AN INVESTIGATION OF THE COMPARABILITY AND ACCURACY OF THREE DIFFERENTIAL ITEM FUNCTIONING (DIF) DETECTION METHODS USING EMPIRICAL AND SIMULATED DATA

by

Ann Elizabeth Harman

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
1995

Approved by

Dissertation Advisor

UMI Number: 9531840

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Dissertation Advisor Lloyd Bond

Committee Members J Hattie

Terry Cooper

March 27, 1995
Date of Acceptance by Committee

March 27, 1995
Date of Final Oral Examination

HARMAN, ANN ELIZABETH, Ph.D. An Investigation of the Comparability and Accuracy of Three Differential Item Functioning (DIF) Detection Methods Using Empirical and Simulated Data. (1995) Directed by Dr. Lloyd Bond. 125 pp.

The purpose of this study was to investigate the comparability and accuracy of three differential item functioning (DIF) detection methods: the Mantel-Haenszel χ^2 approach, the IRT Unsigned Area approach, and the log-linear approach. Based on a review of the professional literature relevant to methodologies used for the detection of differentially functioning test items, two research questions were developed. The first research question addressed the comparability of the DIF indices derived using the three DIF detection methods cited above. The second research question addressed the accuracy of these three DIF detection methods. To investigate these research questions, both empirical and simulated test data were used.

The investigation of the comparability of the DIF detection methods involved six separate analyses of the DIF indices derived from both the empirical and the simulated data. Specifically, for each data set, the analyses focused on the Pearson product-moment correlation coefficients, phi correlation coefficients, and pairwise comparisons of detection rates between each pair of the DIF indices. The investigation of the accuracy of the DIF detection methods involved three additional analyses of the DIF indices derived from the simulated only. Specifically, these analyses focused on the Pearson Product-Moment correlation coefficients, phi correlation coefficients,

and pairwise comparisons of detection rates, number of Type I errors, and number of Type II errors between each of the DIF indices and the actual unsigned area between the item characteristic curves (ICCs) for the simulated test items.

The results of this dissertation study indicate that the Mantel-Haenszel χ^2 approach and the IRT Unsigned Area approach yield highly correlated DIF indices, but have lower agreement rates when flagging items as displaying DIF. Similarly, the Mantel-Haenszel approach and the log-linear approach also yield highly correlated DIF indices, and have moderate agreement rates in flagging DIF items. The IRT Unsigned Area approach and the log-linear approach yield DIF indices which show a low correlation; they also have lower agreement rates with respect to the flagging of items which display DIF.

In terms of the accuracy of the DIF indices, this study found that the Mantel-Haenszel approach and the log-linear approach were both moderately accurate in identifying items displaying uniform DIF, similarly ineffective in detecting nonuniform DIF items, and resulted in few Type II errors. The IRT Unsigned Area approach, however, was highly accurate in identifying both uniform and nonuniform DIF items, but also committed a large number of Type II errors (19 of 47 non-DIF items were flagged). The large number of Type II errors committed by the IRT Unsigned Area approach appear to have resulted from the approaches' oversensitivity to differences in the b-parameters.

ACKNOWLEDGEMENTS

Over the past few years many people have contributed in both tangible and intangible ways to the successful completion of this dissertation and I would like to gratefully acknowledge their contributions and express my deepest appreciation. First, I would thank my chair, Dr. Lloyd Bond, for his support, guidance, patience, and good humor throughout this whole process. Over the last five years he has been a teacher, a colleague, and a friend and I look forward to many more years of working and laughing together. I would also like to thank my committee members, Dr. Terry Cooper and Dr. Rita O'Sullivan. The opportunities I have had to work with and learn from each of them have rewarded me both personally and professionally. Finally, a special thanks to Dr. John Hattie, who gladly joined my committee when I needed him and who went above and beyond the call of duty to support me and ensure that the finished product was one of which I could be proud.

I would also like to thank a good friend and fellow student, Robert Johnson, who has gone through so much of this process with me. The journey is always better when it is made with a friend. Thanks also to Mark Price for his skillful editing of the text. His sharp eye has certainly made it a better read.

I have also received a tremendous amount of support and encouragement from my family and many good friends. First and foremost, I would like to thank my mother, Catherine C. Harman, who is my staunchest supporter, my biggest fan, and my best friend. Without her love, support, encouragement and patience, this would not have been possible. My love and thanks go also to my other family members, Mary and Mark Milam and David and Betty Ann Harman, for their love, support, and encouragement in this and every other area of my life. My love goes also to my two wonderful nieces, Hannah Catherine Milam and Mary Elizabeth Harman, who, for the past year and a half, have brought great joy and many smiles into my life. Throughout the long days and nights of writing this dissertation, their pictures have surrounded me and always bouyed my spirits. I hope someday they will know how much they contributed to this dissertation and the rest of my life.

I would also like to thank my friends and colleagues at the Center for Educational Research and Evaluation: Michelle, Wanda, Barbara, Kristen, John, Ann and Marnie. The daily doses of laughter and encouragement they have provided contributed in no small way to the completion of this process.

And finally, I would like to lovingly dedicate this dissertation to two very special people in my life. First, to my uncle, Thomas G. Church. Throughout my life he has been my role model and his accomplishments have served to inspire my own. And second, to my father, A. David Harman. Although he did not live to see me complete this process, in many ways he will always be with me and I know that he would be proud.

TABLE OF CONTENTS

Page

APPROVAL PAGE ii

ACKNOWLEDGEMENTS iii

CHAPTER

I. INTRODUCTION 1

Notation and Terminology Conventions of this Study 5
Organization of the Remainder of this Study 7

II. REVIEW OF RELEVANT LITERATURE 10

Judgmental Methods for Detecting DIF 11
Sensitivity Reviews 15
Classical Test Theory Methods for Detecting DIF 17
Transformed Item Difficulty 17
Correlational Methods 23
Analysis of Variance (ANOVA) 24
Contingency Table Methods for Detecting DIF 26
Scheuneman's Chi-square 27
Mantel-Haenszel Chi-square 30
Standardization Approach 32
Loglinear Modeling 33
Logistic Regression Model 36
Item Response Theory Methods for Detecting DIF 39
Lord's Chi-square for the Equality of ICC Parameters 41
Signed and Unsigned Area Between ICCs 41
Empirical Research Questions Addressed in this Study 44

III. METHODOLOGY 46

Empirical Data Collection Instrument:
The Graduate Management Admissions Test (GMAT) 47
Empirical Data Reduction Procedures 50
Simulation Data Generation Program: DGEN 50
Simulation Data Generation Procedures 51

Methodology for Investigating the First Research Question 54
Methodology for Investigating the Second Research Question ... 66

IV. RESULTS 68

Results of the Preliminary Factor Analyses 68
Results of Investigation of Research Question 1: Comparability of DIF Indices 72
Results of Investigation of Research Question 2: Accuracy of DIF Indices 93

V. DISCUSSION 99

Summary of Results of Investigation of Research Questions 99
Implications of Results of this Study for the Detection of Differential Item Functioning 110
Implications of Results of this Study for Further Research on the Detection of Differential Item Functioning 111

REFERENCES 113

CHAPTER I
INTRODUCTION

A primary concern for both test developers and test users is the validity or "fairness" of the tests they construct and administer. From the early years of this century when psychological testing first began, the development and use of tests has grown rapidly in many areas of society. Initially, tests were seen by many as objective measures of psychological constructs that could be used to make sound, reliable, and objective judgments. In the 1950s, when concern for the civil rights of racial and ethnic minorities and women began to grow in this country, many looked to psychological testing as a means of ensuring that members of minority groups were given an equal chance at receiving the educational and employment opportunities that they had previously been denied. It was believed that testing in education and employment settings would ensure the fair distribution of these opportunities on the basis of merit alone (Anastasi, 1988).

By the early 1960s, however, this objective had been far from realized for large groups of minorities. Those in the field of psychological testing and measurement were then accused of creating biased tests. Many of these accusations resulted from studies which compared the performances of minority and majority group members on many psychological tests. These studies revealed large and consistent differences in the average scores between the members of these groups with majority group members and males outperforming minority group members and females (Angoff, 1993; Cole, 1981).

The impact these findings had on society in general, and the measurement community specifically, was profound. Test developers could no longer assume that the public would accept without question that the tests they developed measured only the constructs they were intended to measure. The evidence pointed to the fact that many tests might not have been equally valid for all groups of people. As a result, the measurement community was faced with both the challenge and the responsibility of developing new methodologies for comparing the validity of test items across various demographic groups of examinees in order to ensure that any observed between-group differences in average test scores were due to "real" group differences in the construct being measured by the test and not artifactual or due to cultural or gender bias in the test items (Camilli & Shepard, 1994; Cole, 1993; Durovic, 1975). For example, observed between-group differences in the average test scores of males and females on a college level mathematics achievement test would be considered "real" if they were due to differences in high school mathematics preparation. If, on the other hand, the observed between-group differences on the test were due to the use of stereotypical or otherwise offensive language or content or the unequal familiarity or experience of females with the nominal content of the items, then those observed differences would be due to gender bias in the test items.

The development of these new methodologies required the reconceptualization of the conventional definition of validity. The new definition stated that test items measure the construct they purport to measure with the added stipulation that items also validly measure the construct with the same degree of accuracy for all demographic subgroups after matching examinees on the construct or latent trait being measured (Camilli, 1993; Jensen, 1976; Scheuneman, 1979). If the conditions imposed by this new definition held, a fair test item could be defined as one for which the probability of a correct response, after controlling for the construct or latent trait being measured, is equal regardless of the demographic group membership of the examinees. It follows from this definition that if the probability of a correct response is not the same for examinees matched on latent trait and differing only with respect to some demographic characteristic, then the item unfairly disadvantages the members of the lower scoring group (Scheuneman, 1979).

The earliest efforts at identifying and eliminating differential item functioning (DIF) in test items began with the use of judgment-based methods of item examination. These informal, judgment-based procedures focused on the review of test items with respect to the substantive features of the items, such as: the use of stereotypical or otherwise offensive language or content; fair representation of minorities and women; equal familiarity or experience of subgroup members with the nominal content of the items; and the opportunity of subgroup members to learn the item's content or processes (Tittle, 1982). Since these methods were first used as a means for identifying potential sources of DIF, many test developers have formalized and refined these review procedures and mandated their use at all stages of the test development process (Green, Coffman, Lemke, Raju, Hendrick, Loyd, Carlton, & Marco, 1982; Ramsey, 1993; Tittle, 1982).

Judgement-based procedures for detecting DIF were quickly followed by the development of statistical methodologies based on classical test theory. Initially, these methodologies involved the use of correlational and Analysis of Variance (ANOVA) procedures to examine classical test indices such as item difficulties and validity across groups via the ANOVA response-by-group interaction term. These methods, however, were quickly recognized as flawed because they did not control for the construct or latent trait being measured and, therefore, confounded differential impact (i.e., legitimate between-group differences) with DIF (Camilli & Shepard, 1994).

By the early 1980s these classical test theory methods gave way to statistical methodologies based on chi-square (χ^2) techniques and item response theory (IRT) models. Over the course of the 1980s and early 1990s, the χ^2 techniques have been extensively investigated and refined, and the Mantel-Haenszel approach proposed by Holland and Thayer (1988) has emerged as the preferred χ^2 -based DIF detection method for many test developers and other researchers. The methods based on the IRT models, however, are still being investigated and refined, and they continue to be the focus of much debate within the measurement community.

In addition to these more common statistical methodologies, possible log-linear approaches to the detection of DIF were proposed first by Marascuilo and Slaughter (1981) and Mellenbergh (1982). More recently, Green, Crone, and Folk (1989), Kelderman and Macready (1990), and Green (1991) have also suggested the use of log-linear models for both differential distractor functioning (DDF) and DIF analysis and detection, and have proposed tentative models for these purposes. Also, measurement specialists at the Educational Testing Service (ETS) have developed what they have called the Standardization approach for the analysis and detection of DIF and have demonstrated its usefulness in evaluating multiple-choice test items (Dorans, 1989; Dorans & Holland, 1993; Dorans & Kulick, 1986; Dorans, Schmitt & Bleistein, 1992). The results of these studies have shown the DIF indices derived using the Standardization approach to be virtually identical to the indices derived using the Mantel-Haenszel procedure.

Notation and Terminology Conventions of this Study

Throughout this study several notation and terminology conventions will be followed. First, the terms "bias" and "item bias" carry many negative social connotations and have largely been replaced in the professional literature with the value-neutral terms "DIF" and "differential item functioning." Therefore, for the purposes of the discussion presented here, the terms DIF and differential item functioning will be used whenever possible. However, when citing directly from other sources or referring to the work of others, the terms bias or item bias may be used. In these cases the terms DIF, differential item functioning, bias, and item bias are used synonymously and should be interpreted as Jensen (1980) proposed:

> In mathematical statistics, "bias" refers to a systematic under- or over-estimation of a population parameter by a statistic based on samples drawn from the population. In psychometrics, "bias" refers to systematic errors in the predictive validity or construct validity of test scores of individuals that are associated with the individual's [demographic] group membership. ... The assessment of bias is a purely objective, empirical, statistical and quantitative matter entirely independent of subjective value judgements and ethical issues concerning fairness or unfairness of tests and the uses to which they are put. *Psychometric bias is a set of statistical attributes conjointly of a given test and two or more specific subpopulations.* (p. 375)

The term "demographic group" will be used as a generic term to describe racial, ethnic, and gender groups. The examples and references used throughout Chapter II of this dissertation will not distinguish between the three and all procedures can be generalized to any of these demographic groups. Following Holland and Thayer (1988) and for the purpose of continuity the demographic groups of interest will be referred to as either the reference group (i.e., whites or males) or the focal group (i.e., African Americans or females) and will be indexed with either an "r" or an "f" respectively.

For all the DIF detection procedures presented, item responses will be classified as either "right" or "wrong" and will be indexed with either a "1" or a "0" respectively. In addition, where appropriate, the total test score will be assumed to be the criterion for matching examinees on ability, and examinees will be grouped into K ability levels. Finally, the specific item that is the focus of the DIF analysis will be referred to as the studied item.

Organization of the Remainder of this Study

The remainder of this dissertation will be organized into four chapters. First, in Chapter II a review of the professional literature relevant to the detection of differentially functioning test items is presented. The purpose of the chapter is to provide an overview of the dominant DIF detection methodologies that have been developed over the last three decades in order to provide the larger context for the dissertation study described here. For each of the DIF detection methods presented in Chapter II, the discussion focuses on the conceptual definition of differential item functioning adopted by the approach, the statistical procedures applied by the approach in calculating DIF indices and test statistics, the theoretical and practical strengths and weaknesses of the approach, and a discussion of any previous research conducted using the approach found in the professional literature. Chapter II concludes by identifying the two research questions that are the focus of this study.

Chapter III outlines the methodology that was followed while investigating each research question identified through the review of the professional literature presented in Chapter II. Chapter III is divided into six sections. The first section provides a brief description of the Graduate Management Admissions Test (GMAT), the data collection instrument from which the data for the empirical portion of this study was drawn. The second section presents a description of the data reduction procedures that were followed in preparing these empirical data for analysis. The third section provides a brief description of the data generation program, DGEN, which was used to generate the item response data for the simulation portion of this study. The fourth section presents a description of the procedures that were followed in generating the item response data used in the simulation portion of this study. The fifth section provides a discussion of the methodology used to investigate the first research question. And finally, the last section of this chapter provides a discussion of the methodology used to investigate the second research question.

Chapter IV presents the results of the statistical analyses that were performed while investigating the research questions that were defined in Chapter II. The chapter is divided into three sections. The first section presents the results of two preliminary factor analyses. The second section presents the results of the investigation of the first research question. And the last section presents the results of the investigation of the second research question.

Finally, Chapter V provides a discussion of the results of the study. This discussion focuses in three areas. First, a summary and discussion of the results of the study are presented. Second, the implications that the results of the study have for the detection of DIF items are identified and discussed. And finally, the implications that the results of this study have for future research on the detection of DIF are presented.

CHAPTER II
REVIEW OF RELEVANT LITERATURE

The purpose of this chapter is to review and synthesize the professional literature relevant to differential item functioning detection methods. The discussion presented here focuses on the evolution of DIF detection methods from the early 1960s to the present. The methods discussed in this chapter fall into four broad categories: judgmental methods, classical test theory methods, contingency table methods, and methods based on item response theory.

The first section of this chapter focuses on the early use of judgmental methods for the detection of DIF and the present day use of sensitivity reviews to identify and eliminate, in advance, potential sources of DIF from test items. The second section describes DIF detection methods that are based on classical test theory, including the Transformed Item Difficulty (TID) or Delta-Plot method, correlational methods and the Analysis of Variance (ANOVA) method. Contingency table methods for detecting DIF are discussed in the third section. These methods include Scheuneman's Chi-square procedure, the Full Chi-square procedure, the Mantel-Haenszel Chi-square approach, the Standardization approach, and the use of log-linear modeling and logistic regression approaches. Finally, the fourth section presents a discussion of Item Response Theory (IRT) methods for detecting DIF including Lord's Chi-square procedure for testing the equality of item characteristic curve (ICC) parameters, and procedures based on the signed and unsigned area between ICCs.

For each of the DIF detection methods presented, the discussion focuses on four areas: the conceptual definition of differential item functioning adopted by the method; the statistical procedures applied by the method in calculating DIF indices and test statistics; the theoretical and practical strengths and weaknesses of the method; and previous research conducted using the method that can be found in the professional literature. The final section of this chapter defines the research questions that are investigated through this dissertation study.

Judgmental Methods for Detecting DIF

Although the origins of modern psychological testing and measurement can be traced back to the early years of the 20th century and the work of Sir Francis Galton (1822-1911) in England and James McKeen Cattell (1860-1944) in the United States, issues of validity relating to the development and use of psychological tests were of only incidental importance at the time (Anastasi, 1988; Linden & Linden, 1968). It was not until the 1940s that the issue of validity was considered seriously by psychologists and psychometricians. Initially, however, the conception of validity was quite narrowly focused and the codification of validity standards at the time reflected this narrow focus.

The earliest attempts at articulating and formalizing validity standards for psychologists and other professional test developers required only that they demonstrate that the test measured what it claimed to measure for the population of examinees for whom the test was intended (American Psychological Association, 1954; Camilli & Shepard, 1994). To fulfill this requirement, test developers relied on judgment-based evidence to demonstrate the connections between the tests they developed and the psychological constructs that the tests claimed to measure. By the early 1960s when issues of test bias and differential validity first emerged, these judgment-based methods were all that were available to test developers to address the issues.

The earliest attempts at detecting and eliminating bias from test items involved the examination of potential test items by persons who were considered "experts" in identifying sources of racial, cultural, or gender bias (Tittle, 1982). The examination of test items by experts usually involved a review of the potential items which focused on the substantive features of the items. The substantive features on which the reviewers focused were: the use of stereotypical or otherwise offensive language or content; fair representation of minorities and women; equal familiarity or experience of subgroup members with the nominal content of the items; and the opportunity of subgroup members to learn the item's content or processes (Tittle, 1982). Often, those considered to be "experts" were simply members of the various minority groups against whom it was thought the items might be biased. For example, test items were reviewed by racial minorities to determine whether the language, context, content, or other features of the item might be perceived by members of the minority group as stereotypical or offensive in any way (Tittle, 1982).

It was not until the development of statistical methods for the detection of DIF in the early 1970s that test developers fully realized the inadequacies of these judgment-based procedures. It became clear at that time that although the expert review of test items for stereotypical or offensive language and content may remove some of the potential sources of *bias* from those test items, there are often many more subtle features of test items that result in *DIF* which go undetected by expert reviewers (Tittle, 1982). In fact, throughout the history of item bias and DIF detection, researchers have hypothesized about what these subtle features of items might be. They have, over the years, analyzed various item types and, based on their analyses, attempted to either predict or explain the characteristics of items that function differently for matched groups of examinees (McPeek & Wild, 1987; Medley & Quirk, 1974; O'Neill & McPeek, 1993; O'Neill, McPeek, & Wild, 1989; Pearlman, 1987; Rengel, 1986; Schmitt, 1988; Schmitt & Bleistein, 1987). Due to what is apparently the highly idiosyncratic nature of DIF, these attempts have not been very successful. What this research has shown is that although we can define what a differentially functioning test item is via empirical and statistical methods, often we can not recognize one when we see it (Bond, 1993; Engelhard, Hansche, & Rutledge, 1990; Hambleton & Jones, 1992).

