A THEORETICAL AND EMPIRICAL INVESTIGATION OF FACTOR ANALYTICALLY-BASED MATCHING CRITERIA IN DIFFERENTIAL ITEM FUNCTIONING

by

Robert Lewis Johnson

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

Lloyd Bond

Dissertation Advisor

UMI Number: 9544120

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

Grace E. Kissling

May 11, 1995
Date of Acceptance by Committee

May 11, 1995
Date of Final Oral Examination

JOHNSON, ROBERT LEWIS, Ph.D. A Theoretical and Empirical Investigation of Factor Analytically-Based Matching Criteria in Differential Item Functioning. (1995) Directed by Dr. Lloyd Bond. 145 pp.

Modern investigative procedures to detect differential item functioning (DIF) match examinee groups on ability before comparison. The validity of DIF procedures depends, in part, on the unidimensionality of the matching criterion; however, the most popular proxy for ability, examinees' raw scores on the test, consists of items with varying levels of multidimensionality. This study explored the efficacy of two matching criteria--total score and factor score--as tests become increasingly multidimensional. The investigation used empirical data to determine the consistency of flagging items as displaying DIF when matching with total and factor scores in tests that varied in factorial complexity. In addition, in a series of simulations, increasingly complex factor structures were created. In one variation referred to as Factor Structure 1, items loaded either on a first factor (the target factor) or a second factor (the nuisance factor)--but not both. Factor Structure 2 was composed of items that loaded primarily on a first factor and secondarily on a second (or nuisance) factor. Bias was simulated in items associated with the nuisance factor.

The analysis of empirical data revealed less consistency in the flagging designations for the total-score and factor-score matching criteria as the test became more factorially complex. The simulations revealed that the total score matching criterion performed adequately when a test was relatively unidimensional. As a test departed from unidimensionality, the total-score matching criterion was associated with more spuriously flagged items for Factor Structure 1 and Factor Structure 2. In a preliminary study, the rotated, factor-score matching criterion was associated with fewer spuriously flagged items for Factor Structure 1; however, the rotated factor score was associated with more spuriously flagged items for Factor Structure 2. In subsequent analyses, items displaying DIF were removed to create an unrotated, "purified" factor score and a factor-based score. The purified matching criteria correctly identified items with DIF. To simulate the test development process, items associated with the second factor in an empirical data set were removed from the factor-score and factor-based matching criteria. The resulting matching criteria increased the consistency with which items were flagged.

ACKNOWLEDGMENTS

A dissertation is attributed to one author; however, many people contribute to the final creation. I want to acknowledge the contributions of my dissertation committee members. In addition to his guidance in the completion of my dissertation, Lloyd Bond devoted great amounts of patience and time to the development of my understanding of educational measurement. Rita O'Sullivan assimilated me into the education research community through her teaching and her inclusion of me in evaluations. Grace Kissling provided me with a wide range of experience during my participation in statistical consulting. Jim Penny, who made the transition from fellow graduate student to dissertation committee member, engaged me in collegial discussions about differential item functioning and developed simulations that greatly affected my work.

My fellow graduate students in the doctoral program provided support--emotionally and intellectually. Discussions with Barbara Gorney and Ann Harman sharpened my understanding of measurement issues. My thanks to Ann who tirelessly shared her programming abilities. And thanks to Barbara with whom I worked so closely to complete a major research project that we were often referred to as one entity--BarbaraandRobert.

Within the Center of Educational Research and Evaluation the support I received from Wanda Baker, Michelle Parsons, and Marnie Thompson made difficult days possible--and often brought laughter. Anita Hawkins, the center of the Department of Educational Research and Measurement, steered

me through the complexities of the university system; may she continue to side with her graduate students.

Within the measurement community support was extended from Richard Jaeger, Professor at the University of North Carolina at Greensboro; Barbara Plake, Professor at the University of Nebraska at Lincoln; Steve Sireci, Senior Psychometrician at the American Council on Education; and Edward Haertel, Professor at Stanford University.

And returning to home, thanks to Mama (otherwise known as Mildred) who always believed and to Buddy (otherwise known as my other half) who provided the daily support necessary to live while remaining focused on the research at hand.

TABLE OF CONTENTS

Page

CHAPTER I

INTRODUCTION

Test critics point to the differences in the distributions of test scores for ethnic and gender groups as an indicator of test bias. It is possible that the differences in the distributions for African American and White examinees, or male and female examinees, are due to test bias; it is also possible the differences are due to differing ability or achievement levels (whatever their source) for the groups. The challenge for the testing community is to discover strategies that will improve the identification of biased items.

In bias studies the performance of minority groups (referred to as focal groups) is compared to members of a majority group (referred to as the referent group). Historically, the term *bias* was used to describe test items that were judged to disfavor one group over another. In lieu of the term *bias,* current practitioners in the measurement field use the more neutral term, differential item functioning (DIF). An item displays differential functioning if, when controlling for the ability being measured, a member of a group has a greater chance of answering an item correctly when compared to a member in another group of similar ability. Whether an item that displays DIF is biased is left to the judgment of test reviewers. In other words, DIF provides information about the level of the differential performance of the groups. It is left to the judgment of qualified experts to decide whether the differential performance results from psychometrically flawed items that result in an underestimation of the ability or achievement level of particular groups of test takers, or whether the item represents genuine differences in the ability or achievement.

For expository purposes the two terms--bias and DIF--will be used interchangeably in this investigation with the understanding that DIF is not sufficient evidence to determine that an item is biased. Throughout the discourse, the meaning of the terms *DIF* and *bias* will be the same: an item functions differently for comparable members of two groups.

The manner in which an item functions differently for two groups may be *uniform* or *nonuniform*. An item is considered to display *uniform* DIF when the probability of answering the item correctly is consistently different across the ability continuum for comparable members of two groups. An example of uniform DIF is when a member of the referent group *at each ability level* has a greater chance of answering an item correctly than a comparable member of the focal group. An item is considered to display *nonuniform* DIF when the item functions differently for two groups, *and* the group that is advantaged in correctly answering the item varies along the ability continuum. An example of nonuniform DIF occurs when a member of the focal group in the low ability range has a *greater* chance of answering an item correctly than a comparable member of the referent group; and when in the upper ability range, a member of the focal group has a *lower* chance of selecting the correct answer than a comparable member of the referent group.

Modern techniques of DIF detection are based on the practice of comparing examinees with similar ability. The procedure of matching examinees with similar ability prior to examining differences reduces the confounding of true differences with artifactual differences. For the majority of bias investigations, examinees are considered to be of equal ability if they have the same total observed score on the test being investigated. A major assumption in the use of the matching criterion is the unidimensionality of the criterion used in matching examinees. In other words, bias investigation procedures assume that one dominant underlying ability is contributing to an examinee's probability of correctly answering the test question. The degree to which this assumption is violated will determine the validity of any judgments made about whether an item displays DIF.

This investigation seeks to compare the degree to which DIF procedures are robust against the violation of unidimensionality in the matching criterion and to compare total raw scores with factor analytically-based matching criteria. To gain an appreciation of the problem, consider two simple scenarios. In the first scenario, from Bond (1981), a teacher wishes to assess the ability of her eighth grade students to "reason analogically" and develops a verbal analogy test for this purpose. A subset of the items, however, are vocabulary specific in that they contain words in the analogy that are more familiar to examinees raised on a farm than they are to urban students. Because rural students are more familiar with the very words that make up the analogy, the overall test is a "purer" measure of verbal analogical reasoning for them than it is for urban students. Urban students are penalized because of lack of knowledge of rural terms. For the urban examinees the total test score would not be unidimensional in regard to reasoning ability. While the total score for the rural examinees would reflect their underlying ability to reason analogically, the observed score for the urban examinees would be composed of analogical reasoning *and* lack of knowledge of rural terms.

