10-21-2009

A Monte Carlo Study Investigating the Influence of Item Discrimination, Category Intersection Parameters, and Differential Item Functioning in Polytomous Items

Carol Jenetha Thurman
cthurman@gsu.edu

ACCEPTANCE

This dissertation, A MONTE CARLO STUDY INVESTIGATING THE INFLUENCE OF ITEM DISCRIMINATION, CATEGORY INTERSECTION PARAMETERS, AND DIFFERENTIAL ITEM FUNCTIONING PATTERNS ON DETECTION OF DIFFERENTIAL ITEM FUNCTIONING IN POLYTOMOUS ITEMS, by CAROL JENETHA THURMAN, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree Doctor of Philosophy in the College of Education, Georgia State University.

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

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

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

_____
William L. Curlette, Ph.D.
Committee Member

_____
Philo A. Hutcheson, Ph.D.
Committee Member

_____
Date

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

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

AUTHOR'S STATEMENT

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

_____

Carol J. Thurman

NOTICE TO BORROWERS

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

Carol Jenetha Thurman
1270 Colony Place
Marietta, GA 30068

The director of this dissertation is:

Dr. Takako Chris Oshima
Department of Educational Policy Studies
College of Education
Georgia State University
Atlanta, GA 30303-3083

VITA

Carol J. Thurman

ADDRESS: 1270 Colony Place
Marietta, GA 30068

EDUCATION:

Ph.D.	2009	Georgia State University Educational Policy Studies
M.Ed.	1992	Plymouth State University Secondary Mathematics/Computers
B.A.	1980	Harding University Physical Education

PROFESSIONAL EXPERIENCE:

2006-Present	Research Specialist DeKalb County School System, Decatur, GA
2005-2006	Research Associate Georgia State University, Atlanta, GA
2002-2006	Math Curriculum Specialist, Lead PLATO Learning Inc., Atlanta, GA
2001-2002	Middle School Math Teacher Fulton County School System, Sandy Springs, GA
1985-1999	High School and Middle School Math Teacher The American International School, Vienna, Austria

PROFESSIONAL SOCIETIES AND ORGANIZATIONS:

2005-Present	American Educational Research Association
2007-Present	Atlanta Area Evaluation Association

PRESENTATIONS AND PUBLICATIONS:

Benson, G., Curlette, W., Dale, A., Ogletree, S., Segal, D., Taylor, D., Thurman, C. (2008, March). *Evolution of a Professional Development School Approach: Fidelity of Implementation and Teacher-Intern-Professor Groups.* Paper presented at the annual meeting of the American Education Research Association, New York.

Bhagavati, S., Thurman, C. J., (2006, October 27). *A review of the NAEP Report: Comparing private schools and public schools using hierarchical linear*

modeling. Paper presented at the annual meeting of the Georgia Educational Research Association, Savannah, GA.

A MONTE CARLO STUDY INVESTIGATING THE INFLUENCE OF ITEM DISCRIMINATION, CATEGORY INTERSECTION PARAMETERS, AND DIFFERENTIAL ITEM FUNCTIONING PATTERNS ON THE DETECTION OF DIFFERENTIAL ITEM FUNCTIONING IN POLYTOMOUS ITEMS

by
Carol Thurman

ABSTRACT

The increased use of polytomous item formats has led assessment developers to pay greater attention to the detection of differential item functioning (DIF) in these items. DIF occurs when an item performs differently for two contrasting groups of respondents (e.g., males versus females) after controlling for differences in the abilities of the groups. Determining whether the difference in performance on an item between two demographic groups is due to between group differences in ability or some form of unfairness in the item is a more complex task for a polytomous item, because of its many score categories, than for a dichotomous item. Effective DIF detection methods must be able to locate DIF within each of these various score categories.

The Mantel, Generalized Mantel Haenszel (GMH), and Logistic Regression (LR) are three of several DIF detection methods that are able to test for DIF in polytomous items. There have been relatively few studies on the effectiveness of polytomous procedures to detect DIF; and of those studies, only a very small percentage have examined the efficiency of the Mantel, GMH, and LR procedures when item discrimination magnitudes and category intersection parameters vary and when there are different patterns of DIF (e.g., balanced versus constant) within score categories.

This Monte Carlo simulation study compared the Type I error and power of the Mantel, GMH, and OLR (LR method for ordinal data) procedures when variation occurred in 1) the item discrimination parameters, 2) category intersection parameters, 3)

DIF patterns within score categories, and 4) the average latent traits between the reference and focal groups.

Results of this investigation showed that high item discrimination levels were directly related to increased DIF detection rates. The location of the difficulty parameters was also found to have a direct effect on DIF detection rates. Additionally, depending on item difficulty, DIF magnitudes and patterns within score categories were found to impact DIF detection rates and finally, DIF detection power increased as DIF magnitudes became larger. The GMH outperformed the Mantel and OLR and is recommended for use with polytomous data when the item discrimination varies across items.

A MONTE CARLO STUDY INVESTIGATING THE INFLUENCE OF ITEM DISCRIMINATION, CATEGORY INTERSECTION PARAMETERS, AND DIFFERENTIAL ITEM FUNCTIONING PATTERNS ON THE DETECTION OF DIFFERENTIAL ITEM FUNCTIONING IN POLYTOMOUS ITEMS

by
Carol Thurman

A Dissertation

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

Atlanta, GA
2009

ACKNOWLEDGMENTS

I am indebted to so many amazing individuals who encouraged and supported me along the way. First, I would like to thank my husband for his patience, love, and support. I am also thankful to my children, Erica, Kira, and Jonathan for their encouragement. Much thanks to my mother who taught me to value a good education and encouraged me to reach my goals. I would like to thank Robert Hendrick who believed in me and was the first to encourage me to begin this journey. Thank you so very much. I am also very appreciative of the Educational Policy Studies Department at Georgia State University and the many brilliant minds that make it what it is. Dr. Oshima, thank you so very much for your guidance, insights, and support. You are truly amazing. Dr. Furlow, I could not have done this without you. I am truly in awe of your expertise. Thank you from the bottom of my heart. Dr. Hutcheson, thank you for your support from the time I was a green doctoral candidate to the present. Your love for scholarship is contagious! And finally, but certainly not least, I would like to thank Dr. William Curlette for his friendship and support throughout this entire process. I have learned so much under your tutelage. Thank you so much for your advice and willingness to lend a listening ear in those difficult times.

There is a God shaped vacuum in the heart of every man which cannot be filled by any created thing, but only by God, the Creator, made known through Jesus.

1623 Blaise Pascal, mathematician/physicist/religious writer

Soli Deo Gloria! Thank You.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ASA	Average Signed Area
CRC	Category Response Curve
DIF	Differential Item Functioning
GMH	Generalized Mantel-Haenszel
GPCM	Generalized Partial Credit Model
GRM	Graded Response Model
ICC	Item Characteristic Curve
IRT	Item Response Theory
LDFA	Logistic Discriminant Function Analysis
LR	Logistic Regression
OLR	Ordinal Logistic Regression
PCM	Partial Credit Model
PL	Parameter Logistic
MH	Mantel-Haenszel
UCLOR	Unconstrained Cumulative Ordinal Logistic Regression

CHAPTER 1

INTRODUCTION

The detection of differential item functioning in high-stake assessments such as licensure, and credentialing examinations has become an important issue in recent years (Swaminathan & Rogers, 1990). The increased use of these high-stake measures has led to efforts nationwide to promote fairness in testing by constructing assessments that tap into an examinee's deep level of understanding. Consequently, many of these performance measures consist entirely of polytomous items rather than multiple choice items (Wang & Su, 2004; Zwick, Donoghue, & Grima, 1993). Indeed, the use of polytomous item formats nationwide has led to increased attention to the detection of differential item functioning in these items (Bolt, 2002; Chang, Mazzeo & Roussos, 1996; Wang & Su, 2004). Differential item functioning (DIF) occurs when an item performs differently for two contrasting groups of respondents (e.g., males vs females) after controlling for differences in the abilities of the groups (Angoff, 1993). Determining whether the difference in performance on an item between two demographic groups is due to between group differences in ability or some form of unfairness in the item is a more complex task for a polytomous item because of its many score categories, than for a dichotomous item. Because of the number of score levels in a polytomous item, DIF can occur within all of the score categories or within some subsets of score categories within the item, hence requiring testing for DIF at each score level (French & Miller, 1996; Kristjansson, McDowell, & Zumbo, 2005). DIF detection methods capable of detecting

DIF in *all* score categories are essential if issues of fairness in testing are to be adequately addressed.

The Mantel (Mantel, 1963) and the GMH (GMH; Mantel & Haenszel, 1959; Somes, 1986), two direct extensions of the very popular dichotomous DIF detection technique - the Mantel-Haenszel - are two methods that can test for DIF at each score level. The Mantel compares the item means after conditioning on a matching variable while the GMH compares the entire response distribution of the reference and focal groups. The logistic regression (LR) procedure (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990) has also emerged as a popular DIF detection method, as well. The LR is a model based procedure that can provide more specific information on the whereabouts of DIF and the type of DIF that is present. (The LR procedure is oftentimes referred to as ordinal logistic regression (OLR) when the dependent variable is ordered data).

While there has been a marked increase in the use of polytomous assessments in education, there are relatively few studies on the effectiveness of the Mantel, GMH, and the OLR procedures to detect DIF in polytomous item, specifically when an item's difficulty parameters vary among response categories resulting in different patterns of DIF. These patterns of DIF can occur because within polytomous items, transitioning from one response category to the next can increase the likelihood that the transition is more difficult for one group of examinees than the other. For example, in a four category item, DIF might reside in the transition from score category two to score category three but not in the other two response categories. Therefore, it is important to examine what effects differing patterns of DIF have on various DIF detection procedures, particularly

when the discrepancies in item difficulty occur within various score categories. Additionally, the size of DIF occurring within these response categories may have an impact on DIF detection rates. This too, is an area that merits further investigation.

Item discrimination is another factor that has been shown to influence DIF detection rates. Most assessments are developed for the sole purpose of providing information about test takers' differences either on the construct purportedly measured by the test or on some external criterion which the test scores are supposed to predict (Crocker & Algina, 1986). In either case, the parameter of interest must provide information about how well each item effectively discriminates between examinees of high and low ability on the construct the test was developed to measure. The item discrimination parameter is one such factor that can provide this essential piece of information. Therefore, it is important to examine the effect of various item discrimination parameter magnitudes on DIF detection rates for polytomously scored items especially when they occur in an item's different response categories.

Although, there have been a number of studies (Hidalgo & Lopez-Pina, 2004; Rogers & Swaminathan, 1993; Spray & Miller, 1994) that have examined DIF caused by differences in the discrimination (*a*) values for the reference and focal group within conditions, known as non-uniform DIF, this study focused on the conceptually simpler case where within conditions the *a*-values were equal for the reference and focal groups even while other factors varied. This scenario within conditions, in which the *a*-values are equal for both groups, is known as uniform DIF. In this investigation only conditions in which uniform DIF is present was investigated. That is, in this study no DIF was added to the discrimination parameter, rather, in simulating uniform DIF, the *a* parameter was

kept the same for the reference and focal groups at each level of the *b* parameter. The *b* parameter varied, however, causing the item to be more difficult for one of the groups (in most cases, the focal group).

In addition to the impact that item discrimination, item difficulty, DIF patterns of DIF, and size of DIF, can have on DIF detection methods, some studies (e.g., Ankenmann, Witt, & Dunbar, 1999; Wang & Su, 2004) have shown that large differences in group ability can affect DIF detection rates in polytomous items. How large group ability differences impact the Mantel, GMH, and OLR procedures under a variety of study conditions is an area in need of further investigation.

In sum, two major sources of test information, item discrimination and item difficulty, were examined in the context of DIF occurring within response categories, under differing patterns, and under varying DIF magnitudes. That is, this Monte Carlo simulation study compared the Type I error and power of the Mantel, GMH, and OLR procedures to detect DIF for tests that contain only polytomous items under conditions in which variation occurred in (a) the item discrimination parameter values (b) category intersection parameter values (c) DIF magnitudes (d) score categories containing various DIF patterns; and (d) differences in average latent trait between groups. Specifically, this investigation sought to answer the following question: When a test contains only polytomous items, to what extent are the power and Type I error rates of the Mantel, GMH, and OLR affected by the variation in 1) the item discrimination parameter values, 2) category intersection parameter values, 3) DIF patterns within score categories, and 4) average latent trait differences between the reference and focal groups?

CHAPTER 2

REVIEW OF THE LITERATURE

Educational reform efforts, driven by legal and ethical challenges, have led to increased demands on test developers to provide more equitable approaches to testing. To meet these demands for fair testing, a variety of alternatives to the traditional dichotomously scored multiple-choice item have been developed (Potenza & Dorans, 1995; Zwick et al., 1993). These alternatives include item formats with multi-steps (i.e., polytomous items) that provide more opportunities to gather examinee information than their dichotomous counterparts. Cognitive assessments, such as constructed responses and essays, are examples of item formats that can gather detailed information about an examinee's deep level of understanding. In recent years, nationwide testing and assessment programs such as the College Board Advanced Placement tests and the National Assessment of Educational Progress (NAEP) have included polytomously scored items in their assessments (Zwick, Donoghue, & Grima, 1993). In fact, as of 1993, half of all statewide writing assessment programs relied solely on writing samples to assess students' levels of proficiency in grammar, spelling, and sentence construction (Welch & Hoover, 1993).

Many performance assessments, which may include a writing component, are used for selection purposes; because of this trend, in recent years there has been increased attention by test developers to ensure that tests are fair to all applicants (Zumbo, 1999). This phenomenon has led to an increase in the development of performance assessments

that can provide more information on the extent of an examinee's level of understanding (Welch & Hoover, 1993). Indeed, the use of open-ended and constructed-response instruments to assess educational outcomes has greatly increased during the last decade (Ankenmann, Witt, & Dunbar, 1999).

In an attempt to meet the goal of gaining more examinee information, test developers, particularly within the educational assessment arena, are increasingly utilizing testing instruments that can assess information from all item choices rather than from only the two score categories of right or wrong (De Ayala, 1993). Essays and constructed response items where examinees are required to write a lengthy response to a question or statement also require a scoring method that is capable of reflecting the examinee's depth of knowledge. Performance task items (e.g., student portfolios) requiring an examinee to demonstrate his or her understanding of the concept by developing a product would also demand a more complex scoring method other than right/wrong if detailed information is to be gathered on the examinee's level of comprehension. Additionally, many mathematics tests are composed partially, if not entirely, of many problems that are multi-step. Partial credit is often awarded for evidence that the student has understood the problem, has adopted an appropriate strategy, has attempted to solve the problem but has committed a computational error (Masters, 1984). These types of item formats typically require item response models that can represent the relationship between examinee trait level and the probability of responding in a particular category.

Polytomous IRT Models

The increased information on the underlying trait that multiple response categories provide is one of the main reasons for the proliferation of polytomous item formats (Embretson & Reise, 2000; Ostini & Nering, 2006). An item response theory (IRT) framework can be used to understand the relationship between an examinee's item performance and his/her underlying trait. To illustrate this relationship, first a dichotomous IRT model will be used. This relationship can be modeled by a monotonically increasing function known as the item characteristic curve (ICC). The ICC models the probability of a correct response given the examinee's ability and the item's characteristics. The form of an ICC describes how changes in trait level relate to the probability associated with moving from one response category to the next along the entire trait continuum (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). That is, the ICC specifies that as the level of the trait increases, so too does the probability of success on an item. In dichotomous items the relationship between the item characteristic and the underlying trait is modeled by a single monotonically increasing curve, providing information for at most one trait level. Figure 1 shows an ICC for a one parameter logistic model, also known as the Rasch model, the simplest and the most widely used of the IRT models (Hambleton, Swaminathan, & Rogers, 1991). The primary assumption of the one-parameter model is that the item difficulty is the only item characteristic that influences examinee performance. ICCs for the one-parameter logistic model are represented by the following equation:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad (1)$$

where $P_i(\theta)$ is the probability that a randomly chosen examinee with ability, θ, answers item *i* correctly. The natural log base (2.718) is represented by *e*. The item's difficulty parameter, *b*, indicates an ICC's location on the ability scale where the likelihood of a correct response is 0.5. Examinees with higher *b* values have higher probabilities of answering the item correctly than do examinees with lower *b* values regardless of group membership (Hambleton et al., 1991).

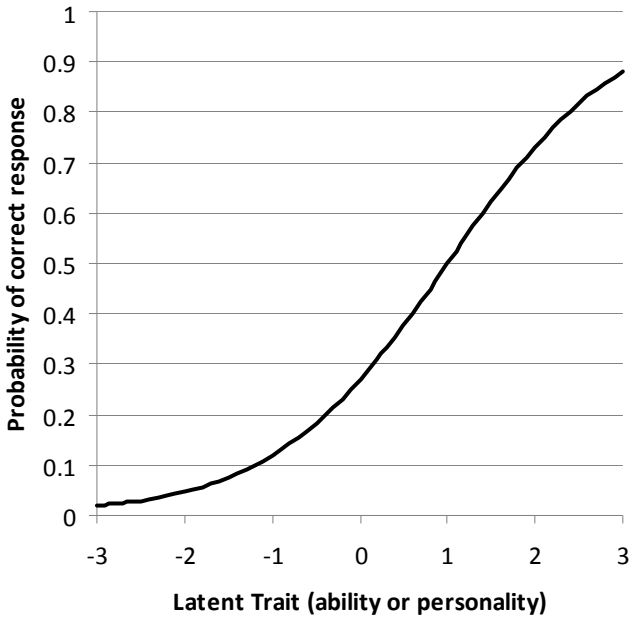

*Figure 1***.** ICC for a dichotomous item where *b*=1.0.

Unlike a dichotomous item, a polytomous item has multiple response categories and must be modeled by multiple curves called category response curves (CRCs). CRCs represent the probability of an examinee, at a given trait level, responding in a particular category (Embretson & Reise, 2000). The CRCs for a polytomous item are located above

the various trait levels, thereby providing multiple pieces of information along the trait continuum. Figure 2 illustrates category response curves for a polytomous item. The CRC of Figure 2 depicts multiple trait levels along the latent trait continuum. It also shows how multiple *b* parameters, located at each category response curve intersection, indicate where on the latent trait continuum a category response becomes more probable for one person than another when their ability levels differ (Embretson & Reise, 2000). A commonly used polytomous model to describe examinee data once the items have been scored is the Generalized Partial Credit Model (GPCM; Muraki, 1992, 1993).

Generalized Partial Credit Model

Muraki's (1992, 1993) Generalized Partial Credit Model (GPCM) is a polytomous IRT model that is a generalization of Master's (1982) Partial Credit Model (PCM). The GPCM, unlike the PCM, allows slope parameters within items to vary (e.g., allows for a discrimination parameter). The GPCM can be used for analyzing test items that award partial credit for the successful completion of at least one of the steps in a multiple step problem. Thus, the GPCM is naturally suited for modeling item responses from cognitive tests (e.g. math problems, essays) where partially correct answers are possible (Embretson & Reise, 2000). The GPCM is also appropriate for rating scale items, such as the Likert scale used in many attitudinal or personality assessments, in which respondents rate their beliefs and where items share a fixed set of rating points (De Ayala, 1993;

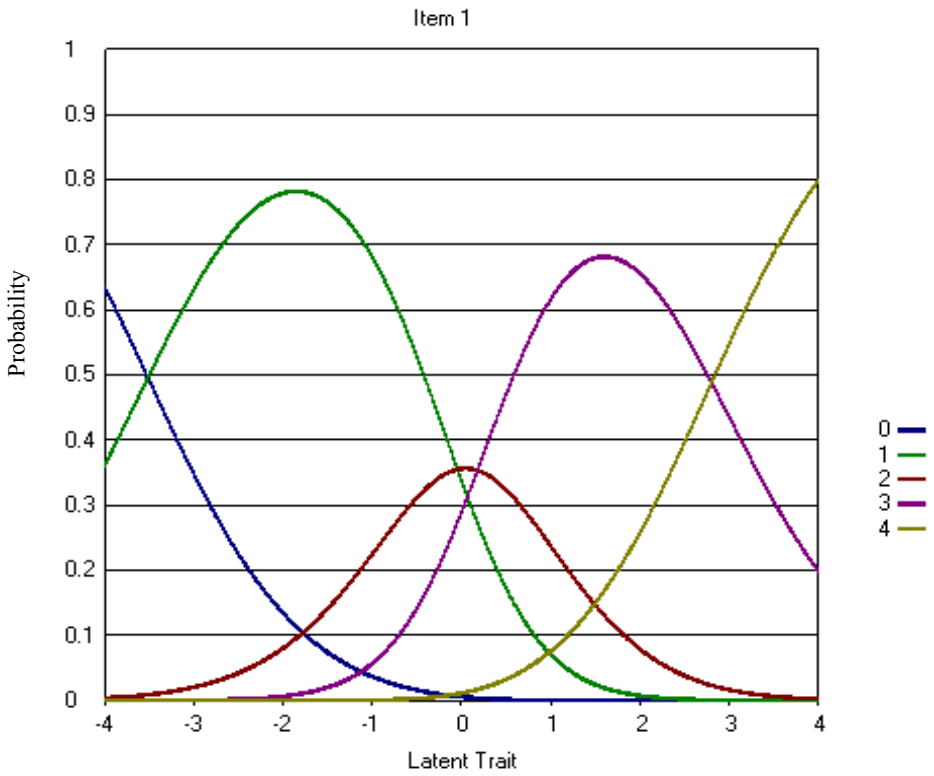

*Figure 2***.** Category response curves for a 5-category polytomous item under the Generalized Partial Credit Model where $\alpha = 0.683$, $b_1 = -3.513$, $b_2 = -0.041$, $b_3 = 0.182$, and $b_4 = 2.808$. The parameter values for this figure were taken from Embretson (2000).

Embretson & Reise, 2000). The GPCM requires that the steps within an item be completed in order, although the steps need not be in order of difficulty or be equally difficult (De Ayala, 1993).

Masters (1984) provided an example of a mathematics problem that illustrated the PCM which can also be applied to the GPCM: "*How many pages are there in a book that requires 2989 digits to number the pages?"* (p. 20). The mathematics problem required an examinee to execute five ordered levels of performance to arrive at the correct solution. One point was awarded for the first step if the examinee demonstrated some evidence of having understood the problem. Another point was awarded when the examinee showed evidence of having adopted a strategy that enabled him/her to work

toward a solution. The third point was acquired when the strategy was pursued to near completion of the problem and the fourth point was awarded for a correct solution. If steps 2, 3, and 4 were answered incorrectly, however, no credit was awarded.

The GPCM, unlike the PCM, does not belong to the family of Rasch models because item slopes may vary. Allowing the slopes to vary assumes that one item can be more effective than another in discriminating among examinees thus providing more insight into the test item characteristics than the PCM (Ostini & Nering, 2006). When the GPCM is used, the probability that an individual with a given trait level, θ, will obtain a category score of x for item i with $m_i + 1$ (from 0 to m_i) categories is given by:

$$P_{ix}(\theta) = \frac{\exp \sum_{j=0}^{x} \alpha_i (\theta - \delta_{ij})}{\sum_{r=0}^{M} [\exp \sum_{j=0}^{r} \alpha_i (\theta - \delta_{ij})]} \quad (2)$$

where the item discrimination parameter or slope is represented by α_i and δ_{ij} is the jth category intersection parameter or item step difficulty for item i. Assuming that the examinee has completed previous steps, the category intersection parameters represent the point on the latent-trait scale where one category response becomes more likely than the previous step (Embretson, 2000).

While the category intersection parameters provide difficulty information, the slope parameter, α_i, can be viewed as "indicating [*sic*] the degree to which categorical responses vary among items as θ level changes" (Embretson, 2000, p. 112). In the GPCM, there is one discrimination parameter, α_i, for each item. Note from Figure 2 that as the slope parameter becomes smaller than 1.0 the CRCs become less peaked, while as

the slope parameter becomes larger than 1.0 the CRCs tend to become more peaked. Figure 3 illustrates an item with a large slope ($\alpha = 1.499$). Note how peaked the curves are relative to the model in Figure 2.

In the GCPM, although the response categories must be ordered, the category intersection parameters, δ_{ij}, need not be. The greater the value of a particular δ_{ij}, the harder a particular step is relative to other steps within an item (Embretson & Reise, 2000). For example, the transition from Category 1 to Category 2 may be more cognitively demanding than the transition from Category 2 to Category 3. Within polytomous items, transitioning from one score category to the next can increase the likelihood that the transition is more difficult for one group of examinees than the other. This difference can exist simply because one group has a greater academic ability than the other; however, when both groups have equivalent abilities but transitioning from one response category to the next is still more difficult for one group than it is for the other, investigations into these differences are imperative. Differential item functioning analysis is one such statistical approach that addresses these kinds of discrepancies.

