ASSESSMENT OF THE MULTIVARIATE OUTLIER APPROACH FOR
DIFFERENTIAL ITEM FUNCTIONING DETECTION: A MONTE CARLO
SIMULATION STUDY

A DISSERTATION
SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE
DOCTOR OF PHILOSOPHY
BY
JULIANNE M. EDWARDS
DISSERTATION ADVISOR: DR. W. HOLMES FINCH

BALL STATE UNIVERSITY
MUNCIE, INDIANA
MAY 2016
ABSTRACT

DISSERTATION: Assessment of the Multivariate Outlier Approach for Differential Item Functioning Detection: A Monte Carlo Simulation Study

STUDENT: Julianne M. Edwards

DEGREE: Doctor of Philosophy

COLLEGE: Teachers College

DATE: May 2016

PAGES: 131

Differential item functioning (DIF) is an important statistical tool that is used to detect potential bias between groups (e.g., males vs. females) in assessment items. The research pertaining to DIF, and the methods that assess DIF have predominantly focused on when only two groups are present (e.g., males vs. females). However, there are times that a psychometrician may want to detect potential bias in items when there are more than two groups. This can occur for large scale assessments when taking into account grouping variables such as ethnicity, SES, and location of the test taker.

Due to the need of methods that can estimate DIF when more than two groups are present, generalized Mantel-Haenszel, generalized logistic regression, Lord’s chi-square, and the multivariate outlier approach were created. However, the multivariate outlier approach has not been assessed in a simulation study to determine how well this method controls for Type I error, and power. Because of this, the current study conducted a simulation study to compare the multivariate outlier approach to generalized Mantel-Haenszel, generalized logistic regression, and Lord’s chi-square in regard to Type I error rates and power.
The current simulation study used a variety of simulation conditions. Specifically, various group sizes, percent contamination, impact, group size ratio, level of DIF, and sample size were used to generate the data. From these conditions, the results provided three main suggestions. First, if power is the only concern, then generalized Mantel-Haenszel should be used, as this method had the highest power, but did not control for Type I error. Second, if power and Type I error are both important to the researcher, then the multivariate outlier approach Orthogonalized Gnanadesikan-Kettenring beta with generalized logistic regression should be used. This is because this method had comparable power to generalized Mantel-Haenszel, but controlled for Type I error better than generalized Mantel-Haenszel. Last, if the concern is only Type I error then the multivariate outlier approach chi-square methods should be used, as Type I error was controlled for in most of the conditions.
TABLE OF CONTENTS

Abstract..........................................................................................................................ii
List of Tables..................................................................................................................vii
List of Figures...............................................................................................................viii
Chapter I: Introduction.................................................................................................1
  Definitions of Key Terms.........................................................................................5
Chapter II: Review of the Literature..............................................................................7
  Classical Test Theory..............................................................................................7
  Item Response Theory.............................................................................................8
    Person and Item Characteristics..........................................................................8
  IRT Models.............................................................................................................13
    Rasch Model.........................................................................................................13
    1 Parameter Logistic Model..................................................................................16
    2 Parameter Logistic Model..................................................................................17
    3 Parameter Logistic Model..................................................................................19
Differential Item Functioning.......................................................................................20
  DIF Detection Methods...........................................................................................21
    Generalized Mantel-Haenszel.............................................................................24
    Generalized Logistic Regression..........................................................................24
    Lord’s Chi-Square Test.........................................................................................26
    Multivariate Outlier Approach............................................................................28
Previous Research.......................................................................................................31
    Two groups..........................................................................................................31
Multiple focal groups.................................................................33
Research Questions........................................................................36
Chapter 3: Methodology..............................................................38
Simulation Conditions...................................................................38
Methods of DIF detection..............................................................38
Number of Groups.........................................................................39
Reference Group Sample Size and Group Size Ratio......................39
Level of Uniform DIF.....................................................................40
DIF Contamination.........................................................................41
Target Item.....................................................................................42
Impact............................................................................................43
Robust Estimators..........................................................................44
IRT and DIF Conditions..............................................................44
Study Outcome...............................................................................44
Repeated Measures ANOVAS......................................................44
Chapter 4: Results..........................................................................50
Two Groups....................................................................................50
Type I Error....................................................................................50
Summary.......................................................................................60
Power............................................................................................61
Summary.......................................................................................74
Three Groups................................................................................75
Type I Error....................................................................................75
Summary....................................................................................................84

Power.....................................................................................................................85

Summary....................................................................................................94

Six Groups................................................................................................................95

Type I Error.............................................................................................................95

Power...........................................................................................................98

Summary..................................................................................................110

Chapter 5: Discussion..............................................................................................111

Study Findings..........................................................................................................111

Difference in Type I error....................................................................................113

Summary..................................................................................................116

Difference in Power............................................................................................116

Summary..................................................................................................118

Limitations and Future Directions........................................................................119

Conclusion................................................................................................................120

References.............................................................................................................124
LIST OF TABLES

Table 1 – Simulation Conditions Considered for the Study..........................................................38

Table 2 – Item Difficulties based on Level of DIF for 40% DIF Contamination.........................43

Table 3 – Description of Each Repeated Measures ANOVA.................................................47- 49

Table 4 – Summary of Study Findings..........................................................................................112

Table 5 – Summary of Strengths and Weaknesses for Each Method........................................121
LIST OF FIGURES

Figure 2.1 – Item Characteristic Curve for Rasch model.................................................................15
Figure 2.2 – Item Information Curve for a Rasch model.................................................................16
Figure 2.3 – Item Characteristic Curve for 2PL model.................................................................18
Figure 2.4 – Item Information Curve for 2PL model.................................................................19
Figure 2.5 – Uniform DIF................................................................................................................21
Figure 2.6 – Nonuniform DIF.........................................................................................................22
Figure 4.1.1 – Three-way interaction of Type I error for two groups, method by level of DIF by impact with the impact of 0........................................................................................................51
Figure 4.1.2 – Three-way interaction of Type I error for two groups, method by level of DIF by impact with the impact of .5........................................................................................................52
Figure 4.1.3 – Three-way interaction of Type I error for two groups, method by level of DIF by impact with the impact of 1........................................................................................................53
Figure 4.2.1 – Three-way interaction of Type I error for two groups, method by sample size by percent contamination with the reference sample size of 500.........................................................55
Figure 4.2.2 – Three-way interaction of Type I error for two groups, method by sample size by percent contamination with the reference sample size of 1000............................................................56
Figure 4.2.3 – Three-way interaction of Type I error for two groups, method by sample size by percent contamination with the reference sample size of 2000............................................................57
Figure 4.3.1 – Three-way interaction of Type I error for two groups, method percent contamination by impact with the impact of 0........................................................................................................58
Figure 4.3.2 – Three-way interaction of Type I error for two groups, method percent contamination by impact with the impact of .5........................................................................................................59
Figure 4.3.3 – Three-way interaction of Type I error for two groups, method percent contamination by impact with the impact of 1.................................................................60
Figure 4.4.1 – Three-way interaction of power for two groups, method by level of DIF by percent contamination with level of DIF of .4.................................................................62
Figure 4.4.2 – Three-way interaction of power for two groups, method by level of DIF by percent contamination with level of DIF of .6.................................................................63
Figure 4.4.3 – Three-way interaction of power for two groups, method by level of DIF by percent contamination with level of DIF of .8.................................................................64
Figure 4.5.1 – Three-way interaction of power for two groups, method by reference sample size by percent contamination with the reference sample size of 500........................................66
Figure 4.5.2 – Three-way interaction of power for two groups, method by reference sample size by percent contamination with the reference sample size of 1000........................................67
Figure 4.5.3 – Three-way interaction of power for two groups, method by reference sample size by percent contamination with the reference sample size of 2000........................................68
Figure 4.6.1 – Three-way interaction of power for two groups, method by reference sample size by level of DIF with the level of DIF of .4.................................................................69
Figure 4.6.2 – Three-way interaction of power for two groups, method by reference sample size by level of DIF with the level of DIF of .6.................................................................70
Figure 4.6.3 – Three-way interaction of power for two groups, method by reference sample size by level of DIF with the level of DIF of .8.................................................................71
Figure 4.7.1. Three-way interaction of power for two groups, method by level of DIF by impact with an impact of 0. .................................................................72
Figure 4.7.2. Three-way interaction of power for two groups, method by level of DIF by impact with an impact of .5. ........................................................................................................73

Figure 4.7.3. Three-way interaction of power for two groups, method by level of DIF by impact with an impact of 1. ........................................................................................................74

Figure 4.8.1 – Three-way interaction of Type I error for three groups, method by level of DIF by percent contamination with a level of DIF of .4. .................................................................76

Figure 4.8.2 – Three-way interaction of Type I error for three groups, method by level of DIF by percent contamination with a level of DIF of .6. .................................................................77

Figure 4.8.3 – Three-way interaction of Type I error for three groups, method by level of DIF by percent contamination with a level of DIF of .8. .................................................................78

Figure 4.9.1 – Three-way interaction of Type I error for three groups, method by sample size by percent contamination with a reference sample size of 500.................................................................79

Figure 4.9.2 – Three-way interaction of Type I error for three groups, method by sample size by percent contamination with a reference sample size of 1000.................................................................80

Figure 4.9.3 – Three-way interaction of Type I error for three groups, method by sample size by percent contamination with a reference sample size of 2000.................................................................81

Figure 4.10.1 – Three-way interaction of Type I error for three groups, methods by sample size by level of DIF with a level of DIF of .4................................................................................82

Figure 4.10.2 – Three-way interaction of Type I error for three groups, methods by sample size by level of DIF with a level of DIF of .6................................................................................83

Figure 4.10.3 – Three-way interaction of Type I error for three groups, methods by sample size by level of DIF with a level of DIF of .8................................................................................84
Figure 4.11.1 – Three-way interaction of power for three groups, methods by reference sample size by level of DIF with a level of DIF of .4. ................................................................. 86

Figure 4.11.2 – Three-way interaction of power for three groups, methods by reference sample size by level of DIF with a level of DIF of .6. ................................................................. 87

Figure 4.11.3 – Three-way interaction of power for three groups, methods by reference sample size by level of DIF with a level of DIF of .8. ................................................................. 88

Figure 4.12.1 – Three-way interaction of power for three groups, methods by level of DIF by percent contamination with a level of DIF of .4. ................................................................. 89

Figure 4.12.2 – Three-way interaction of power for three groups, methods by level of DIF by percent contamination with a level of DIF of .6. ................................................................. 90

Figure 4.12.3 – Three-way interaction of power for three groups, methods by level of DIF by percent contamination with a level of DIF of .8. ................................................................. 91

Figure 4.13.1 – Three-way interaction of power for three groups, methods by reference sample size by percent contamination with the reference sample size at 500. .................................................. 92

Figure 4.13.2 – Three-way interaction of power for three groups, methods by reference sample size by percent contamination with the reference sample size at 1000. .................................................. 93

Figure 4.13.3 – Three-way interaction of power for three groups, methods by reference sample size by percent contamination with the reference sample size at 2000. .................................................. 94

Figure 4.14.1 – Three-way interaction of Type I error for six groups, methods by reference sample size by percent contamination with the reference sample size at 500. .............................. 96

Figure 4.14.2 – Three-way interaction of Type I error for six groups, methods by reference sample size by percent contamination with the reference sample size at 1000. .............................. 97
Figure 4.14.3 – Three-way interaction of Type I error for six groups, methods by reference sample size by percent contamination with the reference sample size at 2000...........................................98

Figure 4.15.1 – Three-way interaction of power for six groups, methods by reference sample size by percent contamination with the reference sample size at 500..................................................100

Figure 4.15.2 – Three-way interaction of power for six groups, methods by reference sample size by percent contamination with the reference sample size at 1000..................................................101

Figure 4.15.3 – Three-way interaction of power for six groups, methods by reference sample size by percent contamination with the reference sample size at 2000..................................................102

Figure 4.16.1 – Three-way interaction of power for six groups, methods by reference sample size by level of DIF with level of DIF at .4..................................................................................104

Figure 4.16.2 – Three-way interaction of power for six groups, methods by reference sample size by level of DIF with level of DIF at .6..................................................................................105

Figure 4.17.1 – Three-way interaction of power for six groups, methods by level of DIF by percent contamination with level of DIF at .4..................................................................................106

Figure 4.17.2 – Three-way interaction of power for six groups, methods by level of DIF by percent contamination with level of DIF at .6..................................................................................107

Figure 4.18.1 – Three-way interaction of power for six groups, methods by percent contamination by impact with impact at 0..................................................................................108

Figure 4.18.2 – Three-way interaction of power for six groups, methods by percent contamination by impact with impact at .5..................................................................................109

Figure 4.18.3 – Three-way interaction of power for six groups, methods by percent contamination by impact with impact at 1..................................................................................110
CHAPTER I
INTRODUCTION

Within our society, assessments are entrenched in many areas of our lives. Assessments are used in psychology, business, and the medical field, just to name a few. But, likely the more prominent area that assessment is found is standardized assessments within the educational system. In the educational system, assessments are used to evaluate the knowledge of students through state standardized and admission standardized tests. These assessments can determine if a teacher is effectively teaching their students, and used to assist universities in determining admission of students. As a result, educational standardized assessments are often high stakes and have many implications. Some of these implications are, teacher evaluation tool for teacher promotion, school evaluation tool for funding, admission to college, and much more. Because assessments are such a force within these systems, particularly in education, it is prudent to assess and ensure the validity of these measurements.

One important component of the validity of an assessment is test fairness. Test fairness refers to the idea that the score on an assessment holds a very specific meaning, and this meaning does not change based on the group that one belongs to (e.g., ethnicity, gender, SES). When test fairness is violated, it indicates that items within the assessment are influenced by an external factor and can no longer accurately measure the construct. Thus, the items are invalid to the construct. As a result, when an assessment includes items that do not demonstrate test fairness, there is a threat to validity (Cohen, Kim, & Wollack, 1996). When test fairness and validity are violated, the construct can not be accurately interpreted, and wrong conclusions may be drawn. However, if items that are not fair are removed from the test bank or modified, the validity of the assessment will be improved (Magis & DeBoeck, 2011). Because of this, it is important to have
the ability to accurately identify items within an assessment that are not fair. One statistical tool that can identify such items is differential item functioning (DIF).

DIF is a statistical tool that is used to assess potential bias that is present in each item. The assessment of DIF is an important aspect of validating an instrument for use with a broad population of individuals from a variety of groups (Zumbo, 1999). DIF is generally defined as a difference in an item response between members of two groups (e.g., males and females) after controlling for ability that is being measured by the instrument (Camilli & Shepard, 1994). In other words, DIF suggests that there may be a secondary factor that is influencing the performance of the item that is identified as having potential bias (Penfield & Camilli, 2007). Specifically, this is indicated by the difficulty of an item differing across the groups (e.g., ethnicity).

To test for DIF, there are a variety of methods that can be used: Mantel-Haenszel (Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990), SIBTEST (Shealy & Stout, 1993), Lord’s chi-square test (Lord, 1980), and the item response theory likelihood ratio test (Thissen, Steinberg, & Wainer, 1988), just to name a few. Even though these methods have shown to be useful, they can only be used in the context when only two groups are present: one reference group (i.e., the group that is expected to have an advantage) and one focal group (i.e., the group that is expected to have a disadvantage). However, there are many areas where researchers would want to assess DIF for when more than one focal group is present (e.g., international assessment studies, SES, ethnicity). For this reason, DIF methods recently have been expanded to test for DIF with more than two groups, as well as new methods created. Therefore, the present study fills a void by assessing DIF methods when multiple focal groups are present with a particular emphasis on multivariate outlier approach (explained below). This
study serves three purposes. First, to determine if the multivariate outlier approach controls for Type I error better than other methods. Second, to compare the power of the multivariate outlier approach to the other methods. Last, to analyze and determine the factors (e.g., sample size, number of groups) and the combination of factors that influence the Type I error rates and power of all methods.