In a second scenario, imagine a mathematics test composed entirely of "word problems." Clearly, the test is a measure of mathematical as well as verbal ability, and an examinee's total test score reflects his or her ability in both areas. If the verbal demands of the test are sufficiently high, and if subgroups of examinees differ in verbal ability, then group differences on test may result from a combination of differences in mathematical ability (the intended construct) as well as verbal ability (an extraneous construct). Several researchers (Camilli & Smith, 1990; Ryan, 1991; Shepard, Camilli, & Williams, 1985) have observed that mathematics items that require reading often function differentially for African American and White examinees. The verbally-loaded math items are often more difficult for African American students than for (mathematically) comparable members of the referent group. In this scenario, the total test score would not be unidimensional in regard to mathematics ability. If the two groups of examinees are equally able in the two abilities, the observed total score can be a valid matching criterion.

To place the development of item bias techniques in context, first the history of public concern over test fairness will be reviewed, and then a review of the literature on methods used to investigate item bias or DIF will follow. The specific focus and research questions of the current investigation will then be described. To wit, using both simulated and real data sets, the sensitivity of total score and a factor analytically-derived score as the matching criteria in DIF studies were investigated under various violations of the assumption of unidimensionality.

CHAPTER II
LITERATURE REVIEW

A Brief History of the Development of Testing and Its Controversies

Controversy has accompanied the testing industry in each stage of its development. The rise of routine and mass testing of American school children originated in the educational establishment's response in the mid-1800s to the growing numbers of students in American public schools (Resnick, 1982). Popular use of the more than 200 achievement tests for elementary and secondary schools prior to World War I attests to this situation. Another early use of achievement tests, and one that continues today unabated, was to compare the "quality" of different schools and school systems. The use of standardized testing to compare schools goes back at least to the end of the 19th century with the spelling surveys of Joseph Rice (Haney, 1981; Resnick, 1982). The most vocal early critic of the use of standardized tests in public schools was (and is) the National Education Association, who in the 1980s called for a ban on all standardized testing in the public schools.

The use of tests for individual diagnosis of "school readiness" and, eventually, for placement in classes for the "educable mentally retarded" began shortly after the turn of the century, when French psychologists Alfred Binet and Theodore Simon developed the famous Binet-Simon scale, a 30-item test designed to identify Parisian students who were unable to benefit from the normal school curriculum (Cunningham, 1986). Lewis Terman of Stanford University adapted a later version of the scale for English-speaking students, the Stanford-Binet; and it quickly became the most widely-used test of "intelligence" in America. The modern era of multiple-choice testing began with two multiple-choice, paper-and-pencil forms of the Stanford-Binet developed by Arthur Otis, known as Army Alpha, and its non-verbal counterpart, Army Beta. These tests were administered to some 1.7 million recruits during World War I to aid the military in placing soldiers in various military jobs (Haney, 1981; Resnick, 1982). The creation and wide-spread use of the army tests increased public debate and promoted skepticism about testing. In the print media of the era, the debate of two individuals, Walter Lippman and Lewis Terman, planted seeds of doubt about the use of tests (Block & Dworkin, 1976). In the 1920s Terman introduced the National Intelligence Test, a group aptitude test, to the public schools; the primary use of the instruments was to create homogenous groups for instruction (Resnick, 1982).

Controversy notwithstanding, the use of standardized tests grew considerably during the 1930s, a decade that witnessed the first publication of Oscar Buros's *Mental Measurement Yearbook.* Numerous tests were developed for use in industry for selection and placement. It was during this era that the College Board introduced the Scholastic Aptitude Test. The development of the optical test scoring system in the mid-1950s by Lindquist provided a technological boost to the popularity of tests. In response to the launching of Sputnik, testing came to the fore in the 1960s with the National Defense Act and the identification of academically talented students.

As was true 25 years earlier, the large-scale use of tests created controversy. Articles in the early 60s questioned the need for so much testing and criticized the level of thinking skills measured by multiple-choice tests (Haney, 1989). The biggest controversy to rock the measurement community was Jensen's (1969) article *How Much Can We Boost IQ and Scholastic Achievement?* in which he argued that score differences between races on intelligence tests might have a genetic basis.

With the advent of the Civil Rights movement, increased attention was paid to the use of tests to select job and school applicants and the resulting implications for minority groups. Typically these concerns centered around whether the selection process was fair. There are a number of competing models of fair selection, all based to some extent upon the regression of criterion performance on test scores. The model of fair selection generally accepted by the measurement community (and in fact endorsed by the 1985 Standards for Educational and Psychological Testing) is the classical or regression model (Cleary, 1968), which specifies that a test is fair if the predictive relationship between test and criterion can be described by a common algorithm (e.g., regression line).

Several alternative models of fair selection have been proposed, among which are Thorndike's Constant Ratio (Thorndike, 1971), Cole's Conditional Probability (Cole, 1973), Linn's Equal Probability (Linn, 1973), the Equal Probability Model of Einhorn and Bass (1971), the culture-modified criterion model of Darlington (1971), and the utility model of Gross and Su (1975). A review and evaluation of these models is beyond the scope of the present investigation. A critical review and critique of the models can be found in Petersen and Novick (1976).

Public skepticism about testing resulted in the enactment of truth-in-testing legislation in 1979; the legislation requires test companies to provide to an examinee the test questions, the examinee's responses, and answer keys. In *The Mismeasure of Man* (Gould, 1981), a doubting public read an historical account of the questionable practices in the research on human intelligence. The growing concern about testing contributed to the formation of the National Center for Fair and Open Testing (FairTest), an organization which functions as a consumer awareness group and challenges current test practices.

During the 1970s and early 1980s, the controversy surrounding the use of tests and their potential for bias and adverse impact on minorities, especially African Americans, was joined in the courts in several celebrated cases, most notably Griggs v. Duke Power Company in the employment arena; Debra P. v. Turlington, Larry P. v. Riles, PASE v. Hannon in education; and the Golden Rule case in professional certification. Griggs v. Duke Power company, a landmark decision in the history of employment testing, established the rule that "the plaintiff carries the burden of establishing a *prima facie* case of discrimination, which, once established, places upon the defendant-employer the burden of demonstrating that test is a 'reasonable measure of job performance' " (Wigdor, 1982).

In Debra P. v. the State of Florida, the Florida state legislature passed a bill that required all students to obtain a minimum passing score on a state-mandated test in order to graduate from high school. Students who could not pass the test were given a Certificate of Attendance, rather than a diploma. If the policy had been implemented, 20% of black students and 2% of white students would have been denied diplomas. The plaintiffs prevailed; the court ruled that the practice could not be instituted immediately, and that the state must show that all students had a reasonable opportunity to acquire the skills and knowledge tested.

In Larry P. v. Riles, the routine use of IQ tests in the selection and placement of students in the state of California into classes for the educable mentally retarded resulted in a disproportionate number of minority children being so placed. The class action suit successfully argued that IQ tests (specifically, the Stanford-Binet and the Wechsler) were not valid for identifying the educational deficits of minority youngsters and the administration of IQ tests to minority students in California was banned.

In a similar suit, PASE v. Hannon, the plaintiff charged that the Wechsler Intelligence Scale for Children, Revised (WISC-R), used in the placement of children into remedial classes, was biased against black children. This case is noteworthy in that the judge, frustrated by the contradictory testimony of opposing expert witnesses, reviewed each item on the test himself, and ruled whether or not it was biased. This case occurred at approximately the same time that measurement specialists began to concentrate their efforts on identifying individual test items that might be biased against minority individuals.

