison of the ET and MH. The findings of this study are somewhat consistent with their findings. When unattainable MH D-DIF values were considered, MH demonstrated larger power with effect size measures than the ET when the focal group sample size was 20 and below. When these values were not included, however, the ET demonstrated superiority. The present study found that as sample size of the focal group increased the difference between the two procedures' power with effect size measures diminished. An examination of the labeling of items as strong or intermediate DIF presented the exact

test as somewhat more advantageous. The exact test had a greater proportion of items classified as strong DIF when compared to MH. Meyer et al.'s applied study found that items were equally likely to be classified as strong DIF by both methods.

Logistic regression's slight superiority regarding power was pervasive. These findings are not consistent with the research completed by Hidalgo and Lopez-Pina (2004), who found MH to be superior when identifying uniform DIF compared with LR. Their study, however, used large sample sizes for both the reference and the focal group. This study resulted in LR consistently displaying a greater effectiveness at detecting uniform DIF when compared to MH. Although logistic regression had higher power than the ET and MH, power in most conditions was still below the acceptable 80%. When the size of the sample was 20 or less, power differences between MH and LR were sizable. As sample sizes increased to above 20, average power differences between the two procedures diminished. Consistent with Swaminathan and Rogers' (1990) study both of the methods' power decreased as focal group sample size decreased.

All three methods were most robust in identifying DIF in item three. This item had an *a* parameter value of 1.8 (see Table 4). According to Kristjansson et al. (2005) this is a highly discriminating item. This is in contrast to the *a* parameter values of items one and two, both of which are categorized as low. This finding supports the work of Kristjansson et al. (2005) on the relationship between power and item discrimination. Variations in the *a* parameter assisted in explaining the large discrepancy in power among the three items. None of the three procedures were able to detect item one with an acceptable rate in contrast all three procedures detected item three with a power rate of

80% or above when the size of the focal group exceeded 40 and DIF magnitude was 0.75 and above.

Item two was defined as moderately difficult based on its *b* parameter value of 0.259; whereas item one was categorized as a difficult item with a *b* parameter value of 0.652, (see Tables 4 and 5). Item three was the least difficult item (*b* parameter value of -0.229). It was expected that fewer students would answer item one correctly when contrasted to the number of students who answered item two and three correctly. Items two and three are also slightly more discriminating than item one. The larger item discrimination parameter indicates them as more effective at distinguishing between those who have the assessed skill or ability and those who do not. The difference between their item parameter values, both *a* and *b*, had an impact on the performance of the three procedures being investigated. The impact on power is visible when the condition with 30% of DIF items is examined. MH's power across all conditions for item one at the $\alpha = .05$ level was 17.5%, for item two it was 26.8%, and for item three, it was 45.4%. The ET's rate for item one was 15.4% compared to 23.9% for item two and 42.6% for item three, and LR's was 19.7% for item one, 28.3% for item two, and 45.1% for item three.

Comparison of investigated methods. The conditions of small reference and small focal group sizes did not impact all methods equally. LR's performance was the most robust of the three methods investigated in this study across conditions, although in most cases, its advantage over the other procedures was rather small. This was particularly true when sample sizes were extreme, 20 or less. As focal group sample sizes increased from 5 to 100, all methods increased in power, and the differences in power between the three

methods also diminished. When effect sizes were taken into account, differences between the three procedures was not as pronounced. When the focal group sizes were extremely small (e.g., 20 and below) complications arose in the calculation of the effect sizes for the MH procedure. Specifically, the effect size for the MH could not be calculated for a number of replications. This did not occur with the other two procedures and presents a distinction between the MH procedure and the other two examined. An investigation of effect sizes did not reveal LR as consistently better than MH or the ET when identifying uniform DIF. When the focal group sample size was 20 or below, the mean power with effect size for conditions with 30% of items with DIF identified LR as more effective. As the size of the focal group increased, however, LR's advantage over the other methods diminished. Focal group sample sizes and changes in DIF magnitude demonstrated a greater influence over performance than the percentage of items embedded with DIF. As the percent of items with DIF increased the power for detecting DIF in item one (the DIF item that was constant in every condition) barely changed. This occurred for item two (20% and 30% of items with DIF) as well.