Although these judgment-based reviews ultimately proved to be inadequate as a DIF detection method in and of themselves, their use during these early years served a number of important purposes. First, they focused public and professional attention on the existence of racial, cultural, and gender bias within the popular culture. Second, they gave professional credibility to the notion that features of a test item, such as language or content which is stereotypical or offensive, can impinge upon an examinee's performance on the item and, thus, is a source of invalidity (McLarty, Noble, & Huntley, 1989; Roid & Wendler, 1983). Finally, although these methods are inadequate on their own, they play an essential role in the development of tests and test items and their use should be continued. This understanding by test developers of the important role judgment-based review procedures play in the development of fair and unbiased tests has lead, over the years, to the incorporation of sensitivity review procedures as a standard part of the test development process (Green, Coffman, Lemke, Raju, Hendrick, Loyd, Carlton, & Marco, 1982; Hunter & Slaughter, 1980; McLarty, Noble, & Huntley, 1989; Ramsey, 1993; Tittle, 1982).

Sensitivity Reviews

Over the last 20 years, many of the large test developers in this country, as well as the professional organizations associated with test development and test use, have formalized sensitivity review guidelines and procedures and incorporated their use into the test development process (American Psychological Association, 1977; Macmillian, 1975; McGraw-Hill, 1968; McGraw-Hill, 1974). The sensitivity review process is a formal, judgment-based process that is an extension of the early judgmental procedures used for the purpose of detecting biased test items. The sensitivity review process, in addition to reviewing items for offensive language or content, also screens items for the use of other words and phrases that have been classified by the test developer as "caution" words and phrases and, as such, are to be avoided whenever possible (Green, Coffman, Lemke, Raju, Hendrick, Loyd, Carlton, & Marco, 1982; Hunter & Slaughter, 1980; Ramsey, 1993).

The Educational Testing Service was among the first of the large test development companies to formalize and mandate the use of sensitivity reviews for all test items as well as test-related publications, nonprinted materials, and research and statistical reports (Ramsey, 1993). Among the words that the ETS sensitivity review process flags as caution words are: backward, barbarian, birthrate, class, colonialism, crime, culturally disadvantaged, developing nation, gangs, ignorant, illegitimate, and inferior (Hunter & Slaughter, 1980; Ramsey, 1993). Although the use of these words, within certain contexts, would not be considered offensive, they have been recognized by ETS to have that potential within other contexts and, therefore, their use in test items and publications is closely monitored.

As the use of sensitivity reviews has become more commonplace, and the list of words and phases that are considered to be potentially offensive has grown, a number of attempts have been made to standardize sensitivity review procedures and to provide review guidelines for test developers as a whole. Among the earliest attempts at a standardization of procedures, Hambleton (1980) developed an Item Bias Review Form which could be used by any test developer to review test items for potential sources of racial, cultural, and gender bias. The form is used to flag test items which do not meet one or more of eight criteria for bias-free language and content. The eight criteria specified by Hambleton (1980) are:

1. Is the item free of offensive sexual, cultural, racial, and/or ethnic content?
2. Is the item free of sexual, cultural, racial, and/or ethnic stereotyping?
3. Is the item free of language that would be offensive to a segment of the examinee population?
4. Is the item free from descriptions that would be offensive to a segment of the examinee population?
5. Will the activities or situations described in the item be familiar to all examinees?
6. Will the words in the item have a common meaning to all examinees?
7. Is the item free of difficult vocabulary and/or sentence structure?
8. Will the item format be familiar to all examinees?

In addition to the Item Bias Review Form developed by Hambleton (1980), a number of other sensitivity review forms have been developed and used over the years to screen test items (Jensen & Beck, 1979; Saario, Jacklin, & Tittle, 1973; Science Research Associates, 1976). Each of these review forms provides test developers with a useful rubric for evaluating test items and determining whether they could be viewed as offensive to racial, cultural, or gender group members.

Classical Test Theory Methods for Detecting DIF

Although these judgment-based procedures could be used to detect language and content that members of various minority groups might find offensive, they were generally inadequate on their own. By the early 1970s test developers had begun to develop more objective, empirically-based methods for identifying test items that functioned differently for matched groups of examinees. The earliest methods were firmly based in classical test theory and included the Transformed Item Difficulty (or Delta-Plot method) developed by Angoff (1972), ANOVA-based procedures that used significant group-by-item interaction to flag DIF items, and correlational methods (Crocker & Algina, 1986).

Transformed Item Difficulty

The Transformed Item Difficulty (TID) method was first used by Angoff (1972) in the early 1970s as a method for detecting DIF. The TID method (also known as the Delta-Plot method) conceptualizes DIF as differential difficulty

(Oosterhof, Atash, & Lassiter, 1984). That is, any item that is relatively more difficult for members of one group than it is for members of the other group is considered to be functioning differently for the two groups (Angoff, 1972; Angoff, 1993; Angoff & Ford, 1973). The TID method provides a graphical representation of item difficulty values based upon Thurstone's (1925) classical Method of Absolute Scaling (Camilli & Shepard, 1994).

The procedures used to create these graphical representations are quite straightforward. First, for each subgroup of examinees, the classical item difficulty or p-value is calculated. This value is simply the proportion of examinees within a subgroup who answered the item correctly. Once these values have been calculated for each item within each subgroup, these values are normalized, typically to a mean of 13 and a standard deviation of 4, but theoretically any mean and standard deviation could be used. (It is from this normalizing or transforming of the classical item difficulties that the method derives its name.) For each item on a test, a pair of transformed item difficulty values, often referred to as deltas, are calculated. These deltas are then plotted on a bivariate graph with the deltas for one group placed along the x-axis and the deltas for the other group placed along the y-axis (Angoff, 1972; Angoff, 1982; Camilli & Shepard, 1994).

This plot, then, is used to identify individual test items that are functioning differently for the two groups. If the item is of exactly equal difficulty for members of the two subgroups, the deltas would fall along a line that extends from the lower left-hand corner of the bivariate graph and raises at a 45° angle toward the upper right-hand of the graph. Using real test data, this exact relationship between deltas for two groups is never achieved. Typically, however, if the two groups of examinees are fairly closely matched on the ability being measured by the test items, the plot of the deltas will form a narrow ellipse around a major axis which lies at nearly a 45° angle extending from the lower left- to the upper right-hand corner of the graph. This type of graph indicates that, for each group, the items have roughly the same rank ordering with respect to the difficulty and, thus, the correlation of the deltas for the two groups is quite high (Angoff, 1982). If one group is of higher ability on the construct being measured by the test items, the deltas for the two groups will still fall narrowly around a nearly 45° line, however, the line will be displaced either vertically or horizontally depending on which is the higher ability group (Angoff, 1982; Camilli & Shepard, 1994). If, on the other hand, the groups come from different populations, or if the items do not have the same meaning for members of the two groups, the deltas will scatter more widely around the major axis indicating a different rank ordering of the item difficulties within each group and, by definition, a lower correlation between the deltas (Angoff, 1982; Angoff, 1993).

Using the plot of the delta values a TID index can be calculated for each item. The TID index for an item is defined simply as the perpendicular distance from the point on the bivariate graph which represents the pair of deltas for an item and the major axis. This TID index has been proposed as an indicator of the amount of DIF being displayed by the items on a test (Angoff, 1982).

Although the Transformed Item Difficulty method has, in the past, had great appeal due to its conceptual simplicity, low cost, the ease with which it can be implemented, and the relatively small sample sizes needed to apply the procedures, it has been criticized by many as being seriously flawed (Hambleton & Rogers, 1991; Harris & Kolen, 1989). The three primary criticisms of the TID method are, first, that the delta values calculated for the two subgroups being compared may not be equally reliable, second, that the method, because it is based on classical item p-values, confounds DIF with legitimate differences in group means, and third, again because the method is based on classical item p-values, DIF is also confounded with legitimate differences in item discriminations (Angoff, 1982; Hunter, 1975; Lord, 1977).

To overcome the problem of unequal reliabilities of the deltas, Cardall and Coffman (1964) and Coffman (1961, 1963) proposed using an arcsine transformation of the p-values in order to control for different item variances (i.e., difficulties). Plake and Hoover (1979) investigated this solution and obtained results which demonstrated that such a transformation is quite effective in equalizing the item variances.

The second criticism (i.e., confounding DIF with legitimate differences in group means) is more difficult to overcome. Cleary and Hilton (1968) and

Angoff and Sharon (1974), have both shown that the use of the within-groups items-by-subjects interaction as an error term virtually always yields results that are statistically significant. This is the case because even when both the sample size and the test length are small, the degrees of freedom will be large, resulting in nearly any effect size being statistically significant regardless of whether the effect size is of any practical significance.

Camilli and Shepard (1994) provided an example which illustrates the third criticism noted above by demonstrating that when the two groups being compared are not of equal ability on the construct being measured by the test items, highly discriminating items (i.e., items that distinguish well between members of the lower ability group and members of the higher ability group) will appear to be functioning differently simply because they do discriminate so well.

In spite of these criticisms of the TID method, and his own acknowledgment of its limitations, Angoff (1982, 1993) continued to defend the TID method and its associated statistics for the evaluation of DIF in test items. He offered a number of remedies for the flaws noted previously. Specifically, he suggested the use of some relevant external criterion measure of the construct being measured by the items as a way to overcome the problem of confounding which results from differences in group means. The criticisms that have been made of this remedy are two-fold. First, although matching the groups on some external criterion prior to applying the TID method would likely reduce the effects of the confounding, it would not eliminate the confounding altogether. Second, in practice, a relevant external criterion on which to match the two groups rarely exists (Camilli & Shepard, 1994).

In addition, Angoff (1982) suggested that the confounding which is the result of different item discriminations can be remedied by making a simple adjustment to the delta values based on the item-test point-biserial correlations. To make this adjustment, he recommended that, prior to constructing the plot of the delta values, each delta value should be divided by its respective item-test point-biserial correlation. Angoff (1982) recommended this as a remedy because, in classical test theory, the item-test point-biserial correlation is an estimate of the item's discriminating power and, by making this transformation, any differences in item discrimination are adjusted out of the deltas. Although Angoff (1982) recommended this transformation, he also notes that point-biserial correlations are generally quite unreliable and, therefore, the adjustment is similarly unreliable. A comparative study by Shepard, Camilli, and Williams (1985) showed that this adjustment actually resulted in a lower rate of agreement between the TID index and other DIF indices based on preferred methodologies.

Angoff (1982) also defended the TID method as a legitimate approach to the analysis of test items for purposes other than the detection of DIF. In support of this position, Angoff (1982) noted that the TID or Delta-Plot method has been used for a variety of purposes other than the detection of DIF, "including the study of cultural and sex differences (Angoff & Ford, 1973; Angoff & Herring, 1976; Breland, Stocking, Pinchak, & Abrams, 1974; Coffman, 1961), equating of scores across groups presumed to be culturally different (Angoff & Modu, 1973; Angoff & Stern, 1973), general score equating (Thurstone, 1925), and the standardization and equating of item difficulties (Thurstone,1947). Additionally, delta-plots have been used to identify miskeyed items, items which have become obsolete, and items having different 'psychological meaning' for different groups." (p. 101-102)

Correlational Methods

The use of correlational methods to detect DIF has also been investigated (Stricker, 1982; Stricker, 1984). One correlational method involved the calculation of classical item difficulties, p-values, for each item within each subgroup. The items were then ranked within each subgroup according to their difficulties and the rank-order correlation of the items for the two subgroups was calculated (Camilli & Shepard, 1994). It was believed that a rank-order correlation close to 1.0 indicated that the relative difficulty of the items across the subgroups was the same and, therefore, the items were measuring similarly for the two groups. If, on the other hand, the rank-order correlation was significantly lower than 1.0, a group membership-by-item difficulty interaction was present.

The other correlational method investigated by Green (1971) and Green and Draper (1972) involved calculating, for each subgroup, the item-test point-biserial correlation. The point-biserial correlations within each subgroup were then classified as either "high" or "low" with high correlations falling in the upper half of all the correlations for the subgroup and low correlations falling in the lower half. Then items which were among the high correlations for one subgroup and among the low correlations for the other subgroup were flagged as biased.

As with the DIF detection methods based on classical test theory discussed previously, these correlational methods are criticized for the same flaws and, therefore, have not been widely used and are not recommended (Hunter, 1975; Camilli & Shepard, 1994).

Analysis of Variance

The use of analysis of variance (ANOVA) for the detection of DIF was widely used throughout the 1970s and into the early 1980s. The popularity of the ANOVA method is probably most directly attributable to two features of the method. First, it is a statistical approach and, therefore, preferred by many over approaches which rely on observational methodologies. And second, ANOVA is a well-known procedure that is easily understood, applied, and explained.

From the ANOVA perspective, DIF is conceptualized as the differential performance (i.e., differential difficulty) by subgroups on a set of test items and, as such, it can be detected via the group-by-item interaction term in the ANOVA model (Dreger & Miller, 1968; Medley & Quirk, 1974; Schmeiser, 1982; Shuey, 1966). DIF analyses using ANOVA are performed by setting up a two-factor, repeated measures ANOVA with examinee subgroup membership as one factor and the test items as the within-groups factor (Camilli & Shepard, 1994). A significant main effect due to groups is interpreted as an indication of average group differences in the construct being measured, while a significant group-by-item interaction is interpreted as an indication of differential difficulty.

In spite of its theoretical and computational appeal, the ANOVA approach to the detection of DIF suffers from the same fundamental weakness that other classical test theory-based methods do. That is, because it relies on the classical p-value or proportion correct score it confounds DIF with legitimate differences in the mean performance of the groups on the items. This weakness has been investigated and discussed in detail by Hunter (1975), Lord (1977), and Camilli and Shepard (1987). Camilli and Shepard (1987) demonstrated, both algebraically and through a simulation study, that when the true group differences in mean performance on the test items is larger than the true DIF of the test items, the ANOVA will attribute a larger proportion of the true DIF to the group main effect and less to the interaction effect. In fact, Camilli and Shepard (1994) conclude, on the basis of their own analyses and the analyses of others, that the "ANOVA should no longer be recommended as a bias detection procedure, even for preliminary screening of items." (p. 34) This well-documented weakness, in conjunction with the recent proliferation of other theoretically preferred methods, has made ANOVA essentially obsolete as a method for the detection of DIF.

Contingency Table Methods for Detecting DIF

By the late 1970s, the inadequacies of DIF detection methods based on classical measures of differential item difficulty were so well established that the search for other, statistically sound and theoretically preferred, methods was well underway. Among the first methodologies to emerge from the research on the detection of DIF were contingency table methods. Scheuneman (1979) was one of the first (and ultimately the most well-known) researchers to offer a contingency table approach. Though her original computation of the statistic, known as Scheuneman's χ^2, contained a flaw, a variation of her method remained in use for many years through a minor correction to the computation of the statistic proposed by Baker (1981) and acknowledged by Scheuneman (1981). This corrected version of Scheuneman's χ^2 is often referred to as the Full χ^2.

The most widely used of the contingency table methods is the Mantel-Haenszel approach. Originally developed by Mantel and Haenszel (1959), the use of the method for the detection of DIF has been popularized in recent years by Holland and Thayer (1988). More recently, a closely related contingency table method, the Standardization approach, has been developed and extensively investigated by researchers at ETS (Dorans, 1989; Dorans & Kulick, 1983; Dorans & Kulick, 1986; Dorans, Schmitt, & Curley, 1988; Rivera & Schmitt, 1988; Schmitt & Bleistein, 1987; Schmitt & Dorans, 1990). Finally, several approaches based on the log-linear modeling of contingency table data and logistic regression have been developed and investigated.

Scheuneman's Chi-Square

Scheuneman (1979) was among the first to propose a χ^2-based procedure for the detection of DIF. According to Scheuneman (1979), "an unbiased item is defined as an item for which the probability of a correct response is the same for all persons of a given ability, regardless of their [demographic] group membership." (p. 145) To test the hypothesis that an item is unbiased according to this definition, Scheuneman proposed a modified χ^2 procedure that is analogous to the item characteristic curve procedures used by the IRT-based methods (Scheuneman, 1979). Scheuneman's procedure proposed that examinees from two demographic groups of interest be matched with respect to total test score and then grouped into K score levels across the range of total test scores with three to five groups as a recommended number. Theoretically, the total number of matched groups can range from one to N+1, where N is the total number of items on the test, but the choice of the number of matched groups is largely dependent upon the amount of available data. In general, a larger number of matched score levels is preferred. The individual observations at each score level are then cross-classified according to the general format depicted in Table 1 below.

Table 1.
Data for the Matched Set of Examinees at the k(th) Score Level

	Score on Studied Item		
	1	0	Total
Reference Group	A_{1rk}	B_{0rk}	$n_{.rk}$
Focal Group	C_{1fk}	D_{0fk}	$n_{.fk}$
Total	$m_{1.k}$	$m_{0.k}$	$T_{..k}$

The element "A" in the table represents the number of reference group members who answered the studied item correctly, while the element "B" represents the number of reference group members who answered the studied item incorrectly. The elements "C" and "D" are all interpreted similarly for the focal group members. The row marginal $n_{.rk}$ represents the total number of reference group members at the k(th) ability level for the studied item. The remaining marginal totals (row and column) are, again, interpreted similarly. Finally, the element $T_{..k}$ represents the total number of respondents at the k(th) ability level for the studied item.

Scheuneman's χ^2 statistic for these K 2x2 contingency tables is expressed as:

$$\chi^2 = \sum_{k=1}^{K} \frac{(F_{ek} - F_{ok})^2}{F_{ek}} + \frac{(R_{ek} - R_{ok})^2}{R_{ek}}$$

where F_{ek} is the expected value for the focal group at the k(th) score level, F_{ok} is the observed value for the focal group at the k(th) score level, R_{ek} is the expected value for the reference group at the k(th) score level, and R_{ok} is the observed value for the reference group at the k(th) score level.

Scheuneman (1979) originally believed that, under the null hypothesis of no DIF, this statistic is distributed as an approximate χ^2 with K-1 degrees of freedom. It was quickly shown that this was not, in fact, the distribution of the statistic (Baker, 1981; Scheuneman, 1981).

Baker (1981), in response to Scheuneman (1979), demonstrated that, under the null hypothesis of no DIF, the expected value of Scheuneman's χ^2 statistic is dependent upon the total number of incorrect responses for each of the K 2x2 tables. He further demonstrated that this dependency could be corrected by including a minor multiplicative factor in the denominator of the statistic (Baker, 1981). This corrected χ^2 statistic is often referred to as the Full χ^2. The mathematical form of the Full χ^2 is:

$$\chi^2 = \sum_{k=1}^{K} \frac{(F_{ek} - F_{ok})^2}{F_{ek}(1 - P_k)} + \frac{(R_{ek} - R_{ok})^2}{R_{ek}(1 - P_k)}$$

In the equation for the Full χ^2 statistic, the term (1- P_k) is the proportion of incorrect responses at the k(th) score level. Scheuneman (1981) acknowledged the need for the correction to the original formulation of the statistic and showed that the appropriate degrees of freedom for the Full χ^2 statistic is K in the case of two demographic groups and K(J-1) in the case of J demographic groups.

The principal weakness of Scheuneman's corrected χ^2 statistic is that although it can be used to identify DIF via a statistical test of the null hypothesis of no DIF, it does not provide an index that can be used to estimate the amount of DIF displayed by the item. The lack of a DIF index associated with the statistical test of the null hypothesis is a significant weakness of Scheuneman's χ^2 method because whenever the statistical test is carried out using the data from a large numoer of examinees, trivially small differences in item functioning will often result in statistical significance and the judgement that items are functioning differentially across groups when, in fact, they are not.