Item Bias and Differential Item Performance

In their book, *Methods for Identifying Biased Test Items*, Camilli and Shepard (1994) traced the origin of the modern investigation of item bias to Eells, Davis, Havighurst, Herrick, and Tyler (1951) study, *Intelligence and Cultural Differences*. Research prior to Eells et al. primarily investigated whether the differences in ability were due to environment or genetics. Eells et al. were the first to investigate systematically the possibility that group differences in performance on tests might be attributed to item content and format rather than the examinee's ability alone.

The investigation of item bias has included groups based on ethnicity, gender, age, disabilities, and geographic areas, but by far the most often compared groups are African American and White examinees (Baghi & Ferrara, 1989, 1990; Ironson & Subkoviak, 1979; Scheuneman & Gerritz, 1990; Spray & Miller, 1992; Zwick & Ercikan, 1989). Investigations of item bias also have included Asian Americans (Schmitt & Dorans, 1990) and Hispanic Americans (Schmitt & Dorans, 1990; Zwick & Ercikan, 1989). More recently, Lai and Saka (1993) investigated the performance of Hawaiian students compared to mainland United States students. Investigation of DIF for groups based on gender have also been numerous (Baghi & Ferrara, 1989; Scheuneman & Gerritz, 1990; Zwick & Ercikan, 1989). Rudner (1978) investigated item bias with hearing-disabled examinees. Ironson and Subkoviak (1979) also investigated differences on item performance for rural examinees versus urban examinees.

The statistical techniques developed for the investigation of DIF have been applied to the Scholastic Aptitude Test (SAT) (Scheuneman & Gerritz, 1990; Schmitt & Dorans, 1990), the American College Testing Mathematics Usage Test (Spray & Miller, 1992), the Graduate Record Examination (Scheuneman & Gerritz, 1990), and the National Assessment of Educational Progress (NAEP) assessments (Zwick & Ercikan, 1989), among others. Recent applications of the various methods have been in the investigation of DIF in performance assessments (Zwick, Donoghue, & Grima, 1993).

The investigation of DIF in the aforementioned groups and instruments has involved sundry techniques for identification of biased items. The techniques can be grouped according to their use of qualitative methods (i.e., judgmental methods) or statistical methods. The two forms of investigation developed in tandem and offer complementary information about item bias. A review of the development of the two methodologies for investigation of item bias will inform the latter discussion.

Development of Judgmental Methods

Early investigations of item bias focused on the role of experts in the identification of biased items. In a comprehensive review of judgmental methods used prior to the 1980s, Tittle (1982) reported the role that judgmental methods had in the stages of the development of a test: test content specification, item writing, and item review; item tryout; item selection; and development of norms and scales. Tittle reported the guidelines developed by the testing industry to train judges and guide their review in an effort to eliminate bias. Early judgmental methods emphasized a review of test items to eliminate stereotypes, increase minority representation in reading passages, and equalize examinee familiarity with the items. To improve expert consistency, guidelines were developed and operationalized by the creation of tally sheets. The tallies served several purposes. They served to determine if various groups (e.g. gender, ethnic) were represented in verbal passages of tests and in reasonable proportions. They also focused the review on the representation of the characters to eliminate stereotypes. Character representation was reviewed in terms of physical attributes, setting of the passage, socioeconomic status, behavior, and consequences of actions.

Early judgmental methods also emphasized a review of test items for the opportunity to learn (Tittle, 1982). In this approach, one source of test bias occurs when tests do not measure what is taught in the classroom. If policy decisions are based on the test results, and that information does not reflect curricular emphases, conclusions drawn from the results may be faulty. The judgmental methods attempted to determine the overlap between test content and curriculum content. The process has been operationalized by surveying teachers to determine the percentage who teach a particular objective or by examining the percentage of objectives of a curriculum represented in a test. In addition to the match between objectives and items, methods were developed to use taxonomies to classify items (e.g. Bloom's), curricular format, content, and skills/processes required.

The importance of the congruence between curriculum and test was highlighted by Bianchini (cited in Tittle, 1982). Bianchini reported that from 1966 to 1970 for the California Miller-Unruh Statewide Testing Program the reading scores of 65% of first grade students on the *Stanford Achievement Test* were in the bottom quartile and the statewide median was at the 38th percentile rank. In 1981, however, the statewide median on the *Cooperative Primary Reading Test* was equal to the national median. Bianchini noted that the vocabulary in the *Cooperative Primary Reading Test* overlapped 55% with the vocabulary in the first grade basals while the vocabulary in *Stanford Achievement Test* only overlapped 19% with the vocabulary in the first grade basals. He argued that the change in the median was not due to easier norms but due to greater overlap between the curriculum and test.

In contrast to judgmental methods, statistical methods for the investigation of item bias are methodologically complex. Plake (1980) compared the agreement between items identified as biased by expert judges with items identified by use of a statistical procedure. The author examined the differential functioning of items when defining group membership by grade level (fifth grade student versus students in other grade levels). Plake used an ANOVA procedure to identify an item by group interaction; items displaying an interaction were considered biased. The judges were asked to select items that would be easier or more difficult for each non-fifth grade group. Plake found that expert judges identified twice as many items as being biased as the statistical method, and the judges frequently predicted bias to be in the opposite direction than it appeared to occur. Englehard (1990), in a study of a teacher certification test, asked 42 teachers to judge which items would function differently for black and white examinees. As proved true for Plake, the judges were unable to indicate the items that would be flagged through use of the Mantel-Haenszel statistical procedure, a chi-square procedure for detection of differential item functioning advanced by Holland and Thayer (1988).

In a study by Hambleton and Jones (1993), the authors sought to improve agreement of identification of biased items using statistical and judgmental methods. To address this concern the authors refined an earlier judgmental review form (Hambleton & Rogers, 1988); in the earlier form the guidelines were divided into two categories: *Stereotyping and Inadequate*

Representation and *Sex, Ethnic, Cultural, Religious, and Class Bias.* In the first category the emphasis was on the traditional sensitivity review criteria. Typical questions addressed issues about whether the test contained material controversial to members of the focal group or depicted minorities in stereotyped occupations. In the second category the criterion emphasized content that would tend to favor the referent group or disfavor the focal group. In this case questions addressed issues about whether the test contained material unfamiliar to members of the focal group or had language specific to a particular group. In the 1993 study Hambleton and Jones refined the criteria for identifying bias in the second category. The authors reported that 5 of the 11 items identified as biased by the judges were identified as displaying DIF by empirical methods. The authors recommended the inclusion of new bias criteria (e.g., avoid negatively worded items) on review forms as features specific to biased items are noted in the literature.

Judgmental procedures may have proven ineffective in identifying items that differentially function for two groups; however, the judgmental methods functioned to make the test content more representative of the examinee population. The review of items in terms of fairness to various groups has become two-pronged reviews: a sensitivity review and a DIF review. A case study by Ramsey (1993) reviewed current judgmental methods--now referred to as sensitivity reviews--as used by the Educational Testing Service (ETS). The goals of ETS sensitivity reviews are to insure that test specifications require material that is representative of minorities and that tests are free of offensive language and stereotypical representation. Stated explicitly by ETS is "the importance attached to sensitivity review does not imply a measurable relationship between material considered offensive by some test takers and the scores of the test takers" (cited in Ramsey, 1993, p. 384). The purpose of the review has become "to create tests that acknowledge and respect diversity through the inclusion of some materials and exclusion of other" (Ramsey, 1993, p. 384).