Differential Item Functioning

Ensuring that tests do not contain differential item functioning (DIF) has become an important part of developing equitable assessments. Methods for detection of DIF have grown, in large part, due to the legal and ethical need to measure respondent performance without bias (Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001). DIF is typically identified using a statistical technique that employs a significance test to determine whether an item functions differently for one group of examinees over another. DIF occurs when individuals from different subgroups who are equivalent on a latent trait

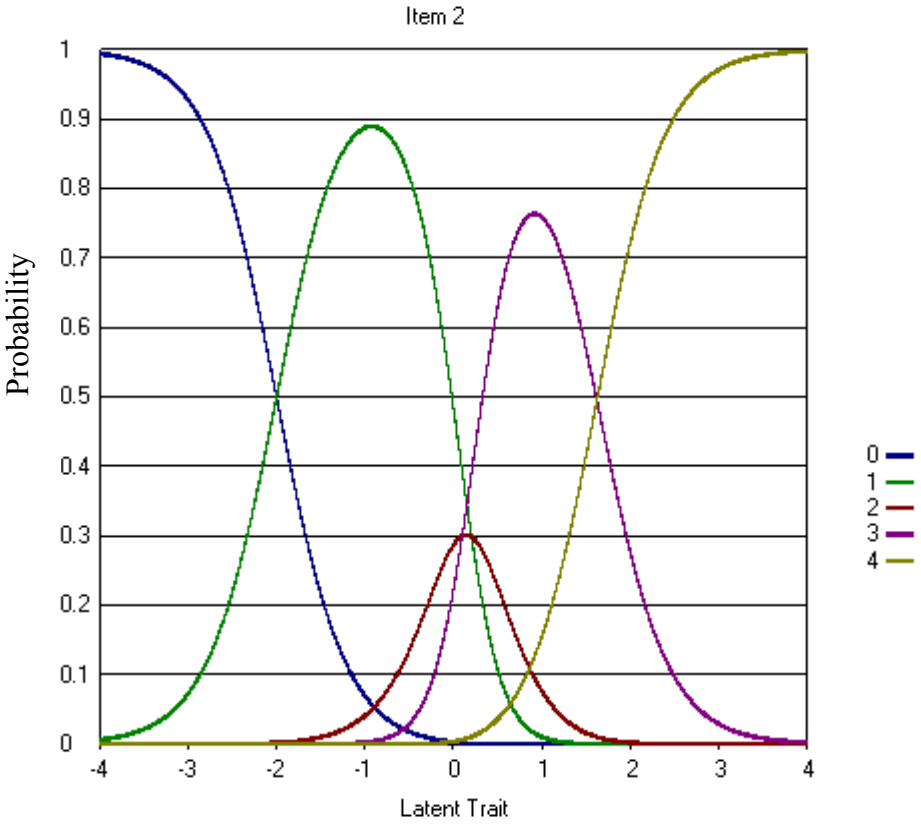

Figure 3. Category response curves for a 5-category polytomous item under the Generalized Partial Credit Model where $\alpha = 1.499$, $b_1 = -1.997$, $b_2 = -0.210$, $b_3 = 0.103$, and $b_4 = 1.627$. The parameter values for this figure were taken from Embretson (2000).

such as ability, show differing probabilities of obtaining the correct response to an item (Hambleton et al., 1991).

DIF analyses compare the item performance of the two groups but the comparison is made only on those members with the same level of ability. For example, if ability is estimated using the total test score, then the difference in item performance of both groups at various score levels would be compared. If those differences between the two groups at various score levels consistently occur across a large portion of the ability continuum, the item is said to function differentially for the two groups, and thus DIF is said to be present (Penfield & Lam, 2000). In DIF analyses, the subgroup under

investigation is referred to as the focal group, and the other, the reference group. The focal group most typically is the minority group of interest (e.g., African-Americans, females, etc.). In any high-stakes context where legal challenges on the issue of fairness arise it is strongly recommended that DIF analyses be conducted in order to provide evidence for items that are potentially biased (Zumbo, 1999). This recommendation is also echoed by *The Standards for Educational and Psychological Testing* (1999) which states:

> When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability and/or linguistic groups in the population of test takers in the content domain measured by the test, test developers should conduct appropriate studies when feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups. (p. 81)

DIF analysis is a way of addressing concerns related to test validity and fairness. Very often differences in the validity of the test at the item level may be interpreted as item bias or result in the item or test being regarded as invalid (Williams, 1997). DIF analyses provide a further means of obtaining evidence that the interpretation of test scores is indeed accurate.

Bias versus DIF

According to Camilli and Shepard (1994), statistical errors and item multi-dimensionality are the two main factors that lead to items being flagged for DIF. Statistical error can take the form of Type I error where items are falsely identified as possessing DIF; item multidimensionality occurs when a test intended to measure only one construct simultaneously measures two or more. For DIF to occur, it is assumed that examinees have been matched on ability for only one construct, the intended construct that the test purports to measure. When statistical evidence points to DIF (i.e., items

functioning differently for groups of examinees who have been matched on ability), then it is important to examine whether or not the discrepancy is due to such factors as Type I error or item multidimensionality. When DIF occurs and there is valid reason to believe that the source of DIF is due to one or more irrelevant constructs being measured by the test, item bias is said to exist. Determining whether or not a test measures one or more irrelevant constructs typically involves review of the items in question by an expert panel to identify items that appear to be more difficult for one group of examinees than another. In the event that an item functions differently for one of the groups, a decision must be made about whether to retain or delete the item from the test. However, without a substantive review of the item by experts, test developers would not know if the reason the item exhibited DIF was due to a construct-relevant or irrelevant dimension of the test. It is important to remember that statistical techniques employed in the detection of DIF provide statistical evidence that determine only whether or not an item functions differently in the two groups. They give no indication of whether the observed DIF constitutes bias (Donoghue & Allen, 1993). That is why it is vital to follow any statistical DIF review with a substantive review of the item by an expert panel.

Bias infers that one group is unfairly advantaged over the other. But groups may differ in their response to an item for reasons other than bias, such as impact. Item impact occurs when examinees from different groups have differing probabilities of responding correctly to an item (this definition differs from that of DIF because DIF only can be said to occur if the groups have been matched on ability). This differing response pattern occurs not because the item unfairly advantages one group over another but because there

are true differences between the groups in the underlying ability that the item is measuring (Zumbo, 1999).

Conceptually DIF is assessed by plotting the ICCs separately for each group under investigation and then comparing them along the trait continuum. Figure 4 is an example of an item that displays substantial DIF with a very large area between the two ICCs. This type of DIF is known as uniform DIF because the two ICCs do not cross, indicating that there is no interaction between ability level and group membership. Nonuniform DIF occurs when there is an interaction between the ability level and group membership. In Figure 5 the ICCs cross, indicating nonuniform DIF. Also, Figure 5 illustrates that Group 1 is favored for those individuals who score at or below the mean (i.e., $\theta \leq 0$) and that Group 2 is favored for those scoring above the mean (i.e., $\theta > 0$); supplying further evidence that nonuniform DIF exists.

DIF Detection Methods

DIF is detected using one of two methods – a parametric approach or a nonparametric approach. The parametric approach assumes a specific IRT model to investigate DIF whereas a nonparametric approach does not. Because parametric approaches rely on specific item response models to investigate DIF, model misspecification is often a problem as even a small amount of misfit may result in unacceptable levels of Type I error (Bolt, 2002). Parametric methods also require large sample sizes of at least 500 each for the reference and focal groups (Narayanan & Swaminathan, 1996; Wang & Su, 2004). In contrast, nonparametric methods are advantageous over parametric methods because they do not assume specific item response models, require large sample sizes, or intensive computation.

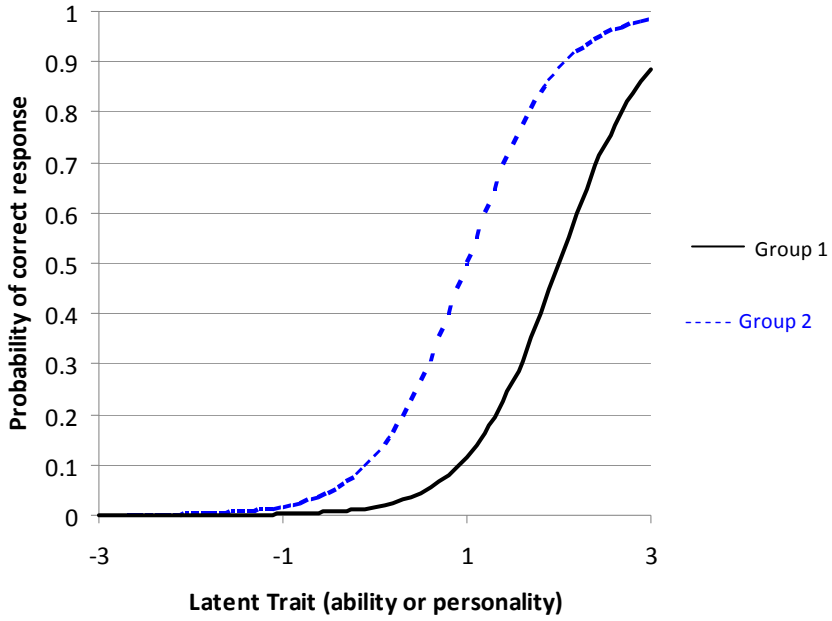

Figure 4. Uniform DIF where a = 1.2, b = 1 for group 1; a=1.2, b = 2 for group 2.

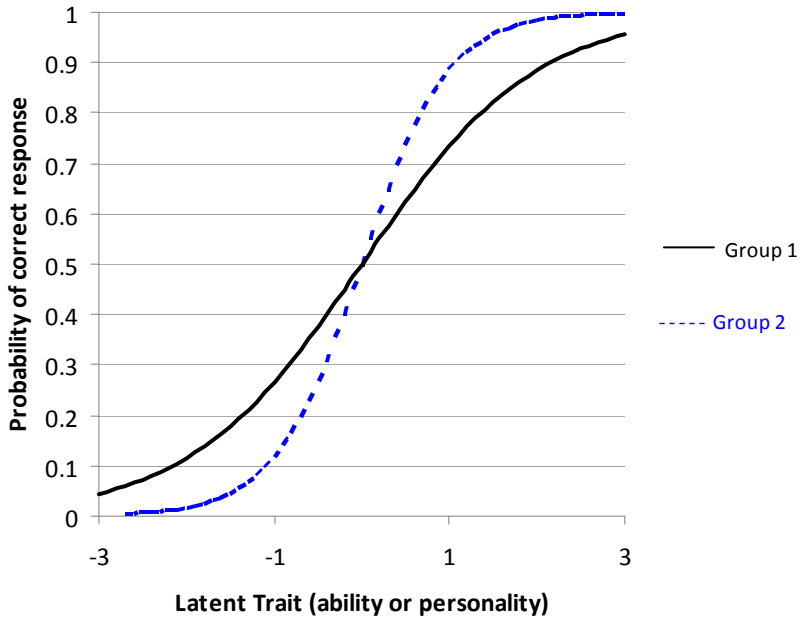

Figure 5. Nonuniform DIF where a = 1.2, b=0 for group 1; a =0.6, b =0 for group 2.

When studying DIF in dichotomous items, the Mantel-Haenszel (MH) method (Holland & Thayer, 1988; Mantel & Haenszel, 1959) is one of the most popular nonparametric DIF detection procedures. While DIF detection is predominantly used in the cognitive context, where answer choices are usually dichotomous, it can also be used in areas where items are typically polytomously scored (Furlow, Fouladi, Gagné, & Whittaker, 2007). For items scored polytomously, one of two direct extensions of the MH method are typically used; either the Mantel method (Mantel, 1963) or the generalized Mantel-Haenszel method (GMH; Mantel & Haenszel, 1959; Somes, 1986). The Mantel procedure was developed for ordered polytomous response data whereas the generalized Mantel-Haenszel method is used when the response categories are treated as nominal data. Detecting DIF in polytomous items can be challenging as DIF can reside within some or all score categories within an item. The Mantel and the GMH are two methods that test for DIF within the various score categories of an item.

A third method for DIF detection with polytomous items is the Logistic Regression (LR) procedure. This method is more commonly known as the Ordinal Logistic Regression (OLR) procedure when the polytomous items are ordered data, such as Likert type item formats (i.e., not important, important, very important). One of the main advantages of the LR method over the GMH is its capacity to detect uniform and nonuniform DIF (French & Maller, 2007; Swaminathan & Rogers, 1990). Because the logistic regression technique is model-based it can test coefficients for significant uniform and nonuniform DIF separately within the same equation (French & Miller, 1996; Kristjansson et al., 2005). In the DIF detection process, once an item is identified as having DIF, it is further classified as having uniform DIF if the probability of a correct

response is the same across all ability levels. If, however, there is an interaction between ability level and group membership, then the item is classified as having nonuniform DIF. For ordered polytomous data, the OLR technique involves recoding ordinal data into $T - 1$ dichotomous sets (where T is the number of score categories).

Matching

Unlike the LR procedure, both the Mantel and the GMH methods rely on significant row mean differences between groups to signal for potential DIF in an item. These row mean differences are based on observed scores. The observed scores serve as the matching criteria that are used to determine if there is a difference in performance on a given item after examinees have been matched on the estimated latent trait or some measure of proficiency, otherwise known as the matching variable. The matching variable could be thought of as a proxy for an individual's performance or ability in the area that is being assessed. According to Mapuranga, Dorans, and Middleton (2008), "matching is a way of establishing score equivalence between groups that are of interest in DIF analyses" (p. 6). When score equivalence between groups is established, DIF analysis is facilitated by enabling relative comparisons between the reference and focal groups.

Types of matching include thin or thick matching. Thin matching uses the total score as the matching variable. Thick matching, however, involves pooling total score levels to form the matching variable. According to Donoghue and Allen (1993), thick matching is the preferred matching technique because 1) estimation of the cell frequencies for each of the levels of the matching variable is more stable and 2) more data can be used because fewer cells have zero frequencies. Equal interval matching is an

example of thick matching because the total score scale is divided into a number of equal widths. In equal interval matching, the researcher calculates several interval width combinations to find the best interval width that will yield the fewest number of cells with missing score data. Overly fine matching is to be avoided as it very often results in elimination of much of the data. In addition to equal interval matching, it is often recommended that extreme scores be pooled into larger widths since there are typically fewer scores at the extremes.

After including all items in the total score, this matching variable then needs to be "purified" (Zumbo, 1999). That is, items that are flagged for DIF are omitted, and the scale or total score is then recalculated. The recalculated total score would then be a "pure" matching criterion for the subsequent DIF analyses of each item that was previously flagged as having DIF. This first stage is called the criterion purification stage (Wang & Su, 2004a). At the second stage, the refined matching score for each subsequent test would include the studied item as well as all of the DIF-free items. This two step purification procedure has been found to increase DIF detection rates by increasing power and reducing Type I error when the proportion of DIF items exceed 10% (Fildago, Mellenbergh, & Muniz, 2000; Holland & Thayer, 1988; Miller & Oshima, 1992).

Dichotomous items: DIF detection when item discrimination and difficulty parameters vary. DIF detection methods for dichotomous items have been extensively researched; however few studies involving dichotomous items have examined the impact on DIF detection when *both* the item difficulty and discrimination parameters vary. Clauser, Mazor, and Hambleton (1991) conducted one of the earliest studies on dichotomous items to explore if varying certain item's discrimination and

difficulty parameters would increase the likelihood that those items would be overlooked by the MH statistic. Only uniform DIF conditions were simulated as the *a* values were equal for the reference and focal groups within conditions, even while other factors varied. (The *a* values, however, were high in some conditions and low in others). These differing *a* values were examined in combination with both between group differences in the *b* parameters and various overall levels of item difficulty.

Five data sets each containing 16 biased items were simulated. Responses were generated for 2000 examinees (i.e., 1000 per group) using a 3 parameter logistic (PL) IRT model. Additionally, to mimic conditions found in practice, item discrimination and difficulty parameters were generated based on estimated values from a Graduate Management Admission Test (GMAT) administration. The *c* (pseudo-guessing) parameters for all items were held at a constant value of 0.20. Sixteen additional items were added to the original 59 items, to create a total of 75 items. Four *a* parameters (.25, .60, .90, 1.25) were crossed with five *b* parameter values (-.2.5, -1.0, 0, 1.0, 2.5). Four levels of difference in the *b* parameter value (DIF) between groups (.25, .50, 1.00, and 1.50) were crossed to produce a total of 80 studied items. These items were then combined 16 at a time with the 59 non-studied items to produce five 75-item tests.

Clauser et al.'s (1991) results indicated that the amount of the DIF, the absolute value of the item discrimination parameter, *a*, and the value of the item difficulty parameter, *b*, influenced the likelihood that an item would be identified as having DIF. Specifically, the probability that the item would be flagged for DIF increased dramatically as the DIF magnitude increased. A similar but less dramatic effect for increases in the discrimination parameter was also noted, though, the absolute value of

the difficulty parameter was found to strongly influence the results, but only for the extreme upper range of the difficulty scale (i.e., *b* values close to +2.50). Under the unequal ability distributions condition, the MH detected fewer items with DIF. Under the equal ability distributions condition, five items of moderate difficulty that went undetected for DIF in the unequal ability distributions condition, were identified as exhibiting DIF. In general, the main effects of the *a* and *b* parameters were found to be partially dependent on the DIF magnitude. That is, when there was no DIF, no difference in difficulty or discrimination was found to impact the MH value. The Clauser et al. findings showed that the MH was most effective with examinees from groups with equal ability distributions, but the results of their investigation also demonstrated that the MH remained useful with groups of considerably different ability.

Another simulation study on dichotomous items that involved manipulation of item discrimination and difficulty parameters was conducted by Donoghue and Allen (1993). This investigation examined the impact that various types of pooling (e.g., thin versus thick matching) had on the MH's ability to detect DIF. Simulated item responses were generated by a 3PL IRT model. DIF in the studied item, the studied item difficulty, and its discrimination were crossed within tests to produce 42 studied items that were added to the core items in each test condition to form a test. Three levels (0.3, 1.0, and 1.5) of the discrimination parameter and 7 levels (-1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5) of the difficulty parameter, b_R, for the reference group were generated. The discrimination parameters were equal for the reference and focal groups at each level of the difficulty parameter. The focal group IRT difficulty parameter, b_F, for the studied item differed

with each value of the reference group difficulty parameter by 0.3 (i.e., no DIF when $b_F = b_R$, and DIF favoring the reference group when $b_F = b_R + 0.3$) .

Both difficulty and discrimination of the studied item were found to have a strong effect on the ability of the MH to detect DIF under thin or thick matching conditions. For very easy items, the mean value for the MH statistic was negative, indicating that the item was more difficult for the focal group, while for hard items the mean value was slightly positive, indicating that the item was somewhat easier for the reference group. Additionally, the findings from this investigation indicated that when the studied item difficulty was increased, the means for non-DIF and DIF items were decreased. Further, the study results indicated that for easy items, increasing the discrimination in the studied item made the between group difficulty differences larger; thus resulting in better DIF detection.

Another study on dichotomous items in which the item discrimination and difficulty parameters were manipulated was conducted by Rogers and Swaminathan (1993). In this investigation, the relative efficacy of the LR and MH procedures under varying conditions was examined. The first part of their study examined the distributions of the test statistics of the OLR and MH procedures. Four conditions were simulated to study the effect that sample size and degree of model-data fit would have on the MH and LR's power to detect DIF. Two levels of model-data fit ("good" fit and "poor" fit) were crossed with two levels of sample size (250 per group and 500 per group). Test data for which the LR model provided "good" fit were generated using the 2PL IRT model whereas a 3PL IRT model was used to generate "poor" fit data. Because the LR method specifies a lower asymptote of 0, when generating the "poor" fit model all *c* (i.e., pseudo-

guessing parameter) values were set at 0.2. Item parameters were chosen for a 40-item test and were selected to produce an approximately standard normal distribution of test scores. For each combination of sample size and data model fit, 100 replications of the data were performed. Because item characteristics can affect the estimation of parameters and hence the distribution of the test statistic for the LR model, five of the 40 items were chosen to vary in level of difficulty and discrimination. The results of the first part of this study revealed that for very easy items, the *c* parameter had an effect only on the very lowest part of the trait scale; subsequently the LR model provided an acceptable fit for the data over nearly all of the range. The study findings also showed that for very difficult items, the *c* parameter affected a much larger part of the trait scale, hence misfit of the LR model was more pronounced. The researchers concluded that this particular problem may not be serious for most achievement tests in which there are few very difficult but highly discriminating items.

The second part of the study by Rogers and Swaminathan (1993) investigated the power of the LR and MH procedures to detect uniform and nonuniform DIF. The item discrimination and difficulty parameters were manipulated to 1) simulate uniform and nonuniform DIF, and 2) determine if this variation would affect parameter estimation, hence DIF detection under the LR procedure. In simulating uniform DIF, the discrimination parameters for the reference and focal groups were kept the same but the item was manipulated to be more difficult for the focal group. Thirty-two conditions were simulated and were obtained by crossing two levels of model-data fit (good or poor fit, simulated as in Study 1 using the 2PL model and the 3PL model), two levels of sample size (250 per group and 500 per group), two levels of test length (40 items and 80 items),

two levels of the shape of the test score distribution (normal and negatively skewed), and two levels of percent of items with DIF (15% including the studied item, and 0% other than in the studied item). Both uniform and nonuniform DIF were simulated within each condition. Four sizes of DIF, corresponding to the area values of .2, .4, .6, or .8 were examined in the uniform and nonuniform conditions. In this investigation, the size of DIF in an item was quantified by the area between the generating ICCs. Area was calculated by using a formula provided by Raju (1988). In simulating uniform DIF, the *a* parameters for the reference and focal groups were kept the same but the *b* parameters for the two groups were different. Sixteen items in the uniform DIF condition were obtained by crossing the level of the *a* (low or high) and *b* parameters (both low, both moderate) for the two groups and the size of the DIF area. Four types of items were studied: (1) low *b*, high *a*; (2) moderate *b*, low *a*; (3) moderate *b*, high *a*; and (4) high *b*, high *a*.

In simulating nonuniform DIF, the researchers kept the *b* parameters for the reference and focal groups the same, but varied the *a* parameters for the two groups. Fifteen items showing nonuniform DIF were created by varying the level of the *b* parameter (low, moderate, high), the level of the *a* parameters for the two groups (both low and high), and the size of the DIF area (.2, .4, .6,.8). Four types of items were studied: (1) low *b*, low *a*; (2) moderate *b*, low *a*; (3) moderate *b*, high *a*; and (4) high *b*, low *a*. In all, 35 items with DIF were constructed. To generate tests with 15% DIF, five items needed for a test length of 40 items or 11 items needed for a test length of 80 were selected from the set of DIF items. These items were kept constant in all of the analyses and were included in the test for the sole purpose of providing the desired degree of test score contamination. DIF statistics were not calculated for these items. Each of the 35

DIF items to be studied was added separately to the test, its DIF statistics calculated; the item was then removed from the test and replaced by another of the items showing DIF. This procedure was used so that DIF could be studied in each item under the same conditions. Similarly, for the condition showing no DIF in all of the non-studied items, each DIF was separately added to the test. Each condition was replicated 20 times, and the percentage of items exhibiting uniform and nonuniform DIF that were detected by the MH and the LR were compared. Item parameter values taken from real data sets were used to generate unbiased items that produced either a normal or skewed test score distribution with normally distributed trait levels for both groups.

For the uniform DIF conditions, the study results demonstrated that the LR and MH procedures were almost equally effective in detecting uniform DIF. The study results also showed that the items with DIF that were more easily detected by both the LR and MH procedures were items of moderate difficulty and high discrimination. For these types of items, the detection rates were as much as 15% greater than for the other item types. For the nonuniform DIF condition, the lowest detection rate for the LR procedure occurred with items of moderate difficulty and low discrimination, and the highest detection rate occurred for items of moderate difficulty and high discrimination. The power of the MH procedure to detect strictly nonuniform DIF was extremely low for items of moderate difficulty. For items of low difficulty, the MH detection rate was still approximately 15% lower than the LR detection rate; but for items of high difficulty, the detection rates were almost identical.

Finally, another more recent study on the MH and LR procedures to detect DIF on dichotomous items when the magnitude of the item discrimination and difficulty

parameters vary was conducted by Hidalgo and Lopez-Pina (2004). In this investigation, a data set containing a reference group and a focal group each with a sample size of 1000 and a normal ability distribution with a mean of 0 and standard deviation of 1 was simulated. Twenty-five tests each containing 59 non-DIF items and 16 items with DIF were simulated. The item responses for these 75 items were simulated by a 3PL model. The *c* parameters were set at 0.20 and the *a* and *b* parameters for the 59 non-DIF items were taken from a previous study by Narayanan and Swaminathan (1996). Four hundred studied items (i.e., 25 tests with 16 DIF items in each test) were generated and randomly assigned to 1 of the 25 tests. For these four hundred items, five levels of difficulty (-1.5, -1.0, 0, 1.0, or 1.5) and four levels of discrimination (0.25, 0.60, 0.90, or 1.25) were chosen. The following conditions were manipulated for the 16 items under investigation: (a) four levels of uniform DIF magnitude (0, 0.30, 0.60, and 1.00) and (b) five levels of nonuniform DIF magnitude (0, 0.25, 0.50, 0.75, and 1.00). In each of these conditions, the differences were generated to favor the reference group over the focal group.