Both power and effect size measures attested to the slight advantage of MH over the exact test. It is possible that the sophistication of technology which makes the exact test a viable option has simultaneously increased the effectiveness of the approximation, MH. One characteristic of the ET was that a large percent of its items identified as practically significant were categorized as having strong DIF. This differed from LR, where the majority and under some conditions all items with practical significance were labeled as intermediate DIF. It was only when the item was highly discriminating, item three, that a majority of items were consistently categorized as having strong DIF for all

three procedures. The ET had a larger percent of practically significant items categorized as category C, strong DIF, compared to MH when the focal group size was below 20. For the same focal group size, however, MH had a higher percent of items categorized as practically significant. A direct comparison between MH and the ET cannot be completely made, because effect sizes were not obtainable for some MH conditions.

Much of the current research on small sample size focuses on distinguishing which of the two, MH or the exact test, is superior. Although the exact test never demonstrated a disparity between measures of practical and statistical significance, MH and LR frequently did. The slight strength of the logistic regression procedure with small sample size found in this study may be of interest. This examination of the three categorical methods revealed that logistic regression is overlooked in the detection of uniform DIF. It had slightly higher power in detecting DIF, illustrating a greater sensitivity to small sample sizes. Practical significance was not as conclusive. While logistic regression displayed higher power compared to the exact test and MH, it did not demonstrate an advantage regarding practical significance. This study did not yield a conclusive answer regarding the comparative performance of the MH test, the exact test, and logistic regression. Further it showed how poorly the three methods worked in general with small focal group sample sizes.

Limitations

This work is a simulation-based study. As is always the case with simulation studies, there is a limit on the number of conditions that can be chosen, due to constraints of time and resources.

The omission of Type I error rates limits the interpretation of the findings. Type I error rates for the MH have been studied, but not with small sample sizes. It would be

beneficial to know if one procedure has a greater likelihood of falsely identifying an item as being induced with uniform DIF. The similarity in performance between logistic regression and MH would greatly benefit by the information provided by Type I error rates. Also, the slight advantage demonstrated by logistic regression may be due to the identification of false positives.

Other study limitations include a test length of 10 items. In condition three (30% DIF items), only seven items were free of DIF. However, the *a* and *b* parameters demonstrated a greater influence over power than test length in this study. Also, in this study, one sample size for the reference group was examined along with uniform DIF and DIF that consistently favored the reference group. There is the possibility that while DIF detection rates were very low in most cases, with a large reference group size, the performance of these methods may have improved.

Implications

The conditions of test administration that lead to extremely small group sizes make the inference of the findings significant. When sample sizes are small and or test length is short, the importance of detecting DIF is amplified. With small focal group sizes under the conditions examined, no method works well. DIF detection methods will rarely detect DIF that is present unless the DIF is large, focal group sizes are large, and the items are more discriminating. Specifically, when the focal group sample size is small DIF detection does not work. The practice of having focal group sizes 40 or below should be avoided when possible. None of the three procedures demonstrated success with identifying items containing DIF when the focal group size was below 40. Each procedure's ability to detect DIF for focal group sizes 60 to 100 was affected by the item discrimination parameter. It can be implied that the three detection methods may be used

if the item is moderately or highly discriminating using the scale employed by Kristjansson, Aylesworth, McDowell, and Zumbo (2005).

This study gave credence to the premise that DIF detection should not be practiced when sample sizes are extremely small. However, if it is used, logistic regression may be a superior method for detecting items containing uniform DIF under extreme conditions. This is significant because large testing companies typically rely primarily on MH (Fidalgo, Ferreres, & Muniz, 2004). The magnitude of DIF and or the population size cannot easily be controlled in real life scenarios. Using a powerful DIF identification procedure is therefore important.

As educators and testing companies are called to provide fair and equitable tests regardless of the population size of the students, testing companies will have to address these concerns. Based on the results of this study more research is needed to determine limitations associated with logistic regression as a method for detecting uniform DIF. Although LR seemed to work best it still did not demonstrate an acceptable power rate (80%) under most conditions. It is notable that in this study, logistic regression had demonstrated the ability to identify the majority of items as having intermediate DIF, but as the items' ability to discriminate increased, the percentages of items identified as large DIF, category C, increased. This differed from the other two procedures for low discriminating items. Nonetheless, items that are induced with DIF are slightly more likely to be identified by logistic regression than by the exact test or MH.

Future Research

A follow-up study to investigate how Type I error rates impact the findings would be informative. More research is needed on the role of item discrimination and power. In this study, a relationship between the *a* parameter and power was demonstrated. All three

procedures demonstrated a higher percent of acceptable power rates for item three, the most discriminating item. As the value of the *a* parameter decreased, so did power. This was observed across all conditions by all three procedures. In an effort to better understand the relationship between item discrimination and DIF detection, a follow-up study could investigate if highly discriminating items have higher Type I error rates.