A sensitivity-review procedure has been formalized by ETS for the review of potential tests prior to being pilot tested. In the review process a test developer submits the instrument to a second party who, in turn, will assign the instrument to a reviewer; test developers are not allowed either to review their tests or to select the reviewer. The reviewer must approve a test before it is allowed into production. Appeal processes have been established for the instances where a general agreement is not reached between the developer and the reviewer.

Before serving as a reviewer, test developers receive one and one-half days of training; every fifth year the reviewer must take a refresher course. During training, reviewers are presented with examples of items ranging from blatantly inappropriate, to questionably inappropriate, to acceptable. Sharing much in common with the criteria used in judgmental methods outlined by Tittle (1982), ETS has outlined six criteria that test materials must meet:

> (a) should be balanced, (b) should not foster stereotypes, (c) should not contain ethnocentric or gender-based underlying assumptions ... (d) should not be offensive when viewed from an examinee's perspective, (e) should not contain material that the subject matter does not demand, and (f) should not be elitist or ethnocentric. (Ramsey, 1992, p. 375)

The modern sensitivity review shares much in common with the earlier judgmental methods. Commonalities include a reliance on experts, a need for training, and identification of relevant criteria.

Informed by the research, the judgmental process no longer claims to identify items that would be more or less difficult for examinees. The identification of items that may be biased in terms of difficulty is the province of statistical techniques. A review of early techniques of identifying differentially functioning items will build a basis for the discussion of modern techniques.

Historical Statistical Techniques Used in the Investigation of DIF

In the late 1960s, statistical indices for the examination of item bias began to focus on the instruments used in selection. Some of the earliest methodologies were the transformed item difficulty index, correlational techniques, and analysis of variance (ANOVA).

Transformed Item Difficulty Index

Early bias studies focused on the differential difficulty of an item as an indicator of bias. In these investigations the classical difficulty index (p, the proportion of examinees who answer an item correctly) was used to examine the different characteristics of items. Items with similar p-values were considered to be free of bias while items with highly discrepant p-values for the groups under investigation were suspect (Eells, 1951 cited in Camilli & Shepard, 1994). Angoff (1972) improved the methodology of differential difficulty and introduced the transformed item difficulty index. This index, also referred to as the delta plot method, converts item proportions correct for each group to normalized z-scores. This is accomplished by first obtaining the percentile corresponding to 1 - *p*. Unlike its derivative *p*-value, this converted score reflects the difficulty of the item. The difficulties for the two groups are graphed in a scatterplot. A 45 degree line, going from the lower left of the plot to the upper right, displays the difference in percentage correct for the two groups. Items that are outliers (i.e., items that deviate from the 45 degree line) are relatively more difficult for one group. An index of bias determines the perpendicular distance of an item from the major axis line of best fit. The technique was found unsatisfactory since when groups differ in mean ability, items that validly discriminate on the basis of the ability or achievement were indicated as biased.

An alternative method--residualized Angoff--was also investigated by Shepard, Camilli, and Williams (1985). In this method the point-biserial correlation of an item is partialled out of the delta index by regressing the original Angoff indices on the combined-group point biserial for the items. The residual delta indices were calculated as the difference between the observed index and the expected delta value based on the item's point-biserial. Shepard et al. reported the modified Angoff index correlated from .59 to .61 with a signed IRT index. In a simulated study the authors found the residualized Angoff index identified 84% of the known bias in the data while chi-square techniques identified 87% of the known bias.

Correlational Techniques

Early studies of item bias used the item discrimination index to gauge if an item was functioning differently for two groups (Green & Draper, 1972). In the Green and Draper study, an item was considered biased if it was in the top half of discriminating items for one group and in the bottom half for the other group. Ironson & Subkoviak (1979) created an unsigned index of item bias by using the absolute value of the difference between the discrimination index for the focal and referent groups. Ironson and Subkoviak found that the item discrimination index correlated poorly with transformed difficulty, chi-square, and item characteristic curve (ICC) indices of bias. They considered the item discrimination technique to be inadequate for the detection of bias.

Analysis of Variance Techniques (ANOVA)

The earliest attempts to identify items that had a differential impact on groups used analysis of variance to test for interactions. In this two-factor ANOVA, examinee group and items served as two factors. Group differences were accounted for in the main effect; while differential difficulty on the items for the focal and referent groups was evidenced in the item by group interaction (Cleary & Hilton, 1968). The ANOVA technique was shown to be ineffective for detection of bias when Camilli and Shepard (1987) in a simulation study demonstrated that even with a large amount of bias built into the items the interaction effect only accounted for single-digit amounts of the variance. The variance due to bias is confounded with the group effect. As a variation of the popular ANOVA method, Angoff and Ford (1973) matched groups on ability and investigated the effect on the item-group interaction. Their finding, that matched groups reduced the interaction, pointed to the necessity of matching on relevant criteria to create comparable groups.

Modern Statistical Techniques Used in the Investigation of DIF

The previous indices of bias share a common problem: each method confounds real differences in ability (mean group differences) with bias. Modern approaches of DIF require that only examinees of similar ability should be compared to determine if group membership has a differential impact on an examinee getting an item correct. Holland and Thayer (1988) wrote: "Basic to all modern approaches to the study of *dif* is the notion of comparing only *comparable* members of F and R in attempting to identify items that exhibit *dif*" (p. 130). Scheuneman (1975) expressed the concept of comparability thus: "An item is unbiased if, for all individuals having the same score on a homogenous subtest containing the item, the proportion of individuals getting the item correct is the same for each population group being considered" (p.2). The current techniques used in the investigation of DIF comprise two categories: item response theory (IRT) and chi-square techniques.

Item Response Theory

In a summary of IRT, Hambleton, Swaminathan, and Rogers (1991) stated:

> Item response theory (IRT) rests on two basic postulates: (a) The performance of an examinee on a test item can be predicted (or explained) by a set of factors called traits, latent traits, or abilities; and (b) the relationship between examinees' item performance can be described by a monotonically increasing function called an *item characteristic function* or *item characteristic curve* (ICC). This function specifies that as the level of the trait increases, the probability of a correct response to an item increases. (p. 7)

The function, or the item characteristic curve (ICC), is formed by a logistic model. As stated earlier the function relates the probability of a correct answer with level of ability. The function is plotted on a Cartesian axis with the x-axis indicating levels of ability, theta (θ), and expressed on a scale similar to z-scores with values ranging between -4.0 to 4.0. The y-axis indicates the probability of getting an item correct, $P(\theta)$.

The one-parameter model or Rasch model forms a logistic curve based on an estimate of ability, theta (θ), and a difficulty parameter (b). The difficulty parameter is defined as the point on the ability scale associated with a .50 probability of getting the item correct. The formula for the model is

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad i = 1, 2, 3, \ldots, n \qquad (1)$$

where

$P_i(\theta)$ is the probability that a randomly chosen examinee with ability θ answers item i correctly,

b_i is the item i difficulty parameter,

n is the number of items in the test,

e is the base of the natural logarithms.

The function of the difficulty parameter in the logistic model is to place the ICC along the ability continuum. More specifically, the ICC of an item with a low value for b_i would be close to the origin on the ability scale (the abscissa); whereas, the ICC of an item with a high value for b_i would be shifted to the right on the ability scale. Possible values of b_i are the same as delineated by theta; however, the value of b_i is generally between -1.5 to 1.5. The one parameter model assumes guessing does not account for variance in the data. The model also carries the assumption that all items are equally discriminating.