Mantel-Haenszel Chi-Square

Since the early 1980s, the most widely used χ^2 procedure for the detection of DIF has been the Mantel-Haenszel χ^2 procedure (Mantel & Haenszel, 1959; Dorans & Holland, 1993; Holland & Thayer, 1988). The primary advantage of the Mantel-Haenszel procedure over Scheuneman's corrected χ^2 is that, in addition to a test of statistical significance, it provides an index that can be used to estimate the amount of DIF displayed by the item. Like Scheuneman's χ^2 statistic, the Mantel-Haenszel approach uses data from K 2x2 tables for matched groups of examinees. Using the proportions of examinees in the reference and focal groups instead of frequencies, the Mantel-Haenszel χ^2 approach provides a statistical test of the null hypothesis of no DIF against the specific alternative hypothesis that a constant odds ratio exists, α_{MH}, which, when different from 1, represents that average amount by which the item is relatively more difficult for members of one group than for members of the other group (Dorans & Holland, 1993). The computational form of the α_{MH} constant-odds ratio is:

$$\alpha_{MH} = \left[\Sigma_k R_{rk} W_{fk} \;/\; N_{tk}\right] / \left[\Sigma_k R_{fk} W_{rk} \;/\; N_{tk}\right]$$

An additional advantage to using the Mantel-Haenszel χ^2 approach for detecting DIF in test items is that it matches examinees on ability at each score level across the range of scores. Therefore, unlike Scheuneman's χ^2 statistic which groups examinees into three to five score levels, the Mantel-Haenszel χ^2 approach does not confound DIF with legitimate differences in mean group performance. In addition, much research has been conducted over the last decade on the use of the Mantel-Haenszel approach to the detection of DIF and the indication has been that it is a theoretically and statistically sound approach which is computationally simple and, as such, has many advantages over most of the other methods discussed previously (Allen & Donoghue, 1991; Baghi & Ferrara, 1989; Baghi & Ferrara, 1990; Camilli & Smith, 1988; Clauser, Mazor, & Hambleton, 1991; Donoghue & Allen, 1993; Donoghue, Holland, & Thayer, 1993; Englehard, et. al, 1990; Hambleton, Clauser, Mazor, & Jones, 1993; Hambleton & Rogers, 1989; Hambleton, Rogers, & Arrasmith, 1986; Mazor, et. al, 1992; Raju, Bode, & Larsen, 1989; Ryan, 1990; Ryan, 1991).

Standardization Approach

The Standardization approach has been developed and extensively investigated by researchers at the Educational Testing Service (Dorans, 1989; Dorans & Kulick, 1983; Dorans & Kulick, 1986; Dorans, Schmitt, & Curley, 1988; Rivera & Schmitt, 1988; Schmitt & Bleistein, 1987; Schmitt & Dorans, 1990). The Standardization approach has been described by Dorans and Holland (1993) as "an IRT-like approach" which compares empirical item response curves using total test score as an estimate of examinee ability. The Standardization approach defines DIF as differences in expected performance on an item for examinees of equal ability from different subgroups (Dorans & Holland, 1993). That is, an item exhibits DIF if equally able examinees from different subgroups do not have the same probability of answering the item correctly (Dorans, Schmitt, & Bleistein, 1992; Wright, 1987).

The Standardization approach is a nonparametric approach which indexes DIF via a weighted difference in proportion correct between focal and reference group members across K score levels. The mathematical form of the standardized P-DIF index is:

$$\text{STD P-DIF} = \Sigma\,\{W_k\,[P_{fk} - P_{rk}]\}\ /\ \Sigma W_k$$

The weighting factor, W_k, at each of the K score levels is typically the number of focal group examinees at score level k, and the term in the denominator is the summation of all of these weighting factors across the K score levels (Dorans, 1989; Dorans & Holland, 1993; Dorans, Schmitt, & Bleistein, 1992).

Several comparative studies have shown that the Standardization and Mantel-Haenszel approaches yield highly correlated indices of DIF (Dorans, 1987; Dorans, 1989; Dorans & Holland, 1992). One primary advantage of the Standardization approach over other methods for detecting DIF is that it has been demonstrated to be quite versatile in investigating other factors which affect subgroup performance on test items such as differential speededness and differential distractor functioning (Dorans & Kulick, 1983; Dorans & Kulick, 1986; Dorans, Schmitt, & Curley, 1988; Rivera & Schmitt, 1988; Schmitt & Bleistein, 1987; Schmitt & Dorans, 1990; Schmitt, Dorans, Crone, & Maneckshana, 1991).

Log-linear Modeling

A method that is closely related to the Standardization method just discussed and the other methods for detecting DIF based on chi-square analyses across K two-way contingency tables is the log-linear approach

(Green, 1991; Green, Crone, & Folk, 1989). Log-linear models are a logical extension of the other contingency table methods discussed. The primary distinction between the log-linear models and the other contingency table methods is that, for each item, the log-linear models are applied to the data after they have been cross-classified into a single three-way contingency table according to group membership, response option, and ability group. Like Scheuneman's χ^2 approach, examinees are matched with respect to total test score and then grouped into K score levels across the range of total test scores with three to five groups as a recommended number. Theoretically, the total number of matched groups can range from one to N+1, where N is the total number of items on the test, but the choice of the number of matched groups is largely dependent upon the amount of available data. In general, a larger number of matched score levels is preferred. Using the cell frequencies from the three-way cross-classification of the data, log-linear models are applied (Bishop, Feinberg, & Holland, 1975; Feinberg, 1990; Green, 1991; Green Crone, & Folk, 1989; Kelderman & Macready, 1990; Knoke & Burke, 1980; Kok, Mellenbergh, & van der Flier, 1985; Marascuilo & Slaughter, 1981; Mellenbergh, 1982). Like an analysis of variance (ANOVA) the log-linear analysis partitions out the three main effects associated with the three classification variables in the model, as well as the three two-way interaction terms associated with each combination of main effects, and the three-way combined effect.

Using the log-linear approach, DIF is operationalized as the two-way interaction effect between group membership and item response. If this interaction term is important in helping to explain the observed cell frequencies that resulted from the three-way cross-classification of the data, then the item is said to be functioning differentially for the different groups. The significance of the group x item response interaction term is calculated by taking the difference between the Likelihood Ratio χ^2 statistic associated with the model that includes the three main effect terms and the three two-way interaction terms and the Likelihood Ratio χ^2 statistic associated with the model that has only the three main effect terms and the two remaining two-way interaction terms, ability group x group and ability group x item response. The resulting statistic is G^2 and it is distributed as a χ^2 with degrees of freedom equal to the degrees of freedom associated with the second model minus degrees of freedom for the first model. This test statistic can be compared to the appropriate χ^2 table to determine its statistical significance (Green, 1991; Green, Crone, & Folk, 1989).

Like the standardization approach, the log-linear approach has the advantage of being quite versatile in investigating other factors which affect subgroup performance on test items such as differential speededness and differential distractor functioning (Green, 1991; Green, Crone, & Folk, 1989).

Logistic Regression

Swaminathan and Rogers (1990) investigated the use of the logistic regression model for characterizing DIF in test items. The logistic regression model is a special case of the log-linear model where individuals are cross-classified by group membership and item response, but ability level is treated as a continuous, not a categorical variable. Swaminathan and Rogers (1990) argued that the logistic regression procedure extended both the Mantel-Haenszel procedure and the log-linear models in two important ways: first, it takes into account the continuous nature of the ability scale and second, it is capable of identifying both uniform and nonuniform DIF.

When investigating DIF in test items, the logistic regression model is used to predict the probability of a correct response to a test item by an examinee given his or her particular ability level. The logistic regression model is:

$$P\left(u_{ik} = 1 \mid \theta_{ik}\right) = \frac{e^{\left(\beta_{0k} + \beta_{1k}\theta_{1k}\right)}}{1 + e^{\left(\beta_{0k} + \beta_{1k}\theta_{1k}\right)}}, \quad i = 1, \ldots, n_k, \; k = 1,2$$

where u_{ik} is the response of the ith examinee at the k(th) score level to the test item, θ_{ik} is the ability level of the i(th) examinee at the k(th) score level, and β_0 and β_1 are the intercept and slope of the regression line, respectively. Using this model, separate regression equations are calculated for the two groups of interest and the parameters of the regression equations are the compared. Within the context of logistic regression, DIF is defined as the unequal probability of a correct response to a test item by members of different groups who have been matched on the construct or latent trait being measured by the item. If the parameters of the two regression equations are the same, then members of the two groups who are at the same level of the construct or latent trait being measured by the item have the same probability of a correct response. If the regression equations have equal intercept parameters, β_0, but different slope parameters, β_1, the curves defined by the two regression equations are parallel but not coincident, thus indicating uniform DIF. If, on the other hand, the regression equations have unequal intercept parameters, β_0, but equal slope parameters, β_1, the curves defined by the two regression equations are not parallel and not coincident, indicating nonuniform DIF.

As noted earlier, Swaminathan and Rogers (1990) argued that the logistic regression procedure for the detection of DIF has a number of advantages over both the Mantel-Haenszel procedure and the more general log-linear models discussed previously. First, the logistic regression model is more general and flexible than the Mantel-Haenszel procedure because it takes into account the continuous nature of the ability scale. In fact, Swaminathan and Rogers (1990) demonstrate algebraically that the Mantel-Haenszel procedure can be characterized as a special case of the logistic regression model where "the ability variable is discrete and no interaction between the group variable and ability is permitted" (p. 365).

A second advantage is that the logistic regression model is capable of identifying both uniform and nonuniform DIF while the Mantel-Haenszel procedure and the general log-linear models are blind to nonuniform DIF. Using simulated data, Swaminathan and Rogers (1990) showed that the logistic regression procedure "is as powerful as the Mantel-Haenszel procedure in detecting uniform DIF and more powerful than the Mantel-Haenszel procedure in detecting nonuniform DIF" (p. 368-369). It should be noted, however, that the simulated data only included nonuniform DIF items resulting from a disordinal interaction between ability level and group membership. That is to say, the item characteristic curves for the nonuniform DIF items crossed in the middle of the ability scale. The results of this study, therefore, may not generalize to nonuniform DIF which results from the ordinal interaction of ability level and group membership where the item characteristic curves cross at either the low or high end of the ability scale.

A final advantage of the logistic regression model noted by Swaminathan and Rogers (1990) is that it provides a model-based approach to the investigation of DIF which allows for the inclusion of curvilinear terms or transformations of the ability variable in the model, In addition, other variables considered relevant or important to the understanding of DIF in the test items being investigated can also be included in the model.

The primary disadvantage noted by Swaminathan and Rogers (1990) to the use of the logistic regression model for the investigation of DIF is cost. Unlike the Mantel-Haenszel procedure which is quick and inexpensive to carry out, the logistic regression procedure is iterative and, therefore, more expensive. Based on their own experiences, Swaminathan and Rogers estimated the logistic regression procedure was 3 to 4 times more expensive than the Mantel-Haenszel procedure in terms of necessary computer resources.

Item Response Theory Methods for Detecting DIF

The most recent advances in the area of DIF analysis have been made in the use of Item Response Theory (IRT) methods to model examinee responses to test items and to identify DIF (Thissen, Steinberg, & Wainer, 1988; Thissen, Steinberg, & Wainer, 1993; Thissen, Wainer, & Steinberg, 1985; Wilson-Burt, Fitzmartin, & Skaggs, 1986). Although the mathematical foundations of IRT models were first described by Lord (1952), it wasn't for another 20 years that these models were applied to the investigation of DIF. According to the IRT approach, examinees responses to an item can be modeled using the logistic function. The logistic function is a monotonically increasing curve that represents the probability of a correct response to the item as a function of ability. These s-shaped curves are called item characteristic curves (ICCs) and can be defined using three parameters: a difficulty parameter, a discrimination parameter, and a pseudo-guessing parameter, referred to as the b, a, and c parameters, respectively.

The value of the difficulty parameter indicates the point along the ability scale, called the theta scale, where examinees at that ability level have a probability of 0.50 of correctly responding to the item. The discrimination parameter is the slope of the ICC at that point on the ability scale where examinees at that ability level have a probability of 0.50 of correctly responding to the item. Finally, the pseudo-guessing parameter is the lower asymptote of the ICC indicating the probability that examinees of extremely low ability will answer the item correctly. From within the IRT framework, then, an item is considered to be displaying DIF if the ICCs for two subgroups do not overlap. This lack of overlap indicates that for subgroup examinees of equal ability on the construct being measured there exists an unequal probability of answering the item correctly.

The inequality of ICCs across subgroups can be measured in two ways. First, Lord (1980) has developed a statistic for testing the equality of the ICC parameters called Lord's χ^2. The other approach often used it to calculate the area between the two ICCs. The first of these methods allows for a statistical test of the null hypothesis of no DIF, while the second is an index of the amount of DIF displayed by the item with respect to the two groups of interest.

Lord's Chi-square for the Equality of ICC Parameters

By definition, if two ICCs overlap, the parameters that define those ICCs must be equivalent. Therefore, the statistic developed by Lord (1980) uses the parameters of the ICCs for the two subgroups to test the null hypothesis of no DIF (i.e., that the two ICCs are equal in all of their parameters) against the alternative that the two ICCs differ with respect to at least one of the parameters.

Lord's χ^2 statistic for testing the equality of ICC parameters is:

$$\chi^2 = (b_r - b_f, a_r - a_f)' S^{-1} (b_r - b_f, a_r - a_f)$$

where S^{-1} is the inverse of the variance-covariance matrix for the item parameters. Lord's χ^2 statistic is distributed as a χ^2 with 2 degrees of freedom.

Signed and Unsigned Area Between ICCs

The signed and unsigned area between ICCs has been developed as an index of the amount of DIF displayed by an item (Rudner, 1977; Rudner, Getson, & Knight, 1980a; Rudner, Getson, & Knight, 1980b; Raju, 1977; Shealy & Stout, 1993; Thissen, Steinberg, & Wainer, 1988; Thissen, Steinberg, & Wainer, 1993; Thissen, Wainer, & Steinberg, 1985). These two indices are closely related and the calculations for each are, therefore, quite similar. The signed area between the ICCs is calculated as the integral:

$$\int_R \left[P_R(\hat{\theta}) - P_F(\hat{\theta})\right] d\,(\theta)$$

The probability of a correct response by a focal group member at ability level theta is subtracted from the probability of a correct response by a reference group member at ability level theta, and these differences in probability are integrated across the ability scale. Using the signed area calculated as described above, if the reference group members have a greater chance than the focal group members of answering the item correctly across the ability scale, the sign on the index will be positive. If the opposite is true, the sign on the index will be negative. If the ICCs cross at some point along the ability scale, the item is said to display nonuniform DIF and the sign may be either positive or negative.

The unsigned area is calculated as the integral:

$$\sqrt{\int_R \left[P_R(\hat{\theta}) - P_F(\hat{\theta})\right]^2 d(\theta)}$$

For the unsigned area, however, the probability of a correct response by the focal group at ability level theta is subtracted from the probability of a correct response by the reference group at ability level theta. To remove the sign, these differences in probability are squared and then integrated across the ability scale. To place the unsigned area on the same scale as the signed area, the square root of the squared area integral is calculated. The unsigned area will always be positive because of the squaring of the individual differences in probabilities. The unsigned area indexes the total area between the two ICCs without regard to which group is advantaged. The unsigned area index is unaffected by nonuniform DIF, and when the DIF displayed is uniform (i.e., one group is consistently advantaged by the item), the signed and unsigned areas will be the same (Camilli & Shepard, 1994).

The IRT approach to DIF detection has been the focus of much attention in the field of DIF analysis, primarily because the mathematical models on which the approach is based have many desirable statistical features. Like all of the models discussed here, the use of the IRT model is appropriate if the normal ogive or logistic function adequately represents the data, and the data are unidimensional (Hambleton, Swaminathan, & Rogers, 1991; Ironson, 1982). If these assumptions are met, many researchers have argued that the three-parameter IRT model is, both theoretically and statistically, the most appropriate method for investigating DIF (Bejar, 1980; Hunter, 1975; Lord, 1980; Petersen, 1977). The advantages and disadvantages of the IRT approach have been extensively investigated and documented in the professional literature (Craig & Ironson, 1981; Hambleton, Swaminathan, & Rogers, 1991; Ironson & Subkoviak, 1979; Rudner & Convey, 1978; Rudner, Getson, & Knight, 1980a; Rudner, Getson, & Knight, 1980b; Shepard, Camilli, & Averill, 1980; Subkoviak, Mack, & Ironson, 1981). The principal advantage of the IRT method is the sample invariant nature of the item and ability parameter estimates. This feature essentially eliminates the confounding of difficulty and discrimination indices, as well as the potential for legitimate differences in average group performance to be mislabeled as DIF (Ironson, 1982). The principal disadvantages of the IRT methods are largely practical in nature. First, it has been argued that the computer programs used to calculate the IRT parameters (e.g., LOGIST) are more expensive to run, in terms of computer time, than are the computer programs used to generate many other DIF indices (Swaminathan & Rogers, 1990). And finally, the sample sizes necessary for reliable parameter estimation are generally quite large.

Empirical Research Questions Addressed in this Study

A review of the professional literature indicates that over the last three decades a variety of methodologies for the detection of differential item functioning have been developed. Many of these methodologies initially appeared to have promise but have since been shown to be statistically unsound. Three of the more recently developed methodologies do not suffer from these same flaws and are currently being used by test developers to screen items for indications of DIF. These methods are the Mantel-Haenszel χ^2 approach, the IRT Unsigned Area approach, and the log-linear approach. Each of these methods, however, have features which may make their application in certain testing and research situations impractical. Therefore, it is of interest to the measurement community to determine, first, the degree to which these three methodologies yield comparable results with respect to the detection of DIF in test items and, second, which, if any, of these methodologies is more accurate in detecting DIF which present in test items.

This dissertation study focuses on two major research questions. The first research question investigated is: To what degree do the Mantel-Haenszel χ^2 approach, the IRT Unsigned Area approach, and the log-linear approach yield comparable indices with respect to the amount of DIF displayed by test items? The second research question investigated is: How accurately does each DIF detection method identify test items with respect to the amount of DIF displayed by each item. The methodology used to investigate these two research questions is presented in Chapter III.

CHAPTER III
METHODOLOGY

At the end of the previous chapter, two research questions were identified through a review of the professional literature relevant to methods used for the detection of differentially functioning test items. Those questions are, first, To what degree do the Mantel-Haenszel χ^2 approach, the IRT Unsigned Area approach, and the log-linear approach yield comparable indices with respect to the amount of DIF displayed by test items?, and second, How accurately does each DIF detection method identify test items with respect to the amount of DIF displayed by each item? The purpose of this chapter is to outline the methodology that was used to investigate these research questions.

This chapter is divided into six sections. The first section provides a brief description of the Graduate Management Admissions Test (GMAT), the data collection instrument from which the data for the empirical portion of this study was drawn. The second section presents a description of the data reduction procedures that were followed in preparing these empirical data for analysis. The third section provides a brief description of the data generation program, DGEN, which was used to generate the item response data for the simulation portion of this study. The fourth section presents a description of the procedures that were followed in generating the item response data used in the simulation portion of this study. The fifth section provides a discussion of the methodology used to investigate the first research question. And finally, the last section of this chapter provides a discussion of the methodology used to investigate the second research question.

Empirical Data Collection Instrument: The Graduate Management Admissions Test

The data for the empirical portion of this study was drawn from a retired form of the Graduate Management Admissions Test (GMAT). The GMAT, developed and administered by the Educational Testing Service (ETS), is a multiple choice standardized test "designed to help graduate schools assess the qualifications of applicants for advanced study in business and management." (Educational Testing Service, 1986, p. 9) Each form of the GMAT consists of eight separately timed sections which measure the examinee's verbal and mathematical skills and abilities. Two of the eight are non-operational sections containing trial items from two of the five areas described below. These items are needed for pretesting and equating purposes only and are not used in calculating the examinee's verbal or quantitative scores.

Three of the operational sections on each form of the GMAT contain items which measure the examinee's verbal skills and abilities using three types of questions: Reading Comprehension questions, Analysis of Situations questions, and Sentence Correction questions. The Reading Comprehension questions measure the examinee's ability to understand, analyze, and apply information and concepts presented in a written format. The Analysis of Situations questions measure the examinee's ability to analyze and evaluate the major aspects of business or management situations. And finally, the Sentence Correction questions measure two aspects of an examinee's language proficiency: correct expression and effective expression (Educational Testing Service, 1986). Each of the three verbal sections contains either 20 or 25 multiple choice items. For each item, the examinee is presented with five response options from which the examinee is to choose the most appropriate option. The three operational sections which make up the verbal subtest contain a total of 75 items which are used in calculating the examinee's verbal score.