The methods that have recently been expanded to assess DIF for multiple focal groups are Generalized Mantel-Haenszel test (GMH; Penfield, 2001), Generalized Logistic Regression (GLR; Magis, Raîche, Béland, & Gérard, 2011), Lord’s Chi-square test (Kim, Cohen, & Park, 1995). Although these methods were demonstrated using real data, there is little research comparing these methods with one another in a simulation study. Finch (2015) did, however, compare GMH, GLR, and Lord’s Chi-square test all in one simulation study. Overall the results of the study found that GMH performs the best. Even so, the multivariate outlier approach was mentioned in Finch’s study, but it was not included in the study. In fact, the multivariate outlier approach has not been examined empirically in a simulation study. Although a simulation study has not assessed the multivariate outlier approach, in a real data example this method appeared to perform well as it produced identical results to Kim et al. (1995; Magis & DeBoeck, 2011). Due to this potential promise from the real data example, this warrants a simulation study to further evaluate the multivariate outlier approach.

The multivariate outlier approach is unique because this method identifies items as containing DIF if the DIF statistic for an item is an outlier of the multivariate distribution. Due to this approach taking a different route, the multivariate outlier approach has potential benefits over the other DIF procedures. The first advantage is that the multivariate outlier approach is a one-step process, which does not require an iterative purification process. Second, it is expected
that the multivariate outlier approach does not have an inflated Type I error when compared to
the other methods (Magis & DeBoeck, 2011). Last, overall Magis and DeBoeck (2011) describe
the multivariate outlier approach as being straightforward and easily implemented. Although
there appears to be advantages to multivariate outlier approach, simulation research needs to be
conducted to determine how the multivariate outlier approach compares to other DIF detection
methods regarding accuracy. By performing a simulation study, this study will be able to assess
the effectiveness of the multivariate outlier approach when compared to the other methods, and
determine if these potential advantages can be accurately valued.

In addition to comparing DIF methods, this study served to analyze factors that can affect
the Type I error rates and power of the methods. In this study, six factors were analyzed. The
first factor is impact. Impact occurs when there is a natural difference in average group abilities.
Second is group size – unequal groups versus equal groups. Group size is considered in this
study as in applied research there are instances when the sample sizes across the groups are equal
or are not equal. Third, sample size was also considered in this study, as prior research has found
that smaller sample sizes yield poorer Type I error rates (Kim, Cohen, & Cohen, 1999). Fourth,
the number of groups was also considered. This is because it is important to assess if multivariate
outlier approach can actually handle the case of when many groups are included. The fifth factor
considered is the number of items that are contaminated with DIF. This factor was included, as it
was expected that as more items contained DIF, multivariate outlier approach would do a poorer
job of identifying the item as containing DIF. This is because the item would no longer be an
outlier. Last, is the level (or strength of DIF). Level of DIF refers to the difference in item
difficulty parameters between the groups.
To compare the DIF methods and factors that can influence Type I error and power, a Monte Carlo simulation study was conducted for the methods GMH, GLR, Lord’s Chi-square test, and the multivariate outlier approach. To compare these methods, 1000 data sets for each condition with 20 dichotomous items focusing on a one parameter logistic item response theory model were generated through Mplus version 7.11 (Muthén & Muthén, 2013). A one parameter logistic IRT model was selected as prior research has already compared GMH, GLR, and Lord’s using a two parameter logistic IRT model. Additionally, item difficulty values were based on a calibrated statewide achievement assessment. The generated data were then used to compare the analyses in R statistical software (R Development Core Team, 2014). This study had a total of 1062 different simulation conditions after crossing all of the manipulated factors. The contamination of item two assessed whether the simulation is testing for Type I error or power. From the results gathered from the simulation study, a repeated measures ANOVA was conducted to compare the Type I error and power rates across the items and conditions.

**Definitions of Key Terms**

**Differential Item Functioning (DIF):** A statistical tool used to detect potential advantage to one group of individuals over another on each item of the test.

**Generalized Logistic Regression (GLR):** Detects DIF for when more than two groups are present that is based on components found in logistic regression.

**Generalized Mantel-Haenszel (GMH):** A DIF detection method that is an extension of the chi-square test of association that is used more than two groups are present.

**Impact:** The differences in ability levels across the groups on the latent trait. That is, when impact is present, one or more groups have a higher natural ability in the content that is being assessed. For example, if DIF is being tested for an English proficiency test with the groups set
to native English speaker and non-native English speaker, the English speaking group should have a higher ability in English.

**Individual Ability:** The number of standard deviations away from the mean a person’s ability is on the measured construct.

**Item Difficulty:** The individual ability an examinee needs in order to have a 50% chance of correctly answering the item.

**Item Response Theory (IRT):** A statistical tool used to examine item characteristics (e.g., item difficulty, item discrimination, item guessing) and individual ability for each individual test item.

**Lord’s Chi-Square Test (Lord’s):** A DIF detection method based in the chi-square test that is used more than two groups are present.

**Multivariate Outlier Approach (MOA):** A DIF detection method that is used for when more than two groups are present. This method identifies an item as DIF if the DIF statistic is an outlier from the rest of the items.

**Nonuniform DIF:** The detection of DIF where the group advantage can change across individual ability.

**Power:** This occurs when a DIF method correctly identified item two as containing DIF.

**Robust Estimators:** A mathematical component selected by the researcher when calculating the multivariate outlier approach.

**Uniform DIF:** The detection of DIF where the group advantage does not change across individual ability.

**Type I Error:** This occurs when a DIF method identified item two as containing DIF when it was not simulated as containing DIF.
CHAPTER II
REVIEW OF THE LITERATURE

Classical Test Theory

Classical test theory (CTT) is a statistical approach that is used to determine the amount of error within a test (Raykov & Marcoulides, 2011). Unlike most statistical tools (e.g., ANOVA, regression), CTT is interested in the nature and characteristics of the measure and not the individual, similar to factor analysis. In CTT it is assumed that a person’s raw score consists of the individual’s true score (i.e., person’s actual ability), and the random error within the measure (DeVellis, 2006). The goal is to use the individuals’ information to determine whether or not the test has a high amount of error (Raykov & Mercoulides, 2011). The equation for CTT is defined as:

\[ X_{ki} = T_{ki} + E_{ki} \]  \hspace{1cm} (1)

where \( X_{ki} \) is the test score for individual \( i \) on test \( k \), \( T_{ki} \) is the true score for individual \( i \) on test \( k \), and \( E_{ki} \) is the error score for individual \( i \) on test \( k \). Unlike the raw score, the true score is a conceptualized score, as it is the average score the individual received across an infinite number of testing sessions (Raykov & Marcoulides, 2011). Once the true score is estimated, CTT determines the amount of error for a given test by taking the individuals true score and subtracting it by the raw score (Crocker & Algina, 1986). This score will tell us how much error is within the test with a number further away from 0 indicating a higher amount of error.

Other components within CTT include item difficulty and item discrimination. For CTT, item difficulty will suggest how difficult an individual item is, which is determined by the percent of individuals answering the item correct (Kline, 2005). Items that are closer to 100% are considered easier items, and items closer to 0% are more difficult items. Item discrimination is
calculated by testing the correlation between the item response and total test score, using a biserial or point-biserial correlation (Raykov & Marcoulides, 2011). Although CTT can calculate item difficulty and item discrimination, these parameters are sample dependent (Hambleton & Jones, 1993). This creates the use of the item parameters difficult, unless the sample is very similar to the population that will be tested. However, it is not always known in a field test how the sample will differ from the population, thus reducing the utility of these item parameters. Because these parameters in CTT could change based upon the samples, psychometricians were interested in creating a method that resulted in constant item and person parameters. Through this need IRT was born (Hambleton & Jones, 1993).

**Item Response Theory**

By understanding CTT, it aids with the understanding of item response theory (IRT) and how IRT builds upon CTT. IRT is an approach, different to CTT, as it examines each individual item, rather than the test as a whole. Specifically, IRT is used in psychometrics to assess the relationship between answering an item correctly based on the unique characteristics (e.g., difficulty and discrimination) of items, and the characteristics of the test takers (Raykov & Marcoulides, 2011). IRT does this by examining both the item characteristics (e.g., item difficulty, item discrimination, guessing) and individual ability (theta) within the mathematical equation. This is to provide an estimation of the likelihood of an individual correctly answering or endorsing each item (Hambleton & Swaminathan, 1985).

**Person and Item Characteristics**

Two of the main components of IRT are individual ability and test characteristics. This is because within IRT, test characteristics and individual ability can be incorporated into the IRT
models, which will later be explained. To begin individual ability must first be explained, as
ability is directly tied into the test characteristics interpretation.

Individual ability is special to IRT, as CTT does not incorporate this parameter into the
CTT equation (de Ayala, 2009). This is one of the main advantages to using IRT, as IRT is able
to specifically see how individual ability impacts the probability of selecting correct answers to
individual items, instead of relying upon a global scale probability. Conceptually, individual
ability in IRT is based on how many items the test taker answers correctly, as well as taking into
consideration the difficulty of each item (Hulin, Drasgow, & Parsons, 1983). For example, if a
test taker correctly answers less than half of the questions, and those items are considered easy
then that individual would have a lower individual ability. However, if these correctly answered
questions were more difficult, then the individual will have a higher ability on a particular test.

Within IRT individual ability is represented by theta \( \theta \) (Hambleton, Swaminathan, &
Rogers, 1991). \( \theta \) is a method of representing a person’s ability in standard deviations, as it
determines how far someone’s ability is from the average ability in terms of standard deviations.
Similar to standard deviation individual ability ranges from \(-\infty \) to \( \infty \), with the typical reported
range from -4 to 4 (Raykov & Marcoulides, 2011). A \( \theta \) of 0 represents an average ability with
any \( \theta \) above 0 suggesting a higher ability and anything lower than 0 represents a poorer ability.
Thus, a \( \theta \) of 1.45 suggests that the individual’s ability level is 1.45 standard deviations above the
mean.

To calculate individual ability there are two main maximum likelihood estimations that
are used, empirical maximum likelihood estimation (MLE) and Newton’s method for MLE (de
Ayala, 2009). In general, MLE methods are used to estimate population parameters from the
sample that is provided. To accomplish this, MLE will use the item parameters and the person
response pattern as guidelines while testing multiple ability levels to determine the ability level that most accurately estimates the individual’s true ability. As a result, both methods assume that the item parameters (e.g., item difficulty, item discrimination, and item guessing) are known, and use the individual’s estimated observed response pattern to estimate individual ability. Newton’s method for MLE differs from the traditional method of MLE by incorporating the standard error of the estimate (SEE), making the overall model more complex and potentially more accurate (de Ayala, 2009).

The first step of MLE and Newton’s method for MLE is to calculate the probability of a response pattern for the collection of items in a measure (de Ayala, 2009). This is defined as:

\[
p(x|\theta, \vartheta) = \prod_{j=1}^{L} p_j(\theta)^{x_j} (1 - p_j(\theta))^{(1-x_j)}
\]

(6)

where \(p(x|\theta, \vartheta)\) stands for the probability of belonging to a response vector \(x\), given \((\theta, \vartheta)\), the individual’s ability level (\(\theta\)), and item parameter matrix \((\vartheta; de Ayala, 2009)\). \(\Pi\) is the product symbol, \(L\) is the total number of items for the instrument, \(j\) is the item number, \(p_j\) stands for \(p(x_j|\theta, \alpha, \delta_j)\), and \(x_j\) is the response to item \(j\). When responses are present the above expression becomes a likelihood function assessing the likelihood of an individual having a particular response vector. This is expressed as:

\[
L(x_j|\theta, \vartheta) = \prod_{j=1}^{L} p_j^{x_j} (1 - p_j)^{(1-x_j)}
\]

(7)

where \(i\) stands for the individual.

And
\[ \ln L(x_i | \theta, \varrho) = \sum_{j=1}^{l} \left( x_{ij} \ln(p_j) + (1 - x_{ij}) \ln(1 - p_j) \right) \] 

(8)

where \( \ln L(x_i | \theta, \varrho) \) is the log likelihood function.

The following equations are used to calculate individual ability, however, MLE and Newton’s MLE differs in how the equations are used (de Ayala, 2009). MLE will conduct a binary search of \( \ln L \) for different ranges of \( \theta 's \). MLE tests these ranges by creating a lower and an upper bound. This bound can comprise of any \( \theta \) levels, but for demonstration purposes our lower bound will be -2 and the upper bound will be 2. MLE then determines the middle point between the bounds (\( \hat{\theta} \)), and determines if \( \ln L \) is above or below. Because the middle point between -2 and 2 is 0, MLE tests to see if \( \ln L \) is above or below \( \hat{\theta} \). If the \( \ln L \) is above \( \hat{\theta} \) then the new lower bound will be set to 0 and the upper bound will be set to 2. This bound is the split resulting in \( \hat{\theta}_1 = 1 \). The \( \ln L \) is then tested against these new bounds. MLE will continue creating new bounds until \( \ln L \) finds the \( \theta \) that has the highest degree of accuracy.

Even though MLE can be used to estimate individual ability it does not provide the standard error of the estimate (SEE; de Ayala, 2009). Unlike MLE, Newton’s method for MLE has the ability to calculate the SEE by applying the maximum likelihood method. Newton’s method uses the \( \ln L \), but instead of looking at where the bounds can no longer be bisected, this method looks at creating smaller right triangles. Right triangles are created until the hypotenuse lays flat on the arch of the curve.

This is accomplished by first estimating the maximum location, and will continue until the vertical line of the triangle (\( \Delta Y \)) equals 0 (de Ayala, 2009). To improve the initial estimated equation the following formula is applied to determine an individual’s maximum \( \theta \):

\[ \theta^{t+1} = \theta^t - \frac{f(\theta^t)}{f'(\theta^t)} \] 

(9)
where \( t \) is the \( t \)-th iteration, \( t = 1 \ldots T \), \( T \) is the maximum number of iterations, \( f(\theta^t) \) is the Equation 5 is implemented until \( \frac{f(\theta^t)}{f'(\theta^t)} = 0 \), so that \( \hat{\theta}^{t+1} = \hat{\theta}^t \); however, because this is difficult to accomplish, when \( \hat{\theta}^{t+1} - \hat{\theta}^t < .001 \) then the solution has converged.

When applying Newton’s method to IRT, equation 5 is replaced with the following formula to incorporate log likelihood into the estimation of the maximum \( \theta \) (de Ayala, 2009):

\[
\theta_i^{t+1} = \theta_i^t - \frac{\partial}{\partial \theta_i} \ln L(x|\theta_i^t) - \frac{\partial^2}{\partial \theta_i^2} \ln L(x|\theta_i^t)
\]

Equation 6 is then altered based on the selected IRT Model (e.g., 3pl), as \( \frac{\partial}{\partial \theta} \ln L(x|\theta_i^t) \) and \( \frac{\partial^2}{\partial \theta^2} \ln L(x|\theta_i^t) \) varies based on the IRT Model. Regardless of the IRT model chosen, the altered Equation 6 is applied until convergence or the maximum number of iterations is reached. Once either occurs, the person’s ability level, \( \theta \), is said to be estimated.

In addition to individual ability, item difficulty and item discrimination are also important to IRT. As previously mentioned, in CTT item parameters are sample dependent, which can create a problem when generalizing these parameters. Unlike CTT, IRT’s approach to estimating item difficulty and item discrimination is not sample dependent. Looking back at item difficulty for CTT, it determined the difficulty of an item based on the proportion of respondents that correctly answered the item (Kline, 2005). However, CTT’s method does not incorporate the ability level for those who correctly answered each item (Lord, 1980). By incorporating individual ability, the interpretation of item difficulty is based on the \( \theta \) that has more than a 50% chance of answering the item correct (Bock, 1997). As a result, the more difficult an item is, the more knowledgeable someone would need, or have a higher \( \theta \) (ability level), in order to be able
to answer the question correct. However, an easier item would not require a high amount of knowledge, and would require a lower $\theta$.