The unanticipated consequence of having focal group sizes extremely small, 20 and below, resulted in an inability to compute effect size measures for some conditions when the MH procedure was used. Although these items were flagged as having statistically significant DIF, no other information could be extracted.

Future research could involve additional design characteristics that might result in better power from DIF detection methods. Specifically, a longer test may counteract the problem of the small focal group sizes and increase the performance of not just MH but LR, and the ET as well. Increasing the size of the reference group may also benefit all three methods.

This study has provided a description and a simulated demonstration of DIF detection by three categorical data analysis procedures. The findings compliment the work previously done by Parshall and Miller (1995) as well as the work done by Swaminathan and Rogers (1990) regarding the strength of MH when compared to the exact test and logistic regression when compared to MH. This study compared all three procedures and their ability to detect DIF when conditions were extreme. None of the three methods performed well by demonstrating acceptable power rates. This study identified possible limitations of MH regarding sample size. This study also recognized

when logistic regression was more effective than MH. The findings from this study are preliminary and would benefit from testing with real or pseudo-real data.

REFERENCES

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley-Interscience.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley-Interscience.

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

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33(3),* 333-353.

Crane, P. K., van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine, 23(2)*, 241–256.

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

Donoghue, J. R., Holland, P.W., & Thayer, D.T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of DIF. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 137-166). Hillsdale, NJ: Erlbaum.

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

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Emons, W.H.M., Sijtsma, K., & Meijer, R.R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, *12*(1), 105-120.

Fidalgo, A. M., Ferreres, D., & Muñiz, J. (2004). Utility of the Mantel-Haenszel procedure for detecting differential item functioning in small samples. *Educational and Psychological Measurement, 64(6),* 925–936.

Fidalgo, A. M., Mellenbergh, G. J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online 2000, 5(3),* 43-53.

Fischer, G. H. (1995). Some neglected problems in IRT. *Psychometrika, 60(4)*, 459–487.

Gómez-Benito, J., & Navas-Ara, M. J. (2000). A comparison of χ^2, RFA and IRT based procedures in the detection of DIF. *Quality and Quantity, 34(1)*, 17–31.

Gierl, M. J., Jodoin, M. G., & Ackerman, T. A. (2000, April). *Performance of Mantel-Haenszel, simultaneous item bias test, and logistic regression when the proportion of DIF items is large.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

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

Hidalgo, M. D. & López-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. & Thayer, D. T. (1986, April). *Differential item performance and the Mantel-Haenszel procedure*. Paper presented at the 67th Annual Meeting of the American Educational Research Association, San Francisco, CA.

J. Douglas (personal communication, March 31, 2003) shared S-Plus programming code.

Kelderman, H., & Macready, G. B. (1990, December). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27*, 307–327.

Kristjansson, E., Aylesworth, R., McDowell, I, & Zumbo, B. D. (2005). A comparison of four methods for detecting DIF in ordered response items. *Educational and Psychological Measurement*, *65*, 935-953.

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

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement,52,* 443-451.

Meyer, J. P., Huynh, H., & Seaman, M. A. (2004). Exact small-sample differential item functioning methods for polytomous items with illustration based on an attitude survey. *Journal of Educational Measurement, 41*, 331–344.

Miller, M. D. & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement, 16(4)*, 381–388.

No Child Left Behind Act of 2001, Public Law 107-110 (2002).

Parshall, C. G. & Miller, T. R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics: A comparison of performance under small-sample conditions. *Journal of Educational Measurement, 32(3)*, 302–316.

Rasch, G. (1993). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: Mesa Press.

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

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

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

Thomas, D. R., & Zumbo, B. D. (1998). *Variable importance in logistic regression based on partitioning an R-squared measure*. Presented at the Psychometric Society Meetings, Urbana, IL.

Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education*. Upper Saddle River, NJ: Merrill.

U.S. Department of Education (2004). Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001. (ERIC Document Reproduction Service No. ED483111)

Whittaker, T. A., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (2003). IRTGEN: A SAS macro program to generate known trait scores and item responses for commonly used item response theory models. *Applied Psychological Measurement, 27(4)*, 299–300.