The results of the investigation by Hidalgo and Lopez-Pina (2004) revealed that, in general, the number of correctly identified DIF items was greater when the LR was used but that a modified MH procedure that was employed showed similar power as the LR. (The modified MH procedure involved splitting the sample into two groups on the ability scale, a high ability level and a low ability level and then implementing the MH procedure separately for the two groups. This was done for the purpose of improving nonuniform DIF detection). The study findings also showed that as the magnitude of uniform and nonuniform DIF increased, so too did the detection rates of the various methods. Additionally, the investigation also revealed that for the nonuniform DIF

conditions, to some extent, depending upon item difficulty, the more discriminating items were slightly less likely to be flagged for DIF. This finding was attributed to the fact that as the *a* parameter increased, the area between the ICCs associated with a given between group difference in *a* parameters decreased. When the difference manipulated in the *a* parameter was 1, generating an item with symmetrical nonuniform DIF, (i.e., differences in only the discrimination parameters), the researchers found that the area between the ICCs of the reference group, calculated using Raju's (1988) formula, decreased as the *a* parameter increased. For example, when $a_{diff} = 1$ and $b_{diff} = 0$, Raju's area measures were 2.606, 0.848, 0.476, and 0.290 when the discrimination parameters for the reference group a_R were 0.25, 0.6, 0.9, and 1.25, respectively. This pattern was also found when the differences between the reference and focal group discrimination parameters were smaller. When the asymmetrical nonuniform DIF magnitude was small ($b_{diff} = 0.3$), a similar pattern to the one found in the symmetrical nonuniform DIF condition was found. In those situations in which the differences in the difficulty parameter was small, the more discriminating items had areas between the ICCs that were smaller than the less discriminating items.

The results of this investigation indicated that when DIF was symmetrical nonuniform the LR procedure had the highest correct DIF detection rates, with 68.75% of DIF items correctly identified compared to 61.25% for the modified MH procedure and 50% for the standard MH procedure. Under the asymmetrical nonuniform DIF condition, the OLR and modified MH procedure showed very similar results (87.9% overall for each procedure) except under conditions with large DIF magnitudes (i.e., 1.0 and 0.75). In those situations, the modified MH procedure was found to be more powerful than the

other two procedures. However, when the DIF magnitude was smaller, LR was more powerful than the standard MH and modified MH procedures. In contrast, for identifying symmetrical nonuniform DIF the LR procedure performed better than the standard MH and modified MH procedures, correctly identifying 68.75% of the DIF items, compared to 61.25% for the modified MH procedure and 50% for the standard MH technique. For uniform DIF conditions, the standard MH procedure performed slightly better than the LR and modified procedures, correctly identifying 55% of the DIF items compared to 53.33% for the LR and 50% for the modified MH procedure.

The results of this investigation found that, overall, because of the small differences in power among the modified MH, and LR procedures, all three methods appeared to be highly comparable.

Polytomous items: DIF detection when item discrimination and difficulty parameters vary. The few studies on the MH and LR procedures, specifically when the item discrimination and difficulty parameters vary, represent a very small percentage of the numerous studies on the efficacy of these two methods to detect DIF in dichotomous items. For polytomous items, the percentage of studies on DIF detection methods is much smaller than that for dichotomous items and of those few studies on polytomous items, very few have examined the efficiency of the Mantel, GMH, and ordinal logistic regression procedures under various study conditions, particularly under conditions manipulating the item discrimination and category intersection parameters (sometimes referred to as difficulty parameters).

Zwick, Donoghue, and Grima (1993) conducted one of the earliest simulation studies involving polytomous data generated by the PCM. Their investigation examined

the efficiency of the Mantel and the GMH in detecting DIF in performance tasks. In each simulated condition, the total number of test items was 25. Twenty-four of these items were DIF-free and were used only to compute the matching score; the 25th item was the studied item. The first 20 items were dichotomous and the last 5 were four-category items. The studied item always had four categories. The factors that were manipulated across the simulated conditions were focal group ability distributions (2 levels) and characteristics of the studied item (27 levels). The item characteristics that were of primary interest included the difficulty parameters, DIF patterns, and DIF magnitude. The studied item characteristics included three sets of reference group parameters, four patterns of DIF (constant, balanced, low-shift, and high-shift), and two non-zero DIF magnitudes (.1 and .25), resulting in 24 types of DIF items. In addition, a null condition in which the studied item had the same parameters for the reference and focal groups (e.g., no DIF) was included for each of the three sets of reference group parameters, resulting in a total of 27 studied items crossed with two ability levels for a total of 54 conditions.

DIF was modeled by starting with a set of reference group parameters and then increasing the item difficulties by a value of .1 or .25. Four patterns of DIF were considered. 1) Constant DIF. In this condition, all of the transitions from a given item score category to the next highest category were assumed to be more difficult for the focal group, and the degree to which they were more difficult remained constant. 2) Balanced DIF. In this condition, the transition from the lowest to the second category was more difficult for the focal group, while the transition from the third category to the highest was easier for the focal group. The remaining transition was the same for the two

groups. 3) Low-shift DIF. In this condition, the transition from the lowest to the second category was more difficult for the focal group. The remaining transitions were the same for both groups. 4) High-shift DIF. In this condition, the transition from the third to the highest category was more difficult for the focal group. The remaining transitions were the same for both groups.

Four ways of computing the matching variable were crossed with these 54 simulation conditions. The four ways of computing the matching variable were determined by whether or not scores on polytomous items were rescaled in computing the matching score (2 levels) and whether or not the studied item was included in the matching score (2 levels). In regards to computing the matching variable, different weights were assigned to the dichotomous and polytomous items. In one condition no rescaling was performed, so that the score range for the dichotomous items was 0-1, and for polytomous items the range was 0-3. In the other condition, the rescaling of the matching variable was performed by dividing the score on the polytomous items by 3 resulting in a score range of 0-1 for both types of items so that now both item formats had the same weight. Another condition that was varied in the computation of the matching variable involved whether or not the studied item was included in the matching score.

One hundred replications were performed for each of the 216 (54 x 4) conditions. In each condition, samples of 500 observations were selected from the reference and focal group distributions, yielding a total sample size of 1000. The reference group distribution was normal (i.e., $N(0,1)$ in all conditions; the focal group distribution was either $N(0, 1)$ or $N(-1,1)$.

Zwick et al.'s (1993) study findings demonstrated that when the reference and focal groups' means are the same, less than ideal procedures (e.g. excluding the studied item from the matching variable) for calculating the matching variable do not have an adverse effect on the Mantel or GMH's ability to detect DIF; however, when the means differ, the method of computation can lead to an increase in Type I error. Additionally, this study showed that scores on polytomous items should not be rescaled when calculating the matching variable.

Although in this investigation, power to detect DIF for the Mantel and GMH procedures was much lower than the widely accepted rate of 80%, the findings merit discussion. Study results regarding DIF patterns revealed that for the constant DIF condition, the Mantel procedure was more powerful than the GMH but that for the balanced DIF condition, the GMH was far superior. In fact, in the balanced DIF condition when the DIF magnitude was 0.25, the rejection rate for the GMH was 25% but only 4% for the Mantel procedure. For the shift-low and shift-high conditions both procedures produced similar rejection rates. For all DIF patterns except the constant pattern, when the DIF magnitude was 0.1, detection rates were extremely low (8% or less). For the constant DIF pattern with a DIF magnitude of 0.1, the rejection rates were approximately 18% and 11%, respectively for the Mantel and GMH methods. The rejection rates for the balanced, shift-low and shift high DIF patterns at the same magnitude of 0.1, ranged from 4.5% to 17%. For a DIF magnitude of 0.25, the rejection rates for the Mantel were 13% for the shift-low condition, 14% for the shift-high, 4% for the balanced condition, and 76% for the constant condition. For the GMH under the same 0.25 DIF magnitude,

rejection rates were 13% for the shift-low condition, 17% for the shift-high condition, 25% for the constant condition, and 60% for the constant condition.

Because the GMH compares the odds that focal group members will be assigned a particular score category to the odds for the reference group, conditional on a matching variable, Zwick et al. (1993) concluded that for data in which the entire response distribution and not just the means is of interest, the GMH might be the best method to use. For most DIF analyses of polytomous items, the researchers concluded that the Mantel (1963) approach which involves comparing the means for two groups, conditional on a matching variable and which takes the ordering of score categories into account, would be more useful.

Another simulation study on polytomous data was conducted by French and Miller (1996). This study evaluated the power of the LR procedure to detect DIF in polytomous items. Several versions of a test containing 25 items were generated. Each item had four score categories, with a total possible score on the item ranging from zero to three. For each of the tests, a single item (i.e., the studied item) was simulated to contain DIF. Item scores were generated using Muraki's (1992) GPCM. Two sample sizes, 500 and 2,000 were used for each group to represent small and large sample sizes. Ability estimates were generated from a standard normal distribution $N(0,1)$. Because LR procedures require a dichotomous dependent variable, polytomous data must be recoded into a number of dichotomous sets, each of which is then ready for a separate regression analysis. Three approaches that are extensions of logistic modeling for polytomous data were used in this study. These methods involved using a different coding scheme and are called the *continuation ratio logits, cumulative logits,* and *adjacent categories* models.

These models use the logit or the ratio of the probability of getting the category correct to the probability of getting the category incorrect (Swaminathan & Rogers, 1990). The three methods were compared for their power in identifying various forms of DIF in dichotomized polytomous data. Each coding scheme produced one less regression than the number of score categories.

The continuation ratio logit coding scheme combines the chi-squared results from separate regressions and adds them to give an overall result, or omnibus test. One disadvantage of this coding scheme is that increasingly smaller amounts of data are isolated across regressions to examine for the presence of DIF. The continuation ratio logits coding scheme involved comparing the zero score category to all other categories combined in the first regression. In the second regression, simulees that received a score of one were compared to those that received scores of two or three. Finally, in the third regression, simulees that received a score of two were compared to those that received a score of three.

The cumulative logits model involves no loss of data in the coding scheme. It simultaneously estimates multiple equations. The number of regression equations it estimates will always be one less than the number of categories in the dependent variable. For example, suppose the dependent variable *Y* for an item has four score categories, three equations will be estimated. Equation one, will model the odds of responding in score category 1 compared to score categories 2, 3, and 4; equation two will model the odds of responding in score categories 1 and 2 compared to score categories 3 and 4; and equation 3 will model the odds of responding in score categories 1, 2, and 3 compared to category 4. For ordinal response data, cumulative logists can be modeled with the

proportional odds model as in the example above. Proportional odds imply that the odds of responding in any of the score categories are the same for both the reference and focal groups. For example, if the regression coefficient for the focal group is significant, that would imply that DIF is present, and that the odds of scoring in a particular category are different for that group.

In the adjacent categories logit model, DIF can be isolated between adjacent categories because each response probability is compared to its neighboring response probability - not to all other score categories as in the cumulative logits model. In the adjacent categories logit model, simulees that received a score of zero were compared to those that received a score of one in the first regression; those simulees that received a score of one were compared to those who received a score of two in the second regression; and those that received a score of two were compared to those that received a score of three in the third regression.

In addition to the two sample size conditions, four other conditions were examined (termed Conditions 1 to 4). In the first three conditions, nonuniform DIF was generated (e.g., differences in the *a* parameter between the reference and focal group), and in the fourth condition, only uniform DIF was generated. These four conditions were applied to only the studied item in each simulation. The remaining 24 items had identical item parameters for both groups. For the 25^{th} item, the differences in the *a* parameters between the focal and reference groups were 0.5, 1.0, and 1.5 for Conditions 1, 2, and 3, respectively. In the fourth condition, only *b* parameters associated with category intersections one and three were varied, such that for examinees of equal abilities receiving a score of one was more difficult for the focal group than the reference group,

and receiving a score category of two was more difficult for examinees in the reference group than in the focal group. It was hypothesized that changes in sample size and item parameters would be expected to affect the power of the LR procedure to detect DIF. Specifically, in regards to changes in item parameters, it was expected that as the differences in the discrimination parameters *a*, between the focal and reference groups from the first to third DIF condition increased, so too would the spread between the ICCs. Therefore, it was hypothesized that DIF in the 25th item would be detected more in the second condition than in the first, and more often in the third condition than in the second.

In the fourth condition, the difficulty parameter *b*, associated with category intersections of one and two were varied, such that receiving a score of one was easier for examinees of equal abilities in the reference group than in the focal group, and receiving a score of two was easier for examinees of equal abilities in the focal group than in the reference group. For the focal group, for each of the three category intersections, the *bs* were as follows: 1 for the first category intersection; -1 for the second category intersection; and 2 for of the third category intersection. For the reference group the *bs* for the first, second, and third category intersections were -2, 1, and 2, respectively.

The differences in the *a* parameters between the focal and reference groups were 0.5, 1.0, and 1.5 for Conditions 1, 2, and 3, respectively. It was hypothesized that the difference in the discrimination parameters between the two groups from the first to the third DIF conditions would result in more spread between the ICCs which would, in turn, lead to DIF being detected more often as the differences increased. In the fourth condition, only *b* parameters associated with score categories one and three were varied,

such that for examinees of equal abilities receiving a score of one was more difficult for the focal group than the reference group, and receiving a score category of two was more difficult for examinees in the reference group than in the focal group. For the focal group, for each of the three category intersections, the *bs* were as follows: 1 for the first category intersection; -1 for the second category intersection; and 2 for of the third category intersection. For the reference group the *bs* for the first, second, and third category intersections were -2, 1, and 2, respectively. Ability estimates were generated from a standard normal distribution $N(0,1)$. Two sample sizes, 500 and 2,000 were used for each group to represent small and large sample sizes. Power was calculated as the proportion of times DIF was correctly identified in the 25th item across the 100 replications of the study. The results of this investigation indicated that, in general, larger sample sizes led to greater power for LR in detecting DIF in polytomous items. For Conditions 1 to 3, larger samples resulted in greater power for detecting DIF. Also, nonuniform and uniform DIF were detected more frequently for each coding scheme as the *a* parameter increased for the focal group. Additionally, as the differences in the item parameters between the focal and reference groups increased, as in going from Condition 1 to 3, the more frequently nonuniform and uniform DIF were detected. For the continuation ratio and cumulative logits coding schemes, when the sample size was large (i.e., 2,000), power to detect nonuniform DIF for each condition was adequate and in some cases strong under all regressions. For example, under Condition 1, the power rates for the first, second, and third regressions were 93%, 96%, and 47% respectively.

The three coding schemes' powers to detect uniform DIF were as effective in Condition 4 as they were in the first three conditions that had nonuniform DIF. DIF was

still detected using LR techniques with the smaller sample size (500) but the power was much weaker. When sample sizes were reduced to 500, power decreased for the first regression under the continuation ratio and cumulative logits coding schemes, and overall for the adjacent categories coding scheme. The adjacent categories coding scheme lost large amounts of data in all three regressions, consequently as expected, was not as powerful in detecting DIF as the other two procedures. Overall, the results showed that with the large sample size (2,000), LR is a good choice for detecting DIF in polytomous items.

Wang and Su (2004b) investigated the performance of the Mantel and GMH on detecting DIF when tests contain polytomous items exclusively, a variety of percentages of DIF items, and various DIF patterns. Even though several studies on dichotomous items have shown that the MH performs poorly (i.e., begins to lose control over Type I error) when the percentage of DIF items in a test is increased to 10% or 15% (Fidalgo, Mellenbergh, & Muniz, 2000; Miller & Oshima, 1992; Narayanan & Swaminathan, 1994, 1996), no studies had examined whether or not those DIF item percent increases would adversely affect the performance of the Mantel and the GMH procedures for polytomous items. Recently, however, some studies have demonstrated that it is the ASA (average signed area- an area measure that depicts the average degree to which a test favors the reference group; the test as a whole advantages the reference group when ASA is positive, the focal group when it is negative, and neither group when ASA is zero) that is more critical than the percentage of DIF items to the type I error of the Mantel and GMH procedures (Wang & Su, 2004a; Wang & Yeh, 2003). Therefore, in addition to examining whether or not the percentage of polytomous DIF items would adversely

affect the power of the Mantel and GMH procedures to detect DIF, Wang and Su's study examined whether the ASA had the same effects on the Mantel and GMH for polytomous items as it did on the MH for dichotomous items The further ASA is from zero, the worse the MH and IRT-based DIF detection methods should perform. The ASA magnitudes in the Wang and Su investigation ranged from 0 to 0.125.

Wang and Su (2004b) manipulated the following eight independent variables in their study: DIF detection methods (two levels), test purification procedure (three levels), item response model (two levels: the PCM, and the GRM), mean ability difference between groups (four levels), test length (three levels), DIF pattern (three levels), magnitude of DIF (two levels), and DIF percentage (six levels). Hence, a total of 5,184 conditions were simulated. The generating item parameter estimates of the reference group were adopted from 10 five-point items of 4^{th}, 8^{th}, and 10^{th}-grade students' responses to the mathematics tests of the Wisconsin Student Assessment System (Kim & Cohen, 1998). The sample sizes of the reference and focal groups were each 500. Members of the reference group were generated from a normal distribution with a mean of zero and a standard deviation of 1 (i.e, N (0, 1)). Members of the focal group were generated from $N(0,1)$, $N(-0.5,1)$, $N(-1,1)$, or $N(-1.5,0)$. Even though several previous studies (e.g., Donoghue et al., 1993; Mullis, Dossey, Owen, & Phillips, 1993) reported that a difference in latent trait means of one standard deviation is typical and realistic between certain reference and focal groups, a difference in latent trait means of one and a half standard deviations was used to explore the boundary of the two methods. Test lengths of 10, 20, and 30 items, representing short, medium, and long tests, respectively

were simulated. Item parameters were for the 10-item tests but were repeated two and three times for the test lengths of 20 and 30 items, respectively.

Three DIF patterns were manipulated: constant, balanced, and constant-item/balanced test. In the constant pattern, the amount of DIF within a polytomous item is held constant across score categories and is unidirectional. In this scenario DIF is usually simulated to favor one group all the time, typically the reference group. For example, all the location parameters within a DIF item of the focal group would be larger than those of the reference group by a constant amount. In the balanced pattern, the magnitude of DIF is balanced across all score categories. For example, in Wang and Su's (2004b) investigation, the balanced pattern was manipulated by allowing the first two location parameters of the focal group to be larger than those of the reference group by an amount *s*, and allowing the last two location parameters to be smaller than those of the reference group by *s*. The constant-item/balanced test pattern exists when all of the DIF patterns are constant within items but balanced within tests. This can sometimes result in a canceling out effect (Wang & Su, 2004a). For example, half of the DIF items could have a positive DIF magnitude, and the other half an equal amount of negative DIF, so that the DIF contamination within tests is cancelled out between groups.

Study findings indicated that under the balanced pattern, both procedures yielded good control over the average Type I error, even when the percentage of DIF items was as high as 40%. Under the constant pattern, however, the Mantel and GMH began to lose control over their average Type I error once the percentage of DIF items reached approximately 30% when the average signed area (ASA) equaled 0.03 and 20% when ASA was 0.05. In the Wang and Su (2004b) investigation, empirical statistical power was

assessed by the proportion of times out of 1,000 replications that an item was correctly identified as possessing DIF. An alpha level of 0.05 was used. Study findings demonstrated that in general, the average power of the Mantel was higher than the GMH under all but the balanced patterns. Study findings also revealed that the Mantel and GMH yielded much higher power under the constant and constant-item/balanced-test pattern than the balanced pattern. Also, varying test length when the PCM was the generating model had no effect on the Mantel and GMH's Type I error and power. However, when the GRM was the generating model, the Mantel and GMH yielded slightly better control over Type I error in the 20 item tests than in the 10 item tests.

Recently, Su and Wang (2005) investigated the power of the Mantel, GMH, and Logistic Discriminant Function Analysis (LDFA) methods to detect DIF in polytomous items. The LDFA is also model based like the LR procedure but uses group membership as the dependent variable rather than item score. Thus, in LDFA, the probability of group membership is estimated from total score and item score. In the Su and Wang (2005) study, responses to dichotomous items were generated under the Rasch (1960) model or the 3PL model and the PCM or the GRM was used to generate polytomous item responses. The simulated test consisted of 20 dichotomous items and five 4-point items. The following eight independent variables were manipulated: DIF detection methods (3 levels), test purification procedure (3 levels), item response model (3 levels: Rasch + PCM, 3PL model + PCM, and 3PL model + GRM), mean ability difference between groups (4 levels), test length (2 levels), DIF pattern (5 levels), magnitude of DIF (3 levels), and DIF percentage (6 levels). Hence, a total of 11,178 conditions were simulated. The sample sizes of the reference and focal groups were each 500. The

members of the reference group were generated from $N(0,1)$. The members of the focal group were generated from $N(0,1)$, $N(-0.5,1)$, $N(-1,1)$, or $N(-1.5,1)$. The tests contained either 25 or 50 items. Five DIF patterns were manipulated: constant, balanced, shift-low, shift-high, and constant-item/balanced test. There were five polytomous items in the 25-item tests. The number of DIF items in those tests was set at 0, 1, 2, 3, 4, or 5. Hence, the percentage of DIF items in the tests were 0%, 4%, 8%, 12%, 16%, or 20%, respectively. All the dichotomous items were DIF free. Because there were only five polytomous items in the 25 item test, a 20% DIF rate meant that all five polytomous items exhibited DIF. The studied item was always included in the matching score, as suggested by Zwick et al. (1993). One hundred replications were conducted for each condition.

The results of this investigation showed that under the constant pattern, all three methods began to lose control over their average Type I error rate once the percentage of DIF items exceeded 12%. The study results also indicated that under the constant pattern, the average power of the Mantel and LDFA methods was similar but higher than that of the GMH. Under the balanced pattern, all three methods had good control over the average Type I error even when the percentage of DIF items was as high as 20%. Additionally, under the balanced pattern, the Mantel and LDFA outperformed the GMH in their power to detect DIF. Under the shift-high and shift-low patterns, the average power of the three methods to detect DIF was roughly the same. Under the constant-item/balanced test pattern, the average power of the Mantel and LDFA methods was similar but higher than that of the GMH and finally, the study findings indicated that the higher the percentage of DIF items, the more inflated the average Type I error became.

Another recent study (Kristjansson et al., 2005) compared the efficiency of the Mantel, GMH, LDFA, and the unconstrained cumulative logits ordinal logistic regression (UCLOLR) to detect DIF using a simulated 26-item test. In this investigation, all items were ordinal with four score levels and the 26th item was the designated studied item. The difficulty and discrimination parameters for the 25 nonstudied items were held constant. Item responses were generated using the GPCM. Thus, both slope and location were estimated. Type I error and power were examined. Kristjansson et al. (2005) manipulated the following four variables in their investigation: 1) presence and type of DIF, 2) studied item discrimination, 3) groups' sample size ratio, and 4) skewness in ability distribution. In total, 96 study conditions were tested (12 studied item levels x 2 sample size ratios x 2 skewness levels x 2 ability differences). Four hundred replications were performed for each study condition.

Three DIF conditions--null DIF, uniform DIF, and nonuniform DIF--were simulated in the studied item. In the null DIF condition, the *a* and *b* parameters were the same for the reference and focal groups. In the uniform DIF condition, *a* parameters for both groups remained the same, but *b* parameters for the focal group were increased by 0.25 at each transition between score levels. This last condition made it more difficult at each transition for the focal group to achieve a higher score. In the nonuniform DIF condition, *b* parameters were equivalent between the two groups, but the *a* parameter for the reference group was higher than that for the focal group. The actual size of the difference varied depending on the studied item discrimination. That is, when the studied item discrimination was 0.8, the *a* parameter was 1.8 for the reference group, when the studied item discrimination was 1.2, the *a* parameter was 2.5 for the reference group and

3.2 for the reference group when the studied item discrimination was 1.6. The values were selected so that the DIF magnitude in the uniform and nonuniform conditions would be approximately equal.

To examine the effect that varying item discriminations would have on DIF detection, three different studied item discriminations (low, $a = 0.8$; moderate, $a = 1.2$; and high, $a = 1.6$) were assessed. Two levels of group ability differences were also evaluated: (a) equal reference and focal group ability distributions (i.e., both groups had a mean of 0 and a standard deviation of 1) and (b) unequal distributions (i.e., mean ability for the focal group was -0.5 and 0 for the reference group; a standard deviation of 1 for both groups).

When assessing the effect of group sample size ratio on DIF detection, Kristjansson et al. (2005) held the total sample size constant at 4,000 and used two levels (the equal and unequal conditions) of group sample size ratio. In the equal condition, the reference and focal groups had sample sizes of 2,000 and in the unequal condition, the sample size of the reference group was four times larger (3,200) than that of the focal group (800).

Two levels of skewness were compared to determine the effect of skeweness on the ability of the Mantel, GMH, LDFA, and UCLOLR to detect DIF. A moderate negative skew (-0.75) in ability distributions for both the focal and reference groups and no skewness were compared.

The results showed that none of the four DIF detection procedures showed any significant departure from the nominal Type I error rate of 0.05. However, for both the Mantel and the LDFA, a slightly increased Type I error rate was related to the interaction

between high item discrimination and group ability differences. Further, all four procedures had excellent power (greater than 0.963) for detecting uniform DIF. However, the GMH and UCLOLR's power to detect uniform DIF was directly related to item discrimination and sample size. Both had higher power for uniform DIF when item discrimination was moderate or high than when item discrimination was low. Additionally, their power to detect uniform DIF was slightly lower when the sample size ratio was 4:1.