IRT also calculates and uses item discrimination differently than CTT (de Ayala, 2009). Unlike CTT, IRT uses a method similar to Newton’s method of MLE by incorporating individual ability into the calculation. Also, item discrimination is used in the IRT model calculations in advanced models. Item discrimination assesses how much an individual item can differentiate among different ability levels (de Ayala, 2009). In essence, item discrimination will suggest if ability levels matter for a particular item, and how quickly the odds of correctly answering an item will increase with ability (de Ayala, 2009). Because of this, item difficulty and item discrimination are often correlated with more difficult items having a higher discrimination.

IRT also has many different models to account for the different parameters that can be included. Even though item difficulty, item discrimination, and item guessing are a part of IRT, every model incorporates a different number of parameters, as some include all three parameters and others include only one parameter. Additionally, models differ based on the outcome that is being estimated. There are two different types of outcomes used in IRT: dichotomous data (e.g., yes/no, correct/not correct), and polytomous data (e.g., likert, multiple choice). However, in this study only dichotomous outcome models will be explained. These models include: the Rasch model, one parameter logistic model, two parameter logistic model, and three parameter logistic model.

**IRT Models**

**Rasch Model.**

A Rasch model is the most basic IRT model, as it does not consider item discrimination or item guessing (Schmidt & Embretson, 2003). As a result, this model will only inform us the
difficulty of particular items, and will not provide any information with regard to how well an item discriminates, or the likelihood an item can be answered by guessing. Even though this model does not provide as much information as more complex models, a Rasch model should always be conducted. This is because it will provide a baseline model, which can be compared to more advanced models to determine if adding more item parameters improves model fit. The formula for the Rasch model is (de Ayala, 2009):

\[ P(x_{ij} = 1|\theta_i, b_j) = \frac{e^{(\theta_i - b_j)}}{1 + e^{(\theta_i - b_j)}} \] (2)

where \(x_{ij}\) is the response \(X\) made by person \(i\) for item \(j\), \(x_{ij} = 1\) refers to a correct response or endorsing the item, \(\theta\) is the ability of person \(i\), \(b\) refers to the difficulty of item \(j\), \(e\) is the base of the natural logarithm \((e = 2.718 \ldots)\), and item discrimination is set to 1. In essence, this model provides an estimate of the likelihood of a correct response taking into consideration both individual ability and item difficulty.

To interpret the results of the Rasch model it is necessary to analyze the item the \(\theta\) indicated in the output, as well as the item characteristic curve (ICC). When analyzing the \(\theta\), the \(\theta\) that is provided is the lowest possible \(\theta\) an individual would need in order to have more than a 50% chance of answering the item correct. The ICC provides an illustration of the interpretation of \(\theta\). As illustrated in Figure 1, the items, represented by each line, show a similar S-shape pattern in the ICC graph for a Rasch model. This is due to the slope not varying between the items, as the items have the same item discrimination. If the discrimination were to vary each item would have a different slope, which would create different S-shape patterns (see Figure 3). Regardless, the ICC in Figure 1 shows the probability one would have to correctly answer the
item based on the person’s ability. For example, looking at item BS238Z in the teal color line, if a person had the ability level of -2, they would have a 20% probability of endorsing the item.

Another graph that is used to interpret the Rasch model is the item information curve IIC illustrated in Figure 2; however, it is not very informative, as the IIC graph is used to demonstrate item discrimination, which is constrained to one for all items. When interpreting the IIC graph, items with a taller peak provide more information (i.e., discrimination) than those items with a flatter line. Again, this is not demonstrated in Figure 2 as the item discrimination is not estimated for each item. However, the IIC graph for the Rasch models shows the ability that each item provides the most information for, which is located at the peak of the slope.

Figure 2.1. Item Characteristic Curve for Rasch model
The next type of IRT model is a one parameter logistic (1pl) model. This model is very similar to the Rasch model, as its ability to estimate item discrimination is limited and does not provide the guessing parameter (de Ayala, 2009). However, the 1pl model is distinct from a Rasch model, as it will estimate item discrimination for the instrument as a whole, rather than setting the values to one as in the Rasch model. Even so, the 1pl model is still limited, as it does not estimate item discrimination for individual items or allow variation amongst the items, compared to more complex IRT models. The formula for the 1pl model is (Hambleton, Swaminathan, & Rogers, 1991):
\[ P(x_{ij} = 1 | \theta, b) = \frac{e^{a_1(\theta_i - b_j)}}{1 + e^{a_1(\theta_i - b_j)}} \]  

(3)

where \( x_{ij} \) is the response \( X \) made by person \( i \) for item \( j \), \( x_{ij} = 1 \) refers to a correct response or endorsing the item, \( \theta \) is the ability of person \( i \), \( b \) refers to the difficulty of item \( j \), \( e \) is the base of the natural logarithm, and \( a \) is the item discrimination.

2 Parameter Logistic

A 2-parameter logistic IRT model (2pl) is the third type of model that is used. Unlike the Rasch and 1pl IRT models, the 2pl IRT model calculates item discrimination for all items (de Ayala, 2009). As a result, you could determine how well each item discriminates between different \( \theta' \)s, and how difficult each item is individually. That is item discrimination will differ for every item. A 2pl model is calculated as (de Ayala, 2009):

\[ P_j = \frac{e^{(1.7a_j(\theta_i - b_j))}}{1 + e^{(1.7a_j(\theta_i - b_j))}} \]  

(4)

where \( P_j \) is the probability of getting item \( j \) correct, \( a_j \) is the discrimination for item \( j \), \( \Theta_i \) is the ability of person \( i \), \( b_j \) is the difficulty for item \( j \).

Similar to a Rasch and 1pl model, to interpret the outcome of this model you would assess the item difficulty and item discrimination. However, unlike the other two models, item discrimination can be interpreted using the ICC and IIC graphs, as these will provide more information about the item discrimination and item difficulty parameters (de Ayala, 2009). Illustrated in Figure 3, the ICC graph shows item lines with different slopes unlike the Rasch model. This is variation in slopes is from incorporating item discrimination into the equation (de Ayala, 2009). If you were to compare the Rasch model ICC (Figure 1) and the 2pl ICC graph (Figure 3) you would notice that the two blue lines (blue and teal) no longer take the same shape.
as before in the 2pl graph. This would suggest that those items have a higher amount of item
discrimination. The 2pl IIC graph also differs from the Rasch model IIC graph. Comparing these
two graphs (Figure 2 and Figure 4) the Rasch IIC has the same shape for all item lines, but the
2pl IIC item lines have different shapes. Again, this is because item discrimination was
incorporated into the equation, which results in slopes of the item lines to differ (de Ayala,
2009). By comparing these two IIC graphs, the Rasch model does not provide any critical
information, but the 2pl model shows that the two items with the taller peak (blue and teal) have
higher item discrimination when compared to the other models.

Ultimately the Rasch and 2pl model can produce different item difficulty parameters due
to the 2pl model incorporating item discrimination into the equation (de Ayala, 2009). However,
like any model based analysis, model fit indices should be compared between the Rasch and 2pl
model to ensure that adding item discrimination increases model fit.

Figure 2.3. Item Characteristic Curve for 2PL model
3 Parameter Logistic

Last, a threeparameter logistic IRT (3pl) model is one of the more advanced forms of IRT as it calculates item difficulty, item discrimination, and item guessing for every item in the model (Kolen & Brennan, 2004). Thus, a 3pl model would be used if you were expecting some items to be answered correctly by guessing, or you have a justification for adding item guessing (de Ayala, 2009). By adding item guessing it can provide more information about each individual item, and it can produce slightly different results from the 2pl model. This is due to item guessing being incorporated into the equation. The 3pl model is calculated by (Kolen & Brennan, 2004):

\[ P_j = c_j + \left(1 - c_j\right) \frac{e^{1.7a_j(\theta - b_j)}}{1 + e^{1.7a_j(\theta - b_j)}} \]  

(5)
where $P_j$ is the probability of getting item $j$ correct, $c_j$ is the guessing for item $j$, $a_j$ is the discrimination for item $j$, $\Theta_i$ is the ability of person $i$, $b_j$ is the difficulty for item $j$.

**Differential Item Functioning**

IRT is an important component of differential item functioning (DIF) as IRT models and person and item characteristics are utilized for the implementation and interpretation of DIF. Specifically, DIF is a statistical tool that is used in conjunction with IRT to suggest if an individual item may provide a potential advantage to one group of individuals over another (e.g., males vs. females; Penfield & Camilli, 2007). These items might include: a question on a math test that requires knowledge of sports (advantage towards males), or a question on a reading passage that requires cultural knowledge (advantage towards one culture). If DIF were present for these questions in general it would suggest that an item provides an advantage for one group, either the focal group (i.e., the group of particular interest in the analysis) or the reference group (i.e., the group that is likely to have the advantage; Holland & Wainer, 1993). In essence, the existence of DIF suggests that there is a secondary factor influencing the performance on an item (Penfield & Camilli, 2007). This is shown, as the groups will have different item parameter values across the groups. Thus, the item difficulty parameter may be lower for the reference group and higher for the focal group. It is also important to note that the groups used for DIF detection are usually predefined based on the participants’ characteristics (e.g., gender, race). Although this is the norm, DIF can still be used when groups are not predefined by using Mixture IRT. Mixture IRT creates latent classes based on the participants’ response patterns, and then uses those clusters to determine if DIF is present for the items (Cohen & Bolt, 2005).

There are two main types of DIF that are commonly used: uniform and nonuniform. Uniform DIF occurs when the differential item performance conditioned on the measured trait is
consistent across the individual’s $\theta$. That is, the DIF does not change for particular $\theta$’s. This is illustrated in Figure 5, as it shows that the DIF is equal across the $\theta$’s for both the reference and focal group. Alternatively, nonuniform DIF can provide more information than uniform DIF, as it will show where the DIF is for the particular groups taking into consideration variations in $\theta$. Thus, the DIF for an item at a $\theta$ of -1.2 may be lower relative to a $\theta$ of one for the focal or reference group. Figure 6 illustrates a nonuniform DIF, as the DIF changes for the groups based on the $\theta$.

![Figure 2.5. Uniform DIF.](image-url)
Figure 2.6. Nonuniform DIF

**DIF Detection Methods**

When testing for DIF, there are a variety of methods that can be used when only two groups are present (e.g., gender). These methods include: Mantel-Haenszel (MH; Holland & Thayer, 1988), logistic regression (LR; Swaminathan & Rogers, 1990), SIBTEST (Shealy & Stout, 1993), Lord’s chi-square test (Lord, 1980), and the item response theory likelihood ratio test (IRT-LR; Thissen, Steinberg, & Wainer, 1988), just to name a few. Even though these methods have generally been shown to accurately assess DIF, they have typically been described in the context of only two groups. Two of these methods, which will not be included in this study, are SIBTEST and IRT-LR. Even though SIBTEST and IRT-LR were not used in this study, these methods will be explained to understand how DIF methods differ in identifying items as DIF.
SIBTEST is a two group (e.g., gender) DIF detection method that was proposed by Shealy and Stout (1993). In this method, DIF is assessed by first standardizing the two groups, so that both groups have a similar distribution. After standardization occurs, SIBTEST detects DIF by estimating what the expected differences are in the score between the two groups (e.g., male and female). This difference indicates the amount of DIF present between the two groups. Thus, if this difference does not equal zero then DIF is present.

The second method, IRT-LR, assesses DIF by comparing the IRT parameters of the two groups simultaneously (Thissen et al., 1988). To accomplish this, the item parameters between the groups for the item being assessed for DIF are constraint to be equal. The rest of the items are allowed to have their item parameters vary. The likelihood ratio test statistic is then used to test if the item parameters between the two groups are equal. If the likelihood ratio test statistic is significant this indicates that DIF is present. In the case of DIF being present, follow up analyses need to be conducted to determine if the DIF is due to the discrimination parameter or the difficulty parameter.

Although two group DIF detection is an important area of research there are many practical areas where researchers will want to assess DIF for multiple groups (e.g., international assessment studies, SES, ethnicity). For this reason, it is important to investigate DIF methods for multiple focal groups to determine under what conditions they accurately assess DIF. The methods that are currently used to assess DIF for more than two groups include: generalized Mantel-Haenszel (GMH; Penfield, 2001), generalize Logistic Regression (GLR; Magis, Raîche, Béland, & Gérard, 2011), Lord’s Chi-square test (Kim, Cohen, & Park, 1995) and the multivariate outlier approach (Magis & DeBoeck, 2011). Each of these methods differ, and
approaches DIF detection from a different angle. However, the multivariate outlier method takes a different approach to detecting DIF when multiple focal groups are present.

**Generalized Mantel-Haenszel.** GMH is a DIF detection method that utilizes the Mantel-Haenszel test statistics. Holland and Thayer (1988) and Narayanan & Swaminathan (1994) introduced Mantel-Haenszel for DIF detection with two groups. GMH is simply an extension of the original Mantel-Haenszel test statistic, but can be used to test for DIF across multiple focal groups by using an extension of the chi-square test of association (Somes, 1986). This is accomplished by setting one group as the reference, and the rest of the groups are set as focals. The chi-square distribution is then used to assess the null hypotheses of no DIF present across the groups. The GMH test statistic is calculated as:

\[ GMH = \left( n_j - \mu_j \right)'V^{-1}\left( n_j - \mu_j \right) \]  

(11)

Where \( n_j \) is the vector of observed number of targeted responses summed across individuals in the reference group with total score \( j \), \( \mu_j \) is the expected number of targeted responses across individuals in the reference group if DIF is not present with total score \( j \), and \( V \) is the covariance matrix of \( n \).

Using equation 11, if the GMH test were significant, the null hypothesis would be rejected. With the rejection of the null hypothesis the Mantel-Haenszel would be used as a follow up analysis. The Mantel-Haenszel is used as a pairwise comparison between the reference groups and each of the focal groups.

**Generalized Logistic Regression.** Logistic regression was a method that was proposed by Swaminathan and Rogers (1990) to assess for uniform DIF, and was later demonstrated by Narayann and Swaminathan (1996) for nonuniform DIF for two groups. The logistic regression method was further extended by Magis, Raîche, Béland, and Gérard (2011) in order to assess
DIF for multiple groups in the form of GLR. Magis et al. (2011) demonstrated the application of GLR to an English proficiency exam that was used for course placement. DIF was assessed using the GLR method, GMH, and Lord’s chi-square test across four groups (i.e., year the individual took the exam). Ultimately, all three methods identified the same items as exhibiting DIF. Although these methods had identical results, Magis et al. (2011) pointed to one distinctive characteristic of GLR. That is, GLR does not require a reference group to be selected, whereas other methods do require this. Due to this potential advantage of GLR, this method will be included in this study.

GLR assesses DIF for multiple groups by including intercepts and slopes for each group, the common intercept across the groups, the common slope across groups, and the predictor variable as the person’s ability. The reference group is also included in the logistic model; however, the reference group intercept and slope are constrained to 0, whereas the focal group parameters are freely estimated. DIF is signaled as being present when the focal group model parameters significantly deviate from 0. Specifically, when the intercept for at least one focal group significantly deviates from 0 (i.e. the constrained value of the reference group intercept) uniform DIF is said to be present. Similarly, when the slope for at least one focal group significantly deviates from 0, nonuniform DIF is present. The GLR model can be defined as:

\[
\ln \left( \frac{\pi_{ig}}{1-\pi_{ig}} \right) = \begin{cases} 
\alpha + \beta S_i & \text{if } g = R \\
(\alpha + \alpha_g) + (\beta + \beta_g)S_i & \text{if } g \neq R 
\end{cases}
\]

Where \( \pi_{ig} \) is the probability of a correct response for person \( i \) in group \( g \), \( \alpha_g \) is the intercept for group \( g \), \( \beta_g \) is the slope for group \( g \), \( S_i \) is the matching score for person \( i \), \( \alpha \) is the intercept common across groups, and \( \beta \) is the Slope common across groups.

Follow up analyses for the GLR method includes the Wald chi-square and the likelihood ratio test. Both methods are commonly used in conjunction with each other as both methods have
their advantages (Magis, Raîche, Béland, and Gérard, 2011). Specifically, the Wald statistics is not as stable when compared to the likelihood ratio test, but the Wald statistic provides a pairwise comparison of the groups creating a more precise detection of DIF (Agresti, 2002; Magis et al., 2011). Because of this, it is recommended that both methods should be used simultaneously (Magis et al., 2011). When the Wald chi-square and likelihood ratio test produce similar results a researcher would use the Wald chi-square results to interpret the results. However, if the results were inconstant between the two methods one would need to proceed with caution when interpreting the DIF results.