The logistic curve formed in the two-parameter model contains the ability (θ) and difficulty parameter (b) of the one-parameter model and incorporates an item discrimination parameter (a) into the calculations. The item discrimination parameter is defined as the slope of the item characteristic curve at the point associated with b_i on the ability scale. The formula for the two-parameter model is given by

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad i = 1, 2, 3, \ldots, n \tag{2}$$

where

D is a correction factor to adjust the logistic function to closely approximate the normal ogive function,

a_i is the item i discrimination parameter.

The item discrimination parameter (a_i) in the logistic model is analogous to the item-total correlation (i.e., the point-biserial) in classical test theory. That is, it is an index of the extent to which the item discriminates between high and low ability examinees. Formally, it is the slope of the ICC at b_i on the theta scale. In the case of an item with a low value for a_i, the slope of the ICC would be nearly flat, or parallel to the abscissa. In contrast, for an item with a high value for a_i, the slope of the ICC would fall rapidly on the left side of the point of inflection (b_i) and rise sharply on the right side of the point of inflection. As was true of the one parameter model, the two-parameter logistic model assumes examinee guessing does not account for variance in the data. This assumption is considered questionable when the items are in a multiple-choice format, thus giving rise to the third logistic model--the three-parameter logistic model.

The three-parameter model incorporates the parameters contained in the previous two models--ability (θ), difficulty parameter (b), item discrimination parameter (a)--and adds a pseudo-chance-level parameter (c). The c parameter is defined as the non-zero lower asymptote for the ICC curve. The three-parameter model is given by

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad i = 1, 2, 3, \ldots, n \qquad (3)$$

The pseudo-chance-level parameter in the logistic model represents the probability of examinees with low ability answering an item correctly (that is, the probability of getting the item right by guessing). In multiple-choice tests, the greater the number of alternatives (distractors), the lower the value of c. By contrast, items with very few alternatives (e.g., true-false items) have a correspondingly high value of c.

When ICCs are plotted separately for the focal and referent groups, the probability of an examinee with a given ability answering an item correctly should be the same regardless of group membership. In other words, the two ICCs should, within measurement error, coincide. The more two ICCs differ for two groups, the greater the amount of DIF. Each parameter--the difficulty parameter, the discrimination parameter, and the pseudo-chance-level parameter--must be the same or the ICCs for the referent and focal groups will not be the same. The greater the amount of area between the two curves, the greater the amount of DIF.

Hambleton, Swaminathan, and Rogers (1991) summarized two approaches to quantifying DIF in IRT: comparison of item parameters and area between ICC curves. The first approach is Lord's chi-square significance test, which compares simultaneously the differences of the groups *a* and *b* parameters (Lord, 1980). The second approach involves calculating the areas between the two curves.

While the use of IRT models for detection of DIF is considered the best procedurally, IRT is too expensive for many testing programs (Camilli & Shepard, 1994; Hambleton & Rogers, 1989). In addition, the sample sizes needed to obtain stable parameter estimates in IRT investigations of bias have been reported at 1,000 for each group--1000 members of the focal group and 1000 members of the referent group (Hills, 1989; Ironson & Subkoviak, 1979). Sample sizes of this magnitude are impossible for small testing programs. Shepard, Camilli, and Williams (1985) also recommended sample sizes of 1,000 per group and reported that *a* and *b* parameters could not be estimated for a sample that contained only 300 focal group members. In less complex IRT models it is possible to obtain stable estimates with fewer examinees; however, even with large samples the *c* parameter is poorly estimated (Baker, 1987; Shepard et al., 1985). Other promising statistical procedures for the investigation of DIF are based on chi-square procedures.

Chi-square Techniques

The practice of comparing examinees of similar ability is an integral part of the chi-square techniques for investigating DIF. Scheuneman (1979) proposed the use of a chi-square procedure as a nonparametric analog of the parameter-based IRT models. Whereas in IRT the ability of examinees is estimated, the chi-square procedure uses the observed total score of examinees to control for ability. By controlling for ability, Scheuneman's model provided the transition from the historical techniques for investigating DIF which confounded group differences with differences in ability. In Scheuneman's approach, the probability of a correct response is examined for each ability level for the focal and referent groups. The author recommended collapsing score intervals such that each cell contains 10 to 20 correct responses. This generally results in three to five ability groups. Scheuneman's procedure was criticized for not using the information contained in the incorrect responses (Baker, 1981); since the distribution using only correct responses is not a true chi-square distribution, Baker correctly suggested that the "full information" chi-square be used.

Two currently used chi-square techniques are the standardization approach (Dorans & Kulick, 1986) and the Mantel-Haenszel approach (Holland and Thayer, 1988). In comparing the performances of two groups on an item, the two methods use 2 X 2 X J contingency tables, where the first 2 indicates the number of groups, the second 2 indicates the number of score levels for the items (1 or 0), and the J indicates the number of score intervals for the test. The format that each table assumes at each level of performance (J) is shown in Table 1.

Table 1

Contingency Table for Comparison of Group Responses at Jth Level of Ability

		Score on Studied Item		
		1	0	Total
Group	Referent	R_{Rj}	W_{Rj}	N_{Rj}
	Focal	R_{Fj}	W_{Fj}	N_{Fj}
	Total	R_{1j}	W_{0j}	N_{tj}

R_{Rj} represents the number of people in the referent group at the *j*th score level who got the item correct, and W_{Rj} represents the number of people in the referent group at the *j*th score level who got the item incorrect. Parallel to the referent groups, R_{Fj} and W_{Fj} represent the number of focal group members with correct and incorrect responses at the *j*th score level, respectively. The total number of correct and incorrect responses at the *j*th level are indicated by R_{1j} and W_{0j}, respectively; and the total number of referent and focal group members at the *j*th level are indicated by N_{Rj} and N_{Fj}, respectively. Finally, the total number of examinees at the jth score level is indicated by N_{tj}.

Standardization Approach

The standardization approach (Dorans, 1989) compares performance for the referent and focal groups by examining the difference in proportion or percent correct at each level of *j*.

$$D_j = P_{Fj} - P_{Rj} \qquad (4)$$

where

$$P_{Fj} = R_{Fj}/N_{Fj} \text{ and } P_{Rj} = R_{Rj}/N_{Rj}$$

For a visual representation of group differences on item performance, the standardization approach utilizes a plot similar to that used in the delta plot method of Angoff. The graph uses conditional percent correct for the referent and focal groups to examine performance of the studied item. The j levels for the test form the abscissa, and the conditional percent correct form the ordinate. In the standardization approach, the expected performance on the studied item at each score level is determined by the referent group.

The standardization procedure uses two indices for flagging items: the standardized *p*-difference (D_{STD}) and the root-mean-weighted-square difference, RMWSD (Dorans, 1989). At each score level, both indices are weighted by the number of members in a standardization group, typically the focal group. The weights are cumulated over the score levels to provide a summary index. The advantage of the use of weights is to concentrate the contribution to the summary score in the score intervals where the greatest number of focal group members occur.

The index D_{STD} ranges from -1.0 to 1.0. The formula for the standardized *p*-difference is:

$$D_{STD} = \frac{\sum_{j=1}^{J} K_j(P_{fj} - P_{rj})}{\sum_{j=1}^{J} K_j} \tag{5}$$

where ($K_j/\Sigma K_j$) is the weight supplied by the standardization group at each level of *j* to weight differences in performance between the referent and focal groups. The value of K is at the discretion of the researcher; however, it most frequently is specified as the number of people in the focal group at each level of j, *or* N_{Fj}. By use of N_{Fj}, greater weight is given to differences of P_{Fj} and P_{Rj} in the score intervals where the majority of members of focal groups are located. With this weighting, D_{STD} becomes:

$$P_f - \hat{P}_f \tag{6}$$

where P_F is the observed performance of the focal group on the item and $\widehat{P}_f$ is the expected performance of the focal group predicted from the referent-group item-test regression curve (Dorans, 1989).