Several issues are unexplored in the above five studies on the performance of the Mantel, GMH, and OLR. First, it is seldom the case that most simulated tests contain only polytomous items. This is the case in the above mentioned studies as most of the simulated tests consisted of a set of dichotomous items and a set of polytomous items. This mixed format was designed to mimic educational assessments that contain a mixture of multiple choice and essay items. However, many educational assessments consist exclusively of polytomous items (e.g., constructed-response tests). Thus, the findings obtained from tests with both dichotomous and polytomous items might not be directly applicable to tests that contain polytomous items exclusively. Even though the Wang and Su (2004b) investigation examined the performance of the Mantel and GMH for a set of polytomous items and the French and Miller (1996) study investigated the performance of the LR for a set of polytomous items, no research to date has compared the performance of all three methods (i.e., the Mantel, GHM, and OLR) on DIF detection when a test contains only polytomous items. It is, therefore, of interest to determine how the Mantel, GMH, and OLR procedures perform when a test contains ordered polytomous items exclusively.

Second, Wang and Su (2004b) manipulated three DIF patterns (constant, balanced, and constant-item/balanced–test) in their investigation and Su and Wang (2005) manipulated five (constant, balanced, shift-low, shift-high, and constant-item/balanced-test) but because the item parameters were generated using the PCM and the GRM, the item discrimination value was not manipulated. It is, therefore, of interest to examine the effectiveness of the Mantel, GMH, and OLR when the item discrimination parameters vary under different DIF patterns.

Third, only a few studies (e.g., Donoghue & Allen, 1993; Hambleton et al., 1993) have examined the efficacy of DIF detection methods, namely the MH, when both the item discrimination and difficulty parameters are manipulated. No relevant research in the literature on polytomous item formats has examined the effects that varying *both* the item discrimination and category intersection parameters would have on the Mantel, GMH, and Ordinal logistic regression procedures to detect DIF.

Fourth, a few researchers have found that large differences in group ability can lead to high Type I error; this effect is further exacerbated when the studied item discrimination is high (Kristjansson et al., 2005). What happens to the Type I error rate of the Mantel, GMH, and OLR procedures under conditions of high item discrimination in the studied item and large group ability differences particularly under different category intersection magnitudes needs further investigation.

Finally, no study to date has examined the effectiveness of the Mantel, GMH, and the OLR on detecting DIF in polytomous items under various DIF pattern conditions when the GPCM is used as the generating parameter model. Whether the GPCM has the same effect as the PCM on the DIF detection rates for polytomous items under certain

study conditions is still an unanswered question. This study will attempt to provide some insights and answers to the above issues and questions. This investigation compared the Type I error and power of the Mantel, GMH, and OLR procedures to detect DIF for tests that contain only polytomous items under a variety of conditions. Specifically, when (a) the item discrimination parameters vary (b) category intersection parameters vary (c) DIF magnitudes vary (d) score categories contain various DIF patterns; (d) there are differences in average latent trait between groups.

CHAPTER 3

METHODOLOGY

A Monte Carlo simulation study was conducted to assess the power and Type I error performance of three DIF detection methods: the Mantel, GMH, and OLR procedures using Muraki's (1992; 1993) Generalized Partial Credit Model as the generating IRT model. Several factors were varied: studied item discrimination, studied item difficulty, DIF magnitude, differences in group ability, and DIF patterns. For each condition, the DIF detection rates of the Mantel, GMH, and OLR were compared. The comparison of these three DIF detection methods was assessed by tallying up the number of times each method correctly identified items with DIF as well as the number of times each method falsely identified an item as exhibiting DIF. DIF detection rates were examined based on statistical significance using an alpha of .05.

Research Questions

The following research question guided this study:

1. When a test contains only polytomous items, to what extent are the Type I error rates and power of the Mantel, GMH, and OLR affected by variation in the item discrimination parameter, category intersection parameter, DIF magnitudes, DIF patterns within score categories, and average latent trait differences between the reference and focal groups?

The Mantel and GMH Statistics

The Mantel, a nonparametric observed score method, is a polytomous extension of the Mantel-Haenszel (MH) method (Mantel, 1963) that takes into account the ordered nature of the response categories when testing for DIF. The Mantel provides a statistic with a chi-square distribution of one degree of freedom when the null hypothesis of no DIF is true (Meyer, Huynh, & Seaman, 2004; Zwick, Donoghue, & Grima, 1993). Calculation is based on item means for groups that have been matched on some measure of proficiency. The GMH is a generalized Mantel-Haenszel statistic used for nominal response data based on group differences across the entire response distribution. The GMH is sensitive to uniform as well as nonuniform DIF because it tests along the entire response distribution, whereas the Mantel has been reported as only being able to consistently detect uniform DIF because it tests differences in mean item scores (Kristjansson et al., 2005). To implement the Mantel or GMH, the data are arranged into a 2 x T x K contingency table, where T is the number of response categories in a polytomous item, and K is the number of levels of the matching variable. One 2 x T table is required at each of the K score levels, as shown in Table 1.

The values $y_1, y_2, \ldots, y_T$ represent the possible T scores of the item. The values n_{RTk} and n_{FTk} represent the number of the reference and focal group members, respectively who receive an item score of y_t at the kth level of the matching variable. The "+" symbol represents the summation over a particular index. The test statistic for the Mantel is represented by

$$\chi^2 = \frac{[\sum_k F_k - \sum_k E(F_k)]^2}{\sum_k Var(F_k)} \qquad (3)$$

Table 1

The kth Level of a 2 x T Contingency Table

	Item Score					
Group	y_1	y_2	y_3		y_T	Total
Reference	n_{R1k}	n_{R2k}	n_{R3k}	....	n_{RTk}	n_{R+k}
Focal	n_{F1k}	n_{F2k}	n_{F3k}	....	n_{FTk}	n_{F+k}
Total	n_{+1k}	n_{+2k}	n_{+3k}	....	N_{+Tk}	n_{++k}

Note. This table was taken from Wang and Su (2004).

where F_k is the sum of the focal group scores at the k^{th} level of the matching variable:

$$F_k = \sum y_1 \, n_{Ftk} \tag{4}$$

where $E\,(F_k)$ is

$$E\,(F_k) = \frac{n_{F+K}}{n_{++k}} \sum y_1 \, n_{+tk}\,; \tag{5}$$

and the Var (F_k) is

$$\frac{n_{R+k} n_{F+k}}{n^2{}_{++k}{}^{(n_{++k}-1)}} [(n_{++k} \sum y_t^2 n_{+tk}) - (\sum_t y_t n_{+tk})^2] \tag{6}$$

The null hypothesis for the Mantel test is that there is no association between the row mean score of the studied group (i.e., the focal group) and the row mean score of the reference group (i.e., the comparison group). A lower row mean score indicates lower performance by a particular group. DIF is present in the studied item whenever there is a

difference in the row mean scores at that particular score level. Under the null hypothesis, the test statistic in Equation 3 has a chi-square distribution with 1 degree of freedom. The null hypothesis for the Mantel is that at a given score level, there is no association between the item score and group membership. If the null hypothesis is rejected, members of the reference and focal groups who have been matched on ability differ in their mean performance on the studied item; consequently, the item is flagged as exhibiting DIF.

The GMH treats the response categories as nominal data. The test statistic for the GMH has a chi-square distribution with *M*-1 degrees of freedom:

$$X^2_{GMH} = \left(\sum A_k - \sum E(A_k)\right)]' \left(\sum V(A_k)\right)^{-1} \left(\sum A_k - \sum E(A_k)\right) \qquad (7)$$

where

$$A'_k = (n_{Rik}, n_{R2k}, \ldots, n_{R(T-1)k}), \qquad (8)$$

$$E(A'_k) = n_{R+k}\, n'_k /n_{++k}, \qquad (9)$$

$$n'_k = (n_{+1k}, n_{+2k}, \ldots, n_{+(T-1)k}), \qquad (10)$$

$$V(A_k) = n_{R+kn_{F+k}} \left(\frac{n_{++k} diag(n_k) - n_k n'_k}{n^2_{++k}(n_{++k-1})} \right) \qquad (11)$$

where diag(n_k) is a (*T*-1) x (*T*-1) diagonal matrix with elements n_k. Whereas A_k, $E(A_k)$ and $V(A_k)$ are scalars in the dichotomous case, A_k, $E(A_k)$ are vectors of length *T*-1 in the polytomous case, corresponding to (any) *T*-1 of the *T* response categories, and $V(A_k)$ is a (*T*-1) by (*T*-1) covariance matrix. Following the notation of Table 1, *R* represents the reference group, and diag (n_k) is a (*T*- 1) x (*T* – 1) diagonal matrix with elements n_k. The

statistic in Equation 7 has a chi-square distribution with *T*-1 degrees of freedom under the null hypothesis for the GMH that there is no conditional association between group membership and response category. If the null hypothesis is rejected, then conditional association is found, thus the item would be found to exhibit DIF.

Ordinal Logistic Regression

The equation for the OLR is as follows:

$$Y = b_0 + b_1\text{TOT} + b_2\,\text{GROUP} + b_3\text{TOT*GROUP}_\text{i} + e_i, \tag{12}$$

where e_i ,(the error term) is normally distributed with a mean of zero and a variance of ($\pi^2/3$). Y, the dependent variable, is the item response (0 or 1) after recoding ordinal data into $K - 1$ dichotomous sets (where K is the number of response categories). For polytomous items, there will be K-1 logistic regression functions for each response category. The independent variables are represented by TOT, the total score; GROUP, group membership (reference or focal); and TOT*GROUP, the interaction between group and total score. In Equation 13 it is seen that the dependent variable is equal to the natural log of a probability of a correct response, p, divided by the probability of an incorrect response, (1-p), where Y is the natural log of the odds ratio; yielding the following equation:

$$Y = \ln\left[\frac{p}{(1-p)}\right] = b_0 + b_1\text{TOT} + b_2\,\text{GROUP} + b_3\text{TOT*GROUP}_\text{i} + e_i. \tag{13}$$

DIF detection using the LR procedure provides a test of DIF conditionally on the relationship between the dependent variable (item response) and the total scale score while simultaneously testing for the presence of both uniform and nonuniform DIF. In testing for the presence of DIF, each model term's (TOT, GROUP, and TOT*GROUP)

contribution to the model is evaluated for improvement of model fit. The item exhibits uniform DIF when the GROUP effect is statistically significant and TOT*GROUP effect is not, whereas the item has non-uniform DIF when the interaction effect of TOT*GROUP is statistically significant (Hidalgo & Lopez-Pina, 2004).

Because the response data in this investigation was ordinal, the cumulative logits model was used and the proportional odds model was employed. Multiple equations were simultaneously estimated. Three regression equations (one less than the number of score categories) were estimated. Equation one, modeled the odds of responding in score category 1 compared to score categories 2, 3, and 4; equation two modeled the odds of responding in score categories 1 and 2 compared to score categories 3 and 4; and equation 3 modeled the odds of responding in score categories 1, 2, and 3 compared to category 4. The null hypothesis for the proportional odds model is that the odds of responding in any of the score categories are the same.

Study Design Conditions

In this study, conditions were simulated in order to investigate power and Type I error of three commonly used DIF detection procedures for polytomous items. Power was investigated in conditions where DIF was present whereas Type I error was examined for false detection of DIF in conditions where the studied item did not include DIF. The power and Type I error of the Mantel, GMH, and OLR procedures were examined under several factors on the ability to detect DIF in a simulated 20-item test. Both the Type I error and power conditions had factors that were held constant and factors that varied. The factors that varied included DIF patterns, DIF magnitude, differences in group ability, studied item discrimination, and studied item difficulty. The non-varying factors

investigated in this study are presented in Table 2 and the varying factors are presented in Table 3.

Factors Held Constant

Generating model. The GPCM was the polytomous IRT model used to generate the data for the reference and focal groups in this study. The GPCM has been used in a number of simulation studies on DIF detection (e.g., Chang et al., 1996; French & Miller, 1996; Kristjansson et al., 2005). The primary reason for this design choice is that under the GPCM, item slope parameters may vary (i.e., items can differ with respect to discriminating power), whereas under the PCM they are not free to vary (i.e., all items have the same discriminating power). Therefore, in order to investigate the power and the extent to which the Mantel, GMH, and OLR procedures maintain control of their Type I error rate under various item discrimination magnitudes, simulated data sets were generated using the GPCM.

Table 2

Fixed Factors in the Study

Factor Category	*Factor*
Generating Model	Muraki's Generalized Partial Credit Model
Number of Replications	1,000
Test Length	20 items
Number of Item Categories	4
Percent of Items with DIF	5%
Ability Distribution Type	Normal
Type of DIF	Uniform
Sample Size	1,000

Table 3

Factors Varied in the Study Design

Factor Category	*Factors*
Patterns of DIF	1. Constant 2. Shift-low 3. Shift-high 4. Balanced
DIF Magnitude	1. 0.10 2. 0.25 3. 0.40
Difference in Group Ability	*1.* $N_R(0, 1)$ and $N_F(0, 1)$ *2.* $N_R(0, 1)$ and $N_F(-0.5, 1)$ *3.* $N_R(0,1)$ and $N_F(-1,1)$
Studied Item Discrimination	1. 0.8 2. 1.2 3. 1.6
Studied Item Difficulty	1. $b_1 = -2, b_2 = 0, b_3 = 2$ 2. $b_1 = -1, b_2 = 0, b_3 = 1$ 3. $b_1 = 0, b_2 = 1, b_3 = 2$ 4. $b_1 = -2, b_2 = -1, b_3 = 0$

Number of replications. Although previous simulation studies (Kristjansson et al., 2005; Rogers & Swaminathan, 1993; Su & Wang, 2005; Wang & Su, 2004b) employed 100 replications, in this investigation, one thousand replications were completed for each condition to ensure the accuracy of the empirical estimations of the sampling distribution characteristics.

Test length. There were 20 items generated under the GPCM. This is a common test length used in studies investigating DIF. This test length also closely approximates that used in studies investigating the effectiveness of various DIF detection methods when the GPCM is the data generating model (e.g., Chang et al., 1996; French & Miller, 1996; Kristjansson et al., 2005).

Number of item categories. Each item was generated to have four score categories (i.e., one point for each correct step) to simulate four ordered levels of performance that an examinee must execute in order to arrive at the correct solution to the problem.

Percent of items with DIF. Five percent of the 20 items in this simulation study contained DIF; therefore, only a single item was assessed for DIF while the other 19 items were simulated to be DIF-free. The item with DIF was always the 20^{th} item.

Ability Distribution. Item parameters were generated using a standard normal distribution. This provided an opportunity to examine results in the context of met distributional assumptions.

Type of DIF. Uniform DIF was the only type of DIF that was generated in this study. Non-uniform DIF was not investigated.

Sample Size. The total simulated sample size was 1,000. That is, there were 500 simulees in the reference group and 500 simulees in the focal group.

Factors Varied

DIF patterns. Four DIF patterns referred to by Zwick et al. (1993) as constant, shift-low, shift-high, and balanced DIF were manipulated. Under the constant pattern, all of the transitions from a given item score category to the next highest category were assumed to be more difficult for the focal group by a constant amount, *s*. The item parameters for the reference and focal groups were determined by:

$$\delta_{miF} = \delta_{miR} + s;\ m = 1, 2, 3. \quad (15)$$

Under the shift-low pattern, the transition from the lowest to the second category was more difficult for the focal group. The remaining transitions were identical for both groups. That is,

$$\delta_{1iF} = \delta_{1iR} + s,\ \delta_{2iF} = \delta_{2iR},\ \delta_{3iF} = \delta_{3iR}. \quad (16)$$

Under the shift-high pattern, the transition from the third to the highest category was more difficult for the focal group. The remaining transitions were identical for both groups. That is,

$$\delta_{1iF} = \delta_{1iR},\ \delta_{2iF} = \delta_{2iR},\ \delta_{3iF} = \delta_{3iR} + s. \quad (17)$$

Under the balanced pattern, the transition from the lowest to the second category was more difficult for the focal group, while the transition from the third to the highest category was easier for the focal group. The remaining transition was the same for both groups. That is,

$$\delta_{1iF} = \delta_{1iR} + s,\ \delta_{2iF} = \delta_{2iR},\ \delta_{3iF} = \delta_{3iR} - s. \quad (18)$$

DIF magnitude. Three non-zero magnitudes (0.10, 0.25, and 0.40) of DIF were investigated in this study. These values of DIF represented the amount of DIF that was simulated to occur within the focal group. These values ranged from small to moderate DIF magnitudes. The magnitude of .25 has been used in several studies (e.g., Chang et al. 1996, Su & Wang, 2005; Zwick et al., 1993), however, it is important to study how smaller and larger magnitudes of DIF may affect DIF detection methods. The generated DIF was added to the category intersection parameters for the items selected to have DIF according to the pattern of DIF that was simulated.

Differences in group ability. Several studies have found that differences in ability distributions, sometimes referred to as *impact*, affect DIF detection rates (e.g., French & Maller 2007; French & Miller, 1996; Narayanan & Swaminathan, 1996; Wang & Su, 2004). To simulate mean latent trait differences between groups, members of the reference group were generated from $N(0, 1)$. There were three levels to the means of the

focal group. Members of the focal group were generated from $N(0,1)$, $N(-0.5,1)$, or $N(-1,1)$. Let

$$\mu_d \equiv \mu_R - \mu_F, \tag{19}$$

where μ_R and μ_F were the mean latent traits of the reference and focal groups, respectively. Consequently, there were three levels of μ_d: 0, 0.5, and 1. Several studies have reported that a difference in mean ability of 1 standard deviation between certain reference and focal groups occurs frequently in real testing situations (Ankenmann et al. 1999; Donoghue et al., 1993; French & Maller, 2007; Su & Wang, 2005). This factor was varied only in the Type I error portion of this study.

Studied item discrimination. Studied item discrimination has been consistently related to the efficacy of DIF detection methods. Research has shown that Type I error rates increase when there is a large difference in ability between groups and the studied item discrimination is high (Chang et al., 1996; Hidalgo & Lopez-Pina, 2004; Zwick et al., 1997)). Also, research has shown that when the studied item discrimination is low, power for uniform DIF is also low but very high when the studied item discrimination is high (Chang et al., 1996). In the present study, three different studied item discriminations were evaluated in the item containing DIF: 0.8,1.2, and 1.6. These parameter values represent a reasonable range of item discrimination values that have been used in previous studies (e.g., French & Miller, 1996; Kristjansson et al., 2005; Rogers & Swaminathan, 1993).

Studied item category intersection parameter magnitude (difficulty). The level of difficulty in the studied item has been shown to influence DIF detection (Clauser et al., 1991; Donoghue & Allen, 1993; French & Miller, 1996; Hambleton et al., 1993; Rogers

and Swaminathan, 1993). In this study a variety of category intersection magnitudes were utilized to reflect the impact of item difficulty on DIF detection. French and Miller used 25 items in their Monte Carlo investigation of the performance of logistic regression for DIF detection when item responses were generated by the GPCM. For these 25 items there were five different sets of values used for b_1, b_2, and b_3. In order to ensure realistic values in the current study, four of these sets of category intersection parameters will be used for the 20th item to make four levels of this study factor. These four levels of b_1, b_2, and b_3 values were equal to -2, 0, and 2; respectively, then -1, 0, and 1; 0, 1, and 2; and finally -2, -1, and 0. These values reflected differing degrees of difficulty across steps of the item and for the item as a whole.

Study Design Overview

This study evaluated the Mantel, GMH, and OLR's power to detect DIF and their associated Type I error rates when the GPCM is the generating IRT parameter model. The power study involved 4 factors that were fully crossed: 4 (DIF patterns) x 3 (DIF magnitudes) x 3 (studied item discrimination) x 4 (studied item difficulty) = 144 fully crossed conditions. There were 1000 replications for each condition.

The Type I error portion of this study involved conditions where no DIF was present. Three factors were fully crossed: 3 (group ability differences) x 4 (studied item difficulty) x 3 (studied item discrimination) = 36 fully crossed conditions. Type I error was only calculated for the 20th item.

Data Generation

Data was generated using the IRTGEN program (Whittaker, Fitzpatrick, Dodd, & Williams, 2003) for SAS, which simulates item responses and trait levels for

dichotomous and polytomous models within the IRT framework. IRTGEN generates item responses by randomly assigning a known theta value from a normal distribution for a simulee. Using this theta value and item parameters for an item, the probability of a simulee responding in each of the four response categories is computed based on the Generalized Partial Credit Model (Muraki, 1992, 1993). These probabilities were then summed, providing cumulative subtotals for each response category. A random number from a uniform distribution was then selected to introduce random error into the simulee's response. If the random number was at or below the cumulative probability for a certain response category, the simulee was awarded that response category score. This procedure was then be repeated for every item and every simulee.

The same generating GPCM item parameters used by French and Miller (1996) were used in this study (see Appendix A) except that five of the items used by French and Miller were removed in order to have the 20 items specified in this study. The difficulty and discrimination parameters for the 19 non-studied items were not manipulated but varied across conditions for the 20^{th} item. DIF was also added solely to the 20^{th} item. Items specified to contain DIF, with the exception of the balanced pattern, had higher category location parameters, according to the specification for the condition under study, for examinees in the focal group indicating that the transition to the category or categories under investigation was more difficult for the focal group; the remaining transitions were identical for both groups. In the balanced pattern, the transition from the lowest to the second category was more difficult for the focal group, while the transition from the third to the highest category was easier for the focal group. The remaining transitions were the same for both groups.

Once DIF was added to the item under investigation, data was generated under the GPCM. Responses for the reference and focal groups were generated separately, and then were combined to create one data set consisting of both reference and focal group responses.

DIF Analyses

SAS 9.1 (SAS Institute, 2001) were used to conduct the DIF analyses. For each simulated data set the Mantel, GMH, and OLR were used for DIF detection. The item being examined for DIF was included in the total score as recommended by Zwick et al. (1993). Prior to DIF detection, matching was performed. A form of thick matching was used because it allowed more cells with non-zero observation frequencies to be used (Donoghue & Allen, 1993). The matching variable was created to have the lowest eight scores pooled together and the highest eight scores pooled together. Once the pooling of the eight lowest and highest scores had been completed, equal intervals of four were used to create the remaining matched groups (Donoghue & Allen, 1993). After matching was done, the GMH, and Mantel statistics (Equations 3, and 7, respectively) were calculated for each of the 1,000 replicated datasets using the PROC FREQ procedure in SAS 9.1. For the OLR procedure, a Chi-squared test for significance of the group (see Equation 12) was performed. Because nonuniform DIF was not examined in this study, the interaction variables were not tested for significance, only the grouping variable. In conditions where DIF was present in the 20^{th} item that item was examined for power. In non-DIF conditions, the 20^{th} item was examined for its Type I error rate across replications. Type I error was the proportion of times out of 1,000 replications where DIF was falsely identified at the 0.05 level of significance. Power was the proportion of times

out of 1,000 replications that DIF was correctly identified at the 0.05 level of significance.

CHAPTER 4

RESULTS

Introduction

The purpose of this Monte Carlo simulation study was to investigate the efficacy of three DIF detection methods, the Mantel, GMH, and the OLR, under various study conditions. A total of 180 unique conditions were simulated. In each condition, 1000 replications were performed, resulting in a total of 180,000 simulated data sets. Tables 4-9 present the power rates for the DIF conditions and Table 10 presents the Type I error rates for the non-DIF conditions. Tables 11-14 present the mean scores for difficult and easy items for the Shift-low and Shift-high conditions. Figures 6-9 depict the effects of item discrimination at each level of item step difficulty. Figures 10 and 11 illustrate the difference in item step difficulties for two conditions within the Shift-low pattern and figures 12 and 13 illustrate the difference in item step difficulties for two conditions within the Shift-high pattern.