**Lord’s Chi-square Test.** The Lord’s chi-square test was originally proposed by Lord (1980) in order to assess if differences were found in item parameters between two groups. To accomplish this, a chi-square test is used to compare the vectors of the item parameters (e.g., item difficulty). To obtain the Lord’s chi-square test statistic the vector of maximum likelihood item parameter estimators \( v_{jk} \), as well as the asymptotic variance and covariance matrix for the vector \( \Sigma_{jk} \) is obtained for the focal and reference group with

\[
v_{jk} = (a_{jk}b_{jk})'
\]

and

\[
\Sigma_{jk} = \begin{pmatrix}
\text{var}(a_{jk}) & \text{cov}(a_{jk}, b_{jk}) \\
\text{cov}(a_{jk}, b_{jk}) & \text{var}(b_{jk})
\end{pmatrix}
\]

Where

- \( a_{jk} \) = Item discrimination for item j in group k
- \( b_{jk} \) = Item difficulty for item j in group k

\( v_{jk} \) and \( \Sigma_{jk} \) can then be used in the Lord’s chi-square statistic, which is described by Lord (1980) as:
\[ \chi^2 = (v_{j1} - v_{j2})' (\Sigma_{j1} + \Sigma_{j2})^{-1} (v_{j1} - v_{j2}) \]  

(15)

The \( \chi^2 \) statistic obtained in equation 15 for each item is then compared to the \( \chi^2 \) distribution with the degrees of freedom (df) set to the number of parameters used in the IRT model (e.g., 2pl model has 2 df). This statistic is assessing the null hypothesis that there are no differences in the item parameters across the groups using the chi-square distribution. That is a significant chi-square test statistic indicates DIF present between the reference and focal group for that item.

Kim, Cohen, and Park (1995) applied the Lord’s chi-square test for the situation when more than multiple focal groups are present. When using this extension the multi-group DIF statistics the Lord’s chi-square test takes the form of the quadratic term \( Q_t \). Additionally, item parameter estimates need to be calibrated on the same scale for all groups in order to comparisons to be made. To do this it is standard to use the reference group as a baseline, and calibrate the focal groups to the reference group. With using the calibrated item parameter estimates for each item \( Q_t \) is calculated as:

\[ Q_t = (C v_{i})' (C \Sigma C')^{-1} (C v_{i}) \]  

(16)

Where \( C \) is the contrast coefficient matrix, \( v_{i} \) is the estimators of the item parameters from the different group

The contrast coefficient matrix is an orthogonal matrix that, for most instances, compares the reference group to each focal groups. This would take form for three groups as:

\begin{bmatrix}
1 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & -1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & -1
\end{bmatrix}

By creating a C matrix it provides the information to test that the item parameters do not differ between the reference and focal groups. For the 2pl three groups case, the null hypothesis is:
\[
H_0 = \begin{bmatrix}
a_R = a_{F1} \\
b_R = b_{F1} \\
a_R = a_{F2} \\
b_R = b_{F2}
\end{bmatrix}
\]

\(a_R\) = item discrimination for the reference group.

\(b_R\) = item difficulty for the reference group.

\(a_F\) = item discrimination for the focal groups one and two.

\(b_F\) = item difficulty for the focal groups one and two.

The estimation of the item parameters for the different groups is defined as \(v_i\):

\[
v_i = (a_{i1}, b_{i1} \ldots a_{ig}, b_{ig})
\]

(17)

Where:

\(a_{ig}\) = Item discrimination values for item i in group g

\(b_g\) = Item difficulty values for item i in group g

The covariance matrix for the item parameter estimates is defined by \(\Sigma_i\) as:

\[
\Sigma_i = \begin{pmatrix}
\text{var}(a_{i1}) & \text{cov}(a_{i1}, b_{i1}) & \cdots & 0 & 0 \\
\text{cov}(a_{i1}, b_{i1}) & \text{var}(b_{i1}) & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \cdots & \vdots \\
0 & 0 & \cdots & \text{var}(a_{ig}) & \text{cov}(a_{ig}, b_{ig}) \\
0 & 0 & \cdots & \text{cov}(a_{ig}, b_{ig}) & \text{var}(b_{i1})
\end{pmatrix}
\]

(18)

Where

\(\text{var}(a_{ig})\) = Discrimination parameter variance for item i in group g.

\(\text{var}(b_{ig})\) = Difficulty parameter variance for item i in group g.

\(\text{cov}(a_{ig}, b_{ig})\) = Discrimination and Difficulty parameters for item i in group g.

**Multivariate Outlier Approach.** The methods described up to this point all require scale purification in order to avoid inflation in Type I error. Purification is an iterative process of removing items identified as containing DIF from the matching scale. This process is necessary...
in the previous DIF methods, as if DIF items are not removed from calculation of the total score there is a reduction in accuracy when detecting DIF in the targeted item. Seeing purification as a potential disadvantage of the current methods used to detect DIF, Magis and DeBoeck (2011) proposed using a multivariate outlier approach to detecting DIF. In the multivariate outlier approach DIF is assessed for the items individually, and is based on their multivariate distribution rather than using a matching score. For this reason, scale purification is not necessary (Magis & DeBoeck, 2011).

The multivariate outlier approach is based upon the notion that when DIF is present for a specific item on a scale, this DIF item will be an outlier when it is compared across the non-DIF items on the scale. In order to use the multivariate outlier approach, a standard DIF statistic is first computed, such as GMH. This statistic is then transformed to the normal distribution so that it can be used with standard outlier detection methods. After the DIF statistic is calculated a modified Mahalanobis distance is calculated using a robust estimate (selected by the researcher) of the covariance matrix. These robust estimators include the Minimum Covariance Determinant (MCD; Rousseeuw & van Driessen, 1999), Orthogonalized Gnanadesikan-Kettenring (OGK; Gnanadesikan & Kettenring, 1972), Stathel-Donoho Robust estimator (SDE; Rocke, 1996), and the M estimator (MEST; Maronna & Yohai, 1995). All of these estimators will be included in this study. The modified Mahalanobis distance is then calculated as:

\[
\tilde{\phi}_j^* = \frac{1}{(j-1)^2} (z_j - \bar{z})' R^{-1} (z_j - \bar{z})
\]

(19)

Where:

J = the number of items

\(z_j\) = DIF statistic for item \(j\)
\( z = \) robust estimate of the mean vector  
\( R = \) robust estimate of the covariance matrix

This statistics is then compared to the quantile of the beta distribution for the alpha level \((Q_{\alpha})\). If the modified Mahalanobis distance is greater than \(Q_{\alpha}\) then item \( j \) is flagged as DIF.

The multivariate outlier approach is different than the other methods, as this method defines items as DIF when the DIF statistic for an item is an outlier from the multivariate distribution. Due to this there are two potential advantages to this approach. First, the multivariate outlier approach does not require an iterative purification process unlike the other approaches. Purification is an iterative process that removes items that are flagged as containing DIF from the pool of test items in order to create a unidimensional pool of items (Lord, 1980). The remaining items are then used to estimate \( \theta \) before the final DIF analysis for a more accurate DIF detection (Clauser, Mazor, & Hambleton, 1993). Even though purification appears as an appropriate step to take to ensure that one does not receive a biased \( \theta \), there is not a consensus in the field. This is due to some research suggesting that purification does not impact the end results (Magis & Facon, 2012), and research has not determined what method is the best method for purification (Hidalgo-Montesinos & Gómez, 2003; Magis & Facon, 2012; Wang & Su, 2004). Second, it is expected that there will not be a Type I error inflation. This is due to Magis and DeBoeck (2012) discovering via a simulation study that the robust outlier approach did not have an inflated Type I error for the case of assess DIF for two groups when compared to other methods. Magis and DeBoeck (2012) suspects that the decrease in Type I error is due to the fact that the multivariate outlier approach does not replace the ability level with the raw score unlike other methods. For this reason, Magis and DeBoeck (2012) expect this to be the case for when multiple focal groups are present. Although the multivariate outlier approach has these potential
advantages this method has not yet been assessed via a simulation study nor compared to the other DIF methods in regard to accuracy. Due to this, the focus of this simulation study will be to compare the multivariate outlier approach to GMH, GLR, and the Lord’s Chi-square test.

**Previous Research**

**Two groups.** DIF research focusing on two groups is more extensive in comparison to DIF with multiple focal groups. These studies have focused on determining what DIF methods (e.g., SIBTEST, IRT-LR, MH, LR) perform best based on Type I error and power and when do those methods perform optimally under different simulation conditions. Lord’s chi-square test for DIF has been assessed in multiple simulation studies under a variety of conditions, but has rarely been compared to the other DIF methods that have been discussed (Cohen & Kim, 1993; Kim, Cohen, & Kim, 1994; Kim & Cohen, 1995; McLaughlin & Drasgow, 1987). Overall research demonstrated that the Type I error rate for Lord’s chi-square only has an inflated Type I error under certain conditions. These conditions include: using a 3PL IRT model, small sample sizes, and shorter tests (Cohen & Kim, 1993; Kim, Cohen, & Kim, 1994; Kim & Cohen, 1995). Even though Lord’s chi-square test performs relatively well, Kim and Cohen (1995) had a concern about the accuracy of this method. This is due to the fact that if the accuracy when estimating the covariance matrix is low then the Lord’s chi-square test will produce inaccurate DIF items.

Other research has focused on comparing LR and SIBTEST. Of the studies that compared LR and SIBTEST, the results have shown LR to have higher power than SIBTEST when detecting nonuniform DIF (Li & Stout, 1996; Narayanan & Swaminathan, 1996). Li and Stout (1996) found that LR has higher power than SIBTEST particularly when the magnitude of DIF was large. However, Narayanan and Swaminathan (1996) results suggested that LR power is
higher when the DIF contamination was lower and that SIBTEST was unaffected by the contamination. In terms of power, both studies found elevated Type I error, however, the elevated Type I error occurred for both LR and SIBTEST in Narayanan and Swaminathn’s study, but was only present for LR in Li and Stout (1996).

Research in IRT-LR is even more limited than the previous methods. The reason being it is time-consuming to calculate IRT-LR and software was not developed until the early 2000’s to easily calculate this statistic (Thissen, 2001). Regardless, research has shown that IRT-LR has an elevated Type I error rate under certain conditions (Cohen, Kim, & Wollack, 1996; Finch, 2005; Wang & Yeh, 2003). The inflated Type I error was triggered from a lower sample with the use of a 3PL model (Cohen, Kim, & Wollack, 1996), and when there is contamination is the anchor (i.e., matching subset) items (Finch 2005; Wang & Yeh, 2003). Finch and French (2007) expanded the research of IRT-LR by comparing it to SIBTEST and LR when detecting uniform DIF under a variety of conditions. Overall, the results of the study found the SIBTEST had the best control for Type I error and power (Finch & French, 2007).

There has been considerable amount of simulation research comparing the Type I error and power between MH and LR. The results across the simulation studies and conditions (i.e., sample size, impact, amount of DIF) showed that overall MH had a lower Type I error rate than LR (de Ayala et al., 2002; DeMars, 2009; Guler & Penfiled, 2009; Herrera & Gomez, 2008; Kim & Oshima, 2012; Li, Brooks, & Johanson, 2012; Rogers & Swaminathn, 1993; Swaminathan & Rogers, 1990; Vaughn & Wang, 2010). Even so, the amount of Type I error was effected by impact, sample size, and group sizes. Some of the interesting findings in these studies occurred with the simulation conditions set to no impact, equal group sizes, and a medium sample size (i.e., 500 - 700). Under these conditions, MH had a lower Type I error in most of the studies, but
both MH and LR had instances when the Type I error exceeded .05. For MH, one study showed that the Type I error rate was greater than .05 (Herrera & Gomez, 2008), whereas for LR there were four studies (Herrera & Gomez, 2008; Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). Additionally, MH did not perform better than LR when there was a large sample (1,000 - 2,500) and an impact of 0. In this simulation condition Vaughn and Wang (2010) and Guler and Penfield (2009) results showed the same Type I error for MH and LR, and MH and LR had a Type I error rate higher than .05 in the three other studies (DeMars, 2009; Herrera & Gomez, 2008; Kim & Oshima, 2012). Two of these three studies found that MH had the higher Type I error (Herrera & Gomez, 2008; Kim & Oshima, 2012). MH also performed better or equal to LR when the group sizes were unequal in all the mentioned studies except for Vaughn and Wang. Even though MH has shown to perform better than LR in regard to Type I error in these studies, LR performs better in regard to power. In terms of power, LR performed better that MH six out of the seven studies (de Ayala et al., 2002; Kim & Oshima, 2012; Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Vaughn & Wang, 2010). The only study that MH had higher power was in Herrera & Gomez.

**Multiple focal groups.** Unlike the research regarding DIF for two groups, the prior research for DIF with multiple focal groups is limited, but still present. Penfield (2001) investigated the utility of three types on Mantel-Haenszel tests: the Mantel-Haenszel procedure, Mantel-Haenszel with a Bonferroni adjusted alpha, and GMH. Penfield’s study compared these three methods in the case for multiple focal groups. This investigation utilized a variety of simulation conditions. Some of these conditions included: equal and unequal group sizes, as well as one, two, three, and four focal groups. Overall, Penfield suggested that GMH had the best
control for Type I error and preserving power. However, GMH demonstrated low power when DIF was present in all focal groups. Regardless of this potential limitation of GMH, this method has shown promise in detecting DIF for when multiple focal groups are present.

As for the Lord’s chi-square method, Kim, Cohen, and Park (1995), applied this method when multiple focal groups are present. This study applied the Lord’s chi-square method to a mathematics exam, which included three groups of 200 university students. The three groups were determined based on whether or not the student could use a calculator. Specifically, one group was not allowed to use a calculator (reference group), and the other two groups used one of the two scientific calculators assigned in the study (focal group). Using this data, the Lord’s chi-square method identified two of the 14 items as containing DIF. Next, pairwise comparisons were conducted to determine where between which group the DIF was present.

Magis, Raîche, Béland, and Gérard (2011) expanded on this research by comparing GLR, GMH and Lord’s chi-square test when assessing uniform and nonuniform DIF on an English proficiency course placement exam. The results of the DIF analyses showed that all methods identified the same items as showing DIF. Although this is the case, Magis et al. (2011) suggests that GLR has advantages over the other methods, as GLR does not require the researcher to select a reference group.

As shown, most of the research for multiple focal groups has not focused on comparing multiple types of methods in one simulation study. However, Finch (2015) compared GMH, GLR, and Lord’s chi-square in one simulation study. This study compared these methods Type I error and power by manipulating the number of groups, sample size, group size ratio, level of DIF, and impact. The results of the simulation study found that of these three methods the Lord’s chi-square method had the lowest power in most of the conditions, but GMH and GLR’s power
were comparable. However, Lord’s had the lowest Type I error, followed by GMH. Lord’s was also the only method that did not have Type I error exceeding .05. From the results, of the three methods mentioned Finch (2015) recommended the use of GMH due to having the highest power, and good control of the Type I error rate.