Similarly, the three remaining operational sections on each form of the GMAT contain items which measure the examinee's mathematical or quantitative skills and abilities using two types of questions: Problem Solving questions and Data Sufficiency questions. The Problem Solving questions measure the examinee's ability to understand verbal descriptions of situations and to solve mathematical problems by applying arithmetic, elementary algebra, or commonly known concepts of geometry. The Data Sufficiency questions measure the examinee's ability to analyze a quantitative problem, to recognize which information is relevant, and to determine at what point there is sufficient information to solve the problem (Educational Testing Service, 1986). Each of the three quantitative sections contains either 20 or 25 multiple choice items from one of these two types of questions. Again, for each item, the examinee is presented with five response options from which the examinee is to choose the most appropriate option. The three operational sections which make up the quantitative subtest contain a total of 65 items. These items are used in calculating the examinee's quantitative score.

From the six operational sections, four subtest scores are calculated: a verbal number-right score, a verbal formula score, a quantitative number-right score and a quantitative formula score. The verbal and quantitative number-right scores are simply the sum of the number of items on each subtest that the examinee answered correctly. For the verbal and quantitative subtests, formula scores are calculated by taking the examinee's number-right score and subtracting from it one-quarter times the number of items the examinee answered incorrectly. This adjustment to the number-right score is a correction for guessing. To this number 0.5 is added and the result is rounded to the nearest whole number to yield the examinee's formula score for the subtest (Educational Testing Service, 1986).

Data from the June 20 and 22, 1987 administrations of the GMAT have been provided to the researcher by the ETS on a public access, computer-readable magnetic tape. The data file provided to the researcher by ETS contains the records for the 68,342 examinees who registered to take the GMAT in June 1987.

Empirical Data Reduction Procedures

For the empirical portion of this study a random sample of 5,000 male and 5,000 female examinees was drawn from among the examinees that took the GMAT in June 1987 and a reduced data file was compiled containing information for these examinees only. The empirical sample was drawn from among examinees whose gender group membership code was available; who identified their racial/ethnic group membership as white/non-Hispanic; who identified their country of citizenship as the United States; who identified their intended degree objective as a Master's in Business Administration (including both MBA and Master of Science in Industrial Management); and who had no missing or miscoded responses to any of the 65 quantitative items. Each record in the reduced data file included: a code identifying the examinee's gender group membership; a vector containing the examinee's responses to the 65 quantitative items; and the examinee's quantitative number-right score.

Simulation Data Generation Program: DGEN

The computer program DGEN, a FORTRAN V program for the generation of dichotomously scored item response data, was used for the simulation portion of this study. The original program was written in 1973 by

Dr. Ron Hambleton and Dr. R. J. Rovinelli, both from the University of Massachusetts, Amherst. The current version of the program was modified in 1992 by Dr. H. Jane Rogers at Teachers College, Columbia University[1].

Simulation Data Generation Procedures

For the simulation portion of this study a random sample of 5,000 male and 5,000 female examinees was generated using the DGEN program. A profile for each of the 10,000 simulated examinees, dichotomously scored item response data for a 65 item test was generated according to the three-parameter logistic IRT model. A data file was compiled and, for each simulated examinee, the data file included a code identifying the examinee's gender group membership and a vector containing the examinee's responses to the 65 simulated test items where a 1 indicated an item to which the examinee responded correctly and a 0 indicated an item to which the examinee responded incorrectly. In order to replicate as nearly as possible the methods used to analyze the empirical data, for each examinee a number-right score for the 65 item test was calculated. This number-right score was used as the matching criteria for the Mantel-Haenszel and log-linear analyses of the simulated data.

Specifications for generating the ability parameters for the simulated examinees, by gender group, were based on an analysis of the distribution of

[1] The DGEN program was provided to this researcher by Dr. Hariharan Swaminathan of the University of Massachusetts, Amherst.

ability estimates for each gender group in the empirical sample. Using the estimates of theta for each gender group (which were output by LOGIST as part of the empirical data analysis), ability parameters for each group of simulated examinees were randomly chosen from normal distributions with the same mean and standard deviation. Similarly, specifications for generating the item difficulty, discrimination, and lower asymptote parameters were based on an analysis of these same item parameter estimates for each group in the empirical sample. Using the results of these analyses to identify the range of a-, b-, and c-parameter values found in the empirical data, the item parameter values for the simulated examinees, by group, were selected from these values are specified within the DGEN program in order to create nine items displaying varying degrees of uniform DIF and nine items displaying varying degrees of nonuniform DIF.

The nine uniform DIF items were created by holding the a- and c-parameters constant for each group at 0.70 and 0.20, respectively, and varying both the value of the b-parameters and the difference between the b-parameters. The value of the b-parameters for these nine items ranged from -2.40 to 1.25, while the difference between the b-parameters for the two groups ranged from 0.21 to 0.90. The combinations of b-parameters and differences between b-parameters were chosen to create unsigned areas between the ICCs that ranged from 0.15 to 0.71.

The nine nonuniform DIF items were similarly created by holding the b- and c-parameters constant across the groups and varying the value of the a-parameters and the difference between the a-parameters. The value of the a-parameters for these nine items ranged from 0.35 to 0.90 for the females and from 0.55 to 1.40 for the males; the difference between the a-parameters for the two groups ranged from 0.20 to 0.50. The combinations of a-parameters and differences between a-parameters were chosen in order to create unsigned areas between the ICCs that ranged from 0.20 to 0.46. The b-parameters for these items ranged from -2.25 to 1.14 so that both ordinal and disordinal interactions between ability level and group membership would be represented. The c-parameters for these items were again held constant at 0.20 except in the case of four items where the c-parameter for the females was adjusted slightly downward (three to 0.18 and one to 0.15) in order to increase the unsigned area.

The remaining 47 items were non-DIF items for which the a-, b-, and c-parameters for each item were identical for the two groups. For these items the a-parameters ranged for 0.20 to 1.10 by increments of 0.30. Similarly, the b-parameters for these items ranged from -4.50 to 3.50 by increments of 1.00. Each of these b-parameter values was paired with each of these a-parameter values, creating 45 items that varied systematically by level of discrimination and level of difficulty. The two remaining non-DIF items both had b-parameter values of 0.00; the a-parameter for the first of these items was set at

0.50 and the other was set at 1.10. Like the 18 DIF items, the c-parameter value for all 47 non-DIF items was held constant at 0.20.

Methodology for Investigating the First Research Question

Analyses of both the empirical and the simulation data were used to investigate the first research question, To what degree do the Mantel-Haenszel χ^2 approach, the IRT Unsigned Area approach, and the log-linear approach yield comparable indices with respect to the amount of DIF displayed by test items? To answer this question using the empirical data drawn from the GMAT, the analysis of these data was divided into three parts, one corresponding to each DIF detection method. Similarly, to investigate this question using the simulation data, the analysis was also divided into three parts, one corresponding to each DIF detection method. However, due to certain constraints imposed by the use of simulated data, slightly different DIF indices were calculated using the simulated data for the log-linear approach than were calculated using the empirical data. Specifically, when using the empirical data, the DIF indices calculated according to the log-linear approach were all based on 10 score levels, but varied on the number of item response classifications used. When the simulation data was used, the DIF indices calculated according to the log-linear approach were all based on a single item response classification, but varied on the number of score levels used. These differences are discussed in greater detail at a later point in this section.

Regardless of the type of data used (empirical or simulation), the focus of all analyses was on, first, the Pearson Product-Moment correlation between each pair of DIF indices, second, the phi correlation coefficient calculated for each pair of DIF indices after items had been "flagged" as either displaying DIF or not, and third, comparison of the detection rates between the pairs of DIF detection methods. The Pearson product-moment correlation coefficients were used as measures of the degree to which the pairs of DIF indices corresponded to each other in terms of magnitude and direction. The phi correlation coefficient and the comparison of the detection rates between the pairs of DIF detection methods was used to determine the degree of agreement between each pair of DIF detection methods in terms of flagging items as displaying DIF.

Prior to conducting any of these DIF analyses, two preliminary analyses were necessary. As discussed in Chapter II, a fundamental assumption of all the DIF detection methodologies used here is that the data are unidimensional, that is, that a single ability underlies the examinees' performances on the set of test items. Although this assumption of unidimensionality can never be strictly met, it is necessary to demonstrate that a single factor dominates the performances of the examinees on the test. To demonstrate that a single dominant component or factor underlies the performances of the examinees on the 65 quantitative items of the GMAT, a principal axis factor analysis was performed. The SPSS and SAS FACTOR procedures were both used to perform two separate factor analyses; the SPSS FACTOR procedure was used to determine the proportion of the *total* variance explained by the first two factors while the SAS FACTOR procedure was used to determine the proportion of the *common* variance explained by the first two factors. Similarly, to demonstrate that a single dominant component or factor underlies the responses of the examinees on the 65 simulation items, a principal axis factor analysis was also performed on these data. Again, the SPSS and SAS FACTOR procedures were both used to perform two separate factor analyses; the SPSS FACTOR procedure was used to determine the proportion of the *total* variance explained by the first two factors while the SAS FACTOR procedure was used to determine the proportion of the *common* variance explained by the first two factors.

For the analysis of the GMAT data, the reduced data file discussed previously was used to create a scored data file to be used to calculate the α_{MH} statistic. For each examinee the scored data file included a vector that represented his or her scored responses to the 65 quantitative items, with a 1 indicating a correct response and a 0 indicating an item to which the examinee responded incorrectly or not at all. The scored response vector was followed by a code identifying the examinee's gender (1 for female and 2 for male) and his or her quantitative number-right score.

Using this scored data file and the SAS macro developed by Harnish (1991), the α_{MH} statistic for each item was calculated. As discussed in Chapter II, the α_{MH} statistic represents the estimate of the common odds ratio for the focal group over the reference group and ranges from a lower bound of 0 to an upper bound of ∞ with a value of 1 indicating equal odds of a correct response for the two groups. For example, a value of the α_{MH} statistic of 0.5 indicates that the reference group is twice as likely as the focal group to answer the item correctly, while a value of 2.0 indicates just the opposite. A separate data file was generated that included the item number and the value of the α_{MH} statistic.

For the second part of the analysis, the reduced data file discussed previously was used to create two additional scored data files to be used to calculate the unsigned area between the focal and reference group ICCs. The first of these new data files included a vector that represented for each female examinee her scored responses to the 65 quantitative items, with a 1 indicating a correct response to the item, a 0 indicating an incorrect response to the item, and a 2 indicating an item to which the examinee did not respond. Similarly, the second of these new data files included a vector that represents for each male examinee his scored responses to the 65 quantitative items, with, again, a 1 indicating a correct response to the item, a 0 indicating an incorrect response to the item, and a 2 indicating an item to which the examinee did not respond.

Each of these new data files was submitted separately to the LOGIST program and the item response function parameters for each item were calculated according to the three-parameter logistic IRT model. These item parameters, along with the item numbers to which they correspond, were output by LOGIST into a data file. These item parameters, generated for males and females separately, were used in conjunction with one another in a SAS program written by the researcher to calculate for each item the unsigned area between ICCs for females and males. These estimates of the unsigned area between the ICCs for each item were added to the data file which contained the α_{MH} statistics calculated previously. In addition, the ICCs for each item was plotted using a MicroSoft Excel spreadsheet program in order to visually depict the item response functions for each group.

For the third part of the analysis, the reduced data file was again used to create two additional scored data files. These data files were used to calculate the G^2 statistic associated with the log-linear models discussed in Chapter II. The first of these files included a vector that for each examinee that represented his or her scored responses to the 65 quantitative items, with a 1 indicating a correct response and a 0 indicating an item to which the examinee responded incorrectly or not at all. The scored response vector was followed by a code identifying the examinee's gender and his or her quantitative number-right score.

The second of these scored data files again included a vector that for each examinee represented his or her scored responses to the 65 quantitative items, with a 1 indicating a correct response, a 0 indicating an incorrect response, and a 2 indicating an item to which the examinee did not respond. The scored response vector was again followed by a code identifying the examinee's gender and his or her quantitative number-right score.

The log-linear analysis began with a univariate analysis of the quantitative number-right scores of all examinees, regardless of gender. Following Green, Crone, and Folk (1989), the results of the univariate analysis were used to divide the sample of examinees into 10 ability groups based on their quantitative number-right scores, with each ability group representing approximately ten percent of the sample. Using these ability groupings, for each of the data files described above, a three-way contingency table was formed using the SAS FREQ procedure. The contingency table provided gender x ability group x response (right, wrong, or omitted) frequencies which were used as the data for the log-linear analysis.

The log-linear portion of the empirical data analysis generated two G^2 statistics for each item: one based on the classification of examinee responses as either right or wrong and one based on the classification of examinee responses as either right, wrong, or omitted. In each case, the G^2 statistics for each item was calculated using the BMDP statistical software for log-linear analyses. For each item within each of the two response formats, two log-linear models were fit: an "expanded" model containing all three two-way interaction terms (gender x ability group (GA), gender x response (GR), and ability group x response (AR)), and a "reduced" model containing only two of the three two-way interaction terms: gender x ability group (GA) and ability group x response (AR). For each item, the G^2 statistics were calculated by subtracting the Likelihood Ratio χ^2 value for the reduced model from the Likelihood Ratio χ^2 value for the expanded model. This G^2 statistic was used as an index of the relative importance of the GR interaction term in the log-linear model. Larger values of G^2 indicating the greater relative importance of knowing the examinees' cross-classification with respect to gender and response (i.e., right or wrong) in explaining the observed distribution of frequencies in the three-way contingency table. Under the null hypothesis of no DIF, the variables gender and response should be independent after controlling for ability group membership. A large G^2 statistic would indicate a dependency between gender and response and, therefore, that the item is functioning differently for males and females. The G^2 statistics for each item within each of the two response formats were added to the data file which contained the α_{MH} statistics and the unsigned area estimates calculated previously.

Finally, for each item four new variables were created: Flag_{MH}, Flag_{EUA}, $\mathrm{Flag}_{G^2RW(10)}$, and $\mathrm{Flag}_{G^2RWO(10)}$. Each of these new variables took on a value of either 1 or 0 depending upon whether the item was "flagged" as either displaying DIF or not, respectively. For each of the three DIF detection methods a seperate flagging criteria was applied. For the Flag_{MH} variable, if the value of the α_{MH} was either greater than or equal to 1.33 or less than or equal to 0.75 the item was flagged as displaying DIF. These values of the α_{MH} log-odds ratio were chosen for flagging items because they represent the point at which one group is 25% more likely than the other to respond correctly to the item. For example, if the α_{MH} log-odds ratio equals 1.33, then the reference group is 25% more likely to respond correctly to the item than the focal group. Conversely, an α_{MH} log-odds ratio equal to 0.75 indicates just the opposite because 0.75 is simply the inverse of 1.33.

For the Flag_{UA} variable, if the value of the unsigned area estimate was greater than or equal to 0.40, again, the item was flagged as displaying DIF. Finally, the two remaining new variables, Flag_{G^2RW} and Flag_{G^2RWO}, were flagged as displaying DIF if the following three conditions were all met: 1) the p-value associated with the Likelihood Ratio χ^2 value for the reduced model was less than or equal to 0.05, indicating that the reduced model did not fit the data; 2) the p-value associated with the Likelihood Ratio χ^2 value for the expanded model was greater than or equal to 0.05, indicating that the expanded model fit the data; and 3) the p-value associated with the G^2 statistic was less than or equal to 0.05, indicating the importance of including the gender x response interaction term in the model.

As mentioned previously, the analysis of the simulation data was also divided into three parts, each corresponding to one of the DIF detection methods that are the focus of this dissertation study. For the first part of the analysis, the simulated data file discussed previously was used to calculate the α_{MH} statistic. Using this data file and the SAS macro developed by Harnish (1991), the α_{MH} statistic for each item was calculated. As in the empirical data analysis, the calculated values of the α_{MH} statistic were so that the α_{MH} statistic consistently represented an estimate of the common odds ratio for the higher scoring group over the lower scoring group for each item, with a value of 1 indicating equal odds of a correct response for each group and values larger than 1 indicating greater amounts of DIF being displayed by the items. A separate data file was generated that included each simulated item number and the value of the α_{MH} statistic associated with that item.

For the second part of the analysis, the simulated data file was used to create two additional data files used to calculate the unsigned area between the focal and reference group ICCs. The first of these data files included a vector that represented for each simulated female examinee her responses to the 65 items, with a 1 indicating a correct response to the item and a 0 indicating an incorrect response to the item. Similarly, the second of these files included a similar vector for each simulated male examinee.

Each of these data files was submitted separately to the LOGIST program and the item response function parameter estimates for each item were calculated according to the three-parameter logistic IRT model. These item parameter estimates, along with the item numbers to which they corresponded, were output by LOGIST into a data file. These item parameter estimates, generated for males and females separately, were used in conjunction with one another in the SAS program used in the empirical portion of this study to calculate for each item the unsigned area between the females' and males' ICCs. These estimates of the unsigned area for each item were added to the data file which contains the α_{MH} statistics calculated previously. In addition, the ICCs for each item were plotted using a MicroSoft Excel spreadsheet program in order to visually depict the item response functions for each group.

For the third part of the analysis, the simulated data file was used to calculate three separate G^2 statistics associated with the log-linear models discussed in Chapter II: one based on five ability groups; one based on ten ability groups; and one based on twenty ability groups. As with the analysis of the empirical data, each of the three log-linear analyses of the simulated data began with a univariate analysis of the number-right scores of all simulated examinees, regardless of gender. The results of the univariate analysis were used to divide the sample of simulated examinees into either five, ten, or twenty ability groups based on their number-right scores, with each ability group representing approximately twenty, ten, or five percent of the sample, respectively. Using these ability groupings, three three-way contingency tables were formed using the SAS FREQ procedure. Each contingency table provided gender x ability group x response (right or wrong) frequencies which were used in the log-linear analyses.

For each of the three ability groupings used in the log-linear portion of the analysis of the simulated data, a single G^2 statistic was calculated for each item based on the classification of examinee responses as either right or wrong. The G^2 statistics for each item were calculated using the BMDP statistical software for log-linear analyses. For each item an "expanded" log-linear model containing each of the three two-way interaction terms (gender x ability group (GA), gender x response (GR), and ability group x response (AR)) was fit to the data. In addition, a "reduced" log-linear model containing only the gender x ability group (GA) and ability group x response (AR) two-way interaction terms will also be fit to the data. For each item, the G^2 statistics were calculated by subtracting the Likelihood Ratio χ^2 value for the reduced model from the Likelihood Ratio χ^2 value for the expanded model. As before, these G^2 statistics were used as indices of the relative importance of the GR interaction term in the log-linear model. The G^2 statistic for each item was added to the data file which contained the α_{MH} statistics and the unsigned area estimates calculated previously.

Finally, for each item four new variables were also created: Flag_{MH}, Flag_{EUA}, $\text{Flag}_{G^2RW(05)}$, $\text{Flag}_{G^2RW(10)}$, and $\text{Flag}_{G^2RW(20)}$. Each of these new variables took on a value of either 1 or 0 depending upon whether the item was "flagged" as either displaying DIF or not, respectively. For each index associated with one of the three DIF detection methods a separate flagging criterion was applied. For the Flag_{MH} variable, if the value of the α_{MH} was either greater than or equal to 1.33 or less than or equal to 0.75 the item was flagged as displaying DIF. For the Flag_{EUA} variable, if the value of the estimated unsigned area was greater than or equal to 0.15, again, the item was flagged as displaying DIF. Finally, the three remaining new variables, $\text{Flag}_{G^2RW(05)}$, $\text{Flag}_{G^2RW(10)}$, and $\text{Flag}_{G^2RW(20)}$, were flagged as displaying DIF if each of the following three conditions were met: 1) the p-value associated with the Likelihood Ratio χ^2 value for the reduced model was less than or equal to 0.05, indicating that the reduced model did not fit the data; 2) the p-value associated with the Likelihood Ratio χ^2 value for the expanded model was greater than or equal to 0.05, indicating that the expanded model did fit the data; and 3) the p-value associated with the G^2 statistic was less than or equal to 0.05, indicating the importance of including the gender x response interaction term in the log-linear model.