The second index, the root-mean-weighted-square difference (RMWSD), provides the additional benefit of accounting for items where crossover of the slopes would cancel any differences in performance as can occur with the standardized *p* difference. As evidenced in Equation 7, the square of each interval is calculated creating an unsigned index (Dorans, 1989):

$$RMWSD = \left[\frac{\sum_{j=1}^{J} K_j (P_{fj} - P_{rj})^2}{\sum_{j=1}^{J} K_j} \right]^{.5} \tag{7}$$

Current test developers use the standardized *p* difference as a flag since the RMWSD flagging criterion was found to be sample-size specific. For flagging items in test-construction practices, ETS has adopted $|D_{STD}| > .10$; and for research purposes, ETS has specified a flagging criterion of $|D_{STD}| > .05$.

Another variation of D_{STD} used to measure DIF is based on an item difficulty metric, the delta (Δ) metric, used at the Educational Testing Service (ETS). The delta metric is calculated as follows:

> A proportion correct is converted to a z-score via a p-to-z transformation using the inverse normal cumulative function, followed by a linear transformation to a metric with a mean of 13 and standard deviation of 4 via:

$$\Delta = 13 - 4[\,\phi^{-1}(p)] \tag{8}$$

such that large values of Δ correspond to difficult items, whereas easy items have small Δ values (Dorans, 1989, p. 227).

D_{STD} is converted to a delta metric by the following formula:

$$\Delta_{STD} = -2.35 \ln \left[\frac{\dfrac{\hat{P}_f}{(1 - \hat{P}_f)}}{\dfrac{P_f}{(1 - P_f)}} \right] \tag{9}$$

Mantel-Haenszel Approach

Another commonly used indicator of DIF, the Mantel-Haenszel approach, uses an odds ratio, which represents the likelihood that referent group members get an item correct exceeds the likelihood for comparable focal group members. As seen in Equation 10, the odds ratio is summed across the test score intervals to provide a summary index of DIF (Dorans, 1989).

$$\alpha_{MH} = \frac{\sum_{j=1}^{J} M_j \, \alpha_j}{\sum_{j=1}^{J} M_j} \tag{10}$$

where

$$M_j = \frac{W_{rj} \, R_{fj}}{N_{tj}}$$

thus

$$\alpha_{MH} = \frac{\sum_{j=1}^{J} \dfrac{R_{rj} \, W_{fj}}{N_{tj}}}{\sum_{j=1}^{J} \dfrac{R_{fj} \, W_{bj}}{N_{tj}}}$$

A common odds ratio of 1 means that after controlling for ability there is no differential performance between the two groups. A common odds ratio of 1.5 would indicate that members of the referent group are one and one-half times more likely to answer an item correctly as comparable members of the referent group. Finally, a common odds ratio of .5 would indicate that members of the referent group are half as likely to answer an item correctly as comparable members of the referent group

Holland (1985) proposed a chi-square test with one degree of freedom for the null hypothesis H_0: α = 1. The hypothesis associated with the test is that there is no relationship between group membership and item response after controlling for ability level. When the Mantel-Haenszel statistic exceeds the table value of chi-square at a specified level of α, it indicates that item performance for the two groups is consistently different. The formula is shown in Equation 11, below (Dorans, 1989).

$$\alpha^2_{MH} = \frac{\left[\sum_{j=1}^{J} R_{rj} - \sum_{j=1}^{J} \mu_j - \frac{1}{2} \right]^2}{\sum_{j=1}^{J} \sigma^2{}_j} \tag{11}$$

where

$$\mu_j = E(R_{rj} \mid \alpha = 1) = \frac{N_{rj}\, R_{tj}}{N_{tj}}$$

and

$$\sigma_j^2 = VAR(R_{rj} \mid \alpha = 1) = \frac{N_{rj}\, N_{fj}\, R_{tj}\, W_{tj}}{N_{tj}^2\,(N_{tj} - 1)}$$

The delta metric is also used to create a variation of the Mantel-Haenszel procedure; the conversion to a Mantel-Haenszel delta metric (D_{MH}) is achieved by the following formula:

$$\text{MH D-DIF} = -2.35 \ln(\alpha_{MH}) \qquad (12)$$

For flagging items, Dorans suggested that a value of $D_{MH} > 1$.

The Mantel-Haenszel procedure is analogous to comparing the area between one-parameter ICC curves for focal and referent group members (Hambleton & Rogers, 1989). The matching total score serves the same function as the latent trait in IRT.

Hills (1989) wrote of the Mantel-Haenszel procedure:

> MH appears to be easily used, easily programmed, has a significance test for use with these small samples, is designed for such small samples, may not require matched samples, and seems more stable than other methods across samples. The statistical test is considered to be very powerful, hence important effects have the most chance of being detected in small samples with this method. (p. 7)

Indeed, the procedure is simple to use and has been incorporated into at least one statistical package (SAS, 1988).

Disadvantages of the chi-square procedures include that they have proven insensitive to nonuniform DIF (Hambleton & Rogers, 1989) and for very large sample sizes the test of significance may not reflect the practical significance of DIF (Hills, 1989).

From DIF Detection to Test Construction

The previous procedures identify items that are differentially functioning for the focal and referent groups; however, the presence of DIF is not conclusive in determining whether an item is biased. In test-development procedures at the Educational Testing Service, once an item is designated as exhibiting DIF, it is classified in the following manner:

Category A: Negligible DIF — MH D-DIF not significantly different from zero or absolute value less than 1.0

Category B: Intermediate DIF — MH D-DIF significantly different from zero and absolute value less than 1.0 and either 1) less than 1.5 or 2) not significantly greater then 1.0.

Category C: Large DIF — MH D-DIF significantly greater than 1.0 and absolute value of 1.5 or more (Zieky, 1993, p. 342).

where statistical significance is at the 5% level for a single item.

Items in Category A are considered to display negligible DIF; items in Category B are considered to display intermediate DIF; and items in Category C are considered to display large DIF (Dorans & Holland, 1993).

This information is incorporated into a set of procedures to guide the selection of items for inclusion in tests. ETS procedures are as follow:

- The content and statistical specifications for the test must be met.
- Large form-to-form variations in DIF in tests made from the same pool of items must be avoided. Test assemblers making more than one test from a pool of items should not use up all of the questions in Category A or the items in Category B with the lowest DIF values in the first tests to be assembled, thereby forcing later tests to have progressively larger DIF values.
- Within the previously mentioned constraints, items from Category A should be selected in preference to items from Categories B or C.
- For items in Category B, when there is a choice among otherwise equally appropriate questions and the equivalence of tests made from the pool can be maintained, items with smaller absolute DIF values should be selected in preference to items with larger values.

- Items from Category C will *not* be used unless they are judged to be fair and essential to meet test specifications.
- If Category C items must be used, the test assembler will document the reason and will explain why the items are not unfairly related to group membership. A reviewer will check to make sure that the use of Category C items was indeed necessary and that the terms are fair. (Zieky, 1993, p. 344)

Comparison of DIF Results Across Methods

Availability of programs, complexity of models, costs, and large sample sizes, have prompted the comparison of bias techniques to determine if less complex models might as efficiently identify differential item functioning. In an empirical study based on data from African American and White examinees, Ironson and Subkoviak (1979) compared the identification of biased items across four methods. The bias investigation techniques that were compared were three relatively simple bias identification models--transformed difficulty, item discrimination, chi-square method--and the more complex bias identification technique--IRT. The authors reported highest correlations between the transformed difficulty, chi-square, and IRT techniques. Ironson and Subkoviak attributed the high level of agreement for IRT, chi-square, and transformed difficulty to the fact that the three methods control for ability level prior to examining differences. In a study of item bias in which the focal group was hearing-impaired examinees, Rudner (1977) found a similar level of agreement for the three methods: transformed difficulty, chi-square, and IRT.