Even though Finch (2015) compared DIF methods for DIF with multiple focal groups, Finch’s study did not include the multivariate outlier approach. The multivariate outlier approach was first presented by Magis and DeBoeck (2011). In their study Magis and DeBoeck (2011) demonstrated the utility of the multivariate outlier approach by applying this method to the data from Kim et al. (1995) study. As a reminder Kim et al. (1995), applied the Lord’s chi-square method to a math exam with three different calculator groups. To assess DIF using multivariate outlier approach, Magis and DeBoeck (2011) used five different methods of robust outlier estimates: Donohn-Stahel estimator, the constrained M estimator, the MCD estimator, and the OGK estimator. These five methods approach detecting outliers slightly different in order to determine which items have DIF present. Amongst these five outlier methods, all identified the same items as having DIF as Kim et al. (1995). From their application study, they suggest that the multivariate outlier approach may be a useful method to detect DIF when multiple focal groups are present. This is due to producing similar results as Kim et al. (1995), the fact that the multivariate outlier approach does not require purification, and robust statistics are common in other fields. Even so, a simulation study needs to be conducted in order to determine if the multivariate outlier approach can accurately identify DIF items. For these reasons, the focus of this study will be to determine how the multivariate outlier approach compares to other DIF methods.
Research Questions

This study will be used to evaluate the performance for four DIF detection methods: GMH, GLR, Lord’s Chi-square test, and the multivariate outlier approach (MOA). However, due to prior research focusing on the first three methods, the multivariate outlier approach will be the main focus of this study. To accurately assess these methods, three research questions with hypotheses and two additional exploratory research questions will be used.

1. Is there a difference in power between GMH, GLR, Lord’s when compared to MOA DIF methods?

Hypothesis: For this research question, due to Finch (2015) finding that GMH had the highest power of all non MOA methods, it is expected that in this study GMH will have the highest power of all of the non MOA methods. For MOA methods, there is no particular expectation for this method, as no previous research has been conducted on MOA.

2. Is there a difference in Type I error between GMH, GLR, Lord’s when compared to MOA DIF methods?

Hypothesis: Due to Magis and DeBoeck (2011) anticipating that MOA should control for Type I error, it is expected that MOA will have the best Type I error out of all of the methods. In regard to GMH, GLR, and Lord’s it is expected that these methods will not always control for Type I error in all conditions. This is consistent with the results of Finch (2015).

3. Is there a difference in power between GMH, GLR, Lord’s when compared to MOA DIF methods as the number of items containing DIF increases?
Hypothesis: It is expected that MOA’s power will reduce as more items contain DIF. This is because as more items are contaminated with DIF, it is expected that MOA will have a difficult time distinguishing all DIF items as an outlier. That is, due to the high number of DIF items, those items will no longer be outlier and thus, will not be identified

4. Does the level of DIF have an impact on power amongst GMH, GLR, Lord’s when compared to MOA DIF methods?

5. Does impact, and/or sample size have an effect on the Type I error and power for GMH, GLR, Lord’s when compared to MOA DIF methods?
CHAPTER III
METHODOLOGY

The present study was designed to assess the effectiveness of DIF detection methods in
the presence of uniform DIF when more than two focal groups are present. To adequately
compare these methods a simulation study was conducted using Mplus version 7.11 (Muthén &
Muthén, 2013) for data generation, and R statistical software (R Development Core Team, 2012)
to compare the analyses on their Type I error and power. Various simulation conditions were
considered for this study (see Table 1).

Table 1. Simulation Conditions Considered for the Study

<table>
<thead>
<tr>
<th>Variable</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Groups</td>
<td>Two, Three, Six</td>
</tr>
<tr>
<td>Reference Sample Size</td>
<td>500, 1000, 2000</td>
</tr>
<tr>
<td>Group Size</td>
<td>Equal Group Sizes</td>
</tr>
<tr>
<td></td>
<td>Unequal Group Sizes</td>
</tr>
<tr>
<td>Level of DIF</td>
<td>0, 0.4, 0.6, 0.8</td>
</tr>
<tr>
<td>DIF Method</td>
<td>Generalized Mantel-Haenszel test,</td>
</tr>
<tr>
<td></td>
<td>Generalized Logistic Regression, Lord’s</td>
</tr>
<tr>
<td></td>
<td>Chi-square test, Multivariate Outlier</td>
</tr>
<tr>
<td></td>
<td>Approach</td>
</tr>
<tr>
<td>DIF contamination</td>
<td>0%, 10%, 20%, 40%</td>
</tr>
<tr>
<td>Impact (mean latent trait group difference)</td>
<td>0, 0.5, 1.0</td>
</tr>
<tr>
<td>Item Two contamination</td>
<td>Yes, No</td>
</tr>
<tr>
<td>Number of Items</td>
<td>20</td>
</tr>
<tr>
<td>Target Item</td>
<td>Item Two</td>
</tr>
<tr>
<td>Number of iterations</td>
<td>1000</td>
</tr>
</tbody>
</table>

Simulation Conditions

Methods of DIF detection. There were four DIF detection methods used in this study:
GLM, GLR, Lord’s Chi-square test, and the multivariate outlier approach. For the Lord’s chi-
square method item parameters were calibrated to match the same scale for the reference group
Additionally, purification was used for GLM, GLR, and Lord’s Chi-Square test. Purification was not used for MOA, as this method does not require purification.

In R, the difR package (Magis, Beland, & Raiche, 2015) was used for GMH (function difGMH), GLR (function genLogistik), and Lord’s (function genLordChi2). The rrcov package (Todorov & Filzmoser, 2009) was used for the multivariate outlier approach with the CovOgk, CovMcd, CovSde, and CovMest functions. These functions served as different multivariate outlier estimates for the three and six group simulation conditions.

**Number of Groups.** The number of groups used in DIF has commonly focused on only two groups; however, this study sought to explore DIF methods that can assess DIF with multiple groups. Due to this, the two groups case was included to provide a baseline of performance for the more traditional methods, which has been commonly studied for the MH, LR, and Lord’s statistics (Finch, 2015; Guler & Penfiled, 2009; Herrera & Gomez, 2008; Kim & Cohen, 1995; Li & Stout, 1996; Thissen, 2001). To expand the current research and assess the methods ability to evaluate DIF with more than two groups, three and six group conditions were included. These two group conditions were selected, as three and six groups are group sizes that can be found in practice with SES and ethnicity. Additionally, previous research has included these group sizes in prior simulation research (Finch, 2015; Kim, Cohen, & Park, 1995; Penfield, 2001).

**Reference Group Sample Size and Group Size Ratio.** In IRT research, sample sizes are important and require a larger sample size than most statistical analyses. This is because a small sample size can result in an IRT model to not converging. For this reason, sample sizes included in this study were selected to assess a wide range of sample sizes that are seen throughout research, and were used to reflect a small, medium, and large sample sizes found in
prior DIF research (Chen & Jiao, 2014; Hauger & Sireci, 2008; Smith, Armes, Richardson, & Stark, 2013; Waiyavutti, Johnson, & Deary, 2012; Wu, King, Witkiewitz, Rac, McMahon, 2012). Additionally, by including such sample size, this study was able to evaluate the methods based on prior research showing that sample sizes can impact Type I error and power (Kim, Cohen, & Cohen, 1994; Vaughn & Wang, 2010; Rogers & Swaminathan, 1993; Kim & Oshima, 2012).

These sample sizes were also used for the equal group size and unequal group size condition; however, the reference group sample sizes dictated the sample sizes for the focal group in the unequal group size condition. Specifically, for the unequal group size condition, the reference group is two times larger than the focal groups. For example, in the six groups unequal group size with the reference group of 2000 condition, each focal group has 1000 cases, thus resulting in a total sample size of 7000. This condition was included in the simulation, as research has shown that unequal group sizes can impact Type I error and power (de Ayala et al., 2002; DeMars, 2009; Finch, 2015; Guler & Penfiled, 2009; Herrera & Gomez, 2008; Kim & Oshima, 2012; Li, Brooks, & Johanson, 2012; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). Additionally, when applying DIF to collected data, it is common to have unequal groups (Chen & Jiao, 2014; Hauger & Sireci, 2008; Mariniello, 2009; Wolf & Leon, 2009; Waiyavutti, Johnson, & Deary, 2012; Wu et al., 2012).

**Level of Uniform DIF.** For this study, data were simulated using wide range of levels of DIF. Level of DIF was used to assess the DIF methods when data had various magnitudes of DIF, similar to effect size. The 0 DIF condition resulted in no changes in the item difficulty parameters for the reference group; however, the other three conditions resulted in altered item difficulty parameters. Using the additional three conditions allows the DIF methods to be
compared when a small, moderate, and large amount of DIF are present, as well as when no DIF is present. Additionally, these conditions were selected, as they have been previously used in research (Finch, 2015; Rogers & Swaminathan, 1993; Penfield, 2001; Rogers & Swaminathan, 1993).

To obtain the altered item difficulty parameters for the items that were simulated as DIF being present, item difficulty parameters were calculated by taking the item difficulty parameters minus -0.28284 for the 0.4 condition, -0.42426 for the 0.6 condition, and -0.6784 for the 0.8 condition. The purpose of doing this is to calculate item difficulty values with different levels of DIF. The item difficulty parameters are provided for all levels of DIF for those items that may have DIF present in the item (see Table 2).

**DIF contamination.** DIF contamination reflects the proportion of items that were simulated with DIF. These four conditions were selected to assess how the DIF methods perform when no items have DIF, when a small (10% - two items), moderate (20% - four items), and large (40% - eight items) number of items are contaminated with DIF. DIF contamination is an important factor as it should directly impact the performance of GMH, GLR, Lord’s chi-square method. This was due to prior research findings that these methods have a higher Type I error and lower power when more items contain DIF (Finch, 2015; Finch & French, 2007; Narayanan & Swaminathan, 1994; Penfield, 2001; Rogers & Swaminathan, 1993). Also, the proportion of items containing DIF should impact the ability of the outlier detection approach to accurately identify DIF items as outliers. For the multivariate outlier approach, it was expected that this method would have a more difficult time detecting items as containing DIF when more items have DIF present. This is because when there are more items containing DIF it reduces the chance of the items as being identified as an outlier.
**Target Item.** Within DIF simulation research, to determine the power and Type I error of the methods, information regarding whether items were flagged as DIF must be obtained. To make this feasible, item two was selected as the target item, and this item’s DIF information was used to assess the power and Type I error of the methods. That is, whether or not item two was correctly or incorrectly identified as an item containing DIF was used to determine the power and Type I error rate. In order to assess both power and Type I error, three conditions were utilized. The first condition was when no items contain DIF, including item two (Type I error). The second condition was when other items contain DIF, but item two does not contain DIF (Type I error). And the last condition was when item two contains DIF along with other items (power). By including this simulation condition Type I error of item two could be assessed when DIF was present in other items. When the second target item condition was present, item 10 was simulated to have DIF present with 40% DIF contamination (see Table 2).
<table>
<thead>
<tr>
<th>Item</th>
<th>0</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>-0.28284</td>
<td>-0.42426</td>
<td>-0.6784</td>
</tr>
<tr>
<td>3</td>
<td>0.86</td>
<td>0.57716</td>
<td>0.43574</td>
<td>0.1816</td>
</tr>
<tr>
<td>4</td>
<td>0.08</td>
<td>0.46716</td>
<td>0.32574</td>
<td>0.0716</td>
</tr>
<tr>
<td>5</td>
<td>-0.72</td>
<td>-1.00284</td>
<td>-1.14426</td>
<td>-1.3984</td>
</tr>
<tr>
<td>6</td>
<td>1.39</td>
<td>1.10716</td>
<td>0.96574</td>
<td>0.7116</td>
</tr>
<tr>
<td>7</td>
<td>-0.91</td>
<td>-1.19284</td>
<td>-1.33426</td>
<td>-1.5884</td>
</tr>
<tr>
<td>8</td>
<td>-0.43</td>
<td>-0.71284</td>
<td>-0.85426</td>
<td>-1.1084</td>
</tr>
<tr>
<td>9</td>
<td>-0.30</td>
<td>-0.58284</td>
<td>-0.72426</td>
<td>-0.9784</td>
</tr>
<tr>
<td>10</td>
<td>-1.19</td>
<td>-1.47284*</td>
<td>-1.61426*</td>
<td>-1.8684*</td>
</tr>
<tr>
<td>11</td>
<td>1.91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>-0.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>-0.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>-0.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>-0.40</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Blank rows indicate no change in item difficulty parameter. * was only used for when item two contamination condition was present.

**Impact.** Because ability levels in groups are not always constant between groups, it was important to include impact in the current study. Specifically, two main conditions were included in the study: 1) ability levels across the groups are equal, and 2) a group has a higher average ability compared to the other groups. When impact was not present, the groups were simulated to have a latent trait ability of 0, but when impact was present one group was simulated to have a latent trait ability of 0 and the other groups had an average ability of 0.5 or 1.0. As an example, when impact of 0.5 was present, the latent variable means were set to 0, 0.5 for two groups, 0, 0.5, 0.5 for three groups, and 0, 0.5, 0.5, 0.5, 0.5, 0.5 for six groups.
**Robust Estimators.** For MOA methods, a variety of robust estimators were used. This was to account for potential differences of Type I error and power amongst the estimators. These robust estimators included: Minimum Covariance Determinant (MCD; Rousseeuw & van Driessen, 1999), Orthogonalized Gnanadesikan-Kettenring (OGK; Gnanadesikan & Kettenring, 1972), Stathel-Donoho Robust estimator (SDE; Rocke, 1996), and the M estimator (MEST; Maronna & Yohai, 1995).

**IRT and DIF Conditions.** After crossing all of the manipulated factors, a total of 1062 different simulation conditions were included in this study. For all combinations of simulation conditions, 1000 data sets were generated using 20 dichotomous items with the 1PL item response theory model. A 1PL model was utilized in this study to expand on previous research by Finch (2015) that used a 2PL model to compare non-MOA methods (GMH, GLR, Lord’s). Uniform DIF was used in this study to include all DIF methods, as GMH method cannot assess nonuniform DIF. Additionally, the item difficulty values have been acquired from a previously calibrated large statewide achievement assessment, ranging from -1.06 and 1.39.

**Study Outcome.** For this study, two outcomes were used: Type I error and power for the DIF detection methods. Type I error was used to assess if the methods detect DIF in item two when in fact DIF was not present. Power was used to assess if the methods accurately detected item two as a DIF item. This was assessed for all conditions where DIF was present, except for when item two was simulated to have no DIF when DIF was present in the other items. This latter condition did not focus on power, as the focus for that condition was if the methods detected DIF for item 2, but when no DIF was present.

**Repeated Measures ANOVAs.** Six repeated measures ANOVAs were used in order to assess the impact of the manipulated study factors on the power and Type I error of DIF
detection in this study. One repeated measures ANOVA was run for each combination of group condition (two, three, and six) by study outcome (Type I Error and power). This separation of the ANOVAs accounts for the different robust estimators used in the two group condition and problems with convergence of the MOA approach when six groups were present. For the two group condition the robust estimators that were used were Mahalanobis distance based on the beta distribution, Mahalanobis distance based on the chi-square distribution, and Mahalanobis distance based on the robust beta distribution. Robust estimators were modified in order to account for multiple groups. For this reason, the robust estimators used for three and six groups are, ogk beta, ogk chi, mcd beta, mcd chi, sde beta, sde chi, m beta, m chi. Additionally, the three and six groups condition was not included in one repeated measures ANOVA due to the six groups condition not converging when the strength of DIF was .8. Thus, the group condition could not be accurately compared as a result of these complications.