The two data files containing the DIF indices and the flagging variables for each item were submitted separately to a SAS program and two correlation analyses were run on each set of indices. First, for each data file the Pearson product-moment correlation was calculated for each pair of DIF indices as a measure of the degree to which the pairs of indices corresponded to each other in terms of magnitude and direction. And second, for each data file the phi correlation coefficient was calculated as a measure of the degree of agreement between each pair of DIF detection methods in terms of flagging items as displaying DIF. The results of these correlation analyses form the basis for assessing the degree to which the Mantel-Haenszel χ^2 approach, the IRT Unsigned Area approach, and the log-linear approach yield comparable indices with respect to the amount of DIF displayed by each item.

Methodology for Investigating the Second Research Question

Only the simulation data were used to investigate the second research question, How accurately does each DIF detection method identify items with respect to the amount of DIF displayed by each item? The focus of this part of the analysis, again, was on the Pearson product-moment correlation coefficient, the phi correlation coefficient, and the DIF detection rate for each methodology. For these analyses, however, the item parameter values used in generating the simulation data were treated as population parameters and the actual unsigned area between the ICCs for the two groups was calculated using the SAS program used to calculate the estimated unsigned area for the previous analyses. The values for the actual unsigned area for each item were then added to the simulation data file which contained the α_{MH} statistics, the unsigned area estimates, the three G^2 statistics, and the five flagging variables calculated previously. One additional flagging variable, Flag $_{AUA}$, was also added to the simulation data file. The variable Flag $_{AUA}$ took on a value of 1 for the 18 DIF items included in the simulation data file and 0 for the remaining 47 non-DIF items.

For this part of the study, the Pearson product-moment correlation coefficient was calculated for each DIF index and the actual unsigned area as a measure of the degree to which each DIF index corresponded to the actual unsigned area between the ICCs in terms of magnitude and direction. The phi correlation coefficient and the detection rates for each of the DIF detection methods were also calculated as measures of the degree of agreement between each DIF detection method and the actual unsigned area in terms of flagging items as displaying DIF. The results of these correlation analyses form the basis for assessing how accurately the Mantel-Haenszel χ^2 approach, the IRT Unsigned Area approach, and the log-linear approach identify test items with respect to the amount of DIF displayed by each item.

CHAPTER IV
RESULTS

This chapter is divided into three sections. The first section presents the results of the preliminary factor analyses performed on both the empirical and the simulated data. The second section presents the results of the investigation of the first research question, To what degree do the Mantel-Haenszel χ^2 approach, the IRT Unsigned Area approach, and the log-linear approach yield comparable indices with respect to the amount of DIF displayed by test items? The final section presents the results of the investigation of the second research question, How accurately does each DIF detection method identify test items with respect to the amount of DIF displayed by each item?

Results of the Preliminary Factor Analyses

Prior to performing the primary analyses that form the basis of this study, two preliminary factor analyses were performed. As discussed in Chapter II, a fundamental assumption underlying all the DIF detection methodologies used here is that the data are unidimensional, that is, that a single ability underlies the examinees' responses to the set of test items. To demonstrate that a single dominant component or factor underlies the responses of the examinees to the 65 quantitative items of the GMAT and the

65 simulated items generated using the DGEN data generation program, two principal axis factor analyses were performed on each of the data sets using both the SPSS and SAS FACTOR procedures.

As indicated in Chapter III, the simulated items were generated such that a single dominant component or factor would underlie the responses of the examinees to the items. In order to confirm that such a factor did, in fact, underlie the examinees' responses to the simulated items, two principal axis factor analyses were performed. The SPSS procedure FACTOR was used to determine the proportion of the *total* variance explained by the first two factors. The SPSS principal axis factor analysis used squared multiple correlations as prior communality estimates and a two factor extraction criteria. The first factor had an eigenvalue of 7.04 and accounted for 10.8% of the total variance in examinees' responses. The second factor had an eigenvalue of 1.95 and accounted for an additional 3.0% of the total variance in examinees' responses.

The SAS procedure FACTOR was used to determine the proportion of the *common* variance explained by the first two factors. Again, using the squared multiple correlations as prior communality estimates, the first factor had an eigenvalue of 6.19 and accounted for 93.7% of the common variance in examinees' responses. The second factor had an eigenvalue of 1.06 and all remaining factors had eigenvalues of less than 1.00. Based on these results, it was concluded that a single dominant factor did, in fact, underlie the responses of the examinees to the 65 simulated items.

For the empirical data the SPSS procedure FACTOR was again used to determine the proportion of the *total* variance explained by the first two factors underlying the examinees' responses to the 65 quantitative items on the GMAT. The SPSS principal axis factor analysis of these data used squared multiple correlations as prior communality estimates and a two factor extraction criteria. The first factor had an eigenvalue of 5.29 and accounted for 8.1% of the *total* variance in examinees' responses. The second factor had an eigenvalue of 2.58 and accounted for an additional 4.0% of the total variance in examinees' responses. All the remaining factors had eigenvalues of 1.73 or less.

The SAS procedure FACTOR was used to determine the proportion of the *common* variance explained by the first two factors. Again, using the squared multiple correlations as prior communality estimates, the first factor had an eigenvalue of 4.43 and accounted for 67.2% of the common variance in examinees' responses. The second factor had an eigenvalue of 1.69 and accounted for an additional 25.7% of the common variance in examinees' responses. All the other factors had eigenvalues of less than 1.00. Based on these results, it was concluded that a single dominant factor also underlied the examinees' responses to the 65 quantitative items of the GMAT.

Perhaps the most convincing evidence, however, of the underlying unidimensionality of the empirical data comes from comparing the first eigenvalue derived from the empirical data with the first eigenvalue derived from the simulated data. As indicated above, the simulated data was specifically generated such that a single dominant factor underlied the examinees' responses to the items. The first eigenvalue derived from the simulated data indicated that the first factor accounted for 10.8% of the total variance in the data. By comparison, the first eigenvalue derived from the empirical data indicated that the first factor underlying the examinees' responses to the 65 quantitative items on the GMAT accounted for 8.1% of the total variance. The comparability of the percentage of the total variance accounted by the first factors underlying each of these data sets further supports the conclusion that the empirical data are unidimensional.

To visually represent the comparability of the underlying unidimensionality of the empirical and the simulated data, Figure 1 below presents an overlay of the scree plots from the two SPSS factor analyses. From these plots it can be seen that for the two data sets the first eigenvalues are appreciably larger than the second eigenvalues and all remaining eigenvalues are trivial.

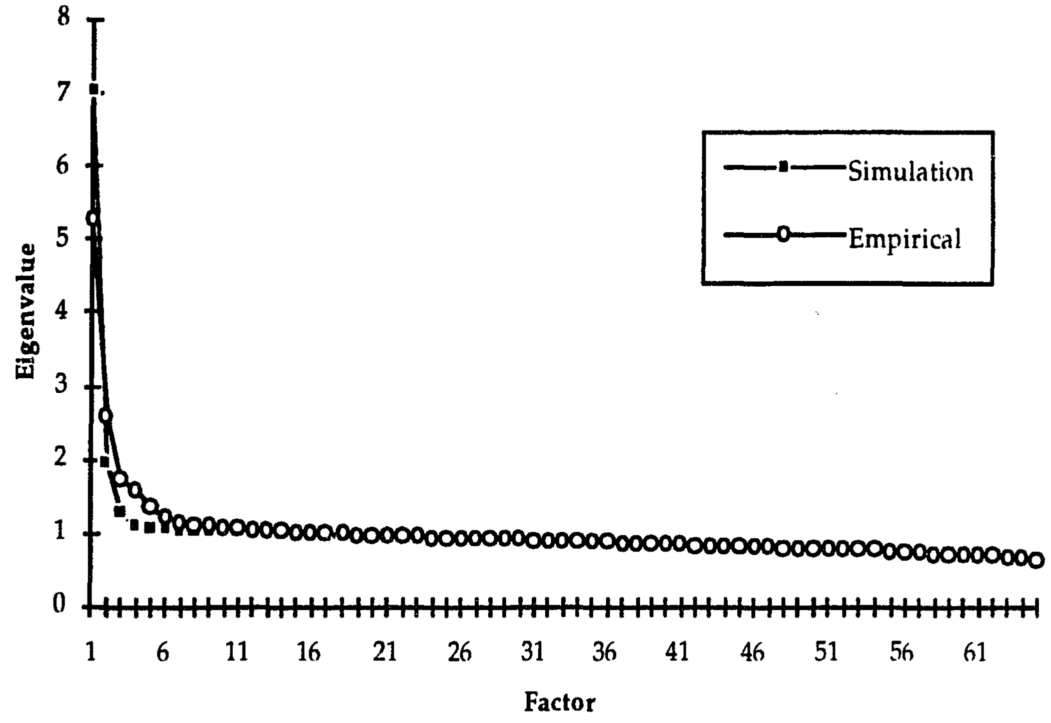

Figure 1. Scree Plots of Eigenvalues from the SPSS Principal Axis Factor Analysis

Results of Investigation of Research Question 1: Comparability of DIF Indices

The focus of the investigation of the first research question involved analyses of both the empirical and the simulated data to determine the comparability of the various DIF detection indices. The analyses of these data were focused in three areas. First, Pearson product-moment correlation coefficients between each pair of DIF indices for both the empirical and the simulated data were used to measure the degree to which each pair of DIF indices corresponded in terms of magnitude and direction. Second, phi coefficients were calculated for each pair of DIF indices after items had been "flagged" as either displaying DIF or not based on the flagging criteria for each DIF detection method described in Chapter II. The phi correlation coefficients between the pairs of DIF indices were used to determine the degree of agreement between each pair of DIF detection methods in terms of identifying items as displaying DIF. Finally, comparisons of the detection rates between the pairs of DIF detection methods were also performed. Like the phi correlation coefficients, the comparisons of the detection rates between the pairs of DIF detection methods were used as a measure of the degree of agreement between each pair of DIF detection methods in terms of identifying items as displaying DIF.

The results of the Pearson product-moment correlation analyses of the DIF indices derived from the empirical data are presented in Table 2 below. The results of this analysis indicate that all of the DIF indices are quite highly correlated with one another with the exception of the correlations between the Estimated Unsigned Area index and the two G^2 indices based on the log-linear model. Although both G^2 indices showed low correlations with the Estimated Unsigned Area index, the addition of a third level to the response classifications (i.e., omitted items) did result in a slight increase in the correlation between the $G^2_{RWO(10)}$ index with the Estimated Unsigned Area index. It should also be noted that the correlation between the $G^2_{RWO(10)}$ index and the α_{MH} index was slightly lower than the $G^2_{RW(10)}$ index and the α_{MH}. This is most likely because the α_{MH} index, like the $G^2_{RW(10)}$ is based on the dichotomous classification of examinees' responses as either right or wrong. (It should also be recalled that the G^2 indices derived from the empirical data varied on the number of response categories used to classify examinees' responses, and held constant the number of score intervals. The G^2 indices derived from the simulated data varied on the number of score intervals into which examinees were grouped, and held constant the number of response categories at two: right and wrong.)

Table 2.
Pearson Product-Moment Correlation Coefficients
Between DIF Indices — Empirical Data

	Estimated Unsigned Area	α_{MH}	$G^2_{RW(10)}$	$G^2_{RWO(10)}$
Estimated Unsigned Area	1.00 0.0*			
α_{MH}	0.80 0.0001	1.00 0.0		
$G^2_{RW(10)}$	0.28 0.0243	0.93 0.0001	1.00 0.0	
$G^2_{RWO(10)}$	0.34 0.0055	0.91 0.0001	0.98 0.0001	1.00 0.0

* p-value > $|R|$ under H_0: $\rho=0$, N=65

Table 3 below presents the results of the correlation analyses of the pairs of DIF indices derived from the empirical data after the items had been flagged as either showing DIF or not based on the flagging criteria for each DIF detection method described previously. Compared to the Pearson product-moment correlation coefficients presented in Table 2, the phi coefficients among all of the DIF indices are appreciably lower. This is not surprising, however, given the reduction in variance caused by the dichotomization of the index variables.

The phi coefficients among the DIF indices ranged from a low of 0.10 to a high of 0.76 compared to a range of 0.28 to 0.98 for the Pearson product-moment correlation coefficients presented in Table 2. Like the Pearson product-moment correlation coefficients, the highest of the phi correlations, 0.76 and 0.75, were between the two log-linear indices and between the $G^2_{RWO(10)}$ index and the α_{MH} index, respectively. Only slightly lower was the correlation between the α_{MH} index and $G^2_{RW(10)}$ index (r_ϕ=0.66). The lowest correlations were among the Estimated Unsigned Area index and the three other DIF indices. For each of the 65 empirical items, Table 4 presents the item parameter estimates for both females and males along with an "X" in the column representing the DIF indices which identified the item as displaying DIF. For greater visual clarity, the data have been sorted by the Estimated Unsigned Area index.

Table 3.
Phi coefficients
Between Flagged Items — Empirical Data

	Flag_{EUA}	Flag_{MH}	$\text{Flag}_{RW(10)}$	$\text{Flag}_{RWO(10)}$
Flag_{EUA}	1.00 0.0*			
Flag_{MH}	0.12 0.3291	1.00 0.0		
$\text{Flag}_{RW(10)}$	0.10 0.4267	0.66 0.0001	1.00 0.0	
$\text{Flag}_{RWO(10)}$	0.11 0.3865	0.75 0.0001	0.76 0.0001	1.00 0.0

* p-value > $|R|$ under H_0: $\rho=0$, N=65

The phi correlations presented in Table 3 can be better understood by looking at a matrix of the individual items and whether each DIF index identified them as displaying DIF. These data are presented in Table 4 below.

Table 4.
Matrix of Individual Item Parameters and Flagging by DIF Index — Empirical Data

	Females			Males			DIF Indices			
Item Number	a	b	c	a	b	c	Estimated Unsigned Area	αMH	GSQRW (10)	GSQRWO (10)
55	0.65	0.99	0.29	0.50	-0.52	0.15	X	X	X	X
59	0.67	0.66	0.23	0.59	-0.48	0.15	X	X	X	X
49	0.64	-1.10	0.15	0.53	-2.00	0.15	X	X	X	X
14	0.63	1.67	0.29	0.79	0.84	0.33	X	X	X	X
10	0.47	1.12	0.09	0.51	0.39	0.15	X	X	X	X
52	0.85	-0.16	0.15	0.90	-0.87	0.15	X	X	X	X
47	0.75	-1.51	0.15	0.69	-2.22	0.15	X	X	X	X
50	0.97	-0.57	0.15	1.08	-1.20	0.15	X	X	X	X
65	0.87	2.93	0.20	1.05	2.31	0.24	X	X	X	X
19	0.64	2.25	0.17	0.75	1.64	0.23	X	X	X	X
2	0.50	-1.67	0.15	0.57	-2.20	0.15	X	X	X	X
53	0.73	-0.52	0.15	0.83	-1.54	0.15	X	X		X
9	0.51	0.40	0.15	0.63	-0.43	0.15	X	X		X
63	0.62	2.07	0.22	0.80	1.23	0.25	X		X	X
57	0.77	1.20	0.15	0.74	0.51	0.13	X		X	X
33	0.66	1.52	0.16	0.52	0.89	0.11	X		X	X
16	0.70	2.28	0.12	0.76	1.45	0.12	X		X	
60	1.08	1.13	0.23	1.21	0.50	0.22	X			X
12	0.36	-0.53	0.15	0.86	0.44	0.49	X			
44	1.66	4.13	0.21	1.02	3.24	0.20	X			
28	0.78	0.57	0.26	0.64	-0.29	0.15	X			
18	0.66	2.55	0.19	0.84	1.82	0.21	X			
40	1.61	2.96	0.21	1.20	2.28	0.18	X			
35	0.36	1.51	0.15	0.52	0.91	0.15	X			
64	0.96	2.63	0.24	0.81	2.09	0.25	X			
13	0.86	0.46	0.15	0.87	-0.04	0.14	X			
6	0.89	-0.26	0.15	0.97	-0.75	0.15	X			
15	0.72	0.92	0.08	0.86	0.44	0.11	X			
39	0.79	0.63	0.06	0.79	0.21	0.06	X			
17	0.67	0.84	0.10	0.95	0.53	0.19	X			
23	0.55	-1.02	0.15	0.57	-0.80	0.15		X	X	X
41	0.82	1.09	0.16	1.06	0.88	0.17		X	X	X
54	0.56	0.19	0.15	0.64	0.02	0.15		X	X	X
21	0.41	-3.90	0.15	0.37	-4.02	0.15		X	X	X
38	0.67	0.20	0.15	0.74	0.09	0.11		X	X	X
30	0.39	0.08	0.15	0.40	0.17	0.15		X	X	X
26	0.49	-1.69	0.15	0.48	-1.74	0.15		X	X	X
11	0.60	-0.88	0.15	0.71	-0.92	0.15		X	X	X
24	0.70	-0.92	0.15	0.71	-0.95	0.15		X	X	X
4	0.66	-1.28	0.15	0.71	-1.27	0.15		X	X	X
46	0.59	-4.22	0.15	0.60	-4.10	0.15		X		X
56	0.43	0.11	0.15	0.54	-0.11	0.15			X	X
7	0.53	-1.16	0.15	0.55	-1.35	0.15			X	X
29	0.33	-1.25	0.15	0.28	-1.41	0.15			X	X

Table 4. (Continued)
Matrix of Individual Item Parameters and
Flagging by DIF Index — Empirical Data

	Females			Males			DIF Indices			
Item Number	**a**	**b**	**c**	**a**	**b**	**c**	**Estimated Unsigned Area**	**αMH**	**GSQRW (10)**	**GSQRWO (10)**
27	0.67	-0.40	0.15	0.78	-0.58	0.15			X	
37	0.55	1.01	0.09	0.66	0.67	0.12				X
8	0.36	-1.28	0.15	0.44	-1.32	0.15				X
1	0.30	-5.28	0.15	0.26	-6.27	0.15				
42	0.62	2.02	0.21	0.45	1.44	0.15				
22	0.70	-2.69	0.15	0.62	-3.21	0.15				
58	0.95	0.42	0.12	0.91	-0.06	0.11				
43	0.82	1.65	0.13	0.74	1.22	0.12				
62	0.86	2.20	0.20	0.85	1.78	0.19				
32	0.73	0.09	0.15	0.74	-0.30	0.15				
45	0.95	2.28	0.14	1.11	1.91	0.15				
20	0.74	1.77	0.07	0.64	1.46	0.07				
5	0.82	-1.42	0.15	0.85	-1.72	0.15				
61	0.91	1.54	0.28	0.78	1.27	0.30				
34	0.21	0.79	0.15	0.23	0.53	0.15				
36	0.96	0.75	0.25	0.98	0.51	0.30				
51	0.41	-2.07	0.15	0.43	-2.31	0.15				
31	0.27	-0.35	0.15	0.30	-0.54	0.15				
48	0.58	-2.76	0.15	0.58	-2.94	0.15				
25	0.56	-2.19	0.15	0.71	-2.06	0.15				
3	0.47	-2.63	0.15	0.50	-2.70	0.15				

The individual comparisons for the data from Table 4 are presented in Tables 5 through 10 below. For each combination of DIF indices, the tables indicate the total number of items that were identified as displaying DIF by both indices, by each index separately, and by neither index in the pair. In addition, the total number of items identified by each DIF detection method individually are indicated along with the percentage of the 65 items that the number represents.