Power Main Effects

DIF patterns. The greatest mean power rates for the GMH, Mantel, and OLR procedures (100% for each) occurred under the constant DIF pattern (see Table 4). This finding is consistent with previous research findings by Wang & Su (2004b) who found that the Mantel and the GMH procedures were more powerful under the constant DIF pattern than under any other pattern. Nine conditions out of 36 conditions in the constant DIF pattern displayed these extremely high power rates for all three methods. The Mantel

Table 4

Power rate across 1000 replications for the constant DIF pattern of the 20th Item

Item Discrimination	*Studied Item Difficulty Values*			*DIF Magnitude*	*DIF Detection Methods*		
					GMH	*Mantel*	*OLR*
0.8	-2	0	2	0.1	16.5	25.0	25.3
	-1	0	1		20.8	28.9	30.2
	0	1	2		21.1	29.1	29.4
	-2	-1	0		15.2	23.8	27.3
	-2	0	2	0.25	73.7	85.3*	87.0*
	-1	0	1		86.7*	94.9*	94.3*
	0	1	2		79.9*	92.1*	90.7*
	-2	-1	0		84.2*	92.7*	92.3*
	-2	0	2	0.4	99.4*	100.0*	100.0*
	-1	0	1		99.9*	100.0*	100.0*
	0	1	2		99.6*	99.9*	99.9*
	-2	-1	0		100.0*	100.0*	100.0*
1.2	-2	0	2	0.1	21.4	32.6	33.8
	-1	0	1		29.5	44.9	45.9
	0	1	2		23.8	37.4	39.7
	-2	-1	0		30.1	43.0	43.3
	-2	0	2	0.25	91.3*	97.7*	97.6*
	-1	0	1		98.2*	99.5*	99.6*
	0	1	2		95.2*	98.8*	98.7*
	-2	-1	0		97.1*	99.6*	99.6*
	-2	0	2	0.4	100.0*	100.0*	100.0*
	-1	0	1		100.0*	100.0*	100.0*
	0	1	2		100.0*	100.0*	100.0*
	-2	-1	0		100.0*	100.0*	100.0*
1.6	-2	0	2	0.1	25.5	39.6	41.1
	-1	0	1		39.4	57.1	59.0
	0	1	2		37.0	51.2	53.4
	-2	-1	0		35.6	50.3	50.2
	-2	0	2	0.25	97.5*	99.1*	99.3*
	-1	0	1		99.8*	100.0*	100.0*
	0	1	2		99.1*	99.8*	99.6*
	-2	-1	0		99.3*	100.0*	100.0*
	-2	0	2	0.4	100.0*	100.0*	100.0*
	-1	0	1		100.0*	100.0*	100.0*
	0	1	2		100.0*	100.0*	100.0*
	-2	-1	0		100.0*	100.0*	100.0*

*Power rate ≥ 80%.

and OLR procedures also had four additional mean power rates of 100% for a total of 13 under the constant DIF pattern. The lowest power to detect DIF for the GMH, Mantel, and OLR procedures were 15.2%, 23.8%, and 25.3%, respectively.

Under the constant DIF pattern, twenty-three out of 36 conditions or 64% displayed power rates for all three procedures that matched or exceeded the widely accepted power rate of 80% under the constant DIF pattern. In 20 of those 23 conditions, the Mantel and OLR procedures had power rates that were in the 90 to100 percent range; in the remaining three conditions the GMH procedure had power rates that were in the 81 to 90 percent range. The mean power rate for the GMH procedure was slightly lower than the Mantel and OLR procedures when the DIF magnitude was 0.25 or 0.4. When the DIF magnitude was 0.1, the GMH had a mean power rate of 26.3% compared to 38.6% and 39.9% for the Mantel and OLR, respectively.

In the shift-low pattern (see Table 5), the greatest mean power rate for the GMH, Mantel, and OLR procedures were 99.8%, 100%, and 100%, respectively. This occurred in one condition. There were a total of 10 conditions or 28% of the shift-low pattern conditions in which all three DIF detection methods had mean power rates of at least 80%. Of those 10 conditions, 7 conditions had mean power rates for all three procedures that were in the 90 percents. The lowest powers to detect DIF for the GMH, Mantel, and OLR procedures occurred in one condition and were 5%, 4.7%, and 4.8%, respectively. Under the shift-low pattern, the GMH consistently had a greater mean power to detect DIF at all levels of DIF magnitude than the Mantel and OLR procedures. When the DIF magnitude was 0.1, the GMH had a mean power rate of 9.6% compared to 8.3% and 9.0% for the Mantel and OLR, respectively. When the DIF magnitude was 0.25, the GMH had a mean power rate

Table 5

Power rate across 1000 replications for the shift-low DIF pattern of the 20th Item

Item Discrimination	*Studied Item Difficulty Values*			*DIF Magnitude*	*DIF Detection Methods*		
					GMH	*Mantel*	*OLR*
0.8	-2	0	2	0.1	7.5	7.7	7.7
	-1	0	1		6.9	5.7	6.5
	0	1	2		10.6	10.8	13.1
	-2	-1	0		5.0	4.7	4.8
	-2	0	2	0.25	79.9*	92.1*	90.7*
	-1	0	1		84.2*	92.7*	92.3*
	0	1	2		91.3*	97.7*	97.6*
	-2	-1	0		98.2*	99.5*	99.6*
	-2	0	2	0.4	38.8	19.3	19.4
	-1	0	1		65.9	39.5	43.1
	0	1	2		83.2*	78.5	84.5*
	-2	-1	0		27.0	14.4	12.7
1.2	-2	0	2	0.1	7.4	6.1	5.5
	-1	0	1		9.9	8.0	9.0
	0	1	2		14.5	12.4	13.9
	-2	-1	0		7.8	7.0	7.7
	-2	0	2	0.25	95.2*	98.8*	98.7*
	-1	0	1		97.1*	99.6*	99.6*
	0	1	2		97.5*	99.1*	99.3*
	-2	-1	0		99.8*	100.0*	100.0*
	-2	0	2	0.4	64.9	27.9	28.5
	-1	0	1		90.9*	67.5	69.8
	0	1	2		97.8*	93.2*	95.3*
	-2	-1	0		45.0	17.2	14.7
1.6	-2	0	2	0.1	6.5	5.3	5.5
	-1	0	1		12.9	7.8	8.3
	0	1	2		19.3	19.3	20.4
	-2	-1	0		6.4	4.7	5.2
	-2	0	2	0.25	31.8	13.5	14.1
	-1	0	1		65.5	39.5	43.1
	0	1	2		83.7*	73.2	79.0
	-2	-1	0		22.2	8.4	9.3
	-2	0	2	0.4	72.2	29.8	30.6
	-1	0	1		97.8*	73.2	75.6
	0	1	2		99.4*	98.0*	98.4*
	-2	-1	0		55.2	15.9	15.0

*Power rate ≥ 80%.

of 78.9% compared to 76.2% and 77.0% for the Mantel and OLR and when the DIF magnitude was 0.4, the GMH had a mean power rate of 68.9% compared to 48.4% and 49.3% for the Mantel and OLR procedures.

Under the shift-high pattern, (see Table 6), the greatest mean power rates for the GMH, Mantel, and OLR procedures occurred in one condition and were 99.7%, 98.3%, and 99.1%, respectively. Only two conditions had mean power rates for all three procedures that exceeded 80%; these conditions had mean power rates that were in the 90 percents. The lowest power rates for the three methods were 5.9%, 4.9%, and 5.4%, respectively. Under the shift-high pattern the GMH performed somewhat better than the Mantel and OLR procedures at all levels of DIF magnitude; even though there were only five out of 36 conditions in which the mean power rate for the GMH was at or above the widely accepted rate of 80%. Additionally, under the Shift-high DIF pattern, the third and most difficult level (0, 1, 2) of the studied item category intersection parameters was consistently associated with the lowest power to detect DIF in the GMH, Mantel, and OLR procedures. These low power rates could be due to the fact that embedding DIF in the last category of a difficult item would primarily affect those few examinees at the upper end of the ability continuum. This would result in less contrast between reference and focal groups throughout the ability continuum, and, therefore, less DIF to detect.

In the balanced pattern (see Table 7), the greatest mean power rates for the GMH, Mantel, and OLR procedures were 100%, 87.2%, and 94.4%, respectively. Only two conditions in the balanced pattern had power rates in which all three DIF detection methods had mean power rates that exceeded 80%. The GMH had mean power rates that

were considerably better than the Mantel and OLR procedures for all conditions under the balanced pattern.

Table 6

Power rate across 1000 replications for the shift-high DIF pattern of the 20th Item

Item Discrimination	Studied Item Difficulty Values			DIF Magnitude	DIF Detection Methods		
					GMH	Mantel	OLR
0.8	-2	0	2	0.1	6.5	6.0	5.6
	-1	0	1		8.0	6.6	6.9
	0	1	2		5.9	4.9	5.8
	-2	-1	0		9.1	10.0	11.6
	-2	0	2	0.25	16.3	11.0	10.2
	-1	0	1		27.0	16.5	16.8
	0	1	2		9.0	6.9	7.2
	-2	-1	0		36.8	32.4	38.0
	-2	0	2	0.4	31.6	14.1	14.4
	-1	0	1		60.1	34.0	36.3
	0	1	2		21.6	8.9	8.7
	-2	-1	0		81.3*	73.5	79.5
1.2	-2	0	2	0.1	7.2	5.9	5.7
	-1	0	1		9.3	7.4	8.9
	0	1	2		6.7	5.6	5.7
	-2	-1	0		12.2	11.0	12.2
	-2	0	2	0.25	17.9	8.3	9.1
	-1	0	1		43.2	25.0	26.3
	0	1	2		16.6	7.6	6.7
	-2	-1	0		69.3	60.0	64.4
	-2	0	2	0.4	41.3	12.4	13.1
	-1	0	1		81.5*	43.7	46.8
	0	1	2		32.7	10.6	9.4
	-2	-1	0		98.3*	92.5*	94.7*
1.6	-2	0	2	0.1	7.7	6.0	6.1
	-1	0	1		13.8	9.9	11.3
	0	1	2		7.5	5.1	5.4
	-2	-1	0		19.5	20.5	21.6
	-2	0	2	0.25	23.5	8.0	8.1
	-1	0	1		57.9	26.9	28.9
	0	1	2		16.0	6.6	7.3
	-2	-1	0		83.5*	71.7	76.3
	-2	0	2	0.4	48.9	14.2	14.4
	-1	0	1		94.9*	58.3	61.9
	0	1	2		38.3	11.4	11.3
	-2	-1	0		99.7*	98.3*	99.1*

*Power rate ≥ 80%.

Table 7

Power rate across 1000 replications, for the balanced DIF pattern of the 20th item

Item Discrimination	Studied Item Difficulty Values			DIF Magnitude	DIF Detection Methods		
					GMH	Mantel	OLR
0.8	-2	0	2	0.1	7.7	4.1	4.6
	-1	0	1		11.3	4.4	4.4
	0	1	2		9.8	8.2	9.2
	-2	-1	0		11.2	7.6	10.1
	-2	0	2	0.25	31.5	4.6	4.7
	-1	0	1		49.7	5.2	5.9
	0	1	2		45.7	18.2	24.9
	-2	-1	0		46.1	18.9	24.6
	-2	0	2	0.4	75.7	5.8	6.1
	-1	0	1		92.4*	6.3	6.3
	0	1	2		89.6*	43.2	58.4
	-2	-1	0		88.3*	41.9	53.4
1.2	-2	0	2	0.1	10.2	5.4	5.9
	-1	0	1		17.1	4.9	5.2
	0	1	2		14.7	9.0	11.6
	-2	-1	0		14.1	9.4	11.2
	-2	0	2	0.25	46.3	4.7	4.9
	-1	0	1		82.6*	3.7	4.4
	0	1	2		74.5	34.4	42.5
	-2	-1	0		73.0	36.5	45.7
	-2	0	2	0.4	90.5*	4.9	5.2
	-1	0	1		100.0*	5.6	94.4*
	0	1	2		99.2*	69.1	79.7
	-2	-1	0		99.2*	70.0	80.1*
1.6	-2	0	2	0.1	10.7	4.3	4.9
	-1	0	1		22.0	5.3	4.7
	0	1	2		21.4	13.6	16.0
	-2	-1	0		21.3	13.2	15.6
	-2	0	2	0.25	57.5	4.7	5.4
	-1	0	1		92.4*	4.7	5.4
	0	1	2		89.6*	47.1	55.2
	-2	-1	0		89.6*	48.1	56.2
	-2	0	2	0.4	97.7*	5.2	5.6
	-1	0	1		100.0*	6.1	5.8
	0	1	2		100.0*	82.9*	91.2*
	-2	-1	0		100.0*	87.2*	92.5*

*Power rate ≥ 80%.

The constant DIF pattern had the greatest number of instances (71) where the GMH, Mantel, and OLR's power to detect DIF was at least 80%. The shift-low pattern had 35 instances; the balanced pattern had 21, and the shift-high pattern, 12. Table 8 summarizes the power rates at or above 80% for all three DIF detection methods under the DIF pattern conditions. All three methods displayed the greatest power to detect DIF under the Constant DIF pattern condition. And across all DIF patterns, the GMH outperformed the Mantel and OLR procedures.

DIF Magnitude. In general, across all levels and for all three DIF detection methods, as the DIF magnitude increased so too did the power to detect DIF. This finding is consistent with previous research (e.g. Clauser et al., 1991). The converse was also true; low power to detect DIF was associated with low DIF magnitudes (see Table 9 & Figures 10-13). This trend was more consistent and pronounced under the constant DIF pattern conditions. Under the constant DIF pattern (see Table 4) when the DIF magnitude was 0.25, in two conditions the Mantel and OLR procedures had mean power rates of 100%. Additionally, under the constant DIF pattern, when the DIF magnitude was 0.4 in all but three of the twelve conditions, mean power rates for all three methods were 100%; in the remaining cases, the Mantel and OLR procedures had mean power rates of 100% whilst the GMH had power rates very close to 100% (99.4%, 99.6%, and 99.9%). Further, under the constant DIF pattern, when the DIF magnitude was 0.1, power to detect DIF for the GMH, Mantel, and OLR was quite low ranging from 4.7 to 57.1 across the four DIF patterns. Further, at all DIF magnitude levels, except when the DIF magnitude value was 0.1, the GMH had the greatest mean power to detect DIF (see Table 9).

Table 8

Count across DIF patterns for power at or above 80% for the GMH, Mantel, and OLR procedures across 1,000 replications

DIF Pattern	*DIF Detection Methods* GMH	Mantel	OLR	*Total Pattern Count*
Constant	23	24	24	71
Shift-low	14	10	11	35
Shift-high	6	2	4	12
Balanced	15	2	4	21
Total	58	38	43	

Table 9

Mean power rates across all conditions

Condition	DIF Detection Method GMH	Mantel	OLR
Item Discrimination			
0.8	47.24	38.20	39.72
1.2	57.75	44.54	47.79
1.6	58.74	42.60	44.11
Item Difficulty			
−2, 0, 2	46.06	30.82	31.08
−1, 0, 1	60.24	39.98	43.38
0, 1, 2	54.22	46.77	49.14
−2, −1, 0	57.79	49.55	51.90
Item Magnitude			
0.1	16.28	17.27	18.24
0.25	67.24	53.64	55.16
0.4	81.25	55.84	59.59

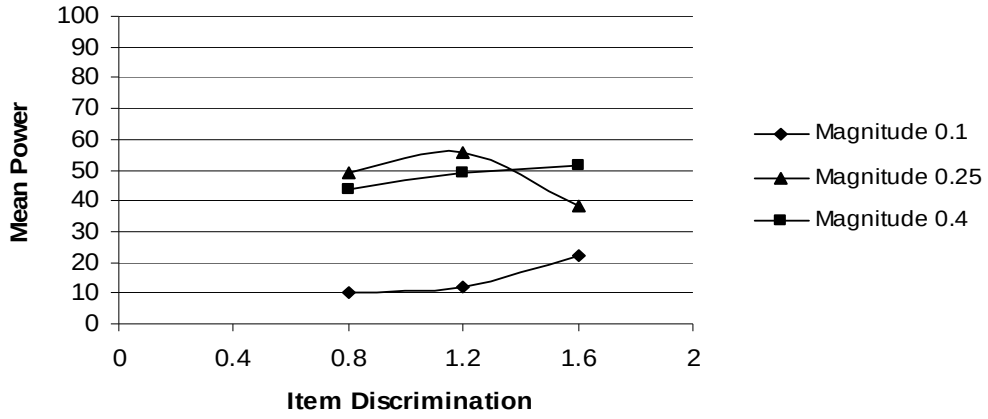

Figure 6. Effects of Item Discrimination at step difficulty (-2, 0, 2).

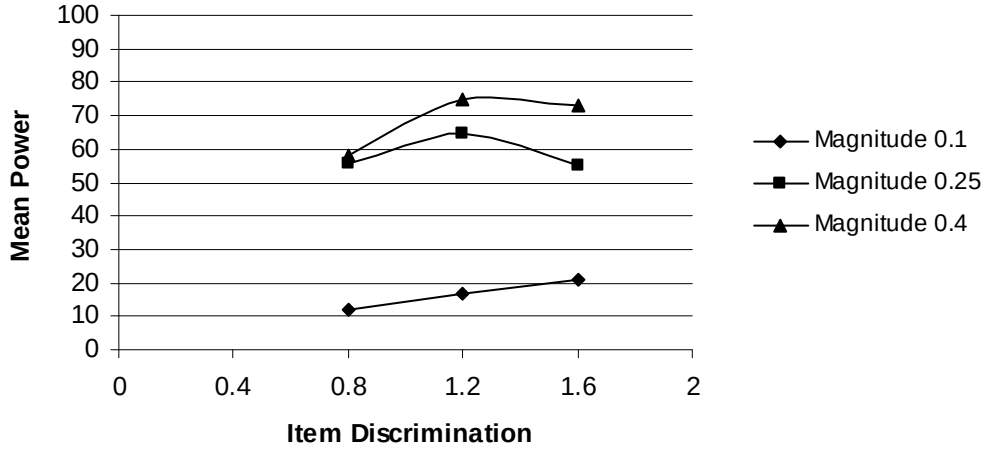

Figure 7. Effects of Item Discrimination at step difficulty (-1, 0, 1).

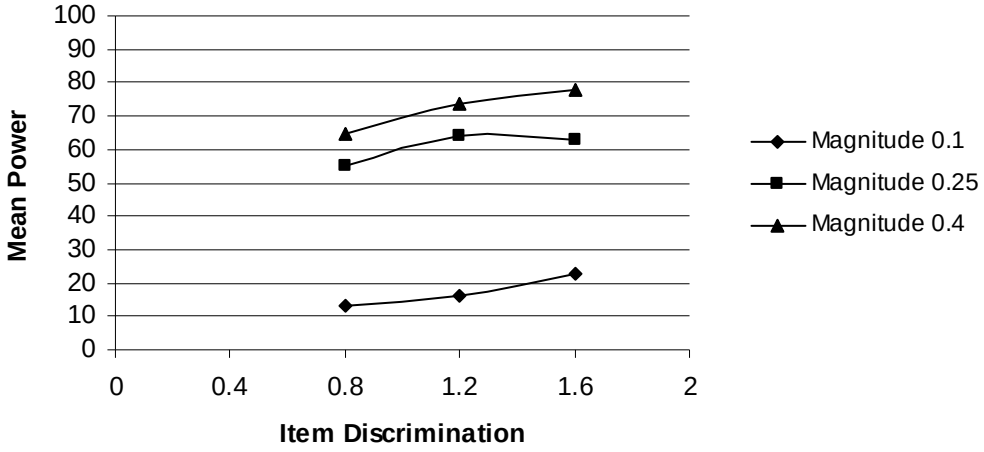

Figure 8. Effects of Item Discrimination at step difficulty (0, 1, 2).

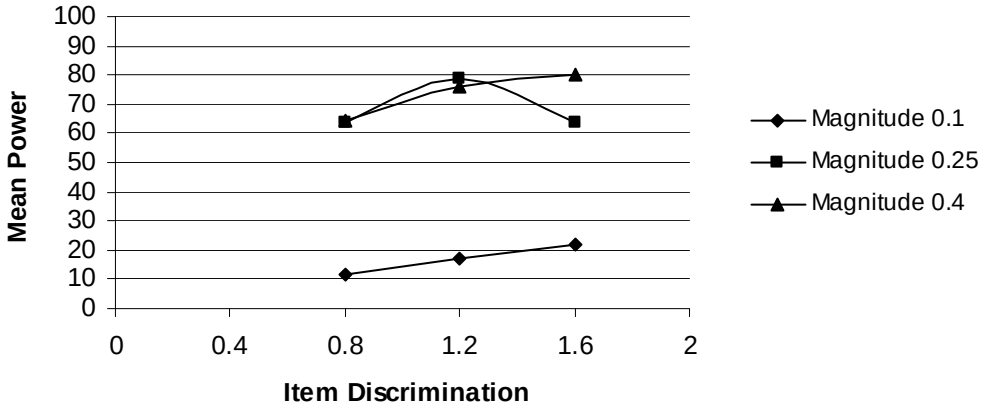

Figure 9. Effects of Item Discrimination at step difficulty (-2,-1, 0).

Studied Item Discrimination. Across all conditions when the item discrimination value was small (0.8) coupled with a DIF magnitude of 0.1, the power rates for all three methods were at their lowest. Even when the studied item discrimination increased, this trend continued as long as the DIF magnitude was 0.1. Under the shift-low pattern, when the studied item discrimination and DIF magnitude values were moderate (i.e., 1.2, and 0.25, respectively) the power rates for the GMH, Mantel, and OLR were at their highest. Table 9 shows the effects of item discrimination on the DIF detection rates. The GMH was the only procedure to consistently show greater power to detect DIF as the item discrimination level increased. Figures 6-9 illustrate the effects of item discrimination at each level of item step difficulty. The figures reveal that at all levels of step difficulty, the mean power to detect DIF increased as item discrimination increased when the DIF magnitude was small (.01). When the DIF magnitude was larger (.25 or .4), although the detection rate increased when the item discrimination increased from .8 to 1.2, there were some cases where the detection rate decreased when item discrimination increased from 1.2 to 1.6.

Studied Item Category Intersection Parameter Magnitude (difficulty). The effect of item difficulty (Hard, 0 1 2 vs. Easy, -2 -1 0) on the detection rate depended on how DIF was embedded. Two conditions (67 and 68) under the shift-low pattern, clearly illustrated the effect that item difficulty had on DIF detection rates (see Figures 10 and 11). In condition 67, the parameter values were as follows: discrimination 1.2, category intersection parameter magnitudes 0, 1, 2, and DIF magnitude 0.4. In condition 68, the parameter values were as follows: discrimination 1.2, category intersection parameter (i.e., difficulty) magnitudes -2,-1,0 and DIF magnitude 0.4. Because of the difficulty

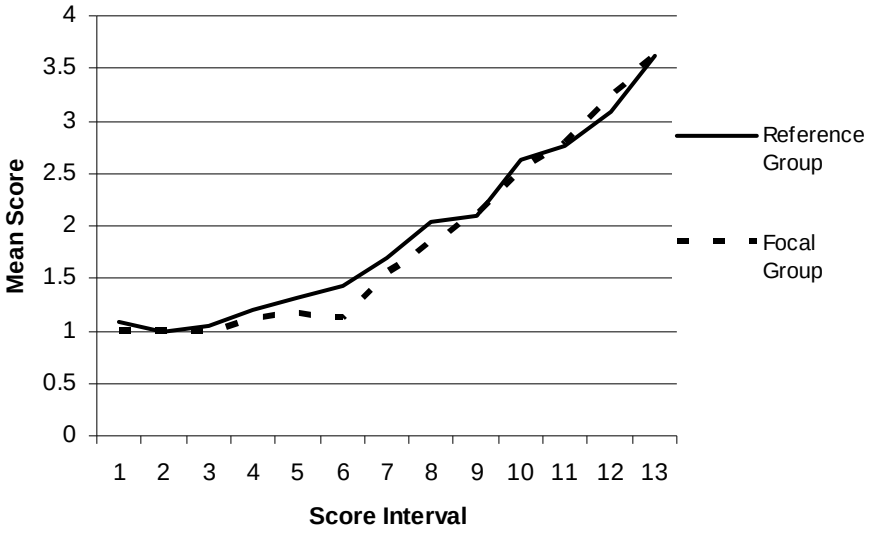

Figure 10. Condition 67. Difficult item (0, 1, 2 item difficulty). Making the item more difficult in the first step (shift-low pattern) affected examinees in all ranges, thus resulting in more DIF detection.