Shepard, Camilli, and Williams (1985) studied the congruence of bias indices for several bias techniques. The procedures compared were the chi-square, the Angoff delta plot, residualized delta plot, pseudo-IRT, and one and three parameter IRT. The unsigned and signed bias indices generated by the three-parameter IRT model were selected as the criteria to evaluate the alternative bias procedures. The agreement between the techniques was measured by calculating Spearman rank-order correlations. The chi-square procedure correlated with the IRT criterion from .50 to .53 for unsigned indices and .57 to .67 for signed indices. These results led the authors to the conclusion that chi-square techniques could be a substitute for IRT models for small samples.

Factors in the Application of Chi-square Techniques

When selecting a method for the investigation of bias, test practitioners must determine which conditions will provide information about the performance of items with fewest Type I or Type II errors. The accuracy of DIF chi-square techniques have been investigated with variations of sample sizes, number of items, and matching variables.

Sample Size

In an earlier section of the *Literature Review* it was reported that large sample sizes are required for IRT estimation procedures. While sample sizes of 1,000 examinees per group are not needed in chi-square procedures, researchers have found certain minimum requirements for sample sizes. When Mantel-Haenszel methods are used, Kubiak and Colwell (1990) suggested a minimum 100 focal group members and a combined sample of 500; however, Hills (1989) suggested as few as 100 focal members with a combined group of 200. Ryan (1991) found MH estimates to be unstable for samples with as few as 141 to 167 focal members. Hoover and Kolen (1984) also reported DIF indices to be unstable with a sample size of 100. When the score distribution reflects the difference of one standard deviation generally found between black and white examinees on cognitive ability measures, Camilli and Smith (1990) found the MH chi-square statistic to be robust with 300 members in the focal group. Others (Engelhard, Anderson, and Gabrielson, 1990; McPeek & Wild, 1986) have found samples of 600 to be inadequate to obtain stable DIF indices. Mazor, Clauser, and Hambleton (1991) examined the ability of MH to detect DIF in samples of 100, 200, 500, 1000, and 2000 examinees for each of the focal and referent groups. As the sample size decreased, the efficiency of the MH decreased from 69% correct identification to 13% correct identification.

Number of Items

In an early study relevant to the current chi-square techniques, the influence of test length on the stability of DIF indices was investigated by Rudner, Getson, and Knight (1980). Two chi-square variations were investigated using five intervals and total possible score intervals (with the restriction that expected cell sizes equal at least 5). The authors found the length of the test did not substantially affect the identification of items as displaying DIF when the number of items on the test was greater than 20.

In a study that primarily investigated the effects of matching criterion on the identification of DIF, Clauser, Mazor, and Hambleton (1991) studied the effect of test length and identification of DIF. The authors used the *New Mexico High School Proficiency Exam*, a 75 item high school proficiency test that measures five life skills areas: knowledge of community resources; consumer economics; government and law; mental and physical health; and occupational knowledge. The items were analyzed for DIF using total test score as the matching criterion and were reanalyzed using as the matching criterion a subtest score composed of 30 to 31 items from the 75 item test. The results of the 75 item test were compared to the results of four subtests of 30 to 31 items. The authors reported that as test length decreased the number of additional items identified as displaying DIF increased.

Matching Variables

To determine whether two groups are of equivalent ability, it is necessary to have a criterion for creating comparable groups. Holland and Thayer (1988) identified three important criteria for measuring comparability: "... (a) measures of the ability for which the item is designed, (b) schooling or other measures of relevant experience, and (c) membership in other groups" (p. 130). The criterion may be external to the test being examined, or the criterion may be internal to the test. Holland and Thayer reported that the predominant matching criterion is the internal criterion of test score.

External Criterion

The use of external criterion as a matching variable was investigated by Hambleton, Bollwark, & Rogers (cited in Hambleton, Clauser, Mazor, & Jones, 1993). In an investigation of DIF in a high school scholarship test, the authors compared the results of using an external criterion--scores on a high school achievement test--with the results using an internal criterion. The study was replicated in three additional subject areas. The finding was that internal and external matching criterion resulted in similar DIF results for the Mantel Haenszel procedure. The authors postulated that the moderate correlations between the criterion measures (ranging from .38 to .52) resulted in similar patterns of identification of DIF. Their research supported the continued use of internal criterion for matching.

Score Intervals

Another variable of consequence when matching examinees on a criterion is the size of the interval which contains "comparable" examinees. Matching can occur for each possible score or a limited number of test score intervals can be created by pooling scores. Scheuneman (1979), in her seminal investigation of the chi-square technique to study DIF, used the total test score as the internal criterion. She proposed the total test score to be appropriate as the matching criterion on a homogenous test; in the case of a more diverse test a subgroup of items containing the item of interest was used. Generally, total test score is used for matching focal and referent examinees in the Mantel-Haenszel procedure.

Pooling score intervals allows an investigator with a small sample to avoid empty cells. Studies have focused on the effect of a limited number of test score intervals on the identification of items displaying DIF. Wright (1986) found that 61 levels for matching were better than six levels. Raju, Bode, and Larsen (1989) recommended a minimum of four levels for matching with the Mantel-Haenszel procedure.

In a Monte Carlo study, Donoghue and Allen (1993) investigated the use of the total test score as the matching variable (thin matching) versus using pooled levels of total test score (thick matching). Nine variations of grouping score intervals, termed thick matching by the authors, were examined. In one variation--equal interval--pairs of total test scores were combined to create the levels on the table. In two variations, percent total and percent focal, score intervals were combined to approximate the quintiles for the total group and focal group, respectively. In three variations, termed Censor (1), Censor (5), and Censor (20), extreme score intervals were collapsed until the minimum number of observations in the collapsed cells were at least one, five, and 20, respectively. Three additional variations of thick matching--Minimum Frequency (1), Minimum Frequency (5), and Minimum Frequency (20)--followed a similar strategy. In contrast to the Censor methods, which pooled only extreme score intervals, the Minimum Frequency methods pooled any score intervals until the interval contained the minimum number of observations as specified in the title of the variation (1, 5, and 20, respectively). In their study the authors compared mean Δ_{MH} and $X^2{}_{MH}$ for 20 replications of the matching methods when no DIF was introduced into the data; in this case the expected mean value of Δ_{MH} is zero and the expected mean value of $X^2{}_{MH}$ is one. The authors also calculated the mean Δ_{MH} and $X^2{}_{MH}$ for the replicated administrations when DIF was simulated for given items. They found the use of thin matching to be superior for long tests (40 items) and large sample sizes (1600). Thick matching techniques were found superior for short tests.

Clauser, Mazor, and Hambleton (1994) also investigated the effects of reducing the number of score intervals. In a Monte Carlo study the authors found that when the number of score intervals is small and the ability distributions of the focal and referent groups are unequal, the number of Type I errors was inflated. As a result of their findings the authors recommend that pooling of score interval not be used when the ability distributions are unequal. The authors theorized that as the score interval increases, the assumption of matching on equal ability is no longer met. The introduction of impact into the matched groups contaminates the matching criterion and results in items being incorrectly identified as displaying DIF.

The effect of creating a more homogenous sample by stratifying using variables based on educational background met with mixed results in a study by Kubiak (1992). In many cases the number of items with DIF remained the same and in some instances more items were identified with DIF.