Table 5.
DIF Detection Rates Using the Estimated Unsigned Area and the α_{MH} — Empirical Data Only

	No. of Items Flagged	Percent
Both	13	20.0
Estimated Unsigned Area Only	17	26.2
α_{MH} Only	11	16.9
Neither	24	36.9
Estimated Unsigned Area	30	46.2
α_{MH}	24	36.9

Table 6.
DIF Detection Rates Using the Estimated Unsigned Area and $G^2_{RW(10)}$ — Empirical Data Only

	No. of Items Flagged	Percent
Both	15	23.1
Estimated Unsigned Area Only	15	23.1
$G^2_{RW(10)}$ Only	14	21.5
Neither	21	32.5
Estimated Unsigned Area	30	46.2
$G^2_{RW(10)}$	29	44.6

Table 7.
DIF Detection Rates Using the Estimated Unsigned Area and $G^2_{RWO(10)}$ — Empirical Data Only

	No. of Items Flagged	Percent
Both	17	26.2
Estimated Unsigned Area Only	13	20.0
$G^2_{RWO(10)}$ Only	16	24.6
Neither	19	29.2
Estimated Unsigned Area	30	46.2
$G^2_{RWO(10)}$	33	50.8

Table 8.
DIF Detection Rates Using the α_{MH} and $G^2_{RW(10)}$ — Empirical Data Only

	No. of Items Flagged	Percent
Both	21	32.3
α_{MH} Only	3	4.6
$G^2_{RW(10)}$ Only	8	12.3
Neither	33	50.8
α_{MH}	24	36.9
$G^2_{RW(10)}$	29	44.6

Table 9.
DIF Detection Rates Using the α_{MH} and $G^2_{RWO(10)}$ — Empirical Data Only

	No. of Items Flagged	Percent
Both	24	36.9
α_{MH} Only	0	0.00
$G^2_{RWO(10)}$ Only	9	13.8
Neither	32	49.2
α_{MH}	24	36.9
$G^2_{RWO(10)}$	33	50.8

Table 10.
DIF Detection Rates Using $G^2_{RW(10)}$ and $G^2_{RWO(10)}$ — Empirical Data Only

	No. of Items Flagged	Percent
Both	27	41.5
$G^2_{RW(10)}$ Only	2	3.1
$G^2_{RWO(10)}$ Only	6	9.2
Neither	30	46.2
Estimated Unsigned Area	29	44.6
$G^2_{RWO(10)}$	33	50.8

The simulated data was also used in the investigation of the first research question. It should be noted that where the G^2 indices derived from the empirical data varied in the number of response categories used to classify examinees' responses, the G^2 indices derived from the simulated data varied in the number of score intervals used to group examinees. The results of the Pearson product-moment correlation analysis of the DIF indices calculated from the simulated data are presented in Table 11 below.

Table 11.
Pearson Product-Moment Correlation Coefficients
Between DIF Indices — Simulated data

	Estimated Unsigned Area	α_{MH}	$G^2_{RW(05)}$	$G^2_{RW(10)}$	$G^2_{RW(20)}$
Estimated Unsigned Area	1.00 0.0*				
α_{MH}	0.49 0.0001	1.00 0.0			
$G^2_{RW(05)}$	0.65 0.0001	0.75 0.0001	1.00 0.0		
$G^2_{RW(10)}$	0.65 0.0001	0.75 0.0001	0.99 0.0001	1.00 0.0	
$G^2_{RW(20)}$	0.58 0.0001	0.74 0.0001	0.94 0.0001	0.94 0.0001	1.00 0.0

* p-value > $|R|$ under H_0: $\rho=0$, N=65

As expected, the highest correlations were again between the log-linear indices. The G^2 index based on five score groups and the G^2 index based on ten score groups showed a nearly perfect correlation, while the G^2 index based on 20 score groups showed only a slightly lower correlation with each of these. The correlations between each of the three G^2 indices and both the α_{MH} index and the Estimated Unsigned Area index are all moderately high ranging from 0.58 to 0.75. The lowest correlation, 0.49, was between the α_{MH} index and the Estimated Unsigned Area index.

Table 12 below presents the results of the phi correlation analyses of the pairs of DIF indices derived from the simulated data after the items had been identified as either displaying DIF or not based on the flagging criteria for each DIF detection method described previously.

Table 12.
Phi coefficients
Between Flagged Items — Simulated data

	Flag$_{EUA}$	Flag$_{MH}$	Flag$_{RW(05)}$	Flag$_{RW(10)}$	Flag$_{RW(20)}$
Flag$_{EUA}$	1.00 0.0*				
Flag$_{MH}$	0.21 0.0913	1.00 0.0			
Flag$_{RW(05)}$	0.0014 0.9913	0.20 0.1112	1.00 0.0		
Flag$_{RW(10)}$	0.075 0.5531	0.49 0.0001	0.53 0.0001	1.00 0.0	
Flag$_{RW(20)}$	0.31 0.0115	0.53 0.0001	0.29 0.0183	0.37 0.0022	1.00 0.0

* p-value > $|R|$ under H_0: $\rho=0$, N=65

As with the empirical data, compared to the Pearson product-moment correlation coefficients presented in Table 11, the phi coefficients among all the DIF indices are, again, appreciably lower. As with the empirical data, these lower correlations can again be attributed, at least in part, to the reduction of variability in the data caused by the dichotomization of the index variables.

The phi coefficients among the DIF indices derived from the simulated data ranged from a low of 0.0014 to a high of 0.54 compared to a range of 0.36 to 0.99 for the Pearson product-moment correlation coefficients. Unlike the Pearson product-moment correlation coefficients, however, the two highest of the phi coefficients, both 0.53, were between the $G^2_{RW(20)}$ index and the α_{MH} index and between the $G^2_{RW(05)}$ index and the $G^2_{RW(10)}$ index. The phi correlation between the α_{MH} and $G^2_{RW(10)}$ was moderate (0.49) while the correlation between $G^2_{RW(20)}$ and the two other log-linear DIF indices, $G^2_{RW(05)}$ and $G^2_{RW(10)}$, were fairly low at 0.29 and 0.37, respectively.

As with the Pearson product-moment correlation coefficients, the lowest phi correlations were among the Estimated Unsigned Area index and all of the other indices. Of the four other indices, the largest calculated correlation coefficient was between the Estimated Unsigned Area and the $G^2_{RW(20)}$ index (0.31), with the correlations between the Estimated Unsigned Area and the α_{MH} index (0.21), the $G^2_{RW(10)}$ index (0.075), and the $G^2_{RW(05)}$ index (0.0014) all being lower.

The phi correlations presented in Table 3 can be better understood by looking at a matrix of the individual items and whether each DIF index identified them as displaying DIF. These data are presented in Table 13 below.

Table 13.
Matrix of Individual Item Parameters and
Flagging by DIF Index — Simulation Data

		Females						Males							DIF Indices				
Item Number	Type of DIF	a	Estim. a*	b	Estim. b*	c	Estim. c*	a	Estim. a*	b	Estim. b*	c	Estim. c*	Actual Unsigned Area	Estimated Unsigned Area Index	αMH	GSQRW (05)	GSQRW (10)	GSQRW (20)
1	Uniform	0.70	0.65	1.56	1.39	0.20	0.16	0.70	0.82	0.91	0.91	0.20	0.22	0.50	X	X	X	X	X
7	Uniform	0.70	0.76	-0.86	-0.85	0.20	0.18	0.70	0.78	-1.41	-1.36	0.20	0.17	0.42	X	X	X	X	X
8	Uniform	0.70	0.72	-0.27	-0.30	0.20	0.18	0.70	0.74	-0.87	-0.96	0.20	0.17	0.47	X	X	X	X	
6	Uniform	0.70	0.73	0.73	0.63	0.20	0.19	0.70	0.73	-0.17	-0.26	0.20	0.17	0.71	X	X		X	X
3	Uniform	0.70	1.01	1.42	1.27	0.20	0.22	0.70	0.89	0.67	0.63	0.20	0.22	0.58	X	X		X	X
10	Nonuniform	0.35	0.37	1.14	1.00	0.20	0.18	0.55	0.66	1.14	1.15	0.20	0.22	0.46	X	X		X	
5	Uniform	0.70	0.83	1.24	1.21	0.20	0.21	0.70	0.81	0.39	0.25	0.20	0.18	0.66	X	X			X
11	Nonuniform	0.60	0.63	-2.25	-2.14	0.20	0.18	1.00	0.96	-2.25	-2.27	0.20	0.17	0.34	X	X			
14	Nonuniform	0.50	0.59	0.72	0.64	0.18	0.17	0.80	0.94	0.72	0.68	0.20	0.21	0.37	X		X		X
34	No DIF	0.20	0.22	-1.50	-1.63	0.20	0.18	0.20	0.18	-1.50	-1.80	0.20	0.17	0.00	X		X		
49	No DIF	0.20	0.23	1.50	1.03	0.20	0.18	0.20	0.20	1.50	1.27	0.20	0.17	0.00	X			X	
4	Uniform	0.70	0.71	-0.83	-0.91	0.20	0.18	0.70	0.71	-1.12	-1.16	0.20	0.17	0.23	X				X
33	No DIF	1.40	1.30	-2.50	-2.70	0.20	0.18	1.40	1.46	-2.50	-2.53	0.20	0.17	0.00	X				
63	No DIF	1.40	2.00	3.50	4.23	0.20	0.20	1.40	1.46	3.50	3.36	0.20	0.20	0.00	X				
27	No DIF	1.10	0.92	-3.50	-4.11	0.20	0.18	1.10	1.12	-3.50	-3.67	0.20	0.17	0.00	X				
47	No DIF	1.10	1.27	0.50	0.38	0.20	0.17	1.10	1.36	0.50	0.57	0.20	0.22	0.00	X				
57	No DIF	1.10	1.25	2.50	2.73	0.20	0.22	1.10	2.00	2.50	2.19	0.20	0.21	0.00	X				
62	No DIF	1.10	1.60	3.50	3.06	0.20	0.20	1.10	0.94	3.50	4.34	0.20	0.20	0.00	X				
18	Nonuniform	0.90	0.95	0.52	0.47	0.18	0.17	1.40	1.65	0.52	0.48	0.20	0.18	0.25	X				
17	Nonuniform	0.85	0.94	-0.47	-0.48	0.20	0.18	1.10	1.16	-0.47	-0.61	0.20	0.14	0.17	X				
31	No DIF	0.80	0.78	-2.50	-2.54	0.20	0.18	0.80	0.95	-2.50	-2.29	0.20	0.17	0.00	X				
51	No DIF	0.80	0.86	1.50	1.44	0.20	0.17	0.80	1.00	1.50	1.31	0.20	0.19	0.00	X				
56	No DIF	0.80	0.74	2.50	2.41	0.20	0.18	0.80	0.98	2.50	2.29	0.20	0.20	0.00	X				
61	No DIF	0.80	0.55	3.50	4.09	0.20	0.20	0.80	1.14	3.50	2.83	0.20	0.20	0.00	X				
16	Nonuniform	0.75	0.75	-1.04	-1.09	0.20	0.18	0.95	1.07	-1.04	-1.01	0.20	0.17	0.17	X				
9	Uniform	0.70	0.83	0.84	0.69	0.20	0.18	0.70	0.66	0.09	-0.01	0.20	0.17	0.59	X				
15	Nonuniform	0.65	0.81	0.18	0.27	0.18	0.22	0.90	1.16	0.18	0.19	0.20	0.22	0.22	X				
13	Nonuniform	0.55	0.60	-0.88	-0.89	0.20	0.18	0.85	0.87	-0.88	-0.85	0.20	0.17	0.37	X				
50	No DIF	0.50	0.62	1.50	1.43	0.20	0.22	0.50	0.47	1.50	1.27	0.20	0.17	0.00	X				
55	No DIF	0.50	0.53	2.50	2.48	0.20	0.19	0.50	0.54	2.50	2.26	0.20	0.18	0.00	X				
60	No DIF	0.50	0.40	3.50	3.38	0.20	0.17	0.50	0.66	3.50	3.05	0.20	0.21	0.00	X				
12	Nonuniform	0.45	0.64	0.94	1.06	0.15	0.21	0.75	0.87	0.94	0.98	0.20	0.23	0.38	X				
29	No DIF	0.20	0.20	-2.50	-2.92	0.20	0.18	0.20	0.25	-2.50	-2.29	0.20	0.17	0.00	X				
39	No DIF	0.20	0.27	-0.50	-0.51	0.20	0.18	0.20	0.18	-0.50	-0.80	0.20	0.17	0.00	X				
44	No DIF	0.20	0.25	0.50	0.43	0.20	0.18	0.20	0.21	0.50	0.15	0.20	0.17	0.00	X				
59	No DIF	0.20	0.41	3.50	2.74	0.20	0.29	0.20	0.22	3.50	2.70	0.20	0.17	0.00	X				

Table 13.
Matrix of Individual Item Parameters and
Flagging by DIF Index — Simulation Data

		Females						Males							DIF Indices				
Item Number	Type of DIF	a	Estim. a*	b	Estim. b*	c	Estim. c*	a	Estim. a*	b	Estim. b*	c	Estim. c*	Actual Unsigned Area	Estimated Unsigned Area Index	αMH	GSQRW (05)	GSQRW (10)	GSQRW (20)
23	No DIF	1.40	0.83	-4.50	-6.21	0.20	0.18	1.40	1.33	-4.50	-4.43	0.20	0.17	0.00		X			
22	No DIF	1.10	0.64	-4.50	-6.90	0.20	0.18	1.10	0.89	-4.50	-5.88	0.20	0.17	0.00		X			
42	No DIF	1.10	1.31	-0.50	-0.48	0.20	0.21	1.10	1.27	-0.50	-0.49	0.20	0.17	0.00			X	X	
65	No DIF	1.10	1.35	0.00	0.00	0.20	0.22	1.10	1.26	0.00	0.03	0.20	0.21	0.00			X	X	
41	No DIF	0.80	0.92	-0.50	-0.52	0.20	0.18	0.80	0.92	-0.50	-0.50	0.20	0.17	0.00			X	X	
45	No DIF	0.50	0.51	0.50	0.35	0.20	0.18	0.50	0.52	0.50	0.39	0.20	0.17	0.00			X		
43	No DIF	1.40	1.54	-0.50	-0.59	0.20	0.13	1.40	1.55	-0.50	-0.54	0.20	0.16	0.00				X	
28	No DIF	1.40	1.39	-3.50	-3.80	0.20	0.18	1.40	1.07	-3.50	-4.17	0.20	0.17	0.00					
38	No DIF	1.40	1.47	-1.50	-1.50	0.20	0.18	1.40	1.49	-1.50	-1.52	0.20	0.17	0.00					
48	No DIF	1.40	1.57	0.50	0.48	0.20	0.19	1.40	1.62	0.50	0.48	0.20	0.20	0.00					
53	No DIF	1.40	1.91	1.50	1.40	0.20	0.21	1.40	1.70	1.50	1.41	0.20	0.20	0.00					
58	No DIF	1.40	2.00	2.50	2.18	0.20	0.20	1.40	1.78	2.50	2.25	0.20	0.20	0.00					
32	No DIF	1.10	1.12	-2.50	-2.54	0.20	0.18	1.10	1.00	-2.50	-2.68	0.20	0.17	0.00					
37	No DIF	1.10	1.36	-1.50	-1.35	0.20	0.18	1.10	1.27	-1.50	-1.44	0.20	0.17	0.00					
52	No DIF	1.10	1.27	1.50	1.37	0.20	0.19	1.10	1.47	1.50	1.34	0.20	0.21	0.00					
21	No DIF	0.80	0.58	-4.50	-5.80	0.20	0.18	0.80	0.77	-4.50	-4.60	0.20	0.17	0.00					
26	No DIF	0.80	0.81	-3.50	-3.59	0.20	0.18	0.80	0.78	-3.50	-3.54	0.20	0.17	0.00					
36	No DIF	0.80	0.82	-1.50	-1.51	0.20	0.18	0.80	0.83	-1.50	-1.52	0.20	0.17	0.00					
46	No DIF	0.80	0.82	0.50	0.44	0.20	0.18	0.80	0.84	0.50	0.44	0.20	0.19	0.00					
2	Uniform	0.70	0.69	-2.29	-2.39	0.20	0.18	0.70	0.70	-2.50	-2.59	0.20	0.17	0.15					
20	No DIF	0.50	0.52	-4.50	-4.42	0.20	0.18	0.50	0.48	-4.50	-4.70	0.20	0.17	0.00					
25	No DIF	0.50	0.48	-3.50	-3.59	0.20	0.18	0.50	0.49	-3.50	-3.74	0.20	0.17	0.00					
30	No DIF	0.50	0.54	-2.50	-2.40	0.20	0.18	0.50	0.55	-2.50	-2.45	0.20	0.17	0.00					
35	No DIF	0.50	0.56	-1.50	-1.37	0.20	0.18	0.50	0.57	-1.50	-1.35	0.20	0.17	0.00					
40	No DIF	0.50	0.51	-0.50	-0.47	0.20	0.18	0.50	0.54	-0.50	-0.57	0.20	0.17	0.00					
64	No DIF	0.50	0.51	0.00	-0.07	0.20	0.18	0.50	0.50	0.00	-0.13	0.20	0.17	0.00					
19	No DIF	0.20	0.21	-4.50	-4.26	0.20	0.18	0.20	0.21	-4.50	-4.54	0.20	0.17	0.00					
24	No DIF	0.20	0.21	-3.50	-3.64	0.20	0.18	0.20	0.21	-3.50	-3.34	0.20	0.17	0.00					
54	No DIF	0.20	0.24	2.50	2.05	0.20	0.18	0.20	0.23	2.50	2.05	0.20	0.17	0.00					

For each of the 65 simulated items, Table 13 presents the item parameter estimates for both females and males along with an "X" in the column representing the DIF indices which identified the item as displaying DIF. For greater visual clarity the data have been sorted by the Estimated Unsigned Area index.

The individual comparisons for these data are presented in Tables 14 through 23 below. For each combination of DIF indices the tables indicate the total number of items that were identified by both indices, by each one separately, and by neither one in the pair. In addition, the total number of items identified by each method individually are indicated along with the percentage of the 65 items that the number represents.

Table 14.
DIF Detection Rates Using the Estimated Unsigned Area and the α_{MH} — Simulated Data Only

	No. of Items Flagged	Percent
Both	8	12.3
Estimated Unsigned Area Only	28	43.1
α_{MH} Only	2	3.1
Neither	27	41.5
Estimated Unsigned Area	36	55.4
α_{MH}	10	15.4

Table 15.
DIF Detection Rates Using the Estimated Unsigned Area and $G^2_{RW(05)}$ — Simulated Data Only

	No. of Items Flagged	Percent
Both	5	7.7
Estimated Unsigned Area Only	31	47.7
$G^2_{RW(05)}$ Only	4	6.2
Neither	25	38.5
Estimated Unsigned Area	36	55.4
$G^2_{RW(05)}$	9	13.8

Table 16.
DIF Detection Rates Using the Estimated Unsigned Area and $G^2_{RW(10)}$ — Simulated Data Only

	No. of Items Flagged	Percent
Both	7	10.8
Estimated Unsigned Area Only	29	44.6
$G^2_{RW(10)}$ Only	4	6.2
Neither	25	38.5
Estimated Unsigned Area	36	55.4
$G^2_{RW(10)}$	11	16.9

Table 17.
DIF Detection Rates Using the Estimated Unsigned Area and $G^2_{RW(20)}$ — Simulated Data Only

	No. of Items Flagged	Percent
Both	7	10.8
Estimated Unsigned Area Only	29	44.6
$G^2_{RW(20)}$ Only	0	0.0
Neither	29	44.6
Estimated Unsigned Area	36	55.4
$G^2_{RW(20)}$	7	10.8

Table 18.
DIF Detection Rates Using the α_{MH} and $G^2_{RW(05)}$ — Simulated Data Only

	No. of Items Flagged	Percent
Both	3	4.6
α_{MH} Only	7	10.8
$G^2_{RW(05)}$ Only	6	9.2
Neither	49	75.4
α_{MH}	10	15.4
$G^2_{RW(05)}$	9	13.8

Table 19.
DIF Detection Rates Using the α_{MH} and $G^2_{RW(10)}$ — Simulated Data Only

	No. of Items Flagged	Percent
Both	6	9.2
α_{MH} Only	4	6.2
$G^2_{RW(10)}$ Only	5	7.7
Neither	50	78.5
α_{MH}	10	15.4
$G^2_{RW(10)}$	11	16.9

Table 20.
DIF Detection Rates Using the α_{MH} and $G^2_{RW(20)}$ — Simulated Data Only

	No. of Items Flagged	Percent
Both	5	7.7
α_{MH} Only	5	7.7
$G^2_{RW(20)}$ Only	2	3.1
Neither	53	81.5
α_{MH}	7	10.8
$G^2_{RW(20)}$	7	10.8

Table 21.
DIF Detection Rates Using $G^2_{RW(05)}$ and $G^2_{RW(10)}$ — Simulated Data Only

	No. of Items Flagged	Percent
Both	6	9.2
$G^2_{RW(05)}$ Only	3	4.6
$G^2_{RW(10)}$ Only	5	7.7
Neither	51	78.5
$G^2_{RW(05)}$	9	13.8
$G^2_{RW(10)}$	11	16.9

Table 22.
DIF Detection Rates Using $G^2_{RW(05)}$ and $G^2_{RW(20)}$ — Simulated Data Only

	No. of Items Flagged	Percent
Both	3	4.6
$G^2_{RW(05)}$ Only	6	9.2
$G^2_{RW(20)}$ Only	4	6.2
Neither	52	80.0
$G^2_{RW(05)}$	9	13.8
$G^2_{RW(20)}$	7	10.8

Table 23.
DIF Detection Rates Using $G^2_{RW(10)}$ and $G^2_{RW(20)}$ — Simulated data Only

	No. of Items Flagged	Percent
Both	4	6.2
$G^2_{RW(10)}$ Only	7	10.8
$G^2_{RW(20)}$ Only	3	4.6
Neither	51	78.5
$G^2_{RW(10)}$	11	16.9
$G^2_{RW(20)}$	7	10.8

Results of Investigation of Research Question 2: Accuracy of DIF Indices

The investigation of the second research question involved analysis of the simulated data only. As before, the analyses of these data also focused in three areas. First, Pearson product-moment correlation coefficients between each DIF index and the Actual Unsigned Area index were used as measures of the degree to which each DIF index corresponded to the Actual Unsigned Area in terms of magnitude and direction. Second, phi coefficients were calculated for each DIF index and the Actual Unsigned Area after items had been "flagged" as either displaying DIF or not based on the flagging criteria for each DIF detection method described previously. The phi coefficients between each DIF index and the Actual Unsigned Area index were used as measures of the accuracy with which each DIF detection method was able to correctly identify the simulated items which displayed DIF. Finally, comparisons of the detection rates of each DIF detection method were also performed. Like the phi coefficients, the comparisons of the detection rates of each DIF detection method were used as measures of the accuracy with which each DIF detection method was able to correctly identify the simulated items which displayed DIF.