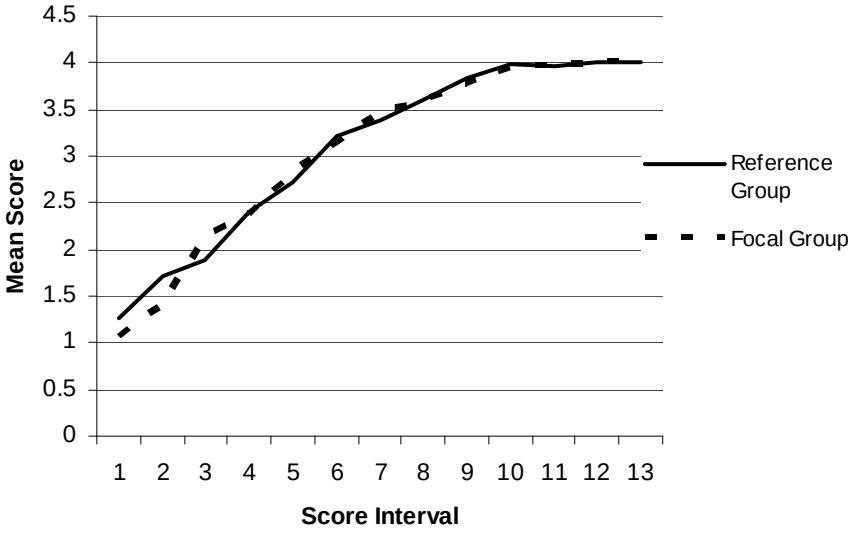

Figure 11. Condition 68. Easy item (-2, -1, 0 item difficulty). Making the first step more difficult (shift-low pattern) primarily affected examinees at the bottom of the range, resulting in less DIF detection.

values associated with condition 67, the item was simulated to be harder than the item in condition 68 which had *b* values of -2, -1, and 0 simulating an item that was easy. The mean scores for the reference and focal groups and the number of reference and focal group examinees in each score interval for these two conditions are presented in Tables 11 and 12. When DIF was embedded in the first category of an already difficult item so that the difficulty parameters became 0.4, 1, and 2 for the focal group versus 0, 1, and 2 for the reference group, the item became more difficult in the first step, affecting examinees in all ranges of the ability continuum, resulting in more DIF detection (see Figure 10). Consequently, the GMH, Mantel, and OLR detection procedures exhibited high power rates (97.8, 93.2, and 95.3, respectively). When DIF was embedded in the first category of a relatively easy item so that the difficulty parameters became -1.6, -1, and 0 for the focal group versus -2, -1, and 0 for the reference group, the item became more difficult in the first step and primarily affected only examinees at the bottom of the range (see Figure 11). This pattern was reversed for the shift-high pattern. Under the shift-high pattern in condition 103, the parameter values were as follows: discrimination 1.2, category intersection parameter magnitudes 0, 1, 2, and DIF magnitude 0.4. In condition 104, the parameter values were as follows: discrimination 1.2, category intersection parameter magnitudes -2,-1,0 and DIF magnitude 0.4. Because of the difficulty values associated with condition 103, the item was simulated to be harder than the item in condition 104 which had *b* values of -2, -1, and 0 simulating an item that was easy. The mean scores for the reference and focal groups and the number of reference and focal group examinees in each score interval for these two conditions are presented in Tables 13 and 14. Figure 12 illustrates that when DIF was embedded in the last category

Table 11

Condition 67: Difficult item (0, 1, 2 item difficulty) for shift-low pattern

Score Interval	N_R	*Mean Score Reference Group*	N_F	*Mean Score Focal Group*
1	11	1.09	17	1
2	12	1.00	25	1
3	35	1.06	33	1
4	54	1.20	47	1.1
5	73	1.31	47	1.17
6	62	1.42	79	1.13
7	74	1.70	62	1.55
8	49	2.04	55	1.86
9	42	2.10	44	2.11
10	29	2.62	36	2.53
11	29	2.76	31	2.81
12	22	3.09	13	3.23
13	8	3.63	11	3.64

Table 12

Condition 68: Easy item (-2, -1, 0 item difficulty) for shift-low pattern

Score Interval	N_R	*Mean Score Reference Group*	N_F	*Mean Score Focal Group*
1	11	1.27	14	1.07
2	14	1.71	22	1.41
3	25	1.88	18	2.11
4	33	2.39	29	2.38
5	53	2.72	57	2.82
6	70	3.21	64	3.14
7	76	3.38	82	3.46
8	60	3.60	66	3.58
9	53	3.83	55	3.78
10	43	3.98	36	3.94
11	32	3.97	28	3.96
12	17	4.00	17	4.00
13	13	4.00	12	4.00

Table 13

Condition 103. Difficult item (0, 1, 2 item difficulty for shift-high pattern

Score Interval	N_R	*Mean Score Reference Group*	N_F	*Mean Score Focal Group*
1	9	1.00	22	1.00
2	22	1.05	22	1.09
3	30	1.00	19	1.05
4	56	1.18	43	1.12
5	58	1.28	58	1.17
6	72	1.47	72	1.53
7	61	1.67	66	1.65
8	51	1.94	56	2.00
9	50	2.38	43	2.30
10	35	2.66	52	2.52
11	26	2.85	24	3.00
12	19	3.68	15	3.07
13	11	3.45	8	3.25

Table 14

Condition 104. Easy item (-2, -1, 0 item difficulty for shift-high pattern

Score Interval	N_R	*Mean Score Reference Group*	N_F	*Mean Score Focal Group*
1	10	1.50	12	1.50
2	23	1.61	20	1.65
3	21	2.14	26	1.92
4	44	2.39	31	2.35
5	39	2.82	57	2.70
6	68	3.07	64	2.98
7	48	3.48	79	3.13
8	74	3.54	51	3.51
9	52	3.67	54	3.65
10	48	3.90	41	3.73
11	39	3.92	38	3.92
12	20	4.00	17	4.00
13	14	4.00	10	4.00

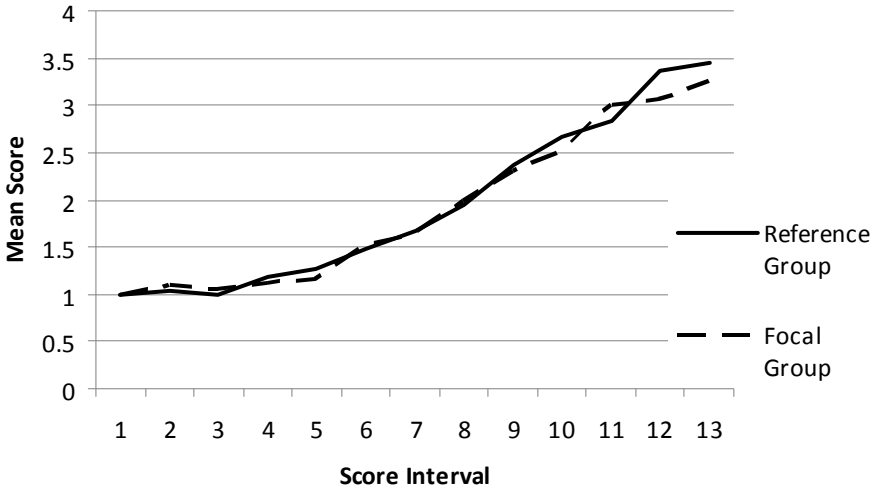

Figure 12. Condition 103. Difficult Item (0, 1, 2 item difficulty). Making the last step even more difficult (Shift-high pattern) affected only very high ability examinees.

of an already difficult item so that the difficulty parameters became 0, 1, and 2.4 for the focal group versus 0, 1, and 2 for the reference group, the item became more difficult in the last step, affecting only examinees in the upper ability range. Because only examinees in the upper ability range were affected, less contrast existed between examinees in the reference and focal groups over the entire ability continuum, except at the upper extreme, thus making it harder to detect DIF. Consequently, the GMH, Mantel, and OLR detection procedures exhibited low power rates (32.7, 10.6, and 9.4, respectively) as there was little DIF to detect.

Figure 13 illustrates that when DIF was embedded in the last category of a relatively easy item so that the difficulty parameters became -2, -1, and 0.4 for the focal group versus -2, -1, and 0 for the reference group, the last step became more difficult and affected examinees in all ability ranges. This made it easier to detect DIF (i.e., more DIF present), resulting in high power rates (98.3, 92.5, and 94.7, respectively) for the GMH,

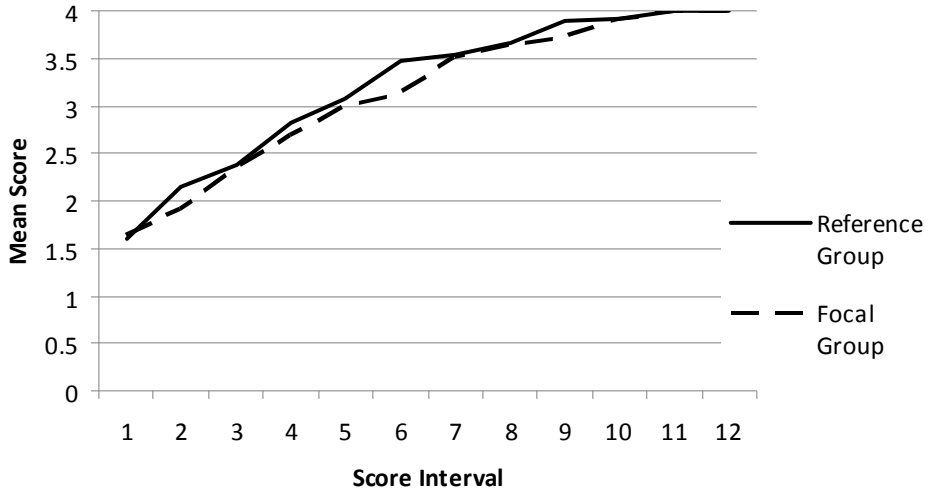

Figure13. Condition 104. Easy item (-2, -1, 0 item difficulty). Making the last step more difficult (Shift-high pattern) affected examinees in all ability ranges.

Mantel, and OLR procedures. This finding is consistent with that of Donoghue and Allen (1993), even though their study examined DIF in dichotomous items and did not include polytomous items. These researchers found that for easy items, increasing the discrimination in the studied item made between group differences larger, thus resulting in better DIF detection for the Mantel-Haenszel method.

The effect of the spread of item difficulty (-2 0 2 vs. -1 0 1) on the detection rates was consistent throughout conditions (see Figures 6 and 7). The first level (-2, 0, 2) of the studied item category intersection parameter was consistently associated with the lowest power to detect DIF in all methods. Embedding DIF on an extreme value (e.g., -2 or 2) would affect only a small number of examinees (i.e, either the most able or least able examinees) resulting in less DIF to detect; thus, yielding lower power rates.

Type I Error. The results for the Type I error rates are displayed in Table 10 for all conditions including the three levels of the group ability difference factor. For the

Table 10

Type I error rates across 1,000 replications for the 20th item

Item Discrimination	Studied Item Difficulty Values			Group Ability Difference	DIF Detection Methods		
					GMH	Mantel	OLR
0.8	-2	0	2	0	4.7	4.6	4.8
	-1	0	1		5.3	5.1	4.9
	0	1	2		5.3	4.9	5.3
	-2	-1	0		4.6	3.9	3.9
	-2	0	2	-0.5	4.8	5.8	6.8
	-1	0	1		5.1	4.1	3.7
	0	1	2		5.5	4.4	4.8
	-2	-1	0		5.4	5.3	5.2
	-2	0	2	-1	5.6	4.6	4.9
	-1	0	1		4.5	4.6	4.7
	0	1	2		4.4	4.9	5.9
	-2	-1	0		4.2	5.7	5.8
1.2	-2	0	2	0	5.6	7.2	7.9*
	-1	0	1		5.5	4.8	5.3
	0	1	2		4.8	4.7	5.0
	-2	-1	0		3.1	3.6	4.2
	-2	0	2	-0.5	4.7	4.9	4.3
	-1	0	1		6.4	7.1	6.2
	0	1	2		5.3	5.5	5.2
	-2	-1	0		6.0	5.7	5.6
	-2	0	2	-1	6.0	7.6*	6.4
	-1	0	1		6.8	8.4*	6.5
	0	1	2		3.9	5.9	5.3
	-2	-1	0		6.4	7.5	6.4
1.6	-2	0	2	0	5.5	5.0	4.8
	-1	0	1		4.4	5.2	5.5
	0	1	2		4.9	5.4	5.0
	-2	-1	0		5.0	4.8	5.3
	-2	0	2	-0.5	7.2	6.3	5.8
	-1	0	1		7.0	7.1	7.7*
	0	1	2		6.4	7.6*	6.5
	-2	-1	0		6.7	7.5	6.4
	-2	0	2	-1	8.6*	10.5*	6.9
	-1	0	1		7.4	11.6*	10.0*
	0	1	2		5.6	8.3*	6.1
	-2	-1	0		10.7*	13.7*	10.4*

*Type I error rate not meeting Bradley's (1978) liberal robustness criterion.

evaluation of Type I error conditions, Bradley's (1978) liberal robustness criterion was used. That is, each DIF detection procedure was interpreted as providing adequate control of Type I error if the estimated Type I error rate was within the range of

$$\alpha_{nominal} \pm .5\ \alpha_{nominal} \quad (20)$$

or .025 to .075, for this study where alpha was 0.05.

The Type I error rates were all close to the nominal rate of .05 in all conditions in which there were no mean latent trait differences between the reference and focal groups and when the item discrimination value was moderate or high (i.e., 1.2 or 1.6). In most conditions where the focal group had a lower mean than the reference group, Type I error rates exceeded the nominal rate of 0.05 but fell within Bradley's (1978) liberal robustness criterion of .025 to .075 range for this study. Also, for conditions in which the item discrimination value was high, and there were mean latent trait differences between groups, the GMH, Mantel, and OLR procedures began to lose control over their average Type I error. This phenomenon became even more pronounced when the groups differed in ability by as much as one standard deviation. When the groups differed by as much as one standard deviation and the item discrimination was high, the type I error rates exceeded Bradley's liberal robustness criterion at all levels of item step difficulty. Nine conditions had Type I error rates that fell outside of Bradley's criterion. The GMH had only one condition that had a type I error rate outside of Bradley's liberal robustness criterion range, the OLR had four, and the Mantel had seven conditions where the type I error rates did not meet Bradley's liberal robustness criterion. The different levels of the category intersection parameters did not appear to have an effect on the Type I error rates except when the item discrimination value was high (1.6) and the difference in group

ability was one standard deviation. In that scenario, the Type I error rates for the three methods departed substantially from the nominal rate of .05, to yield 10.7, 13.7, and 10.4 for the GMH, Mantel, and OLR procedures, respectively.

CHAPTER 5

DISCUSSION

Summary

The purpose of this study was to investigate the power and Type I error rates for the GMH, Mantel, and OLR procedures when there is variation in (a) item discrimination parameter values, (b) category intersection magnitudes, (c) DIF magnitudes, (d) DIF patterns within score categories, and (e) average latent traits between the reference and focal groups.

The results of this Monte Carlo simulation study indicated that the GMH generally outperformed the Mantel and OLR procedures across various DIF patterns in its ability to detect DIF. Table 8 presents a count across all DIF patterns for power at or above 80%. Consistent with previous research (Clauser et al., 1991; Hidalgo & Lopez-Pina, 2004), the main effects on the discrimination and category intersection parameters were found to be partially dependent on DIF magnitude. That is, as DIF magnitude increased so too did the power to detect DIF. This was true for all three methods. As could be expected, in conditions where the item discrimination and DIF magnitude were low, power to detect DIF was at its weakest in all three methods. The results of this investigation also supported previous research findings (e.g., Kristanjsson et al. 2005) that the GMH showed higher power to detect uniform DIF when the item discrimination was moderate or high than when it was low. Additionally, even though Rogers and Swaminathan's (1993) study examined only dichotomous items, their research results

support this research finding in that their investigation revealed that items of both moderate difficulty and high discrimination were more easily detected for DIF by the Mantel-Haenszel and LR procedures.

The study findings of this investigation clearly showed that a small DIF magnitude of 0.1 did not impact DIF detection rates and, perhaps, could be viewed as trivial DIF within the context of these particular study conditions. This investigation also revealed that uniform DIF detection is directly related to item discrimination. In general, the highest mean power rates occurred at the highest level of the studied item discrimination and power to detect DIF generally increased as the item discrimination level increased. This finding might seem problematic for practitioners as test developers desire items that can effectively discriminate between examinees of high ability and low ability on the construct or criterion of interest. Furthermore, highly discriminating items would more than likely contribute to higher instrument reliability. However, it is important to note that highly discriminating items that are flagged for DIF do not necessarily imply that the item is biased; they merely indicate that the items in question are functioning differently for the two groups that have been matched on ability. In reality, a substantial review of the item or items that exhibit DIF would need to follow a statistical DIF analyses to determine whether or not the items were biased.

The results of this investigation regarding Type I error rates were consistent with previous research findings (e.g., Zwick et al., 1993) which indicated that a difference in group means can lead to an increase in Type I error rates, depending on the method of DIF detection used. Additionally, study findings revealed that when the item discrimination value was high and the group means differed by as much as one standard

deviation, inflation of Type I error occurred. This increase in Type I error in conditions where no DIF is present, indicated that impact (true differences in group ability) was flagged as DIF.

All three DIF detection methods used in this investigation were equally easy to implement and use. However, the GMH showed the greatest power to detect DIF in most conditions. This is inconsistent with previous research findings (Su &Wang, 2005; Wang & Su, 2004b) that indicated that the Mantel had higher DIF detection rates than the GMH under the constant, and balanced patterns and performed roughly the same as the Mantel under the shift-low, and shift-high patterns. Zwick et al.'s (1993) also concluded from their research study that for most DIF analyses the Mantel would be a better method than the GMH. These differences in research findings could be attributed to the use of the GPCM as the generating parameter model in this investigation compared to the use of the PCM in the Su and Wang, and Wang and Su investigations. Additionally, the above mentioned researchers' studies did not examine the effect that varying the item discrimination would have on DIF detection rates. This investigation clearly showed that an item's discrimination value can impact DIF detection rates. From a practioner's point of view, it seems clear that test developers need not only be concerned with the amount of DIF in an item but also with the item's discrimination value when developing assessments or selecting which DIF detection method to use. In the present climate of high-stakes testing in the educational arena, due in large part to the enactment of the *No Child Left Behind Act of 2001*, cut-off marks on many standardized assessments determine a child's and school's academic standing, locally, state-wide and nationally. Even a small amount of DIF in an item could be problematic; therefore, it is incumbent

upon test developers to select the best DIF detection procedure to use when developing instruments. This study revealed that the GMH is a suitable DIF detection method to use for tests that are comprised of polytomous data.

Although, this investigation involved only uniform DIF, the GMH is sensitive to both uniform and non-uniform DIF because it tests along the entire response distribution; this is another advantage of the GMH. The GMH procedure is, therefore, recommended as the DIF detection method of choice to detect uniform DIF when the data is polytomous, and item discrimination is suspected to vary across items.

Limitations and Future Research

This investigation examined DIF in only one polytomous item, however, further research is needed to assess DIF in polytomous items when a test contains multiple DIF items. It is more than likely that in a real testing situation, a test would have more than one DIF item; how the GMH, Mantel, and OLR procedures would perform under similar study conditions with more than one studied item would be of interest to test developers. Additionally, this study simulated a single test with 20 items. Wang & Su (2004b) in their investigation simulated three tests - a short test of 10 items, a medium test of 20 items, and a long test of 30 items. Wang & Su examined DIF patterns and the effect that varying the test length would have on the GMH and Mantel when the PCM was the generating model. Their study results indicated that varying the test length had no effect on the GMH's Type I error and power. Wang & Su's study findings, however, indicated that when the Graded Response Model (GRM) was the generating model, the Mantel and GMH yielded slightly better control over Type I error in the 20 item test than in the 10 item test. Future research is needed to determine if test length will have an effect on Type

I error and power rates when the GPCM is the generating parameter model. Also, research is needed to determine how the GMH and OLR procedures perform when nonuniform DIF is simulated (i.e. when DIF is added to the item discrimination parameter) in polytomous items.

References

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

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

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement, 36*(4), 277-300.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous dif detection methods. *Applied Measurement in Education, 15*(2), 113-141.

Bradley, J. V. (1978). Robustness? *The British Journal of Mathematical and Statistical Psychology, 31,* 144-152.

Camilli, G., Shepard, L.A. (1994). *Methods for Identifying Biased Items.* Thousand Oaks, CA: Sage Publications, Inc.

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

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1991). *Examination of various influences on the Mantel-Haenszel statistic* (Laboratory of Psychometric and Evaluative Research Report No. 210). Amherst, MA: University of Massachusetts, School of Education.

Crocker, L., & Algina. (1986). *Introduction to classical & modern test theory.* Belmont, CA: Wadsworth.

De Ayala, R.J., (1993). An introduction to polytomous item response theory models. *Measurement and Evaluation in Counseling and Development, 25,* 172-189.

Donoghue, J. R., & Allen, N. L. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* (3), 233-251.

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel

procedure for detecting DIF. *Journal of Educational Statistics, 18*(2), 131-154.

Dorans, N. J., & Schmitt, A.P. (1993). Constructed response and differential item), functioning: A programmatic perspective. In R.E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135-165). Hillsdale, NJ: Lawrence Erlbaum.

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

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

French, B., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67,* 373-393.

French, A. W., & Miller, T.R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*(3), 315-332.

Furlow, C. F., Fouladi, R. T., Gagne, P., & Whittaker, T. A. (2007). A monte carlo study of the impact of missing data and differential item functioning on theta estimates from two polytomous rasch family models. *Journal of Applied Measurement, 8*(4), 388-403.

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

Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practices,* 3-14.

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

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

Hambleton, R.K., Clauser, E. B., Mazor, K. M., & Jones, R. W. (1993). *Advances in the detection of differentially functioning test items*. ERIC Document Reproduction Service No. ED356 264)

Hidalgo, D. M., Lopez-Pina, J. A., (2004). Differential item functioning detection and effect size: a comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64,* 903-914.

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

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

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58,* 690-700.

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

Mapuranga, R., Dorans, N. J., & Middleton, K. (March, 2008). A review of recent developments in differential item functioning. Paper presented at the National Council on Measurement in Education, New York.

Masters, G.N. (1984). Constructing an item bank using partial credit scoring. *Journal of Educational Measurement, 21* (1), 19-32.

Masters, G.N. (1988b). The analysis of partial credit scoring. *Applied Measurement in Education, 1,* 279-297.

Masters, G.N. (1988a). Measurement models for ordered response categories. In R. Laneheine & J. Rost (Eds.), *Latent traits and latent class models* (pp. 11-29). New York: Plenum.

Meyer, J. P., Huynh, H., & Seaman, M. A. (2004). Exact small-sample differential item functioning methods for polytomous items with illustration based on an attitude survey. *Journal of Educational Measurement, 41*(4), 331-334.

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

Mullis, I., Dossey, J., Owen, E., & Phillips, G. (1993). *NAEP 1992 mathematics report card for the nation and the states.* Washington, DC: National Center for Education Statistics.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17,* 351-363.

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. *Journal of Educational Measurement, 36,* (3), 217-232.

Narayanan, P., & Swaminathan, H., Identification of items that show nonuniform DIF (1996) *Applied Psychological Measurement, 20* (3), 257-274.

No Child Left Behind Act.20 U.S.C. § 6301. (2001).

Ostini, R., & Nering, M.L. (2006). *Polytomous item response theory models*. CA: Sage.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: review and recommendations. *Educational Measurement: Issues and Practice, 19*(3), 5-15.

Potenza, M. T., & Dorans, J. N. (1995). DIF assessment for polytomously scored items: a framework for classification and evaluation. *Applied Psychological Measurement, 19* (1), 23-37.

Raiford-Ross, T. (2007). The impact of multidimensionality on the detection of differential bundle functioning using SIBTEST. Unpublished doctoral dissertation, Educational Policy Studies, Georgia State University.

Rogers, J. H., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17* (2), 105-116.

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

Somes, G. W. (1986). The generalized Mantel-Haenszel statistic. *The American Statistician, 40,* 106-108.

Spray, J., & Miller, T. (1994). *Identifying nonuniform DIF in polytomously scored test items*(*ACT* Rep. No. 94-1).

Stiggins, R.J. (1991). Facing the challenges of a new era of educational assessment. *Applied Measurement in Education, 4*, 263-273.

Su, Y., & Wang, W. (2005). Efficiency of the Mantel, Generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education, 18* (4), 313-350.

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

Thissen, D. (1991). *MULTILOG User's Guide.* Chicago: Scientific Software.

Wang, W., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*, 479-498.

Wang, W., & Su, Y. (2004a). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education, 17*, 113-144.

Wang, W., & Su, Y. (2004b). Factors influencing the mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28*(6), 450-480.

Welch, C., & Hoover, H.D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. *Applied Measurement in Education, 6*(1), 1-19.

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

Williams, V.S.L. (1997). The “unbiased” anchor: bridging the gap between DIF and item bias. *Applied Measurement in Education, 10*(3), 253-267.

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

Zwick, R., Donoghue, J. R., and Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233-251.

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

APPENDIXES

APPENDIX A

Item Parameters for Data Generation

Item Parameters for the First 19 Items

Item	*a*	b_1	b_2	b_3
1.	0.50	-2.00	1.00	2.00
2.	0.50	-1.00	0.00	1.00
3.	0.50	0.00	1.00	2.00
4.	0.50	-2.00	-1.00	0.00
5.	0.75	-2.00	0.00	2.00
6.	0.75	-1.00	0.00	1.00
7.	0.75	0.00	1.00	2.00
8.	0.75	-2.00	-1.00	0.00
9.	1.00	-2.00	0.00	2.00
10.	1.00	-2.00	1.00	2.00
11.	1.00	0.00	1.00	2.00
12.	1.00	-2.00	-1.00	2.00
13.	1.25	-2.00	0.00	2.00
14.	1.25	-2.00	1.00	2.00
15.	1.25	-1.00	0.00	1.00
16.	1.25	-2.00	-1.00	0.00
17.	1.50	-2.00	0.00	2.00
18.	1.50	-2.00	1.00	2.00
19.	1.50	-1.00	0.00	1.00

Note. These item parameters came from French and Miller (1996).