Each repeated measures ANOVA included either Type I error or power as the dependent variable with the methods serving as the repeated condition (see Table 3). For both Type I error and power, the values were obtained by averaging the Type I error and power across the 1000 iterations for each combination of conditions. The between-subjects effects were sample size, group size ratio, level of DIF, DIF contamination, and impact. For each between subject effects two-way interactions (e.g., method*sample size), and three-way interactions were included into the model (e.g., method*sample size*impact). Four-way interactions were not included, as the results did not add anything meaningful that was not already included in the three-way interaction. Additionally, only three-way interactions are interpreted in the results, as all between-subjects effects are explained within the the three-way interactions. This removes the need to interpret the two-way interactions. The within-subjects effect was the method.
Additionally, Greenhouse-Geisser was used, as the assumption of sphericity was violated for all ANOVAs. When sphericity is violated it indicates that the variances of differences across the methods are not equal. Because of this, the Greenhouse-Geisser was used to correct for the unequal variances. Specifically, Greenhouse-Geisser corrects the degrees of freedom in the repeated measures ANOVA in a conservative manner.
<table>
<thead>
<tr>
<th>Outcome Variable</th>
<th>Number of Groups</th>
<th>Within-Subjects Methods</th>
<th>Between-Subjects Variables</th>
</tr>
</thead>
</table>
• Group Size Ratio (Equal Group Sizes, Unequal Group Sizes)  
• Level of DIF (.4, .6, .8)  
• Percent Contamination (0%, 10%, 20%, 40%)  
• Impact (0, .5, 1) |
• Group Size Ratio (Equal Group Sizes, Unequal Group Sizes)  
• Level of DIF (.4, .6, .8)  
• Percent Contamination (10%, 20%, 40%)  
• Impact (0, .5, 1) |
• Group Size Ratio (Equal Group Sizes, Unequal Group Sizes)  
• Level of DIF (.4, .6, .8)  
• Percent Contamination (10%, 20%, 40%)  
• Impact (0, .5, 1) |
<table>
<thead>
<tr>
<th>Outcome Variable</th>
<th>Number of Groups</th>
<th>Within-Subjects Methods</th>
<th>Between-Subjects Variables</th>
</tr>
</thead>
</table>
• Group Size Ratio (Equal Group Sizes, Unequal Group Sizes)  
• Level of DIF (.4, .6, .8)  
• Percent Contamination (10%, 20%, 40%)  
• Impact (0, .5, 1) |
• Group Size Ratio (Equal Group Sizes, Unequal Group Sizes)  
• Level of DIF (.4, .6)  
• Percent Contamination (10%, 20%, 40%)  
• Impact (0, .5, 1) |
<table>
<thead>
<tr>
<th>Outcome Variable</th>
<th>Number of Groups</th>
<th>Within-Subjects Methods</th>
<th>Between-Subjects Variables</th>
</tr>
</thead>
</table>
• Group Size Ratio (Equal Group Sizes, Unequal Group Sizes)  
• Level of DIF (.4, .6)  
• Percent Contamination (10%, 20%, 40%)  
• Impact (0, .5, 1) |

*Note.* *Item two does not contain DIF, ** Item two contains DIF
CHAPTER IV
RESULTS
Two Groups

Type I Error

The results of the ANOVA revealed that three three-way interactions were statistically significant: method by level of DIF by impact, method by N by percent contamination, method by percent contamination by impact. There were other statistically significant interactions, but these did not have an effect size ($\eta^2$) of less than 0.1. For this study, all interactions with effect sizes less than 0.1 were not included in the interpretation because it was a very small effect size and accounted for less than 10% of the variance.

The interaction with the highest effect size was method by DIF by impact, $F(6.822, 204.67) = 12.049, p < .001, \eta^2 = .287$. Most of the MOA methods had a Type I error rate less than .05, regardless of level of DIF and impact (see Figure 1). The exceptions are robust MOA with GMH and robust MOA with GLR. However, robust MOA with GMH Type I error rate for the target item fell under .05 at level of DIF of .4 when impact was 0 or .5 (see Figure 1.1 and Figure 1.2), but fell under .05 at .6 when impact was one (see Figure 1.3). There were also two interesting aspects of this interaction. First, for GMH, GLR, and Lord’s, Type I error increased as DIF increased from DIF of 0 to DIF of .6. There was then a decrease in Type I error from DIF of .6 to .8. However, this pattern changed when impact was one (see Figure 1.3). Instead of the decreased Type I error from DIF of .6 to .8, the increase in Type I error continued. Thus, for this instance there was a continual increase in Type I error across all DIF levels. Second, unlike GMH, GLR, and Lord’s all MOA methods had a decrease in Type I error across all DIF level regardless of impact.
Figure 4.1.1. Three-way interaction of Type I error for two groups, method by level of DIF by impact with the impact of 0.
Figure 4.1.2. Three-way interaction of Type I error for two groups, method by level of DIF by impact with the impact of .5.
Next, was the three-way interaction of method by sample size by percent contamination, $F(6.822, 204.67 = 6.909, p < .001, \eta^2 = .187$. MOA methods had a Type I error rate of .05, except for robust MOA with GMH and robust MOA with GLR in some instances (see Figure 2). Robust MOA with GMH Type I error for the target item fell below .05 when the reference sample size was 1000 and percent contamination was greater than 10%, and when the reference sample size was 2000 and percent contamination was greater than 0%. Additionally, even though there was a general decrease in Type I error for the target item across percent contamination for MOA methods, this was not the case for GMH, GLR, and Lord’s. Instead, there was an increase

Figure 4.1.3. Three-way interaction of Type I error for two groups, method by level of DIF by impact with the impact of 1.
in Type I error for the target item past .05 across percent contamination. However, there was a greater increase of Type I error across percent contamination for GMH, GLR, and Lord’s as the sample size increase. For Lord’s, the Type I error rate for the target item increased past .05 when percent contamination was greater than 0%. The only exception was when the sample size was 500. In this instance, the Type I error rate increased past .05 when percent contamination was greater than 10%. GMH’s Type I error rate for the target item was always greater than .05, except for when the reference sample size was 2000 and the DIF contamination was 0%. Unlike Lord’s and GMH, GLR never had a Type I error rate for the target item less than .05.
Figure 4.2.1. Three-way interaction of Type I error for two groups, method by sample size by percent contamination with the reference sample size of 500.
Figure 4.2.2. Three-way interaction of Type I error for two groups, method by sample size by percent contamination with the reference sample size of 1000.
The final statistically significant interaction associated with the Type I error rate when the number of groups was two was the three-way interaction of method by percent contamination by impact, \( F(6.822, 204.67 = 6.529, p < .001, \eta^2 = .179 \). Following a similar theme of the other three-way interactions, most MOA methods had a Type I error rate below .05, except for robust MOA with GMH and robust MOA with GLR in most cases (see Figure 3). Robust GMH Type I error rate for the target item was less than .05 except for when percent contamination was greater than 10% regardless of impact. Unlike robust GMH, robust GLR only had one instance when
Type I error for the target item was less than .05, impact of one and percent contamination of 20. There was also a decrease in Type I error in most MOA methods as percent contamination increases. This was not the case for GMH, GLR, and Lord’s. For these three methods, the theme was a general increase in Type I error as percent contamination increased. In addition, there was an even greater effect for this increase across percent contamination as impact increased.

Figure 4.3.1. Three-way interaction of Type I error for two groups, method percent contamination by impact with the impact of 0.
Figure 4.3.2. Three-way interaction of Type I error for two groups, method percent contamination by impact with the impact of .5.
Summary. Overall, the results of the three-way interactions of Type I error rates for the number of groups of two found that for GMH, GLR, and Lord’s, Type I error rate did not fall below .05 in every instance. Instead, the results indicated that there was an increase in Type I error rates for detecting DIF for the target item as percent contamination increased or as DIF magnitude in the nontarget items increased. Unlike GMH, GLR, and Lord’s, most MOA methods showed little change for Type I error and the Type I error of the target item stayed
below .05. In fact, the only two MOA methods that had Type I error rates higher than .05 was robust GMH and robust GLR.

**Power**

The repeated measures ANOVA of power for the number of groups two found four statistically significant three-way interactions with an effect size greater than .1: method by DIF by percent contamination, method by sample size by percent contamination, method by sample size by DIF, method by DIF by impact. The three-way interaction with the highest effect size was method by DIF by percent contamination, $F(15.953, 478.593) = 30.371$, $p < .001$, $\eta^2 = .503$. GMH and GLR had the highest power regardless of level of DIF and percent contamination (see Figure 4). However, GMH and GLR did not control the Type I error for the target item as well as MOA methods for the Type I error significant interactions. Also, for the MOA methods, robust GLR generally had the highest power. Robust GLR power did fall below other MOA methods when DIF was .8 and percent contamination was 40%. Also, even though robust GLR had the highest power, it did not control for Type I error as well as other MOA methods. It did, however, control for Type I error better than the non-MOA methods GMH and GLR. Thus, interpretation of power rates for these statistics must be done with caution. For all of the methods, the interaction showed that as percent contamination increased, power decreased, regardless of the level of DIF. However, for all robust MOA methods as the DIF level increased there was a greater decrease in power between percent contaminations 20% and 40%. Unlike the robust MOA methods, beta GMH and chi GMH had a greater decrease between percent contaminations of 10% and 20% when DIF was either .4 or .8.
Figure 4.4.1. Three-way interaction of power for two groups, method by level of DIF by percent contamination with level of DIF of .4.
Figure 4.4.2. Three-way interaction of power for two groups, method by level of DIF by percent contamination with level of DIF of .6.
Figure 4.4.3. Three-way interaction of power for two groups, method by level of DIF by percent contamination with level of DIF of .8.

The interaction of method by sample size by percent contamination had the next highest effect size, $F(15.953, 478.593) = 24.132, p < .001, \eta^2 = .446$. GMH and GLR had the highest power regardless of percent contamination and sample size (see Figure 5). However, GMH and GLR did not control for Type I error for the target item as well as MOA methods. Robust GLR usually had the highest power of the MOA methods. The exception was when the sample size was 2000 and the DIF contamination was greater than 10%. Although robust GLR had the highest power of the methods, it did not control for Type I error the best out of the MOA
methods. However, compared to non-MOA methods GMH and GLR, robust GLR controlled for Type I error of the target item best. In all of the methods, as the percent contamination increased power decreased. However, the decrease in power was greater as sample size and percent contamination increased for robust MOA methods. Specifically, the decrease was found between the two percent contaminations of 20% and 40%. Unlike the robust MOA methods, the MOA methods beta GMH and chi GMH, had a greater decrease in power as sample size increased, but the decrease occurred when percent contamination was between 10% and 20%. In regard to non-MOA methods, there was a decrease in power as percent contamination increased, but sample size did not have a large effect on power. Instead, as sample size increased, the power of the methods would be greater.
Figure 4.5.1. Three-way interaction of power for two groups, method by reference sample size by percent contamination with the reference sample size of 500.
Figure 4.5.2. Three-way interaction of power for two groups, method by reference sample size by percent contamination with the reference sample size of 1000.
The next three-way interaction for power when number of groups was two was method by sample size by DIF, $F(15.953, 478.593) = 12.746, p < .001, \eta^2 = .298$. GMH and GLR had the highest power compared to the other methods (see Figure 6). However, GMH and GLR did not control Type I error for the target item as well as the MOA methods, therefore these higher power rates must be interpreted with caution. Of the MOA methods, robust GLR had the highest power in most instances. Robust GLR did have lower power than other MOA methods for the condition of DIF at .8 and sample size of 2000. Even though robust GLR had the highest power
of the MOA methods, it did not control for Type I error the best, but robust GLR controlled Type I error better than the non-MOA methods GMH and GLR. In regard to the effect of sample size, there was a consistent increase of power for all methods. That is, as the sample size increased, power increased, regardless of the level of DIF. However, there was a greater impact of the increase in power across sample size for GMH and GLR, but only when DIF was either .4 or .6.

Figure 4.6.1. Three-way interaction of power for two groups, method by reference sample size by level of DIF with the level of DIF of .4.
Figure 4.6.2. Three-way interaction of power for two groups, method by reference sample size by level of DIF with the level of DIF of .6.
Figure 4.6.3. Three-way interaction of power for two groups, method by reference sample size by level of DIF with the level of DIF of .8.

The last significant three-way interaction with an effect size greater than .1 for power when the number of groups was two was method by level of DIF by impact, $F(15.953, 478.593) = 8.667, p < .001, \eta^2 = .224$. The results showed that GMH and GLR have greater power than the other methods (see Figure 7). However, these methods did not control for Type I error of the target item better than MOA methods, and thus interpretation of these higher power rates must be made with caution. Additionally, in most instances robust GLR had the highest power of the MOA methods. Robust GLR did not have the highest power of the MOA methods when impact
was 0 and DIF was .8. Instead, robust GMH had the highest power. Also, even though robust GLR had the highest power, it did not control for Type I error as well as the other MOA methods. However, robust GLR did control for the Type I error of the target item better than GMH and GLR. In regard to the interaction, there was a consistent increase in power as DIF increases. This was not the case for the condition of impact of 1. In this instance DIF of .6 created either a sharp increase or decrease in power for some methods. The methods that had an increase include GMH, GLR, Lord’s, and MOA GLR robust, but a decrease for MOA GMH beta and MOA GMH chi-square.

Figure 4.7.1. Three-way interaction of power for two groups, method by level of DIF by impact with an impact of 0.
Figure 4.7.2. Three-way interaction of power for two groups, method by level of DIF by impact with an impact of .5.
Figure 4.7.3. Three-way interaction of power for two groups, method by level of DIF by impact with an impact of 1.

**Summary.** For power when the number of groups was two, the results indicated that overall GMH and GLR had the highest power for the target item. However, these methods did not control for Type I error as well as the other methods. The results also showed that robust GLR had the highest power of the MOA methods, but again, did not control for Type I error as well as other MOA methods. Additionally, the results showed that robust MOA methods power were affected by higher percent contaminations. When higher percent contaminations occurred
robust MOA methods would decrease in power dramatically. Sample size also impacted power rates of the target item with higher sample sizes yielding higher power rates.

**Three Groups**

**Type I Error**

The repeated measures ANOVA for Type I error when the number of groups was three found four statistically significant three-way interactions with an effect size greater than 0.1: methods by DIF by percent contamination, methods by sample size by percent contamination, methods by sample size by DIF. The first interaction with the highest effect size was method by level of DIF by percent contamination, $F(9.506, 287.874) = 40.342, p < .001, \eta^2 = .574$. The results of the interaction found that chi-square MOA methods using GMH and Lord’s controlled Type I error best (see Figure 8). For the MOA method chi-square Lord’s, Type I error did not exceed .05 regardless of the robust estimator used. However, the MOA method chi-square GMH exceeded the .05 threshold when percent contamination was 40% and DIF was .8. Also, unlike the chi-square MOA methods, beta MOA methods did not control for Type I error well. The interaction also demonstrated that for GMH, GLR, and Lord’s, as percent contamination increased, the increase in Type I error of the target item across DIF was greater. This was particularly the case for GLR and Lord’s, as the Type I error for the target item exceeded .45 as DIF increases. The opposite effect was found for beta MOA methods for GLR and Lords. Instead of an increase in Type I error rate for the target item, there was either a decrease or no change for Type I error.
Figure 4.8.1. Three-way interaction of Type I error for three groups, method by level of DIF by percent contamination with a level of DIF of .4.
Figure 4.8.2. Three-way interaction of Type I error for three groups, method by level of DIF by percent contamination with a level of DIF of .6.
Figure 4.8.3. Three-way interaction of Type I error for three groups, method by level of DIF by percent contamination with a level of DIF of .8.

The next interaction to be interpreted was methods by sample size by percent contamination, \( F(9.506, 287.874) = 28.231, p < .001, \eta^2 = .485 \). The methods that controlled the Type I error rate best are chi-square MOA methods GMH and Lord’s (see Figure 9). The Type I error rate for the target item for chi-square MOA methods GMH and Lord’s did not exceed .05 regardless of the simulation conditions or robust estimator used. Unlike chi-square MOA methods, beta MOA methods did not control for Type I error in most conditions. The interaction also found that the Type I error across sample size was greater as percent contamination increased for GMH, GLR, and Lord’s. However, GLR and Lord’s demonstrated a larger increase
in Type I error than GMH, as the Type I error rate exceeded .45 as sample size and percent contamination increased. Type I error had a different effect for beta MOA methods GMH and Lord’s. The interaction showed that as percent contamination increased the decrease in Type I error was greater. However, Type I error for those methods were never greater than .05.

Figure 4.9.1. Three-way interaction of Type I error for three groups, method by sample size by percent contamination with a reference sample size of 500.
Figure 4.9.2. Three-way interaction of Type I error for three groups, method by sample size by percent contamination with a reference sample size of 1000.
Figure 4.9.3. Three-way interaction of Type I error for three groups, method by sample size by percent contamination with a reference sample size of 2000.