Table 24 below presents the Pearson product-moment and phi coefficients between each DIF index and the Actual Unsigned Area.

Table 24.
Pearson and Phi coefficients Between
Each DIF Index and the Actual Unsigned Area

	Pearson			Phi
	Uniform	Non-Uniform	Overall	Overall
Estimated Unsigned Area	0.82 0.0071*	0.57 0.1099*	0.68 0.0001†	0.36 0.0037†
α_{MH}	0.56 0.1174	-0.01 0.9784	0.60 0.0001	0.50 0.0001
$G^2_{RW(05)}$	0.75 0.0192	0.43 0.2456	0.77 0.0001	0.15 0.2328
$G^2_{RW(10)}$	0.75 0.0194	0.44 0.2333	0.77 0.0001	0.27 0.0291
$G^2_{RW(20)}$	0.59 0.0927	0.22 0.5710	0.71 0.0001	0.56 0.0001

*p-value > | R | under H_0: $\rho=0$, N=9

†p-value > | R | under H_0: $\rho=0$, N=65

The Pearson product-moment correlation coefficients clearly indicate a moderately strong relationship between each of the DIF indices and the Actual Unsigned Area indicating that each of the DIF indices investigated are reasonably good measures of the difference between two item characteristic curves.

The phi correlations are much weaker in most cases. The largest correlations are between the $G^2{}_{RW(20)}$ index and the α_{MH} index and the Actual Unsigned Area index. These correlations indicate that of the five DIF indices investigated, these two DIF indices and their associated flagging criteria most accurately identify DIF items.

The individual comparisons of detection rates for each of the DIF detection indices are presented in Tables 25 through 29 below. For each comparison, the table presents the number of uniform and nonuniform DIF items that the DIF detection method was able to correctly identify. In addition, each table also presents the number of Type I errors made by the DIF detection method.

Table 25.
DIF Detection Rate and Type I Errors
Using the Estimated Unsigned Area

Type of DIF	No. of Items on Test	No. of Items Flagged	No. of Type I Errors
Uniform	9	8 (88.9%)	1 (5.6%)
Nonuniform	9	9 (100.0%)	0 (0.0%)
No DIF	47	19 (40.4%)	N/A

Table 26.
DIF Detection Rate and Type I Errors Using α_{MH}

Type of DIF	No. of Items on Test	No. of Items Flagged	No. of Type I Errors
Uniform	9	6 (66.7%)	1 (33.3%)
Nonuniform	9	2 (22.2%)	7 (77.8%)
No DIF	47	2 (40.4%)	N/A

Table 27.
DIF Detection Rate and Type I Errors Using $G^2_{RW(05)}$

Type of DIF	No. of Items on Test	No. of Items Flagged	No. of Type I Errors
Uniform	9	3 (33.3%)	6 (66.7%)
Nonuniform	9	1 (11.1%)	8 (88.9%)
No DIF	47	5 (10.6%)	N/A

Table 28.
DIF Detection Rate and Type I Errors Using $G^2_{RW(10)}$

Type of DIF	No. of Items on Test	No. of Items Flagged	No. of Type I Errors
Uniform	9	5 (55.6%)	4 (44.4%)
Nonuniform	9	1 (11.1%)	8 (88.9%)
No DIF	47	5 (10.6%)	N/A

Table 29.
DIF Detection Rate and Type I Errors Using $G^2_{RW(20)}$

Type of DIF	No. of Items on Test	No. of Items Flagged	No. of Type I Errors
Uniform	9	6 (66.7%)	3 (33.3%)
Nonuniform	9	1 (11.1%)	9 (100.0%)
No DIF	47	0 (0.0%)	N/A

The data from these tables indicate that in one sense the Estimated Unsigned Area index was the most accurate DIF detection method in that it correctly identified the greatest number of both the uniform and nonuniform DIF items. However, the Estimated Unsigned Area approach also had the largest Type II error rate of all of the DIF detection methods investigated with 19 of the 47 non-DIF items (40.4%) being incorrectly identified as displaying DIF. It is this high Type II error rate that is reducing the phi correlation between the Estimated Unsigned Area index and the Actual Unsigned Area noted above. The second highest Type II error rate was associated with both the $G^2_{RW(05)}$ and the $G2_{RW(10)}$ indices which each incorrectly identified 5 of the non-DIF items (10.6%) as displaying DIF.

Of the four remaining DIF indices, the $G^2_{RW(20)}$ and the α_{MH} indices yielded comparable results in terms of accurately identifying uniform DIF items, number of Type I errors, and number of Type II errors. The least accurate of the DIF indices in terms of correct identification of the uniform and nonuniform DIF items, number of Type I errors, and number of Type II errors were the $G^2_{RW(05)}$ and the $G^2_{RW(10)}$ indices. Both of these indices resulted in low detection rates, particularly with respect to the nonuniform DIF items, high Type I error rates, and moderately low Type II error rates. Each of these four methods failed to detect more than two of the nonuniform DIF items.

CHAPTER V
DISCUSSION

In the previous chapter the results of the empirical and simulation data analyses were presented. The purpose of this chapter is, first, to highlight and discuss the results presented in Chapter IV and, based on those findings, address the two research questions that were the focus of this study. Next, the implications of the results of this study for the detection of differential item functioning as a general measurement issue will be discussed. Finally, the implications of the study have for further research on the detection of differential item functioning will be presented.

Summary of Results of Investigation of Research Questions

The first research question focused on the comparability of the three DIF detection methods: the Mantel-Haenszel χ^2 approach, the IRT Unsigned Area approach, and the log-linear approach. To investigate this question, several DIF indices were derived using each of these approaches with both empirical and the simulated tests. In all, six separate analyses were performed using these DIF indices in order to address the first research question. Three of the analyses used the empirical data from the GMAT while the remaining three analyses used the simulated data. As discussed previously, for both "tests" the following DIF indices were derived: the α_{MH} index, the Estimated

Unsigned Area index, and the $G^2_{RW(10)}$ index. In addition, several other DIF indices based on the log-linear model were derived using only the simulated test. These indices were the $G^2_{RW(05)}$ index and the $G^2_{RW(20)}$ index.

For the first part of this analysis, Pearson product-moment correlation coefficients were calculated using the DIF indices derived, first, from the empirical data and, second, from the simulated data. The results of these analyses were presented in Tables 2 and 11 in Chapter IV. The Pearson product-moment correlation coefficients between pairs of DIF indices were used as one measure of the comparability of the indices. The results of these analyses indicated that the Mantel-Haenszel approach and IRT-based Estimated Unsigned Area approach yielded comparable DIF indices. The correlations between the DIF indices derived using these two approaches were r=0.80 (N=65, p<0.0001) for the empirical data and r=0.49 (N=65, p<0.0001) for the simulated data. As discussed previously, the reduction in the correlation between the two indices for the simulated test is attributable to the reduction in the variability in the data resulting from the 47 non-DIF items.

The Pearson product-moment correlation coefficients between all of the DIF indices associated with the log-linear approach ($G^2_{RW(10)}$ and $G^2_{RWO(10)}$ derived using the empirical data and $G^2_{RW(05)}$, $G^2_{RW(10)}$, and $G^2_{RW(20)}$ derived using the simulated data) and the Estimated Unsigned Area index associated with the IRT-based approach yielded mixed results. The correlations between these DIF indices derived using the empirical data were low, ranging from

0.28 to 0.34. Conversely, the correlations among the three DIF indices associated with the log-linear model derived using the simulated data were all moderately low and essentially identical, ranging from 0.58 to 0.65. The differences between the two sets of results are likely due to the somewhat artificial nature of the simulated data in that the 47 non-DIF items in the simulated test were all perfectly coincident. As a result, the G^2 indices which reflect the importance of the ability group x gender interaction term in the log-linear model are expected be very small. Similarly, the parameter estimates for the simulated items calculated by LOGIST and used to estimate the unsigned area between the ICCs should reflect only small random errors of estimation. The results from the empirical sample, therefore, reflect the more realistic relationship between the indices, namely, that the G^2 indices investigated are not comparable to the Estimated Unsigned Area index.

The Pearson product-moment correlation coefficients between all of the DIF indices associated with the log-linear approach ($G^2_{RW(10)}$ and $G^2_{RWO(10)}$ for the empirical data and $G^2_{RW(05)}$, $G^2_{RW(10)}$, and $G^2_{RW(20)}$ for the simulated data) and the DIF index associated with the Mantel-Haenszel approach yielded essentially identical results. The correlations between the DIF indices derived using the empirical test were high, ranging from 0.91 to 0.93. The correlations between these DIF indices derived using the simulated test were also moderately high and nearly identical, ranging from 0.74 to 0.75. These findings suggest that when two response classifications are used (i.e., right and wrong) and the number of score intervals increases toward the number of score values (i.e., N+1), the G^2 index based on the log-linear model and the α_{MH} index yield essentially identical results.

Finally, a comparison of the two G^2 indices derived from the empirical data showed a nearly perfect correlation of 0.98. The correlations between the G^2 indices derived using the simulated data were also quite high, ranging from 0.94 to 0.99. The high correlations among these pairs of DIF indices, taken together, highlight two important findings of the present study. First, the results from the analysis of the empirical data indicate that the addition of a third classification category (i.e., omitted) to the response variable did not significantly change the DIF indices derived from the empirical data. Second, the results from the analysis using the simulated data also indicate that an increase in the number of score levels into which examinees were grouped did not significantly change the relationship among the DIF indices. In general, the results of these analyses indicate that both G^2 indices are highly correlated to the α_{MH} index, but not with the Estimated Unsigned Area index. On the other hand, the α_{MH} index was highly correlated with both G^2 indices and also with the Estimated Unsigned Area index.

For the second part of this analysis, phi coefficients were calculated, again using the indices derived, first, from the empirical data and, second, from the simulated data. The results of these analyses were presented in Tables 3 and

12 in Chapter IV. Like the Pearson correlations just discussed, the phi coefficients between pairs of DIF indices were also used as a measure of the comparability of the indices, but this time in terms of flagging items as displaying DIF. The results of these analyses indicate that the Mantel-Haenszel approach and the IRT-based Estimated Unsigned Area approach do not yield similar results. The correlations between the DIF indices derived using these two approaches were low for both the empirical (r_ϕ=0.12) and the simulated test (r_ϕ=0.21).

The phi coefficients between all of the DIF indices associated with the log-linear approach ($G^2_{RW(10)}$ and $G^2_{RWO(10)}$ derived using the empirical data and $G^2_{RW(05)}$, $G^2_{RW(10)}$, and $G^2_{RW(20)}$ derived using the simulated data) and IRT-based Estimated Unsigned Area index also indicated a low correspondence between items flagged as showing DIF by each of these approaches. The correlations between the DIF indices derived using the empirical data were very low and quite similar to the correlation between the α_{MH} index and the Estimated Unsigned Area index derived from the empirical data. The correlations between the DIF indices derived using the simulated data were also quite low, ranging from a low of 0.0014 to a high of 0.31. These correlations are similar to the correlations found with the empirical data. These analyses indicate that neither the Mantel-Haenszel approach nor the log-linear approach yield comparable results in terms of flagging items as displaying DIF.

The phi coefficients between all of the DIF indices associated with the log-linear approach ($G^2_{RW(10)}$ and $G^2_{RWO(10)}$ derived using the empirical data and $G^2_{RW(05)}$, $G^2_{RW(10)}$, and $G^2_{RW(20)}$ derived using the simulated data) and the Mantel-Haenszel approach also yielded much higher correlations. The correlations between the DIF indices derived using the empirical data were between 0.66 and 0.75. These correlations are considerably higher than the correlation between the α_{MH} index and the Estimated Unsigned Area index or the G^2 indices and the Estimated Unsigned Area index. The correlations between the DIF indices derived using the simulated data were low, ranging from 0.20 to 0.53.

Finally, a comparison of the two G^2 indices derived from the empirical data showed a moderately high correlation of 0.75. The correlations between the G^2 indices derived using the simulated data were all low, ranging from 0.29 to 0.53. These results indicate that although the G^2 indices were highly correlated in their raw form, the application of the flagging criteria discussed previously resulted in some discrepancies in terms of the items that are flagged using each index.

The third part of the analyses involved several comparisons of the individual items and whether or not each was or was not flagged by the various approaches. Tables 4 and 13 presented matrix displays of the items and whether or not each item was flagged by the various approaches for both the empirical and the simulated data, respectively. In addition, for the empirical data Tables 5 through 10 presented pairwise comparisons of the aggregate number of items that were identified in common by the two approaches being compared, by each approach separately, and by neither of the approaches. Tables 14 through 23 presented the same comparisons for the simulated data. Through these analyses, some additional clarification of the results of the two correlation analyses can be gained.

Inspection of these tables, particularly Table 13, indicates that the generally low phi correlations between the Estimated Unsigned Area index and all of the other indices for both the empirical and the simulated data is attributable largely to the larger number of items that were flagged by the Estimated Unsigned Area index, but not by the other indices. Table 14 shows that the Estimated Unsigned Area index flagged 36 of the 65 simulated items. Of those 36, only 17 were DIF items; the remaining 19 were non-DIF items. Comparing the parameter estimates for the items that the Estimated Unsigned Area index flagged, and the ones that it did not, suggests that the Estimated Unsigned Area index is overly sensitive to differences in the b-parameters and, as a result, flags items which should not be flagged. This tendency for the Estimated Unsigned Area index to flag items with larger differences in the b-parameters is also reflected in the empirical data.

Based on the results of these analyses, the following conclusions can be drawn regarding the first research question, To what degree do the Mantel-Haenszel χ^2 approach, the IRT Unsigned Area approach, and the log-linear approach yield comparable indices with respect to the amount of DIF displayed by test items? First, the Estimated Unsigned Area approach and the Mantel-Haenszel approach yield moderately comparable indices with respect to the amount of DIF displayed by test items for both the empirical and the simulated data. Similarly, the Mantel-Haenszel approach and the log-linear approach also yield comparable indices with respect to the amount of DIF displayed by test items for both the empirical and the simulated data. The two approaches which did not yield comparable indices for both tests used in this study were the IRT Unsigned Area approach and the log-linear approach. The Pearson correlations derived from the empirical data were much lower that the correlations derived from the simulation data.

Second, the results of this study indicate that the three approaches were much less comparable after the flagging criteria had been applied to the indices. Based on the phi coefficients between the pairs of DIF indices, the indices associated with the log-linear model and the index associated with the Mantel-Haenszel approach were moderately comparable in terms of flagging items as displaying DIF. However, comparisons of the DIF indices associated with the IRT Unsigned Area approach and the DIF indices associated with the other two approaches were much less comparable, even to the point of being completely uncorrelated, as in the case of the Estimated Unsigned Area index and the $G^2_{RW(05)}$ index (r_ϕ=0.0014).

The second research question investigated in this dissertation study focused on the accuracy of the three DIF detection methods: the Mantel-Haenszel χ^2 approach, the IRT Unsigned Area approach, and the log-linear approach. To investigate this question, the DIF indices derived from the simulated test were used. In all, three separate analyses were performed in order to address the second research question.

The first part of the analysis used the Pearson product-moment correlation coefficients between each of the DIF indices derived from the simulated data and the Actual Unsigned Area between the ICCs as a measure of the accuracy with which each detection method identified the amount of DIF displayed by each item. The Pearson product-moment correlation coefficients were calculated for the uniform and the nonuniform DIF items separately and overall. The results of these analyses indicated that all of the approaches were moderately to highly correlated with the Actual Unsigned Area for the nine uniform DIF items, with correlation coefficients ranging from a low of 0.56 for the Mantel-Haenszel approach to a high of 0.82 for the IRT Unsigned Area approach. The results for the nine nonuniform DIF items varied much more widely, with correlation coefficients ranging from a low of -0.01 for the Mantel-Haenszel approach to a high of 0.57 for the IRT Unsigned Area approach. Overall, the correlation coefficients based on all 65 simulated items indicated that the three approaches yielded indices that were all moderately highly correlated with the Actual Unsigned Area, with correlation coefficients ranging from a low of 0.60 for the Mantel-Haenszel approach to a high of 0.77 for the $G^2_{RW(05)}$ and $G^2_{RW(10)}$ indices which are associated with the log-linear model.

The second part of the analysis used phi coefficients between each of the DIF indices derived from the simulated data and the Actual Unsigned Area between the ICCs, after the flagging criteria had been applied to the indices, as another measure of the accuracy with which each detection method identified whether or not the items displayed DIF. Unlike the Pearson product-moment correlation coefficients, only the overall phi correlation coefficients could be calculated for this part of the analysis. This was the case because if the items were disaggregated into uniform and nonuniform DIF items after the flagging criteria had been applied, no variability would exist for the Actual Unsigned Area index, and as a result, the phi coefficients could not be calculated. The results of the overall correlation analysis indicated that the approaches varied widely, with phi correlation coefficients ranging from a low of 0.15 for the $G^2_{RW(05)}$ index associated with the log-linear approach to a high of 0.56 for the $G^2_{RW(05)}$ index also associated with the log-linear approach. The important result to note is that the Estimated Unsigned Area index, which had the highest Pearson correlation coefficients with both the uniform and nonuniform DIF items, had a phi correlation of only 0.36 with the Actual Unsigned Area once the flagging criteria were applied. The reason for this disparity can be seen from the third part of the analysis.

The third part of the analysis involved individual comparisons of each approach's detection rate by type of DIF (i.e., uniform or nonuniform), Type I error rate, and Type II error rate. These individual comparisons were presented in Tables 25 through 29 in Chapter IV. The results of these comparisons clearly indicate that although the Estimated Unsigned Area index had the highest detection rate with 17 of the 18 DIF items being correctly identified, it also had the highest Type II error rate with 19 of the 47 non-DIF items being incorrectly identified. At the other extreme, the $G^2_{RW(20)}$ index correctly identified 6 of the 9 uniform DIF items, only 1 of the nonuniform DIF items, and all of the non-DIF items, that is, the $G^2_{RW(20)}$ index made no Type II errors. All of the other DIF indices had varying levels of success in identifying the nine uniform DIF items, relatively little success in identifying the nine nonuniform DIF items, and made comparably few Type II errors.