APPENDIX B

SAS Code

```
OPTIONS linesize=72;
%macro dissertation (A=,SD1=,SD2=,SD3=,MAG1=,MAG2=,MAG3=, focabil=,
seed=, seed1=, seed2=, seed3=, cond=);

%do i=1 %to 1000; *number of replications;
proc printto log='e:\Dissertation\Diss Code\Results\disslog.txt' new;

proc printto print='e:\Dissertation\Diss Code\Results\outputdisslog.txt' new;
options mprint;

*reading in item parameters for reference and focal groups;
data params;
infile 'e:\Dissertation\Diss Code\GPCMREF.txt' missover;
input A SD1 SD2 SD3;

data ref20;
A=&A;
SD1=&SD1;          *creating 20th item for reference group;
SD2=&SD2;
SD3=&SD3;

data foc20;
A=&A;
SD1=&SD1 + &MAG1;   *creating 20th item for focal group;
SD2=&SD2 + &MAG2;
SD3=&SD3 + &MAG3;

data d2; set params ref20;
data DFdiff1;set params foc20;

*PROC IML;
%INCLUDE 'e:\Dissertation\Diss Code\IRTGEN.SAS'; *This has the first seed in the
IRTGEN prog;
%IRTGEN(MODEL=GPC, DATA=D2, OUT=D3, NI=20, NE=500);
```

```
%INCLUDE 'e:\Dissertation\Diss Code\IRTGEN1.SAS';
%IRTGEN(MODEL=GPC, DATA=DFdiff1, OUT=D3F, NI=20, NE=500);*This has
seed1 in the IRTGEN1 prog;
QUIT;

DATA D3; SET D3;
FOC=0;
DATA D3F; SET D3F;
FOC=1;
RUN;

data Theta; SET D3 D3F;

**********GMH**********
********;
data Intervals; set Theta;
Totscore=sum(of R1-R20); *making matching variable(total score);

if totscore >=20  AND totscore =<27 THEN equint =1; *creating interval variables for
every simulee;
else if totscore >= 28  AND totscore =<31 THEN equint = 2;
else if totscore >= 32 AND totscore =<35 THEN equint = 3;
else if totscore >=36  AND totscore =<39 THEN equint = 4;
else if totscore >=40  AND totscore =<43 THEN equint =5;
else if totscore >=44  AND totscore =<47 THEN equint = 6;
else if totscore >=48  AND totscore =<51 THEN equint = 7;
else if totscore >=52  AND totscore =<55 THEN equint = 8;
else if totscore >=56  AND totscore =<59 THEN equint =9;
else if totscore >=60  AND totscore =<63 THEN equint = 10;
else if totscore >=64  AND totscore =<67 THEN equint = 11;
else if totscore >=68  AND totscore =<71 THEN equint = 12;
else if totscore >=72  AND totscore =<80 THEN equint = 13;
run;



proc sort DATA=Intervals; by equint; *puts all totalscores into their intervals;

ods output CMH=GMH20_ODS;*saves the sig values to a data set for procedures w/o
output statements;
ods listing; *shows output in output window;
Proc Freq data=Intervals;
Tables equint*FOC*R20/CMH; *******GMH***********;
run;
ods trace off;
```

```
data gmh20; set GMH20_ODS;
rep=&i;
keep R20gmh R20gmhp rep;
R20gmh=value;
R20gmhp=prob;
if statistic=3;
run;

**********************************MANTEL**************************
*********;

ods output CMH=Mantel1_ODS;
ods listing;
Proc Freq data=Intervals;
tables equint*foc*R20/cmh;
run;
ods trace off;

data Mantel1; set Mantel1_ODS;
rep=&i;
keep R20mantel R20mantelprob rep;
R20mantel=value;
R20mantelprob=prob;
if statistic=2; *picks up Mantel info from output;
run;


**************************************************OLR***************
**********************;

PROC LOGISTIC data=Intervals;
MODEL R20 = totscore foc;
ods output parameterestimates=logistic_ods;
run;

data OLR; set logistic_ods;
keep waldchisq probchisq rep;
if variable='FOC';
rep=&i;


DATA ALLPATTERNS; MERGE GMH20 MANTEL1 OLR; BY REP;
length pattern $9;
A=&A;
M1=&MAG1;
M2=&MAG2;
```

```
M3=&MAG3;
cond=&cond;

if &MAG1=&MAG2=&MAG3 THEN pattern = 'constant';
else if &MAG2=0 AND &MAG3=0 THEN pattern='shiftlow';
else if &MAG1=0 AND &MAG2=0 THEN pattern='shifthigh';
else if &MAG2=0 AND &MAG1 NE &MAG3 THEN pattern='balanced';


proc append base=ALLREPS DATA=ALLPATTERNS;
run;

%end;
%mend dissertation;


**********Constant DIF
condition**********;

%dissertation (A=1.36, SD1=-2, SD2=0, SD3=2, MAG1=0.1, MAG2=0.1, MAG3=0.1,
focabil=0, seed=51009, seed1=80077, seed2=48877, seed3=18663, cond=1);
%dissertation (A=1.36, SD1=-1, SD2=0, SD3=1, MAG1=0.1, MAG2=0.1, MAG3=0.1,
focabil=0, seed=28877, seed1=31445, seed2=50041, seed3=61099, cond=2);
%dissertation (A=1.36, SD1=0,  SD2=1, SD3=2, MAG1=0.1, MAG2=0.1, MAG3=0.1,
focabil=0, seed=42167, seed1=77921, seed2=96301, seed3=89579,cond=3);
%dissertation (A=1.36, SD1=-2, SD2=-1,SD3=0, MAG1=0.1, MAG2=0.1, MAG3=0.1,
focabil=0, seed=85475, seed1=63553, seed2=10365, seed3=51085,cond=4);

%dissertation (A=2.04,SD1=-2, SD2=0, SD3=2, MAG1=0.1, MAG2=0.1, MAG3=0.1,
focabil=0, seed=48663, seed1=32639, seed2=81525, seed3=91921,cond=5);
%dissertation (A=2.04,SD1=-1, SD2=0, SD3=1, MAG1=0.1, MAG2=0.1, MAG3=0.1,
focabil=0, seed=69011, seed1=91567, seed2=17955, seed3=46503,cond=6);
%dissertation (A=2.04,SD1=0,  SD2=1, SD3=2, MAG1=0.1, MAG2=0.1,
MAG3=0.1,focabil=0, seed=92157, seed1=14577, seed2=98427, seed3=15011,cond=7);
%dissertation (A=2.04,SD1=-2, SD2=-1,SD3=0, MAG1=0.1, MAG2=0.1,
MAG3=0.1,focabil=0, seed=72905, seed1=39975, seed2=93093, seed3=46573,cond=8);

%dissertation (A=2.72,SD1=-2, SD2=0, SD3=2, MAG1=0.1, MAG2=0.1,
MAG3=0.1,focabil=0, seed=93969, seed1=40961, seed2=36857, seed3=91977,cond=9);
%dissertation (A=2.72,SD1=-1, SD2=0, SD3=1, MAG1=0.1, MAG2=0.1,
MAG3=0.1,focabil=0, seed=32363, seed1=91245, seed2=12765, seed3=61129,cond=10);
%dissertation (A=2.72,SD1=0,  SD2=1, SD3=2, MAG1=0.1, MAG2=0.1,
MAG3=0.1,focabil=0, seed=65795, seed1=20591, seed2=72295, seed3=27001,cond=11);
%dissertation (A=2.72,SD1=-2, SD2=-1,SD3=0, MAG1=0.1, MAG2=0.1,
MAG3=0.1,focabil=0, seed=62765, seed1=56349, seed2=42595, seed3=83473,cond=12);
```

%dissertation (A=1.36,SD1=-2, SD2=0, SD3=2, MAG1=0.25, MAG2=0.25,
MAG3=0.25,focabil=0, seed=97265, seed1=85393, seed2=29515,
seed3=28277,cond=13);
%dissertation (A=1.36,SD1=-1, SD2=0, SD3=1, MAG1=0.25, MAG2=0.25,
MAG3=0.25,focabil=0, seed=85689, seed1=92737, seed2=10281,
seed3=42751,cond=14);
%dissertation (A=1.36,SD1=0, SD2=1, SD3=2, MAG1=0.25, MAG2=0.25,
MAG3=0.25,focabil=0, seed=94305, seed1=96423, seed2=74103,
seed3=51259,cond=15);
%dissertation (A=1.36,SD1=-2, SD2=-1,SD3=0, MAG1=0.25, MAG2=0.25,
MAG3=0.25,focabil=0, seed=78171, seed1=49127, seed2=55293,
seed3=77341,cond=16);

%dissertation (A=2.04,SD1=-2, SD2=0, SD3=2, MAG1=0.25, MAG2=0.25,
MAG3=0.25,focabil=0, seed=67245, seed1=46473, seed2=82765,
seed3=81263,cond=17);
%dissertation (A=2.04,SD1=-1, SD2=0, SD3=1, MAG1=0.25, MAG2=0.25,
MAG3=0.25,focabil=0, seed=17453, seed1=76393, seed2=30995,
seed3=81647,cond=18);
%dissertation (A=2.04,SD1=0, SD2=1, SD3=2, MAG1=0.25, MAG2=0.25,
MAG3=0.25,focabil=0, seed=31273, seed1=77233, seed2=48237,
seed3=70997,cond=19);
%dissertation (A=2.04,SD1=-2, SD2=-1,SD3=0, MAG1=0.25, MAG2=0.25,
MAG3=0.25,focabil=0, seed=46901, seed1=19731, seed2=58731,
seed3=13363,cond=20);

%dissertation (A=2.72,SD1=-2, SD2=0, SD3=2, MAG1=0.25, MAG2=0.25,
MAG3=0.25,focabil=0, seed=30883, seed1=67917, seed2=44407,
seed3=84673,cond=21);
%dissertation (A=2.72,SD1=-1, SD2=0, SD3=1, MAG1=0.25, MAG2=0.25,
MAG3=0.25,focabil=0, seed=64809, seed1=58745, seed2=23219,
seed3=39667,cond=22);
%dissertation (A=2.72,SD1=0, SD2=1, SD3=2, MAG1=0.25, MAG2=0.25,
MAG3=0.25,focabil=0, seed=52267, seed1=69445, seed2=59533,
seed3=18103,cond=23);
%dissertation (A=2.72,SD1=-2, SD2=-1,SD3=0, MAG1=0.25, MAG2=0.25,
MAG3=0.25,focabil=0, seed=82651, seed1=38005, seed2=14513,
seed3=19885,cond=24);

%dissertation (A=1.36,SD1=-2, SD2=0, SD3=2, MAG1=0.4, MAG2=0.4,
MAG3=0.4,focabil=0, seed=86385, seed1=18317, seed2=40027, seed3=20849,cond=25);
%dissertation (A=1.36,SD1=-1, SD2=0, SD3=1, MAG1=0.4 , MAG2=0.4,
MAG3=0.4,focabil=0, seed=15179, seed1=69179, seed2=92477, seed3=59931,cond=26);
%dissertation (A=1.36,SD1=0, SD2=1, SD3=2, MAG1=0.4, MAG2=0.4,
MAG3=0.4,focabil=0, seed=18663, seed1=56865, seed2=38867, seed3=94595,cond=27);

```
%dissertation (A=1.36,SD1=-2, SD2=-1,SD3=0, MAG1=0.4,  MAG2=0.4,
MAG3=0.4,focabil=0, seed=37937, seed1=90229, seed2=83149, seed3=67689,cond=28);

%dissertation (A=2.04,SD1=-2, SD2=0, SD3=2, MAG1=0.4,  MAG2=0.4,
MAG3=0.4,focabil=0, seed=56873, seed1=20655, seed2=26575, seed3=74087,cond=29);
%dissertation (A=2.04,SD1=-1, SD2=0, SD3=1, MAG1=0.4,  MAG2=0.4,
MAG3=0.4,focabil=0, seed=53537, seed1=22987, seed2=87589, seed3=66969,cond=30);
%dissertation (A=2.04,SD1=0,  SD2=1, SD3=2, MAG1=0.4,  MAG2=0.4,
MAG3=0.4,focabil=0, seed=72695, seed1=56869, seed2=70659, seed3=81305,cond=31);
%dissertation (A=2.04,SD1=-2, SD2=-1,SD3=0, MAG1=0.4,  MAG2=0.4,
MAG3=0.4,focabil=0, seed=99547, seed1=22209, seed2=44819, seed3=17617,cond=32);

%dissertation (A=2.72,SD1=-2, SD2=0, SD3=2, MAG1=0.4,  MAG2=0.4,
MAG3=0.4,focabil=0, seed=25417, seed1=58727, seed2=35797, seed3=82271,cond=33);
%dissertation (A=2.72,SD1=-1, SD2=0, SD3=1, MAG1=0.4,  MAG2=0.4,
MAG3=0.4,focabil=0, seed=71341, seed1=93965, seed2=80059, seed3=56307,cond=34);
%dissertation (A=2.72,SD1=0,  SD2=1, SD3=2, MAG1=0.4,  MAG2=0.4,
MAG3=0.4,focabil=0, seed=30015, seed1=25331, seed2=44013, seed3=90655,cond=35);
%dissertation (A=2.72,SD1=-2, SD2=-1,SD3=0, MAG1=0.4,  MAG2=0.4,
MAG3=0.4,focabil=0, seed=51851, seed1=23495, seed2=71585, seed3=97735,cond=36);


**********Shift-Low DIF
condition**********;

%dissertation (A=1.36, SD1=-2, SD2=0, SD3=2, MAG1=0.1, MAG2=0,
MAG3=0,focabil=0, seed=50001, seed1=46557, seed2=58151, seed3=59193,cond=37);
%dissertation (A=1.36, SD1=-1, SD2=0, SD3=1, MAG1=0.1, MAG2=0,
MAG3=0,focabil=0, seed=81817, seed1=98947, seed2=86645, seed3=76797,cond=38);
%dissertation (A=1.36, SD1=0,  SD2=1, SD3=2, MAG1=0.1, MAG2=0,
MAG3=0,focabil=0, seed=44137, seed1=18059, seed2=40801, seed3=84637,cond=39);
%dissertation (A=1.36, SD1=-2, SD2=-1,SD3=0, MAG1=0.1, MAG2=0,
MAG3=0,focabil=0, seed=80287, seed1=69975, seed2=32427, seed3=61607,cond=40);

%dissertation (A=2.04,SD1=-2, SD2=0, SD3=2, MAG1=0.1, MAG2=0,
MAG3=0,focabil=0, seed=32081, seed1=34095, seed2=36207, seed3=39911,cond=41);
%dissertation (A=2.04,SD1=-1, SD2=0, SD3=1, MAG1=0.1, MAG2=0,
MAG3=0,focabil=0, seed=17983, seed1=12565, seed2=60045, seed3=15053,cond=42);
%dissertation (A=2.04,SD1=0,  SD2=1, SD3=2, MAG1=0.1, MAG2=0,
MAG3=0,focabil=0, seed=65255, seed1=85977, seed2=20847, seed3=31595,cond=43);
%dissertation (A=2.04,SD1=-2, SD2=-1,SD3=0, MAG1=0.1, MAG2=0,
MAG3=0,focabil=0, seed=42607, seed1=96067, seed2=12659, seed3=41135,cond=44);

%dissertation (A=2.72,SD1=-2, SD2=0, SD3=2, MAG1=0.1, MAG2=0,
MAG3=0,focabil=0, seed=48413, seed1=15475, seed2=84855, seed3=93161,cond=45);
```

%dissertation (A=2.72,SD1=-1, SD2=0, SD3=1, MAG1=0.1, MAG2=0,
MAG3=0,focabil=0, seed=20969, seed1=96189, seed2=88267, seed3=45585,cond=46);
%dissertation (A=2.72,SD1=0, SD2=1, SD3=2, MAG1=0.1, MAG2=0,
MAG3=0,focabil=0, seed=84115, seed1=16439, seed2=18425, seed3=63213,cond=47);
%dissertation (A=2.72,SD1=-2, SD2=-1,SD3=0, MAG1=0.1, MAG2=0,
MAG3=0,focabil=0, seed=92259, seed1=10367, seed2=30421, seed3=64835,cond=48);

%dissertation (A=1.36,SD1=-2, SD2=0, SD3=2, MAG1=0.25, MAG2=0,
MAG3=0,focabil=0, seed=70663, seed1=25555, seed2=33611, seed3=29841,cond=49);
%dissertation (A=1.36,SD1=-1, SD2=0, SD3=1, MAG1=0.25, MAG2=0,
MAG3=0,focabil=0, seed=19655, seed1=41151, seed2=47363, seed3=19661,cond=50);
%dissertation (A=1.36,SD1=0, SD2=1, SD3=2, MAG1=0.25, MAG2=0,
MAG3=0,focabil=0, seed=84903, seed1=21069, seed2=81825, seed3=74917,cond=51);
%dissertation (A=1.36,SD1=-2, SD2=-1,SD3=0, MAG1=0.25, MAG2=0,
MAG3=0,focabil=0, seed=74461, seed1=90511, seed2=20285, seed3=44947,cond=52);

%dissertation (A=2.04,SD1=-2, SD2=0, SD3=2, MAG1=0.25, MAG2=0,
MAG3=0,focabil=0, seed=64161, seed1=15227, seed2=19509, seed3=44919,cond=53);
%dissertation (A=2.04,SD1=-1, SD2=0, SD3=1, MAG1=0.25, MAG2=0,
MAG3=0,focabil=0, seed=82517, seed1=65855, seed2=76655, seed3=86679,cond=54);
%dissertation (A=2.04,SD1=0, SD2=1, SD3=2, MAG1=0.25, MAG2=0,
MAG3=0,focabil=0, seed=91291, seed1=88863, seed2=20103, seed3=53389,cond=55);
%dissertation (A=2.04,SD1=-2, SD2=-1,SD3=0, MAG1=0.25, MAG2=0,
MAG3=0,focabil=0, seed=30613, seed1=33703, seed2=18593, seed3=39615,cond=56);

%dissertation (A=2.72,SD1=-2, SD2=0, SD3=2, MAG1=0.25, MAG2=0,
MAG3=0,focabil=0, seed=25625, seed1=75601, seed2=28551, seed3=29975,cond=57);
%dissertation (A=2.72,SD1=-1, SD2=0, SD3=1, MAG1=0.25, MAG2=0,
MAG3=0,focabil=0, seed=12777, seed1=77919, seed2=34693, seed3=96909,cond=58);
%dissertation (A=2.72,SD1=0, SD2=1, SD3=2, MAG1=0.25, MAG2=0,
MAG3=0,focabil=0, seed=35509, seed1=36103, seed2=38917, seed3=85963,cond=59);
%dissertation (A=2.72,SD1=-2, SD2=-1,SD3=0, MAG1=0.25, MAG2=0,
MAG3=0,focabil=0, seed=58629, seed1=99505, seed2=60697, seed3=77775,cond=60);

%dissertation (A=1.36,SD1=-2, SD2=0, SD3=2, MAG1=0.4, MAG2=0,
MAG3=0,focabil=0, seed=56941, seed1=32307, seed2=54613, seed3=16379,cond=61);
%dissertation (A=1.36,SD1=-1, SD2=0, SD3=1, MAG1=0.4 , MAG2=0,
MAG3=0,focabil=0, seed=64951, seed1=55157, seed2=40719, seed3=90707,cond=62);
%dissertation (A=1.36,SD1=0, SD2=1, SD3=2, MAG1=0.4, MAG2=0,
MAG3=0,focabil=0, seed=98253, seed1=95725, seed2=94953, seed3=22851,cond=63);
%dissertation (A=1.36,SD1=-2, SD2=-1,SD3=0, MAG1=0.4, MAG2=0,
MAG3=0,focabil=0, seed=42791, seed1=73211, seed2=48501, seed3=90449,cond=64);

%dissertation (A=2.04,SD1=-2, SD2=0, SD3=2, MAG1=0.4, MAG2=0,
MAG3=0,focabil=0, seed=59583, seed1=97809, seed2=45709, seed3=87338,cond=65);

```
%dissertation (A=2.04,SD1=-1, SD2=0, SD3=1, MAG1=0.4,  MAG2=0,
MAG3=0,focabil=0, seed=57491, seed1=73115, seed2=18629, seed3=90725,cond=66);
%dissertation (A=2.04,SD1=0,  SD2=1, SD3=2, MAG1=0.4,  MAG2=0,
MAG3=0,focabil=0, seed=38935, seed1=96773, seed2=16631, seed3=30405,cond=67);
%dissertation (A=2.04,SD1=-2, SD2=-1,SD3=0, MAG1=0.4,  MAG2=0,
MAG3=0,focabil=0, seed=21581, seed1=21457, seed2=16153, seed3=78919,cond=68);

%dissertation (A=2.72,SD1=-2, SD2=0, SD3=2, MAG1=0.4,  MAG2=0,
MAG3=0,focabil=0, seed=37169, seed1=50001, seed2=91227, seed3=44657,cond=69);
%dissertation (A=2.72,SD1=-1, SD2=0, SD3=1, MAG1=0.4,  MAG2=0,
MAG3=0,focabil=0, seed=43937, seed1=21885, seed2=46515, seed3=37449,cond=70);
%dissertation (A=2.72,SD1=0,  SD2=1, SD3=2, MAG1=0.4,  MAG2=0,
MAG3=0,focabil=0, seed=81899, seed1=10493, seed2=68379, seed3=18039,cond=71);
%dissertation (A=2.72,SD1=-2, SD2=-1,SD3=0, MAG1=0.4,  MAG2=0,
MAG3=0,focabil=0, seed=33309, seed1=16705, seed2=35101, seed3=81953,cond=72);

**********Shift-High DIF
condition**********;

%dissertation (A=1.36, SD1=-2, SD2=0, SD3=2, MAG1=0, MAG2=0,
MAG3=0.1,focabil=0, seed=21199, seed1=84979, seed2=66999, seed3=78095,cond=73);
%dissertation (A=1.36, SD1=-1, SD2=0, SD3=1, MAG1=0, MAG2=0,
MAG3=0.1,focabil=0, seed=70331, seed1=70225, seed2=94851, seed3=96131,cond=74);
%dissertation (A=1.36, SD1=0,  SD2=1, SD3=2, MAG1=0, MAG2=0,
MAG3=0.1,focabil=0, seed=63175, seed1=46891, seed2=64995, seed3=81223,cond=75);
%dissertation (A=1.36, SD1=-2, SD2=-1,SD3=0, MAG1=0, MAG2=0,
MAG3=0.1,focabil=0, seed=23167, seed1=33339, seed2=14367, seed3=68335,cond=76);

%dissertation (A=2.04,SD1=-2, SD2=0, SD3=2, MAG1=0, MAG2=0,
MAG3=0.1,focabil=0, seed=57047, seed1=17403, seed2=14349, seed3=42559,cond=77);
%dissertation (A=2.04,SD1=-1, SD2=0, SD3=1, MAG1=0, MAG2=0,
MAG3=0.1,focabil=0, seed=46949, seed1=83197, seed2=87025, seed3=20795,cond=78);
%dissertation (A=2.04,SD1=0,  SD2=1, SD3=2, MAG1=0, MAG2=0,
MAG3=0.1,focabil=0, seed=51111, seed1=39117, seed2=66321, seed3=31935,cond=79);
%dissertation (A=2.04,SD1=-2, SD2=-1,SD3=0, MAG1=0, MAG2=0,
MAG3=0.1,focabil=0, seed=47539, seed1=89303, seed2=92431, seed3=46583,cond=80);

%dissertation (A=2.72,SD1=-2, SD2=0, SD3=2, MAG1=0, MAG2=0,
MAG3=0.1,focabil=0, seed=38391, seed1=70765, seed2=60627, seed3=61337,cond=81);
%dissertation (A=2.72,SD1=-1, SD2=0, SD3=1, MAG1=0, MAG2=0,
MAG3=0.1,focabil=0, seed=98275, seed1=49323, seed2=87637, seed3=53381,cond=82);
%dissertation (A=2.72,SD1=0,  SD2=1, SD3=2, MAG1=0, MAG2=0,
MAG3=0.1,focabil=0, seed=10119, seed1=74211, seed2=27889, seed3=53363,cond=83);
%dissertation (A=2.72,SD1=-2, SD2=-1,SD3=0, MAG1=0, MAG2=0,
MAG3=0.1,focabil=0, seed=81973, seed1=51281, seed2=96783, seed3=14267,cond=84);
```

%dissertation (A=1.36,SD1=-2, SD2=0, SD3=2, MAG1=0, MAG2=0,
MAG3=0.25,focabil=0, seed=38329, seed1=38351, seed2=41575,
seed3=28609,cond=85);
%dissertation (A=1.36,SD1=-1, SD2=0, SD3=1, MAG1=0, MAG2=0,
MAG3=0.25,focabil=0, seed=79401, seed1=65831, seed2=16275,
seed3=58353,cond=86);
%dissertation (A=1.36,SD1=0, SD2=1, SD3=2, MAG1=0, MAG2=0,
MAG3=0.25,focabil=0, seed=44167, seed1=66523, seed2=15059,
seed3=45021,cond=87);
%dissertation (A=1.36,SD1=-2, SD2=-1,SD3=0, MAG1=0, MAG2=0,
MAG3=0.25,focabil=0, seed=81959, seed1=20979, seed2=89917,
seed3=63445,cond=88);