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

APPENDIXES

APPENDIX A

Hypergeometric Distribution and Mantel-Haenszel Sample Calculation

Given the following 2 x 2 contingency table:

	Item Correct	Item Incorrect	Row Total
Reference Group	y_j	x_j	m_j
Focal Group	y_j'	x_j'	m_j'
Column Total	n_j	n_j'	N_j

	Correct Response	Incorrect Response	Total
Reference Group	6	2	8
Focal Group	4	4	8
Total	10	6	16

The hypergeometric function below

$$P\left(y_j\right)=\frac{\binom{m_j}{y_j}\binom{m_j'}{n_j-y_j}}{\binom{N_j}{n_j}}$$

is computed as

$$P(6)=\frac{\binom{8}{6}\binom{8}{10-6}}{\binom{16}{10}} = \frac{\binom{8}{6}\binom{8}{4}}{\binom{16}{10}} = \frac{\left[\frac{8!}{6!(2!)}\right]\left[\frac{8!}{4!(4!)}\right]}{\frac{16!}{10!(6!)}} =$$

$$\frac{\left(\frac{40,320}{1,440}\right)\left(\frac{40,320}{576}\right)}{\frac{2.09\,\mathrm{x}\,10^{13}}{2.62\,\mathrm{x}\,10^{9}}} = .24$$

The Mantel-Haenszel statistic below

$$X^2_{\mathrm{MH}} = \frac{\left[\left(\frac{y_j x_j' - y_j' x_j}{N_j}\right)\right]^2}{Var}$$

is computed as

$$X^2_{\mathrm{MH}} = \frac{\left[\frac{(6\cdot 4)-(2\cdot 4)}{16}\right]^2}{\frac{10\cdot 6\cdot 8\cdot 8}{15\cdot 16^2}} = .95$$

APPENDIX B

Stat Exact SAS Code

```
%macro ss;
%do i=1 %to 200;
proc printto log='c:\phd documents\se\logstuff2.txt' new;
proc printto print='c:\phd documents\se\outputstuff2.txt' new;
data run1111; (data file containing values for a particular condition)
infile "c:\phd documents\datafiles2\2run3117\gphr&i..txt";
input r1-r10 theta foc;
score = r1 + r2 + r3 + r4 + r5 + r6 + r7 + r8 + r9 + r10;
data rd;
set run1111;
(create thick matching categories)
if score >= 0 and score <= 2 then tally = 0; if score = 3 then tally = 3;
if score = 4 then tally = 4;
if score = 5 then tally = 5;
if score = 6 then tally = 6;
if score >= 7 and score <= 10 then tally = 7;
proc sort;
by tally;
```

```
run;

(create a table for each score category)
proc stratify data=rd out=d1 sc=3 disp_acc=5;
ho/ex; ro foc; co r1; stratum tally;
proc stratify data=rd out=d2 sc=3 disp_acc=5;
ho/ex; ro foc; co r2; stratum tally;
proc stratify data=rd out=d3 sc=3 disp_acc=5;
ho/ex; ro foc; co r3; stratum tally;
proc stratify data=rd out=d4 sc=3 disp_acc=5;
ho/ex; ro foc; co r4; stratum tally;
proc stratify data=rd out=d5 sc=3 disp_acc=5;
ho/ex; ro foc; co r5; stratum tally;
proc stratify data=rd out=d6 sc=3 disp_acc=5;
ho/ex; ro foc; co r6; stratum tally;
proc stratify data=rd out=d7 sc=3 disp_acc=5;
ho/ex; ro foc; co r7; stratum tally;
proc stratify data=rd out=d8 sc=3 disp_acc=5;
ho/ex; ro foc; co r8; stratum tally;
*proc print data=d8;
proc stratify data=rd out=d9 sc=3 disp_acc=5;
ho/ex; ro foc; co r9; stratum tally;
*proc print data=d9;
```

```
proc stratify data=rd out=d10 sc=3 disp_acc=5;
ho/ex; ro foc; co r10; stratum tally;


data all; set d1 d2 d3 d4 d5 d6 d7 d8 d9 d10;
rep=&i;
keep col value rep;
if item = 'XCTPVAL2'; ( identifies what value from the tables is needed)
proc print data=all;
proc append base=final data=all;
run;
%end;
%mend ss;
%ss;
data theend; set final;
file 'c:\phd documents\se\serun3117.txt';
format VALUE 7.5;
put col value rep;
run;
```