The final statistically significant three-way interaction with an effect size greater than .1, was method by sample size by DIF, $F(9.506, 287.874) = 28.231, p < .001, \eta^2 = .342$. Of the methods, the chi-square MOA methods for GMH and Lord’s controlled for Type I error the best (see Figure 10). The control for Type I error for those methods were always below .05 regardless of the sample size, DIF, and robust estimator used. Unlike chi-square MOA methods, beta MOA methods do not control for the Type I error rate for target item well. In all instances the Type I error rate was greater than .05. The interaction also showed that the increase in Type I error was
larger across DIF as sample size increases for GMH, GLR, and Lord’s. However, the increase was greater for GLR and Lord’s compared to GMH, as the increase resulted in Type I error rates greater than .35. Unlike these methods, for beta MOA methods GMH, GLR, and Lord’s there was a decrease in Type I error as sample size increase.

Figure 4.10.1. Three-way interaction of Type I error for three groups, methods by sample size by level of DIF with a level of DIF of .4.
Figure 4.10.2. Three-way interaction of Type I error for three groups, methods by sample size by level of DIF with a level of DIF of .6.
Three-way interaction of Type I error for three groups, methods by sample size by level of DIF with a level of DIF of .8.

**Summary.** The results for Type I error when the number of groups was three found three significant interactions. Within those interactions there were two main themes throughout. The first theme was that chi-square MOA GMH and Lord’s controlled for Type I error better than any other method. There was only one case when chi-square GMH had a Type I error rate greater than .05, and chi-square Lord’s never had a Type I error greater than .05. The second theme was that there was an increase in Type I error for the Target item for GMH, GLR, and Lord’s, as the conditions got more extreme. However, the increase in Type I error was greater of GLR and Lord’s when compared to GMH.
Power

The repeated measures ANOVA for power when the number of groups was three found three three-way interactions significant and an effect size of greater than .1: method by sample size by DIF, method by DIF by percent contamination, method by sample size by percent contamination. Of the interactions, method by sample size by DIF had the highest effect size, $F(10.707, 321.225) = 29.146, p < .001, \eta^2 = .493$. Overall, GMH had the highest power in most conditions (see Figure 11). However, GMH did not have the highest power when DIF was .4 and sample size was 500. Also, even though GMH had the highest power, it did not control for Type I error as well as chi-square MOA methods. In fact, in many conditions the Type I error rate for GMH was greater than .05. Of the MOA methods, OGK beta with GLR had the highest power, but was not greater than non-MOA GMH. The methods that did control for Type I error the best, chi GMH methods and chi GL methods, had poor power. These methods had an average power from .09 and .21. The interaction also shows that there was an increase in power as sample size and magnitude of DIF increases. However, the increase in power became smaller as DIF increased for GMH, GLR, Lord’s, chi-square MOA GLR, and all beta MOA methods. In contrast, the increase in power across sample size for chi-square methods GMH MOA and Lord’s MOA methods was greater as DIF increased.
Figure 4.11.1. Three-way interaction of power for three groups, methods by reference sample size by level of DIF with a level of DIF of .4.
Figure 4.11.2. Three-way interaction of power for three groups, methods by reference sample size by level of DIF with a level of DIF of .6.
The interaction of method by DIF by percent contamination, $F(10.707, 321.225) = 29.396, p < .001, \eta^2 = .391$ was statistically significant and had an effect size greater than .1. The method with the highest overall power was GMH, however, GMH did not control for Type I error (see Figure 12). Also, GMH did not have the highest power in all conditions. GMH did not have the highest power when percent contamination was 10% and DIF was .4. Instead, OGK beta with GLR had higher power. When GMH did not have the highest power, the power rate was .740, but the power for OGK beta with GLR was .792. Also, the chi-square methods for the MOA approaches had the lowest power of the MOA methods, but the lowest power was chi-
square MOA Lord’s as the power was always less than .1. Although the chi-square MOA methods had lower power, MOA GLR had more comparable power to beta MOA methods. Within the interaction, for GMH and GLR there was a greater increase in power across DIF as percent contamination increases. Also, for MOA methods the increase in power diminished when percent contamination was 40%. The MOA method with the lowest power was chi-square Lord’s.

Figure 4.12.1. Three-way interaction of power for three groups, methods by level of DIF by percent contamination with a level of DIF of .4.
Figure 4.12.2. Three-way interaction of power for three groups, methods by level of DIF by percent contamination with a level of DIF of .6.
Finally, the three-way interaction of method by sample size by percent contamination, \( F(10.707, 321.225) = 12.559, p < .001, \eta^2 = .295 \) was statistically significant with the effect size greater than .1. The results were similar to those for the interaction of method by DIF by percent contamination (see Figure 13). Specifically, the method with the highest power was GMH in most conditions. GMH did not have the highest power when the conditions were sample size of 500 and percent contamination at 10%. In that condition, OGK beta with GLR had higher power, as GMH power was .767 and OGK beta with GLR’s power was .815. Even though GMH did
have the highest power in most conditions, it did not control for Type I error as well as MOA methods, meaning that its power rates must be interpreted with caution. Also, despite the fact that MOA chi-square had the lowest power, the power for MOA GLR chi-square methods were more comparable to the other methods, with chi-square MOA Lord’s having the lowest power. Found in the interaction for the methods GMH and GLR as sample size increases the decrease in power across percent contamination reduces. The opposite effect was found in the MOA methods. As sample size increased, the decrease in power across percent contamination was greater.

Figure 4.13.1. Three-way interaction of power for three groups, methods by reference sample size by percent contamination with the reference sample size at 500.
Figure 4.13.2. Three-way interaction of power for three groups, methods by reference sample size by percent contamination with the reference sample size at 1000.
Summary. The results for power when the number of groups was three, found that overall GMH and MOA ogk beta GLR had the highest power. Even so, these two methods did not control for Type I error as well as the other methods. The results also found that there were three significant interactions. From these interactions there were two main theme. First, the interactions showed that percent contamination had a considerable effect on MOA methods. With percent contamination increases there is a large decrease in power for MOA methods. This decrease in power was particularly found between 20% contamination and 40% contamination.
Second, the results also showed for all methods there is an increase in power as sample size increases. However, this increase was greater for MOA methods.

Six Groups

Type I Error

The repeated measures ANOVA for Type I error when the number of groups was six found one three-way interaction significant with an effect size greater than .1: method by sample size by percent contamination. The statistically significant interaction was methods by sample size by percent contamination, \( F(8.290, 153.373) = 4.063, p < .001, \eta^2 = .180 \). The results demonstrated that MOA Lord’s chi square methods ogk, mcd, and sde controls the best in regard to Type I error across sample size and percent contamination (see Figure 14). However, chi-square MOA Lord’s mcd and sde did not control for Type I error of the target item when the condition was reference sample size of 2000 and percent contamination of 40%. In this condition, chi-square MOA Lord’s mcd had a Type I error rate of .051, and chi-square MOA Lord’s sde had a Type I error rate of .055. Unlike chi-square MOA Lord’s methods ogk, mcd, and sde, most of the other methods in all conditions within this interaction the Type I error rate was above .05. This was particularly true for MOA beta methods. The MOA beta method with the highest Type I error rate was beta MOA GLR method ogk with the highest Type I error rate of .806. Also, the interaction showed that for GMH, GLR, and Lord’s as sample size increased, the increase in Type I error across percent contamination was greater. This increase in Type I error rate was especially seen when the reference sample size was 2000. This was not the case for the MOA methods. Instead, the Type I error rate was generally consistent with minor changes across percent contamination.
Figure 4.14.1. Three-way interaction of Type I error for six groups, methods by reference sample size by percent contamination with the reference sample size at 500.
Figure 4.14.2. Three-way interaction of Type I error for six groups, methods by reference sample size by percent contamination with the reference sample size at 1000.
Figure 4.14.3. Three-way interaction of Type I error for six groups, methods by reference sample size by percent contamination with the reference sample size at 2000.

Power

The repeated measures ANOVA for power when the number of groups was six found three three-way interactions significant with effect size values in excess of 0.1: method by sample size by percent contamination, method by sample size by DIF, and method by DIF by percent contamination. The interaction with the highest effect size was method by sample size by percent contamination, $F(12.882, 238.309) = 28.806, p < .001, \eta^2 = .609$. The results found that the MOA method OGK beta with GLR had one of the highest power rates (see Figure 15). There
were, however, conditions for which GMH had the highest power. These conditions included reference sample size of 1000 with percent contamination at 40%, reference sample size of 2000 regardless of percent contamination. When the reference sample size was 1000 and percent contamination was 40% the power for beta MOA GLR with OGK was .799, but the power for GMH was .852. Also, when the reference sample size was 2000 the power rates were comparable between the two methods, except for when the percent contamination was 40%. In this condition beta MOA GLR with OGK had a power of .810, but GMH had a power of .982. Although MOA GLR with OGK and GMH had the highest power, these methods did not control for Type I error. Of the two methods GMH had the best control of Type I error, but was never below .05. The interaction also shows that as the reference sample size increases there was an increase in power regardless of the percent contamination. However, when taking sample size and percent contamination into consideration there was a decrease in power for all methods as percent contamination increases. Additionally, as sample size increase, the decrease in power as percent contamination increases was greater. This was particularly when looking at the power raters from 20% and 40%.
Figure 4.15.1. Three-way interaction of power for six groups, methods by reference sample size by percent contamination with the reference sample size at 500.
Figure 4.15.2. Three-way interaction of power for six groups, methods by reference sample size by percent contamination with the reference sample size at 1000.
Three-way interaction of power for six groups, methods by reference sample size by percent contamination with the reference sample size at 2000.

The second significant interaction was methods by sample size by DIF, $F(12.882, 238.309) = 33.74$, $p < .001$, $\eta^2 = .477$. MOA method OGK beta with GLR consistently had one of the highest power rates, as well as GMH (see Figure 16). Beta MOA GLR with OGK had the highest power when DIF was .4 and the reference sample size was 500 and 1000, and when DIF was .6 and the reference sample size was 500. In the three other conditions, GMH had the highest power. Although beta MOA GLR with OGK and GMH had the highest power, neither method controlled the Type I error, as discussed above. Of these two methods, MOA GLR with
OGK had the worst Type I error. The MOA beta methods showed a steady increase in power across sample sizes for both DIF conditions (see Figure 16). The results for GMH, GLR, and Lord’s differed slightly from the MOA beta methods. Even though GMH, GLR, and Lord’s had a large increase in power across the sample size for both DIF conditions, for the DIF condition of .60, GMH and GLR did not have as large of an increase in power compared to the DIF condition .40. But for Lord’s, there was a greater increase in power for DIF of .6 than .4. MOA chi methods also had a similar pattern to Lord’s. That is, even though there was an increase in power across sample size, the increase was more prominent for the condition of DIF .60. However, the increase was not as large as Lord’s.
Figure 4.16.1. Three-way interaction of power for six groups, methods by reference sample size by level of DIF with level of DIF at .4.
Figure 4.16.2. Three-way interaction of power for six groups, methods by reference sample size by level of DIF with level of DIF at .6.

The third, statistically significant interaction involved method by DIF by percent contamination, $F(12.882, 238.309) = 24.594, p < .001, \eta^2 = .399$. The results demonstrated that MOA method OGK beta with GLR had the highest power compared to the other methods (see Figure 17). The one exception to this was for percent contamination of 40% and DIF of .60. In this condition, GMH had the highest power with the power of .92, but beta MOA GLR with OGK had a power of .80. Although MOA GLR with OGK and GMH had the highest power, both methods did not control for Type I error. For all of the methods there was a decrease in
power across percent contamination. However, for all MOA methods, as DIF contamination increased the power was greater when percent contamination in 10%, but then there was a large decrease in power.

Figure 4.17.1. Three-way interaction of power for six groups, methods by level of DIF by percent contamination with level of DIF at .4.
Figure 4.17.2. Three-way interaction of power for six groups, methods by level of DIF by percent contamination with level of DIF at .6.

The final statistically significant interaction with an effect size greater than 0.1 was method by percent contamination by impact, $F(12.882, 238.309) = 3.422, p < .001, \eta^2 = .156$. Overall, MOA method OGK beta with GLR had the highest power (see Figure 18). The only exception was when percent contamination was 40% when impact was 0 and .05. In this case, GMH had the highest power. Specifically, when impact was 0 and percent contamination was 40% beta MOA GLR with OGK power was .808 and GMH had a power of .823, and when impact was .5 and percent contamination was 40% beta MOA GLR with OGK power was .798 and power for
GMH was .810. The interaction also demonstrated that for all methods, as percent contamination increases, power decreases. This was especially the case for MOA methods. This decrease in power across percent contamination was more prevalent as impact decreases.

Figure 4.18.1. Three-way interaction of power for six groups, methods by percent contamination by impact with impact at 0.
Figure 4.18.2. Three-way interaction of power for six groups, methods by percent contamination by impact with impact at .5.
Summary. The results for power when the number of groups was six found that overall GMH and MOA ogk beta GLR had the highest power. However, these GMH and MOA ogk beta GLR did not control for Type I error as well as other methods. The results of power with the number of groups of six found similar results to when the number of groups was three. Like the results for when the number of groups was three, percent contamination had a predominant influence on power for MOA methods. As percent contamination increased the power for MOA methods had a decrease in power. The results were also similar to when the number of groups was three, as the results showed that as sample size increased power also increased.
CHAPTER V

DISCUSSION

This study examined and compared the Type I error and power rates for DIF detection MOA methods to GMH, GLR, and Lord’s chi-square, through a simulation study. To address the goals of the study, uniform DIF was simulated using 20 items with either 0%, 10%, 20%, or 40% of the items contaminated with DIF. In addition to assessing the percent of items contaminated with DIF, other commonly found conditions within real testing data were included in the simulation study, thus providing researchers and measurement professionals with guidance regarding when particular DIF detection methods might work best. The additional variables that were manipulated included the number of groups, sample size, group size ratio, level of DIF, and impact. Repeated measure ANOVAs were then used to compare the simulation conditions and determine which impacted Type I error and power of the methods. The effect size $\eta^2$ was also used to assist in determining the main effects and interactions that had the largest effect on the Type I error and power rates.

Study Findings

Overall, the results of this study revealed that when taking into consideration both Type I error and power there is not one best method from among those examined (see Table 4). Some methods exhibited relatively high power but poor Type I error control, whereas other methods yielded low power rates but good Type I error control. As a result, it is necessary to discuss the findings for Type I error and power separately to gain a better perspective of where the MOA methods can best be used.
Table 4. Summary of Study Findings

<table>
<thead>
<tr>
<th>Research Question</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Is there a difference in power between GMH, GLR, Lord’s when compared to MOA DIF methods?</td>
<td>GMH had the highest overall power compared to the other methods, and OGK beta with GLR had the highest power of MOA methods. MOA chi-square methods had the worst power.</td>
</tr>
<tr>
<td>2. Is there a difference in Type I error between GMH, GLR, Lord’s when compared to MOA DIF methods?</td>
<td>MOA chi-square methods controlled for Type I error, while the rest of the methods had Type I error rate above .05. MOA beta methods did not always control for Type I error, particularly as the number of groups increased. GMH, GLR, and Lord’s only controlled for Type I error in a few instances.</td>
</tr>
<tr>
<td>3. Is there a difference in power between GMH, GLR, Lord’s when compared to MOA DIF methods as the number of items containing DIF increases?</td>
<td>As the number of items containing DIF increased, the power for MOA methods had a strong decreased. The power for GMH, GLR, and Lord’s had a slight decrease as the number of items containing DIF increased.</td>
</tr>
<tr>
<td>4. Does the level of DIF have an impact on power and Type I error amongst GMH, GLR, Lord’s when compared to MOA DIF methods?</td>
<td>For Type I error, level of DIF had an effect on GMH, GLR, and Lord’s. As level of DIF increased, Type I error increased. Level of DIF did not have a pronounced effect on Type I error for MOA methods. For power, level of DIF had an effect on all methods, except for chi-square MOA methods. As level of DIF increased, power increased. There was no change in power for chi-square MOA methods.</td>
</tr>
<tr>
<td>5. Does impact and/or sample size have an effect on the Type I error and power for GMH, GLR, Lord’s when compared to MOA DIF methods?</td>
<td>For Type I error, as impact and sample size increased, Type I error also increased for GMH, GLR, and Lord’s. Impact and sample size did not have an effect on Type I error for all MOA methods. For power, as sample size increased, power increased for all methods. Impact did not have an effect on power rates for the methods, unless impact was included in a three-way interaction.</td>
</tr>
</tbody>
</table>
Difference in Type I error

The results demonstrated that the chi-square MOA methods controlled the Type I error rates the best, regardless of the simulation conditions. In fact, for chi-square MOA methods, the Type I error was in control across all simulation conditions. This, however, was not the case for the beta MOA methods. Although the Type I error rates were relatively consistent across conditions for these methods, it was above the nominal .05 rate in nearly all cases. This result is consistent with Magis and DeBoeck’s (2012) results for the robust approach, which is simply the MOA method for only when the number of groups was two, and the expectation of the MOA method controlling for Type I error stated by Magis and DeBoeck (2011). Similar to this study, Magis and DeBoeck’s (2012) simulation study for the robust approach showed that the chi-square MOA method controlled for Type I error in all conditions; however, the robust beta GMH controlled for Type I error in most conditions, unlike in the current study. For this reason, it is surprising that beta and robust beta MOA methods had such high rates of Type I error, especially for when the number of groups was three or six.