%dissertation (A=2.04,SD1=-2, SD2=0, SD3=2, MAG1=0, MAG2=0,
MAG3=0.25,focabil=0, seed=84067, seed1=37949, seed2=84463,
seed3=17937,cond=89);
%dissertation (A=2.04,SD1=-1, SD2=0, SD3=1, MAG1=0, MAG2=0,
MAG3=0.25,focabil=0, seed=64297, seed1=57015, seed2=10573,
seed3=72163,cond=90);
%dissertation (A=2.04,SD1=0, SD2=1, SD3=2, MAG1=0, MAG2=0,
MAG3=0.25,focabil=0, seed=38857, seed1=12143, seed2=65651,
seed3=86355,cond=91);
%dissertation (A=2.04,SD1=-2, SD2=-1,SD3=0, MAG1=0, MAG2=0,
MAG3=0.25,focabil=0, seed=44133, seed1=45799, seed2=21999,
seed3=24413,cond=92);

%dissertation (A=2.72,SD1=-2, SD2=0, SD3=2, MAG1=0, MAG2=0,
MAG3=0.25,focabil=0, seed=24813, seed1=37621, seed2=15665,
seed3=17361,cond=93);
%dissertation (A=2.72,SD1=-1, SD2=0, SD3=1, MAG1=0, MAG2=0,
MAG3=0.25,focabil=0, seed=16487, seed1=39147, seed2=61023,
seed3=60563,cond=94);
%dissertation (A=2.72,SD1=0, SD2=1, SD3=2, MAG1=0, MAG2=0,
MAG3=0.25,focabil=0, seed=59089, seed1=15765, seed2=53115,
seed3=66499,cond=95);
%dissertation (A=2.72, SD1=-2, SD2=-1,SD3=0, MAG1=0, MAG2=0,
MAG3=0.25,focabil=0, seed=11977, seed1=92237, seed2=92063,
seed3=33941,cond=96);

%*dissertation* (A=1.36,SD1=-2, SD2=0, SD3=2, MAG1=0, MAG2=0,
MAG3=0.4,focabil=0, seed=16815, seed1=76463, seed2=81249, seed3=46609,cond=97);
%*dissertation* (A=1.36,SD1=-1, SD2=0, SD3=1, MAG1=0 , MAG2=0,
MAG3=0.4),focabil=0, seed=64535, seed1=62825, seed2=24369,
seed3=83035,cond=98);
%dissertation (A=1.36,SD1=0, SD2=1, SD3=2, MAG1=0, MAG2=0,
MAG3=0.4,focabil=0, seed=12151, seed1=43997, seed2=47075, seed3=15035,cond=99);

```
%dissertation (A=1.36,SD1=-2, SD2=-1,SD3=0, MAG1=0,  MAG2=0,
MAG3=0.4,focabil=0, seed=97161, seed1=62757, seed2=71945,
seed3=25549,cond=100);

%dissertation (A=2.04,SD1=-2, SD2=0, SD3=2, MAG1=0,  MAG2=0,
MAG3=0.4,focabil=0, seed=26445, seed1=21361, seed2=83991,
seed3=32305,cond=101);
%dissertation (A=2.04,SD1=-1, SD2=0, SD3=1, MAG1=0,  MAG2=0,
MAG3=0.4,focabil=0, seed=92351, seed1=83765, seed2=32989,
seed3=26759,cond=102);
%dissertation (A=2.04,SD1=0,  SD2=1, SD3=2, MAG1=0,  MAG2=0,
MAG3=0.4,focabil=0, seed=73823, seed1=20801, seed2=41035,
seed3=71013,cond=103);
%dissertation (A=2.04,SD1=-2, SD2=-1,SD3=0, MAG1=0,  MAG2=0,
MAG3=0.4,focabil=0, seed=60397, seed1=34537, seed2=31335,
seed3=88815,cond=104);

%dissertation (A=2.72,SD1=-2, SD2=0, SD3=2, MAG1=0,  MAG2=0,
MAG3=0.4,focabil=0, seed=17869, seed1=49071, seed2=73923,
seed3=15263,cond=105);
%dissertation (A=2.72,SD1=-1, SD2=0, SD3=1, MAG1=0,  MAG2=0,
MAG3=0.4,focabil=0, seed=18611, seed1=29789, seed2=62570,
seed3=42865,cond=106);
%dissertation (A=2.72,SD1=0,  SD2=1, SD3=2, MAG1=0,  MAG2=0,
MAG3=0.4,focabil=0, seed=55657, seed1=26113, seed2=25651,
seed3=86367,cond=107);
%dissertation (A=2.72,SD1=-2, SD2=-1,SD3=0, MAG1=0,  MAG2=0,
MAG3=0.4,focabil=0, seed=36693, seed1=27195, seed2=56891,
seed3=97473,cond=108);

**********Balanced DIF
condition**********;

%dissertation (A=1.36, SD1=-2, SD2=0, SD3=2, MAG1=0.1, MAG2=0, MAG3=-
0.1,focabil=0, seed=16489, seed1=16553, seed2=35083, seed3=18735,cond=109);
%dissertation (A=1.36, SD1=-1, SD2=0, SD3=1, MAG1=0.1, MAG2=0, MAG3=-
0.1,focabil=0, seed=85205, seed1=26123, seed2=45349, seed3=29891,cond=110);
%dissertation (A=1.36, SD1=0,  SD2=1, SD3=2, MAG1=0.1, MAG2=0, MAG3=-
0.1,focabil=0, seed=14361, seed1=99447, seed2=83325, seed3=71899,cond=111);
%dissertation (A=1.36, SD1=-2, SD2=-1,SD3=0, MAG1=0.1, MAG2=0, MAG3=-
0.1,focabil=0, seed=74353, seed1=45393, seed2=48223, seed3=17247,cond=112);

%dissertation (A=2.04,SD1=-2, SD2=0, SD3=2, MAG1=0.1, MAG2=0, MAG3=-
0.1,focabil=0, seed=51125, seed1=39339, seed2=31601, seed3=19687,cond=113);
%dissertation (A=2.04,SD1=-1, SD2=0, SD3=1, MAG1=0.1, MAG2=0, MAG3=-
0.1,focabil=0, seed=18749, seed1=68607, seed2=25471, seed3=67107,cond=114);
```

%dissertation (A=2.04,SD1=0, SD2=1, SD3=2, MAG1=0.1, MAG2=0, MAG3=-0.1,focabil=0, seed=11163, seed1=78675, seed2=17095, seed3=45233,cond=115);
%dissertation (A=2.04,SD1=-2, SD2=-1,SD3=0, MAG1=0.1, MAG2=0, MAG3=-0.1,focabil=0, seed=20203, seed1=15475, seed2=41001, seed3=83531,cond=116);

%dissertation (A=2.72,SD1=-2, SD2=0, SD3=2, MAG1=0.1, MAG2=0, MAG3=-0.1,focabil=0, seed=28865, seed1=48373, seed2=35931, seed3=68645,cond=117);
%dissertation (A=2.72,SD1=-1, SD2=0, SD3=1, MAG1=0.1, MAG2=0, MAG3=-0.1,focabil=0, seed=73817, seed1=23153, seed2=59649, seed3=46751,cond=118);
%dissertation (A=2.72,SD1=0, SD2=1, SD3=2, MAG1=0.1, MAG2=0, MAG3=-0.1,focabil=0, seed=68995, seed1=47689, seed2=79375, seed3=68833,cond=119);
%dissertation (A=2.72,SD1=-2, SD2=-1,SD3=0, MAG1=0.1, MAG2=0, MAG3=-0.1,focabil=0, seed=41867, seed1=89203, seed2=93911, seed3=88525,cond=120);

%dissertation (A=1.36,SD1=-2, SD2=0, SD3=2, MAG1=0.25, MAG2=0, MAG3=-0.25,focabil=0, seed=66345, seed1=81651, seed2=84081, seed3=34405,cond=121);
%dissertation (A=1.36,SD1=-1, SD2=0, SD3=1, MAG1=0.25, MAG2=0, MAG3=-0.25,focabil=0, seed=17639, seed1=34327, seed2=80377, seed3=54339,cond=122);
%dissertation (A=1.36,SD1=0, SD2=1, SD3=2, MAG1=0.25, MAG2=0, MAG3=-0.25,focabil=0, seed=12515, seed1=22923, seed2=14777, seed3=57375,cond=123);
%dissertation (A=1.36,SD1=-2, SD2=-1,SD3=0, MAG1=0.25, MAG2=0, MAG3=-0.25,focabil=0, seed=15957, seed1=70625, seed2=32523, seed3=30429,cond=124);

%dissertation (A=2.04,SD1=-2, SD2=0, SD3=2, MAG1=0.25, MAG2=0, MAG3=-0.25,focabil=0, seed=56087, seed1=14951, seed2=71795, seed3=43805,cond=125);
%dissertation (A=2.04,SD1=-1, SD2=0, SD3=1, MAG1=0.25, MAG2=0, MAG3=-0.25,focabil=0, seed=50245, seed1=74301, seed2=25299, seed3=94617,cond=126);
%dissertation (A=2.04,SD1=0, SD2=1, SD3=2, MAG1=0.25, MAG2=0, MAG3=-0.25,focabil=0, seed=52689, seed1=35909, seed2=58861, seed3=81073,cond=127);
%dissertation (A=2.04,SD1=-2, SD2=-1,SD3=0, MAG1=0.25, MAG2=0, MAG3=-0.25,focabil=0, seed=56613, seed1=98227, seed2=12133, seed3=52799,cond=128);

%dissertation (A=2.72,SD1=-2, SD2=0, SD3=2, MAG1=0.25, MAG2=0, MAG3=-0.25,focabil=0, seed=32261, seed1=78547, seed2=82163, seed3=72811,cond=129);
%dissertation (A=2.72,SD1=-1, SD2=0, SD3=1, MAG1=0.25, MAG2=0, MAG3=-0.25,focabil=0, seed=41961, seed1=70735, seed2=98931, seed3=35165,cond=130);
%dissertation (A=2.72,SD1=0, SD2=1, SD3=2, MAG1=0.25, MAG2=0, MAG3=-0.25,focabil=0, seed=28725, seed1=51805, seed2=85001, seed3=60383,cond=131);
%dissertation (A=2.72,SD1=-2, SD2=-1,SD3=0, MAG1=0.25, MAG2=0, MAG3=-0.25,focabil=0, seed=51275, seed1=34971, seed2=84387, seed3=99533,cond=132);

%dissertation (A=1.36,SD1=-2, SD2=0, SD3=2, MAG1=0.4, MAG2=0, MAG3=-0.4,focabil=0, seed=91511, seed1=78095, seed2=14645, seed3=28225,cond=133);
%dissertation (A=1.36,SD1=-1, SD2=0, SD3=1, MAG1=0.4 , MAG2=0, MAG3=-0.4,focabil=0, seed=92277, seed1=60859, seed2=13261, seed3=22717,cond=134);

%dissertation (A=1.36,SD1=0, SD2=1, SD3=2, MAG1=0.4, MAG2=0, MAG3=-0.4,focabil=0, seed=44437, seed1=25499, seed2=98289, seed3=85653,cond=135);
%dissertation (A=1.36,SD1=-2, SD2=-1,SD3=0, MAG1=0.4, MAG2=0, MAG3=-0.4,focabil=0, seed=50501, seed1=21597, seed2=34191, seed3=92325,cond=136);

%dissertation (A=2.04,SD1=-2, SD2=0, SD3=2, MAG1=0.4, MAG2=0, MAG3=-0.4,focabil=0, seed=23541, seed1=34925, seed2=70925, seed3=85065,cond=137);
%dissertation (A=2.04,SD1=-1, SD2=0, SD3=1, MAG1=0.4, MAG2=0, MAG3=-0.4,focabil=0, seed=907251, seed1=75567, seed2=50585, seed3=19585,cond=138);
%dissertation (A=2.04,SD1=0, SD2=1, SD3=2, MAG1=0.4, MAG2=0, MAG3=-0.4,focabil=0, seed=164081, seed1=156641, seed2=950121, seed3=643641,cond=139);
%dissertation (A=2.04,SD1=-2, SD2=-1,SD3=0, MAG1=0.4, MAG2=0, MAG3=-0.4,focabil=0, seed=304051, seed1=574911, seed2=731151, seed3=186291,cond=140);

%dissertation (A=2.72,SD1=-2, SD2=0, SD3=2, MAG1=0.4, MAG2=0, MAG3=-0.4,focabil=0, seed=316241, seed1=389351, seed2=967731, seed3=166311,cond=141);
%dissertation (A=2.72,SD1=-1, SD2=0, SD3=1, MAG1=0.4, MAG2=0, MAG3=-0.4,focabil=0, seed=190661, seed1=744261, seed2=139311, seed3=789191,cond=142);
%dissertation (A=2.72,SD1=0, SD2=1, SD3=2, MAG1=0.4, MAG2=0, MAG3=-0.4,focabil=0, seed=215811, seed1=214571, seed2=161531, seed3=422381,cond=143);
%dissertation (A=2.72,SD1=-2, SD2=-1,SD3=0, MAG1=0.4, MAG2=0, MAG3=-0.4,focabil=0, seed=913401, seed1=74301, seed2=446571, seed3=556121,cond=144);

**********TYPE I ERROR**********;

%dissertation (A=1.36, SD1=-2, SD2=0, SD3=2, MAG1=0, MAG2=0, MAG3=0,focabil=0, seed=275041, seed1=653901, seed2=500011, seed3=912271, cond=145);
%dissertation (A=1.36, SD1=-1, SD2=0, SD3=1,MAG1=0, MAG2=0, MAG3=0,focabil=0, seed=465151, seed1=374491, seed2=115081, seed3=371691, cond=146);
%dissertation (A=1.36, SD1=0, SD2=1, SD3=2,MAG1=0, MAG2=0, MAG3=0,focabil=0, seed=218851, seed1=824861, seed2=637981, seed3=309861,cond=147);
%dissertation (A=1.36, SD1=-2, SD2=-1,SD3=0, MAG1=0, MAG2=0, MAG3=0,focabil=0, seed=329911, seed1=976561, seed2=439371, seed3=603361,cond=148);

%dissertation (A=2.04,SD1=-2, SD2=0, SD3=2, MAG1=0, MAG2=0, MAG3=0,focabil=0, seed=636211, seed1=180391, seed2=856361, seed3=796261,cond=149);
%dissertation (A=2.04,SD1=-1, SD2=0, SD3=1,MAG1=0, MAG2=0, MAG3=0, focabil=0, seed=683791, seed1=358111, seed2=674121, seed3=522101,cond=150);
%dissertation (A=2.04,SD1=0, SD2=1, SD3=2,MAG1=0, MAG2=0, MAG3=0, focabil=0, seed=351011, seed1=819531, seed2=818991, seed3=104931,cond=151);

%dissertation (A=2.04,SD1=-2, SD2=-1,SD3=0,MAG1=0, MAG2=0, MAG3=0,
focabil=0, seed=202061, seed1=350061, seed2=839461, seed3=167031,cond=152);

%dissertation (A=2.72,SD1=-2, SD2=0, SD3=2,MAG1=0, MAG2=0,
MAG3=0,focabil=0, seed=333091, seed1=194741, seed2=763841,
seed3=642021,cond=153);
%dissertation (A=2.72,SD1=-1, SD2=0, SD3=1,MAG1=0, MAG2=0, MAG3=0,
focabil=0, seed=180021, seed1=124261, seed2=119031, seed3=332781,cond=154);
%dissertation (A=2.72,SD1=0, SD2=1, SD3=2, MAG1=0, MAG2=0,
MAG3=0,focabil=0, seed=669991, seed1=780951, seed2=578021,
seed3=407421,cond=155);
%dissertation (A=2.72,SD1=-2, SD2=-1,SD3=0, MAG1=0, MAG2=0,
MAG3=0,focabil=0, seed=152241, seed1=381401, seed2=21191,
seed3=849791,cond=156);

%dissertation (A=1.36,SD1=-2, SD2=0, SD3=2,MAG1=0, MAG2=0, MAG3=0,
focabil=-.5, seed=303621, seed1=702251, seed2=948511, seed3=961311,cond=157);
%dissertation (A=1.36,SD1=-1, SD2=0, SD3=1,MAG1=0, MAG2=0, MAG3=0,
focabil=-.5, seed=848461, seed1=649951, seed2=812231, seed3=703311,cond=158);
%dissertation (A=1.36,SD1=0, SD2=1, SD3=2,MAG1=0, MAG2=0, MAG3=0,
focabil=-.5, seed=631751, seed1=468911, seed2=987821, seed3=329061,cond=159);
%dissertation (A=1.36,SD1=-2, SD2=-1,SD3=0, MAG1=0, MAG2=0,
MAG3=0,focabil=-.5, seed=143671, seed1=683351, seed2=164861,
seed3=112211,cond=160);

%dissertation (A=2.04,SD1=-2, SD2=0, SD3=2, MAG1=0, MAG2=0,
MAG3=0,focabil=-.5, seed=316621, seed1=333391, seed2=839741,
seed3=156561,cond=161);
%dissertation (A=2.04,SD1=-1, SD2=0, SD3=1, MAG1=0, MAG2=0,
MAG3=0,focabil=-.5, seed=155201, seed1=141531, seed2=204921,
seed3=935261,cond=162);
%dissertation (A=2.04,SD1=0, SD2=1, SD3=2, MAG1=0, MAG2=0,
MAG3=0,focabil=-.5, seed=859001, seed1=237921, seed2=231671,
seed3=474981,cond=163);
%dissertation (A=2.04,SD1=-2, SD2=-1,SD3=0, MAG1=0, MAG2=0,
MAG3=0,focabil=-.5, seed=236321, seed1=174031, seed2=143491,
seed3=425591,cond=164);

%dissertation (A=2.72,SD1=-2, SD2=0, SD3=2, MAG1=0, MAG2=0,
MAG3=0,focabil=-.5, seed=870251, seed1=207951, seed2=439721,
seed3=570471,cond=165);
%dissertation (A=2.72,SD1=-1, SD2=0, SD3=1, MAG1=0, MAG2=0,
MAG3=0,focabil=-.5, seed=831971, seed1=120501, seed2=298201,
seed3=265041,cond=166);

```
%dissertation (A=2.72,SD1=0,  SD2=1, SD3=2, MAG1=0, MAG2=0,
MAG3=0,focabil=-.5, seed=663211, seed1=319351, seed2=469491,
seed3=993241,cond=167);
%dissertation (A=2.72,SD1=-2, SD2=-1,SD3=0, MAG1=0, MAG2=0,
MAG3=0,focabil=-.5, seed=166941, seed1=511111, seed2=839441,
seed3=729581,cond=168);

%dissertation (A=1.36,SD1=-2, SD2=0, SD3=2, MAG1=0, MAG2=0,
MAG3=0,focabil=-1, seed=992541, seed1=465831, seed2=424161,
seed3=859221,cond=169);
%dissertation (A=1.36,SD1=-1, SD2=0, SD3=1, MAG1=0, MAG2=0,
MAG3=0,focabil=-1, seed=893031, seed1=240101, seed2=174081,
seed3=924311,cond=170);
%dissertation (A=1.36,SD1=0,  SD2=1, SD3=2, MAG1=0, MAG2=0,
MAG3=0,focabil=-1, seed=475391, seed1=135741, seed2=135741,
seed3=154181,cond=171);
%dissertation (A=1.36,SD1=-2, SD2=-1,SD3=0, MAG1=0, MAG2=0,
MAG3=0,focabil=-1, seed=319261, seed1=299921, seed2=606271,
seed3=613371,cond=172);

%dissertation (A=2.04,SD1=-2, SD2=0, SD3=2, MAG1=0, MAG2=0,
MAG3=0,focabil=-1, seed=533811, seed1=383911, seed2=707651,
seed3=253881,cond=173);
%dissertation (A=2.04,SD1=-1, SD2=0, SD3=1, MAG1=0, MAG2=0,
MAG3=0,focabil=-1, seed=144221, seed1=493231, seed2=876371,
seed3=919621,cond=174);
%dissertation (A=2.04,SD1=0,  SD2=1, SD3=2, MAG1=0, MAG2=0,
MAG3=0,focabil=-1, seed=533631, seed1=826741, seed2=789801,
seed3=982751,cond=175);
%dissertation (A=2.04,SD1=-2, SD2=-1,SD3=0, MAG1=0, MAG2=0,
MAG3=0,focabil=-1, seed=954521, seed1=101191, seed2=742111,
seed3=278891,cond=176);

%dissertation (A=2.72,SD1=-2, SD2=0, SD3=2, focabil=-1, MAG1=0, MAG2=0,
MAG3=0,seed=897281 , seed1=967831, seed2=417441, seed3=142671,cond=177);
%dissertation (A=2.72,SD1=-1, SD2=0, SD3=1, focabil=-1, MAG1=0, MAG2=0,
MAG3=0,seed=270221, seed1=819731, seed2=512841, seed3=337321,cond=178);
%dissertation (A=2.72,SD1=0,  SD2=1, SD3=2, focabil=-1, MAG1=0, MAG2=0,
MAG3=0,seed=896321, seed1=415751, seed2=286091, seed3=199241,cond=179);
%dissertation (A=2.72,SD1=-2, SD2=-1,SD3=0, focabil=-1, MAG1=0, MAG2=0,
MAG3=0,seed=583531, seed1=383291, seed2=5469001, seed3=383511,cond=180);

proc sort data=ALLREPS; by rep;
data typeIgmh; set allreps; by rep;

if R20gmhp ne . and R20gmhp lt .05 then sig20gmh=1;
```

```
else if R20gmhp ge .05 then sig20gmh=0;

if R20mantelprob ne . and R20mantelprob lt .05 then sig20man=1;
else if R20mantelprob ge .05 then sig20man=0;

if probchisq ne . and probchisq lt .05 then sig20LR=1;
else if probchisq ge .05 then sig20LR=0;

ods html body= 'e:\Dissertation\Diss Code\Results\outputdisslog.xls';
Title 'DIF DETECTION';
proc sort data=typeIgmh; by cond;

proc freq; tables sig20gmh sig20man sig20LR; by cond;
run;

ods html close;
```

Note. The scaling constant (D=1.7) must be multiplied to the discrimination parameters before being passed to the IRTGEN macro program (Whittaker et al, 2003).

APPENDIX C

Relevant Literature Overview

Study		Generating Model(s)	DIF detection methods	Factors Examined	Findings
Dichotomous	Clauser et al. (1991)	3PLM	MH	*Discrimination Difficulty Ability distribution	Items more likely to be flagged for DIF as the DIF magnitude increased MH most effective with groups from equal ability distributions
	Donoghue & Allen (1993)	3 PLM	MH	DIF magnitude Ability distribution Pooling (thin vs thick)	Very easy items more difficult for focal group Hard items easier for reference group For easy items, increasing the discrimination in the studied item made between group differences larger, resulting in better DIF detection
	Rogers & Swaminathan (1993)	2 PLM 3 PLM	MH LR	Sample size (250 & 500) Degree of model-fit	For very easy items, LR fit data well over all but the very lowest part of trait scale For very difficult items, LR misfit was more pronounced
				Discrimination Difficulty	Items of moderate difficulty and high discrimination, more easily detected for DIF Items of high

Study		Generating Model(s)	DIF detection methods	Factors Examined	Findings
					difficulty, MH & LR had almost identical detection rates
	Hidalgo & Lopez-Pina (2004)	3 PLM	MH Modified MH LR	Discrimination Difficulty	DIF detection rates for all methods increased, as magnitude of uniform & nonuniform DIF increased LR had better DIF detection rates for symmetrical nonuniform conditions Overall, all 3 methods performed similarly
Polytomous	Zwick et al. (1993)	PCM	Mantel GMH	Ability distribution Difficulty DIF patterns DIF magnitude	For most DIF analyses, Mantel is better method to use
	French & Miller (1996)	GPCM	LR	Sample Size (500 & 2,000) Coding Schemes 4 Conditions	Continuation ratio logits model & cumulative logits model better for detecting uniform and nonuniform DIF than adjacent categories model Overall, with the large sample size, LR is a good choice for polytomous DIF detection
	Wang & Su (2004b)	PCM Graded Response Model (GRM)	Mantel GMH	Test purification Ability distribution Test length % DIF DIF patterns ASA	Mantel & GMH more powerful under the constant and constant-item/balanced test pattern
	Su & Wang	PCM	Mantel	Ability	Under constant

Study		Generating Model(s)	DIF detection methods	Factors Examined	Findings
	(2005)	GRM	GMH LDFA	distribution DIF pattern DIF magnitude %DIF	pattern all three methods begin to lose control over Type I error Under the balanced pattern, all three methods had good control over Type I error Under the shift-high and shift-low patterns, the average power of the three methods to detect DIF was roughly the same. Under the constant-item/balanced test pattern, the average power of the Mantel and LDFA methods was similar but higher than that of the GMH The higher the percentage of DIF items, the more inflated the average Type I error became
	Kristjansson et al. (2005)	GPCM	Mantel GMH LDFA UCLOLR	Presence & type of DIF Discrimination Sample size ratio Skewness of ability distribution	None of the four DIF detection procedures showed any significant departure from the nominal Type I error rate of 0.05 All four procedures had excellent power (greater than 0.963) for detecting uniform DIF GMH and UCLOLR's power to detect uniform DIF was better when

Study		Generating Model(s)	DIF detection methods	Factors Examined	Findings
					item discrimination was moderate or high

*Only uniform DIF simulated