APPENDIX C

Mantel-Haenszel SAS Code

```
OPTIONS PAGENO=1 LINESIZE=100 CENTER FORMDLIM='=';
%macro ss;
%do i=1 %to 200;
proc printto log='c:\PhD Documents\mh\LOGSTUFF2.txt' new;
PROC PRINTTO PRINT='C:\PhD Documents\mh\OUTPUTSTUFF2.TXT' NEW;
DATA TEMP;
infile "c:\PhD Documents\datafiles2\2run3417\gphr&i..txt";
input r1-r10 theta foc;
if foc=1 then newfoc=0; (identification of reference and focal group)
        else newfoc=1;
score = r1 + r2 + r3 + r4 + r5 + r6 + r7 + r8 + r9 + r10;
DATA EQUINT;
SET TEMP;
IF score >= 0 AND score =< 2 THEN tally = 0; (create thick matching)
IF score =3 THEN tally = 3;
IF score =4 then tally =4;
if score =5 then tally = 5;
```

```
if score =6 then tally = 6;

if score >= 7 and score <= 10 then tally = 7;

RUN;

/***************  FINISHED EFFECT SIZE CALCULATION *****************/

ods output cmh=gmh;

ods output commonrelrisks=dir;

ods listing;

proc freq;

tables tally*newfoc*R1-R10/CMH ;

run;

data dd; set gmh;

if statistic=3;

proc sort; by table;

data ee; set dir;

if StudyType='Case-Control';

proc sort; by table;

data comb; merge dd ee; by table;

rep=&i;

PROC APPEND BASE=ALL DATA=comb;

run;

%end;

%mend ss;

%ss;
```

```
data keepitall ; set all;
file 'c:\PhD Documents\mh\mhrun3417.txt';
put table value prob rep;
data sig; set keepitall;
if prob lt .05 ;
proc print data=sig; run;
proc print data=keepitall; run;
proc freq; tables table; run;
```

APPENDIX D

Logistic Regression SAS Code

```
%macro ss;
%do i=1 %to 200;
proc printto log='c:\PhD Documents\log_1dif\logstuff2try.txt' new;
proc printto print='c:\PhD Documents\log_1dif\outputstuff2try.txt' new;
data run1111;
infile "c:\PhD Documents\datafiles2\2run1417\gphr&I..txt";
input R1-R10 THETA FOC;
SCORE = R1 + R2 + R3 + R4 + R5 + R6 + R7 + R8 + R9 + R10;
if foc=1 then newfoc=0;
        else newfoc=1;
run;
DATA D2;
SET RUN1111;
IF SCORE >= 0 AND SCORE <= 2 THEN TALLY = 0;
IF SCORE =3 THEN TALLY = 3;
IF SCORE =4 THEN TALLY = 4;
IF SCORE =5 THEN TALLY = 5;
IF SCORE =6 THEN TALLY = 6;
```

```
IF SCORE >= 7 AND SCORE <= 10 THEN TALLY = 7;
PROC SORT;
BY TALLY;
RUN;
ods output globaltests=log11;
PROC logistic data=d2;
model R1 = tally /rsq;
data logsig11; set log11;
        if test='Likelihood Ratio';
ods output globaltests=log12;
ods output rsquare=rslog;
PROC logistic data=d2;
model R1 = tally newfoc/rsq;
data logsig12; set log12;
        if test='Likelihood Ratio';
data logsigcomb11; set logsig11 ;
        chi1=chisq;
        df1=df;
data logsigcomb12; set logsig12 ;
        chi2=chisq;
        df2=df;
```

(Use of the R^2 value to categorize strong, moderate, or weak DIF)

```
data comb1; set logsigcomb11; set logsigcomb12; set rslog;
```

```
        keep chi1 df1 chi2 df2 dfn chi chip rsquare rep;
        chi=chi2-chi1;
        dfn=df2-df1;
        chip=1-(probchi (chi, dfn));
        rsquare=cvalue1;
        rep=&i;
proc append base=all data=comb1;
run;
run;
%end;
%mend ss;
%ss;
data final; set all;
         if chip lt .01 then sig=1;
                else sig=0;
        if sig=1 & rsquare ge .130 then sigprac=1;
                else sigprac=0;
        proc freq; tables sig sigprac; run;
data d3; set final;
file "c:\PhD Documents\log_1dif\logrun1417try.txt";
put chi1 df1 chi2 df2 dfn chi chip rsquare rep sig sigprac;
run;
```