In contrast to the MOA methods, which tended to control Type I error rates across simulation conditions, such was not the case for GMH, GLR, and Lord’s. Instead, the Type I error rates for these methods were impacted by several interactions of the study conditions. To keep these findings concise, the results for each group condition are explained briefly, unless the result was similar across multiple group conditions. When similar results occurred, only one explanation is provided. Following these explanations, reasons for why these results occurred are also explored.

The first surprising result was that the Type I error rates for GMH, GLR, and Lord’s was not controlled below the nominal .05 level in many cases, regardless of the number of groups.
Although it was expected that GMH, GLR, and Lord’s would not control for Type I error as well as MOA methods, it was surprising how poor its Type I error control was in most conditions. Previous research that investigated these three methods in a simulation study using a 2PL model did find that GMH, GLR, and Lord’s had an inflated Type I error rate in some situations (Finch, 2015). Finch (2015) also reported that GLR had the highest Type I error inflation, whereas Lord’s yielded the lowest error rate. This outcome is consistent with the findings of this study. However, what was inconsistent between Finch (2015) and this study was the degree of the inflation of Type I error. One potential reason the result may have differed in this regard is due to this study using a 1PL and Finch (2015) using a 2PL model. This is because inclusion of differing item discrimination parameter values by group, which were used by Finch (2015), can have an effect on other item and person parameter estimation (de Ayala, 2009). For this reason, it explains one potential reason for the differing results. Another possible reason for the differing results is that Finch (2015) generated a matching subset scale that was purified. This could have caused different results because no other items contained DIF. However, in the current study multiple items contained DIF. Ultimately, this produced different results because there is a change in the conditions.

Another common result across group conditions was that the Type I error for GMH, GLR, and Lord’s increased as conditions became more extreme; however, the conditions in which Type I error would increase was different. One similarity amongst group results is that as percent contamination increased, Type I error also increased, but this increase in Type I error became greater as sample size increased. Thus, there was an increase in percent contamination when the reference sample size was 500, but the slope of the increase was greater as the sample size increased. The results also revealed two interesting findings for GMH, GLR, and Lord’s for
the when the number of groups was two. The first finding is that for these three methods, Type I error increased as level of DIF and impact increased. The second finding was that there was an increase in Type I error for GMH, GLR, and Lord’s as percent contamination increased. Additionally, as impact increased the increase of Type I error across percent contamination was larger. When the number of groups was three, the results found that 1) Type I error increased as sample size and level of DIF increased; however, this increase in Type I error was more prominent across level of DIF as sample size increased, and 2) as level of DIF increased Type I error increased, and this increase became greater as percent contamination increased.

A possible reason for this increase in Type I error is because as conditions become more extreme, particularly impact, it is more difficult for these methods to estimate DIF. Many studies focusing on MH and LR have found that when impact is present there is an increase in Type I error (Meredith & Millsap, 1992; Roussous & Stout, 1996; Uttara & Millsap, 1994; Zwick, 1990). DeMars (2010) states that this increase in Type I error may be because when impact is present the test takers that received a particular score may not have the same person ability. This is due to the groups having a different mean ability. Other studies have also found that higher levels of sample size, level of DIF, and percent contamination have lead to inflated Type I error rates for when DIF is estimated for when the number of groups was two (de Ayala et al., 2002; DeMars, 2009; Guler & Penfiled, 2009; Herrera & Gomez, 2008; Kim & Oshima, 2012; Li, Brooks, & Johanson, 2012; Rogers & Swaminathn, 1993; Swaminathan & Rogers, 1990; Vaughn & Wang, 2010). Simulation research for estimating DIF for more than two groups has also indicated that more extreme conditions can lead to high Type I error rates. In addition, research focused only on GMH found that a higher sample size and in the presence of impact
Type I error rates greater than .05 was found. Similar results were found by Finch (2015) who discovered that GMH, GLR, and Lord’s Type I error rate increased as sample size increased.

**Summary.** Overall the result for Type I error yielded three primary findings. First, Type I error rates for beta MOA methods were inflated, however Type I error was not inflated for chi-square MOA methods. Second, GMH, GLR, and Lord’s all had inflated Type I error as all conditions (e.g., percent contamination, level of DIF, and sample size) increased to more extreme levels. The last finding was that performance of GMH, GLR, and Lord’s was more strongly impacted by the interaction of conditions compared to MOA methods. Instead, MOA methods generally had a consistent Type I error rate. These interaction of conditions included impact by DIF, reference sample size by percent contamination, impact by percent contamination, level of DIF by percent contamination, and sample size by level of DIF.

**Difference in Power**

The results presented in Chapter IV demonstrated that overall, GMH had the highest power rates across most conditions, and that of the MOA methods, OGK beta with GLR had the highest power in most conditions regardless of the number of groups. However, neither method controlled for Type I error as well as did the other methods. This finding for power is consistent with similar findings found in Magis and DeBoeck (2012) and Finch (2015). Magis and DeBoeck (2012) found that MH for two groups had the highest rates of power, with robust beta methods power comparable to MH. Also, even though GMH and OGK beta with GLR had the best power overall, there were interaction of conditions that effected the power of all of the methods. To understand these interactions, each was explored in order to provide an explanation of them. The explanation of the results will be similar to how Type I error results were explained.
Unlike with Type I error, the power rates of the MOA methods were affected the most by the simulation conditions, rather than GMH, GLR, and Lord’s. Instead, GMH, GLR, and Lord’s had only slight increases or decreases in power across the conditions, but there were no dramatic shifts in the rates, regardless of simulation settings. MOA methods, on the other hand had drastic changes in power in some conditions. Many of these changes in power across conditions were consistent across group size. Specifically, there were three findings that were the same for all three group conditions. The first finding was that the increase in power reduced across sample size as level of DIF increased. This finding is not surprising, as Finch (2015) found similar results, whereby as the magnitude of DIF increased there is less of an increase in power across reference group sample size. The reason these results were found in the current study and Finch (2015) is likely because as the level of DIF increased the power increased for the reference group sample size. Consequently, with a higher beginning power rate, there is less of an increase to be had.

Another result that was similar across the group conditions was that power rates for all MOA methods were influenced by percent of items that were contaminated by DIF. Specifically, power for all MOA methods decreased as percent contamination increased. Furthermore, this decrease in power was greater as the magnitude of DIF increased. Similar results were found for sample size and percent contamination. That is, as sample size increased the decrease in power across percent contamination was greater for all MOA methods. This result is not surprising because as percent contamination increases, it is more and more difficult for a DIF item to be considered an outlier. Looking back at how MOA identifies items as containing DIF, a modified Mahalanobis distance is calculated for the DIF item statistics using a robust estimate of the covariance matrix. The Mahalanobis distance is then compared to the quintile of the beta
distribution for the alpha level ($Q_\alpha$), with a Mahalanobis distance greater than the alpha level indicating the item as DIF. However, the covariance matrix of the robust estimators is included into the Mahalanobis distance calculation to the negative first power, thus inversing the robust estimator. As a result, a greater covariance matrix of the robust estimators, which would occur when more items contain DIF, will create smaller Mahalanobis distance. Thus, the decreased Mahalanobis distance is less likely to be greater than the alpha level identifying the item, failing to identify the item as DIF.

One last power result that was unique to the when the number of groups was two was an increase in power as the level of DIF increased across levels of impact. However, there was a decrease in power when level of DIF was .6 and impact was one for MOA GMH beta and MOA GHM chi-square, but power increased between .6 and .8 level of DIF. This result is unique as similar results were not found in previous research. Instead, based other findings with level of DIF found in this study, and findings with level of DIF found in Finch (2015), it would be expected that the increase would continue across level of DIF.

**Summary.** The results for power demonstrated that power rates for the MOA methods were more strongly impacted by percent contamination than were GMH, GLR, and Lord’s. For a greater percent contamination, power for the MOA methods was lower, when compared to that of GMH, GLR, and Lord’s. But again, this does not come as a surprise, as it would be more difficult to identify an item as an outlier as more items are in fact outliers (i.e., contain DIF). The second main finding was that GMH had the highest power of all the methods, and OGK beta with GLR had comparable power unless percent contamination was high. Last, an increase in sample size had a positive impact on power for all methods, as would be expected.
Limitations and Future Directions

Like any study, there are limitations of the current work that are worth noting so that future research can address these limitations. The first such limitation is that this study used a 1PL model to generate the data. Although, the 1PL model is widely used in psychometrics, the MOA method should be compared to GMH, GLR, and Lord’s using a 2PL and/or 3PL model, as well. By using a different model, the Type I error and power rates of the MOA method may change due to the additional parameters of item discrimination and item guessing within the IRT model. Because of these potential changes in Type I error and power by using a 2pl or 3pl model, these models should be explored to fully understand how well MOA functions in commonly used IRT models.

A second limitation of this study is that it examined the performance of DIF for an assessment with 20 items and did not vary this number across other conditions. Naturally, not all assessments will have 20 items, and the number of items can have an impact on Type I error rates and power, although it may be a small effect (Scott et al., 2009). Thus, further research needs to be conducted to determine the Type I error rates and power for assessments with the number of items less than and greater than 20.

Third, the unequal group sizes were limited to one condition with the unequal groups 50% less than the reference group. Within assessment data, the data will each group can have varying degrees of sample sizes, and the focal groups will not always all be less that 50% of the reference group. Because of this further unequal group size conditions should be further examined to determine their effect on Type I error and power.

The last limitation is that nonuniform DIF was not included in this study. This limits the usability of the results of this study to only when psychometricians are estimating uniform DIF,
as these results cannot be applied to nonuniform DIF. This is due to the fundamental differences between uniform and nonuniform DIF. In uniform DIF, there is a consistent advantage given to one particular group. However, nonuniform DIF functions under the idea that the group that has the advantage can change based on the ability level. Because this study did not include nonuniform DIF, the accuracy of nonuniform MOA has yet to be studied under a simulation study, thus limiting the usability of this method. For this reason, future research should be conducted to determine how well MOA estimates nonuniform DIF. This would assist psychometricians understanding of MOA, and knowledge of when MOA should be used instead of GLR and Lord’s when estimating nonuniform DIF.

**Conclusion**

Due to the mixed results for Type I error and power rates reported above, there does not appear to be one best method for detecting uniform DIF using a 1PL model. However, three guidelines can be provided to determine when a particular method should be used based off the strengths and weakness of the methods found in this study. In addition, a review of the strengths and weakness may be beneficial to further understand the provided guidelines (see Table 5).
Table 5. Summary of Strengths and Weaknesses for Each Method

<table>
<thead>
<tr>
<th>Methods</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMH</td>
<td>Best power, conditions do not negatively impact power much.</td>
<td>Low Type I error, Type I error becomes worse as conditions become more extreme.</td>
</tr>
<tr>
<td>GLR</td>
<td>Good power, conditions do not negatively impact power much.</td>
<td>Low Type I error, Type I error becomes worse as conditions become more extreme.</td>
</tr>
<tr>
<td>Lord’s</td>
<td>Okay power, conditions do not negatively impact power much.</td>
<td>Low Type I error, Type I error becomes worse as conditions become more extreme.</td>
</tr>
<tr>
<td>MOA GMH Chi-square</td>
<td>Controls for Type I error, conditions do not effect Type I error or power</td>
<td>Poor Power</td>
</tr>
<tr>
<td>MOA GMH Beta</td>
<td>Good power, good Type I error for two groups.</td>
<td>Poor Type I error control for three and six groups, power effected by number of items contaminated with DIF.</td>
</tr>
<tr>
<td>MOA GLR Chi-Square</td>
<td>Controls for Type I error, conditions do not effect Type I error or power.</td>
<td>Poor Power</td>
</tr>
<tr>
<td>MOA GLR Beta</td>
<td>Good power, OGK beta had the best power of MOA methods, good Type I error for two groups.</td>
<td>Poor Type I error control for three and six groups, power effected by number of items contaminated with DIF.</td>
</tr>
<tr>
<td>MOA Lord’s Chi-square</td>
<td>Controls for Type I error, conditions do not effect Type I error or power.</td>
<td>Poor Power</td>
</tr>
<tr>
<td>MOA Lord’s Beta</td>
<td>Good power, good Type I error for two groups.</td>
<td>Poor Type I error control for three and six groups, power effected by number of items contaminated with DIF.</td>
</tr>
</tbody>
</table>

First, if power is the main concern then GMH is perhaps the optimal option for researchers investigating DIF. This is because overall GMH had the best power and was not affected by extreme conditions (e.g., 40% contamination, .8 level of DIF). However, GMH did not have low
Type I error and did not control for Type I error rates in most conditions. Because of this, if both power and Type I error are a concern then GMH would not be the best option.

A researcher and a psychometrician would be interested in this recommendation when the validity of the measure is of most important, and the budget for creating this measure is higher. This is because due to the poor Type I error rate, items will be identified as containing DIF when in fact they do not. Ultimately, this can lead to higher production costs and time due to items unnecessarily being removed or revised from this misidentification. These higher production costs could be minimized by selecting a DIF method with better Type I error control.

This leads to the second guideline: if the researcher is concerned about both power and Type I error then OGK beta with GLR for more than two groups, or robust GLR for only two groups, should be considered for use in uniform DIF detection. This is because in comparison, OGK beta with GLR had the highest overall power of the MOA methods and demonstrated better Type I error rate control than did GMH in most instances, although Type I error rates were still greater than .05 in some conditions.

This guideline should be used by a researcher or psychometrician when both the validity of the measure, time to create the measure, and the budget are of concern. This is because by having a better control of Type I error rate than GMH, fewer items will be misidentified as containing DIF. Consequently, this will not increase the cost and time of producing the measure by having to remove items and creating new items. The misidentification of items may also be a concern if the researcher is creating a measure for a construct that has certain criteria. For example, if one was creating a measure in regard to depression, it may be difficult to create new items based on the diagnosis criteria. Particularly, if the item being replaced did not actual have potential bias. However, caution should be used with using OGK beta with GLR should be used
if the number of groups is greater than three or if it is expected that more extreme conditions will be found in the data. First, caution should be used when there are more than three groups if Type I error is of concern because Type I error was higher for OGK beta with GLR than GMH. Also, in terms of power, when extreme conditions occurred with percent contamination (i.e., 40% contamination).

Third, if only Type I error rate control is the concern, then any chi-square MOA method is the best option. This is because all chi-square MOA methods controlled for Type I error the best, and the Type I error was never above .05. However, the chi-square MOA methods had very low power, so if power is of any concern then the chi-square MOA methods are likely not to be a good option. Even so, GLR chi-square MOA methods had the highest power when compared other GMH chi-square MOA and Lord’s chi-square MOA methods. A researcher or psychometrician will be interested in this recommendation if they are only concerned with time and/or budget. This is because these items will rarely misidentify an item as containing DIF. Thus, not creating the need to write new items for the measure. This in turn will assist with the constraint of time and budget. However, caution should be used because these items do have poor power. In turn, this may create the measure cost effective, but have poor validity.
References