Based on the results of these analyses, the following conclusions can be drawn regarding the second research question, How accurately does each DIF detection method identify test items with respect to the amount of DIF displayed by each item? First, the results of the investigation of the second research question showed mixed results with respect to the accuracy with which the various DIF indices were able to correctly identify the 18 simulated DIF items. All of the indices were moderately successful in identifying the nine uniform DIF items, with the Estimated Unsigned Area index being the most accurate. Conversely, with the exception of the Estimated Unsigned

Area index, all of the DIF indices had difficulty in identifying the nine nonuniform items. Therefore, these results indicate that all of the approaches could be used to identify uniform DIF in items when it exists, but only the Estimated Unsigned Area index can be used to identify nonuniform DIF in items. Mitigating these results, however, is the large number of Type II errors made using the Estimated Unsigned Area index. The results of these analyses, when considered together, clearly indicate that the $G^2_{RW(20)}$ index was the most accurate of the DIF detection approaches.

Implications of Results of this Study for the Detection of Differential Item Functioning

The results of this investigation have several implications for the detection of differential item functioning in general. First, the results clearly indicate that the Estimate Unsigned Area index used in the analyses was extremely sensitive to difference in the b-parameter estimates. Further studies should be conducted in order to further explore this finding and to try to determine whether there exists a threshold difference in b-parameter estimates beyond which items were likely to be flagged as displaying DIF.

Second, although the use of log-linear models for the detection of DIF have been proposed by a number of researchers over the past 15 years, these models have not yet been as thoroughly investigated as other DIF detection procedures. The results of this study suggest that there is a need for further research into the appropriate application of these models to the detection of differential item functioning. One of the most critical issues around which further research is needed has to do with the minimum number of score levels needed when grouping examinees according on the ability of interest. Previous research has indicated that approximately five score levels are sufficient for matching examinees (Green, Crone, & Folk, 1989) The results of this study, however, suggest that for the 65 item tests used here, fewer than 20 score groups was not sufficient to reasonably approximate the results of other DIF detection methods. In general, the results of this study indicate that the finer the interval used in matching the examinees, given the amount of data available, the better.

Finally, although the evidence from this study is limited, it appears that the use of the third response level for classifying items to which an examinee did not respond, does not significantly enhance the power of the log-linear model to more accurately identify differentially functioning items. The usefulness of the omitted response category may have been masked by the small number of score groups used with the empirical data. Further studies are warranted to either support or refute the findings of this study with respect to the classification of omitted items.

Implications of Results of this Study for Further Research on the Detection of Differential Item Functioning

Based on the results of this study a number of avenues for further research in this area are indicated. The present study used a test of relatively short length (65 items) and a very large sample size (10,000 examinees) Greater understanding of the comparability and accuracy of the log-linear models for detecting differential item functioning could be gained through further investigation of these models when both the test length and sample size are varied.

As indicated previously, the results of the present study appear to indicate that the use of the third response level for classifying items to which an examinee did not respond did not significantly enhance the power of the log-linear model to more accurately identify differentially functioning items. Similar studies to determine whether this finding generalizes to response classifications other than the right, wrong, or omitted classifications used here would also be important. Some preliminary work has already been done in this area with respect to differential distractor functioning (e.g., Green, Crone, & Folk, 1989), but other possible applications should also be investigated, such as multiple-choice tests in which all of the response options represent correct responses, but vary in the level of sophisticated understanding of the material required by the examinees.

REFERENCES

Allen, N. L., & Donoghue, J. R. (1991, April). Applying the Mantel-Haenszel procedure to complex sample data. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic technique. Washington, DC: Author.

American Psychological Association. (1977). Guidelines for nonsexist language in APA journals. Washington, DC: Author.

Anastasi, A. (1988). Psychological testing, 6th edition. New York: Macmillan Publishing.

Angoff, W. H. (1982). The use of difficulty and discrimination indices in the identification of biased test items. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 96-116). Baltimore MD: Johns Hopkins University Press.

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.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-105.

Angoff, W. H., & Herring, C. L. (1976). Study of the appropriateness of the Law School Admission Test for Canadian and American students (Report No. LSAC-71-1). In Law School Admission Council, Reports of LSAC sponsored research: Volume II, 1970-1974. Princeton, NJ: Law School Admission Council.

Angoff, W. H., & Madu, C. C. (1973). Equating the scales of the Prueba de Aptitude Academica and the Scholastic Aptitude Test (Research Report No. 3). New York: College Entrance Examination Board.

Angoff, W. H., & Sharon, A. T. (1974). The evaluation of differences in test performance of two or more groups. Educational and Psychological Measurement, 34, 807-816.

Angoff, W. H., & Stern, J. (1971). The equating of the scales for the Canadian and American Scholastic Aptitude Tests (Project Report 71-24; CEEB RDR 71-72, No. 4). Princeton, NJ: Educational Testing Service.

Baghi, H., & Ferrara, S. F. (1989, March). A comparison of IRT, Delta Plot, and Mantel-Haenszel techniques for detecting differential item functioning across subpopulations in the Maryland Test of Citizenship Skills. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Baghi, H., & Ferrara, S. F. (1990, February). Detecting differential item functioning using IRT and Mantel-Haenszel techniques: Implementing procedures and comparing results. Paper presented at the Annual Meeting of the Eastern Educational Research Association, Clearwater, FL.

Baker, F. B. (1981). A criticism of Scheuneman's item bias technique. Journal of Educational Measurement, 18, 59-62.

Bejar, I. I. (1980). Biased assessment of program impact due to psychometric artifacts. Psychological Bulletin, 87, 513-524.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). Discrete multivariate analysis. Cambridge, MA: MIT Press.

Bond, L. (1993). Comments on the O'Neill and McPeek paper. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 277-280). Hillsdale, NJ: Lawrence Erlbaum Associates.

Breland, H. M., Stocking, M., Pinchak, B. M., & Abrams, N. (1974). The cross-cultural stability of mental test items: An investigation of response patterns for ten sociocultural groups (Project Report 74-2). Princeton, NJ: Educational Testing Service.

Camilli, G. A. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 397-413). Hillsdale, NJ: Lawrence Erlbaum Associates.

Camilli, G. A. & Shepard, L. A. (1987). The inadequacy of ANOVA for detecting test bias. Journal of Educational Statistics, 12, 87-99.

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

Camilli, G. A., & Smith, J. K. (1988, April). Comparison of the Mantel-Haenszel test with a randomized and a jackknife test for detecting biased items. Journal of Educational Statistics, 15(l), 53-67.

Cardall, C., & Coffman, W. E. (1964). A method for comparing the performance of different groups on the items in a test (Research Bulletin RB-64-61). Princeton, NJ: Educational Testing Service.

Clauser, B., Mazor, K. & Hambleton, R. K. (1994). Effects of score group width on DIF with the Mantel-Haenszel procedure. Journal of Educational Measurement, 31(1), 67-78.

Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. Educational and Psychological Measurement, 28, 61-75.

Coffman, W. E. (1961). Sex differences in responses to items in an aptitude test. Eighteenth Yearbook of the National Council on Measurement in Education (p. 117-124).

Coffman, W. E. (1963). Evidence of cultural factors in responses of African students to items in an American test of scholastic aptitude (Research and Development Reports). New York, NY: College Entrance Examination Board.

Cole, N. S. (1981). Bias in testing. American Psychologist, 36, 1067- 1077.

Cole, N. S. (1993). History and development of DIF. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 25-29). Hillsdale, NJ: Lawrence Erlbaum Associates.

Craig, R., & Ironson, G. H. (1981, April). The validity and power of selected item bias techniques using an a priori classification of items. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart, & Winston.

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.

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

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. Applied Measurement in Education, 2, 217-233.

Dorans, N. J., & Kulick, E. (1983a). Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach (Research Report No. 83-9). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kulick, E. (1983b). Assessing unexpected differential item performance of Oriental candidates on SAT Form CSA6 and TSWE Form E33: November 1980 Administration (Unpublished Statistical Report No. SR 83-106). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approaches to assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 23, 355-368.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. Journal of Educational Measurement, 29(4), 309-319.

Dorans, N. J., Schmitt, A. P., & Curley, W. E. (1988, April). Differential speededness: Some items have DIF because of where they are, not what they are. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

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

Dreger, R. M., & Miller, K. S. (1968). Comparative psychological studies of negroes and whites in the United States: 1959-1965. Psychological Bulletin (Monograph Supplement, 70(3), Part 2).

Durovic, J. J. (1975, October). Test bias: An objective definition for test items. Paper presented at the meeting of the Northeastern Educational Research Association, Ellenville, NY. (ERIC Document Reproduction Service No. ED 128 381).

Educational Testing Service. (1986). The official guide for GMAT review. Princeton, NJ: Author.

Engelhard, G. Jr., et. al (1990). An empirical comparison of Mantel-Haenszel and Rasch procedures for studying differential item functioning on teacher certification tests. Journal of Research and Development in Education, 23(3), 172-179.

Engelhard, G., Jr., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. Applied Measurement in Education, 3(4), 347-360.

Fienberg, S. E. (1991). The analysis of cross-classified categorical data. Cambridge, MA: The MIT Press.

Green, D. R. (1971). Racial and ethnic bias in test construction. Monterey, CA: CTB/McGraw-Hill.

Green, B. F. (1991, November). Differential item functioning: Techniques, findings, and prospects. Paper presented at the conference, "Modern theories for measurement: Issues and practices." University of Ottawa, Ottawa, Canada.

Green, D. R., Coffman, W. E., Lemke, J. M., Raju, N. S., Hendrick, F. A., Loyd, B. H., Carlton, S. T., & Marco, G. L. (1982). Methods used by test publishers to debias standardized tests. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD: The Johns Hopkins University Press.

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. Journal of Educational Measurement, 26(2), 147-160.

Green, D. R., & Draper, J. F. (1972, September). Exploratory studies of bias in achievement tests. Paper presented at the annual meeting of the American Psychological Association, Honolulu. (ERIC Document Reproduction Service No. ED 070 794)

Hambleton, R. K. (1980, April). Review methods for criterion-referenced test items. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Hambleton, R. K., Clauser, B. E., Mazor, K. M., & Jones, R. W. (1993). Advances in the detection of differentially functioning test items (Laboratory of Psychometric and Evaluative Research Report No. 237). Amherst, MA: University of Massachusetts, School of Education.

Hambleton, R. K., & Rogers, H. J. (1991). Evaluation of the plot method for identifying potentially biased test items. In P. L. Dann, S. H. Irvine, & J. M. Collins (Eds.), Advances in computer-based human assessment (pp. 307-330). Boston, MA: Kluwer Academic Publishers.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT and Mantel-Haenszel methods. Applied Measurement in Education, 2(4), 313-334.

Hambleton, R. K., Rogers, H. J., & Arrasmith, D. (1986, August). Identifying potentially biased test items: A comparison of the Mantel-Haenszel statistic and several item response theory methods. Paper presented at the meeting of the American Psychological Association, Washington, DC.

Hambleton, R. K., & Jones, R. W. (1992, April). Comparison of empirical and judgmental methods for detecting differential item functioning. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

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

Harnisch, D. L. (1991, February). Techniques for assessing differential item performance on achievement tests. Paper presented at Sixteenth Annual SAS Users Group International Conference, New Orleans, LA.

Harris, D. J., & Kolen, M. J. (1989). Examining the stability of Angoff's delta item bias statistic using the bootstrap. Educational and Psychological Measurement, 49(1), 81-87.

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

Hunter, J. E. (1975, December). A critical analysis of the use of item means and item test correlations to determine the presence or absence of content bias in achievement test items. Paper presented at the National Institute of Education conference on test bias, MD.

Hunter, R. V., & Slaughter, C. D. (1980). ETS test sensitivity review process. Princeton, NJ: Educational Testing Service.

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 117-160). Baltimore, MD: Johns Hopkins University Press.

Ironson, G. H. (1983). Using item response theory to measure bias. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 155-174). Vancouver, BC: Educational Research Institute of British Columbia.

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. Journal of Educational Measurement, 18, 209-225.

Jensen, A. R. (1976). Test bias and construct validity. Phi Delta Kappan, 58, 340-346.

Jensen, A. R. (1980). Bias in mental testing. New York: Free Press.

Jensen, M., & Beck, M. D. (1978). Gender balance analysis of the Metropolitan Achievement Tests. Measurement and Evaluation in Guidance, 12, 25-34.

Kelderman, H., & Macready, G. B. (1990). The use of log linear models for assessing differential item functioning across manifest and latent examinee groups. Journal of Educational Measurement, 27(4), 307-327.

Knoke, D., & Burke, P. J. (1980). Log-linear models. Beverly Hills, CA: Sage Publications.

Kok, F. G., Mellenbergh, G. J., & van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. Journal of Educational Measurement, 22(4), 295-303.

Kulick, E., & Dorans, N. J. (1984, April). The standardization approach to assessing unexpected differential item performance. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Linden, K. W., & Linden, J. D. (1968). Modern mental measurement: A historical perspective. Boston, MA: Houghton Mifflin Company.

Lord, F. M. (1952). A theory of test scores. Psychometric Monograph, No. 7.

Lord, F. M. (1977). A study of bias using item characteristic curve theory. In N. H. Poortinga (Ed.), Basic problems in cross-cultural psychology (pp. 19-29). Amsterdam: Swits & Vitlinger.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Macmillan (1975). Guidelines for creating positive sexual and racial images in educational materials. New York: Author.

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.

Marascuilo, L. A., & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on chi-square statistics. Journal of Educational Measurement, 18, 229-248.

Mazor, K., Clauser, B., & Hambleton, R. K. (in press). The effect of sample size on the functioning of the Mantel-Haenszel statistic. Educational and Psychological Measurement.

Mazor, K. M., et al. (1992, April). Identification of non-uniform differential item functioning using a variation of the Mantel-Haenszel procedure. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

McGraw-Hill (1968). Recommended guidelines for multiethnic publishing in McGraw-Hill Book Company publications. New York: Author.

McGraw-Hill (1974). Recommended guidelines for equal treatment of the sexes in McGraw-Hill Book Company publications. New York: Author.

McLarty, J. R., Noble, A. C., & Huntley, R. M. (1989). Effects of item wording on sex bias. *Journal of Educational Measurement*, *26*, 285-293.

McPeek, W. M., & Wild, C. L. (1987, August). *Characteristics of quantitative items that function differently for men and women*. Paper presented at the meeting of the American Psychological Association, New York.

Medley, D. M., & Quirk, T. J. (1974). The application of a factorial design to the study of cultural bias in general culture items on the National Teacher Examination. *Journal of Educational Measurement*, *11*, 235-245.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*, 105-118.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum Associates.

O'Neill, K. A., McPeek, W. M., & Wild, C. L. (1989, March). *Characteristics of GRE verbal test items that show differential item functioning for Black and White examinees*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Oosterhof, M., Atash, M. N., & Lassiter, K. L. (1984). Facilitating identification of item bias through delta plots. *Educational and Psychological Measurement*, *44(3)*, 619-627.

Pearlman, M. A. (1987, April). *Trends in women's total score and item performance on verbal measures*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Plake, B. S., & Hoover, H. D. (1979). An analytical method of identifying biased items. *Journal of Experimental Education*, *48*, 153-154.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *33*, 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14(2)*, 197-207.

Raju, N. S., Bode, R. K., & Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. Applied Measurement in Education, 2, 1-13.

Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 367-388). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rengel, E. (1986, August). Agreement between statistical and judgmental item bias. Paper presented at the meeting of the American Psychological Association, Washington, DC.

Rivera, C., & Schmitt, A. P. (1988). A comparison of Hispanic and White students' omit patterns on the Scholastic Aptitude Test (Research Report No. 88-44). Princeton, NJ: Educational Testing Service.

Roid, G. H., & Wendler, C. L. W. (1983, April). Item bias detection and item writing technology. Paper presented at the meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Rudner, L. M. (1977). An approach to biased item identification using latent trait measurement theory. Paper presented at the meeting of the American Educational Research Association, New York.

Rudner, L. M., & Convey, J. J. (1978, March). A evaluation of selected approaches for biased item identification. Paper presented at the annual meeting of the American Educational Research Association, Toronto.

Rudner, L. M., Getson, P. R. & Knight, D. L. (1980a). Biased item detection techniques. Journal of Educational Statistics, 5, 213-233.

Rudner, L. M., Getson, P. R. & Knight, D. L. (l980b). A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 17, 1-10.

Ryan, K. E. (1990, April). The performance of the Mantel-Haenszel procedure. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA.

Ryan, K. E. (1991). The performance of the Mantel-Haenszel procedure across samples and matching criteria. Journal of Educational Measurement, 28(4), 325-337.

Saario, T. N., Jacklin, C. N., & Tittle, C. K. (1973). Sex role stereotyping in the public schools. Harvard Educational Review, 40, 386-416.

Scheuneman, J. D. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Scheuneman, J. D. (1981). A response to Baker's criticism. Journal of Educational Measurement, 16, 143-152.

Scheuneman, J. D. (1986, April). Differential item performance: Use of computer simulation to evaluate indices. In W. H. Angoff (Chair), Differential item performance: Methodological and measurement issues. Symposium conducted at the meeting of the American Educational Research Association, San Francisco.

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. Applied Measurement in Education, 2 255-275.

Schmeiser, C. B. (1978, March). The impact of black and white English test content on black and white student performance and selected test characteristics. Paper presented at the annual meeting of the American Educational Research Association, Toronto.

Schmitt, A. P. (1988). Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the Scholastic Aptitude Test. Journal of Educational Measurement, 25, 1-13.

Schmitt, A. P., & Bleistein, C. A. (1987). Factors affecting differential item functioning for Black examinees on Scholastic Aptitude Test analogy items (Research Report No. 87-23). Princeton, NJ: Educational Testing Service.

Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. Journal of Educational Measurement, 27, 67-81.

Schmitt, A. P., Dorans, N. J., Crone, C. R., & Maneckshana, B. T. (1991). Differential speededness and item omit patterns on the SAT (Research Report No. 91-50). Princeton, NJ: Educational Testing Service.

Science Research Associates. (1976). Fairness in educational materials: Exploring the issues. Chicago, IL: Author.

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential test functioning. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 197-240). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.

Shepard, L. Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

Shuey, A. (1966). The test of Negro intelligence (2nd edition). New York: Social Science Press.

Stricker, L. J. (1982). Identifying test items that perform differently in population subgroups: A partial correlation index. Applied Psychological Measurement, 6, 261-273.

Stricker, L. J. (1984). The stability of a partial correlation index for identifying items that perform differentially in subgroups. Educational and Psychological Measurement, 44(4), 831-837.

Subkoviak, M. J., Mack, J. S., & Ironson, G. H. (1981, April). Item bias detection procedures: Empirical validation. Paper presented at the American Educational Research Association, Los Angeles.

Subkoviak, M. J., Mack, J. S., Ironson, G. H., & Craig, R. D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. Journal of Educational Measurement, 21, 49-58.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27(4), 361-370.

Thissen, D. Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 147-167). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Wainer, H., & Steinberg, L. (1985, October). Studying differential item performance with item response theory. Paper presented at the Military Testing Association meeting, San Diego, CA.

Thurstone, L. L. (1925). A method of scaling educational and psychological tests. Journal of Educational Psychology, 16, 263-278.

Thurstone, L. L. (1947). The calibration of test items. American Psychologist, 2, 103-104.

Tittle, C. K. (1982). Use of judgmental methods in item bias studies. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 31-63). Baltimore, MD: The Johns Hopkins University Press.

Wilson-Burt, C., Fitzmartin, R. D., & Skaggs, G. (1986, April). Baseline strategies in evaluating IRT item bias indices. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Wright, D. J. (1987). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. In A. P. Schmitt & N. J. Dorans (Eds.), Differential item functioning on the Scholastic Aptitude Test (Research Memorandum No. 87-1). Princeton, NJ: Educational Testing Service.