Multiple Matching Criterion

When examinees were matched on two criteria, Ryan (1991) found no significant improvement in the stability of the MH statistics; she did find that more items were shifted into the *A* classification (negligible DIF) as used by ETS. Unlike Ryan, McPeek and Wild (1990) found that matching on multiple criteria--analytical and verbal test scores--reduced DIF in logical and analytical reasoning items on the GRE.

Unidimensionality of Matching Criterion

The validity of matching on total score in the investigation of DIF depends on the ability of the total score to represent the underlying ability being measured by the test. If the total score is composed of items that are affected by extraneous factors, then the validity of the matching is questionable. Hambleton et. al (1993) wrote: "When individual items measure more than one ability, or when items measuring different abilities are part of one test, the adequacy of matching criterion may be compromised, leading to errors in the identification of DIF" (p.24). The investigation of the validity of the total score as a matching criterion has taken several directions.

True Score

In the use of observed score an assumption is made that an examinee's observed score reflects the examinee's true abilities. The question arises whether matching on observed score is adequate for the detection of DIF when the true score for examinees' differs from the observed scores. Spray and Miller (1992) investigated the use of observed score when the true abilities of the examinees are incongruent and found that the detection of DIF was not seriously affected if tests are relatively free of DIF.

Purified Criterion

Drawing on a study of techniques using the Rasch model and the Mantel-Haenszel procedure, Holland and Thayer (1988) suggested that in the analysis of an item for DIF the studied item must be included in the criterion. Inclusion of the item in the analysis does not mask the existence of DIF; however, the inclusion of other items displaying DIF will mask differential functioning of the studied item. They offer a two-step procedure for identification of potentially biased items:

> *Step 1:* Refine the matching criteria by eliminating items based on a preliminary *dif* or impact analysis
>
> *Step 2:* Use as the matching criterion the total score on all items left in the refined criterion plus the studied item--even if it is then omitted from the criterion of all other items when they are studied in turned [sic] (p. 141).

The effect of inclusion of the studied item on the identification of DIF was investigated by Clauser, Mazor, and Hambleton (1993). The authors used a three-parameter IRT model to simulate 2,000 examinee responses: 1,000 focal and 1,000 referent. Two types of ability distributions were studied: equal distributions and unequal distributions. In the former, the distribution of test scores for the referent and focal group were standardized to a mean of 0.0 and a standard deviation of 1.0. To create an unequal distribution of total scores for the focal group similar to those seen in testing situations, the mean was -1.0 and the standard deviation was maintained at 1.0. Also manipulated in the investigation were test length--lengths studied were 20, 40, and 80 items--and percent of items displaying DIF--percent of items displaying DIF were 0, 3, 8, and 20. DIF was simulated by increasing the *b* parameter for the focal group by 0.6.

Clauser, Mazor, and Hambleton (1993) found the two-step procedure recommended by Holland and Thayer (1988) to be superior or equal to the single-step procedure in identifying items with simulated DIF. The two-step process lowered Type I error. In addition, they reported that with unequal distributions, as the number of score levels were reduced, thus contaminating the matching variable, more items were incorrectly identified as displaying DIF--higher Type I error.

Zwick, Donoghue, and Grima (1993) also studied the impact of removing the studied item when investigating differential item functioning in performance assessments. The authors found that for two variations of the Mantel-Haenszel technique--Mantel and the generalized Mantel-Haenszel procedure--were useful in examining DIF for polytomously scored items. They also reported increased Type I errors increased when the studied item was omitted.

Multidimensionality

In an article that addressed multidimensional IRT issues, Ackerman (1992) demonstrated that matching focal and referent group members on total test scores proved to be inadequate when the test items measure both a valid latent ability and a nuisance variable. He noted that measurement practitioners need examine the conditional distribution of the nuisance ability at each level of the valid ability. If the distribution of the nuisance ability differs for the two groups, then there exists a potential for bias. He cited Pine (1977) to define an item as being unbiased if all examinees with the same "intended-to-be-measured" ability have an equal probability of getting the item correct. By contrast, item impact, according to Ackerman, occurs when two groups differ on abilities that are considered to be a valid part of the test construct.

Other authors have noted "nuisance" factors. Dorans, Schmitt, and Bleistein (1992) noted that matching on a total score that is contaminated by speededness would affect STD P-DIF. The authors recommended that to avoid the spurious detection of DIF it may be necessary to remove the speed component from the matching criterion. Bleistein and Schmitt (1987) found that number of items flagged as exhibiting DIF is related to the unidimensionality of the matching criterion; that is, the more nearly unidimensional the matching criterion, the fewer the number of items flagged. Camilli and Smith (1990), Ryan (1991), and Shepard et al. (1985) have observed DIF being displayed in verbally-loaded mathematics items that required a the "nuisance" ability of reading. Ryan noted that a study of DIF in mathematics items while controlling for reading would be of interest.

Ackerman proposed the use of a "validity sector" that has a specified measurement direction as described by Shealy and Stout (1989). Items that lie outside the validity sector and that are closer to a nuisance factor in the factor space would not be used for creating homogenous groupings. The validity sector contains items that load heavily on the target ability (factor) intended to be measured by the test developer; thus, the test developer must identify the items that closely align with the intended construct. Matching on ability would use the number correct from items identified as associated with the intended ability. Ackerman emphasized that matching for ability on two-dimensional data would create groups that are not homogenous and would give spurious results in a DIF analysis. If the data were unidimensional, then the matching on raw scores would not be problematical.

Clauser, Mazor, and Hambleton (1991) recommended that the matching criterion used in the Mantel-Haenszel procedure be approximately unidimensional. In their aforementioned study the authors used items from a high school proficiency test, the *New Mexico High School Proficiency Exam*. From the five life skills areas measured by the test--knowledge of community resources; consumer economics; government and law; mental and physical health; and occupational knowledge--the authors selected items that required differing abilities for solution. The abilities required by the items included reading; mathematical calculation; interpretation of tables, charts, or maps; and prior knowledge. The items were analyzed for DIF using total test score. Items were reanalyzed by matching on the test score created by pooling items from the same category (e.g. reading, prior knowledge). In essence, the authors created a subtest score for the matching criterion using items that Ackerman would argue fit in the validity sector for the intended measure.

Clauser, Mazor, & Hambleton (1991) found that the choice of criterion--subtest or total test score--had an influence on the classification of items as displaying DIF with the MH method. When the items that displayed DIF in the original analysis were analyzed in the context of similar items, nearly a third (32%) were no longer identified as displaying DIF. Clauser et. al recommended that to avoid Type I errors, test developers should screen items with similar items. The authors hypothesized that the results may be due to changes in dimensionality of the regrouped tests.

In a review of the research conducted at the Laboratory of Psychometric and Evaluative Research, Hambleton, Clauser, Mazor, and Jones, (1993) formulated guidelines for the review of items for DIF. The authors found the Mantel-Haenszel procedure to be effective in the identification of DIF under certain conditions:

> The criterion used for matching examinees must be approximately unidimensional. Both Ackerman (1982) and Clauser, Mazor, and Hambleton (1991a) have shown that substantial Type I error may result from violations of this assumption. If this assumption is in question for the test as a whole, the test may be broken down based on item content. MH analysis may then be carried out on approximately unidimensional subtests. (p. 31)

The results of Clauser et. al are similar to those of Bleisten and Schmitt (1987), who found that with a unidimensional matching criterion fewer items are flagged for DIF.

If an item is multidimensional in regard to the abilities that it measures, the previous studies indicate that the item should be eliminated when forming the matching criterion. A judgmental process can be used to determine which items align with the validity sector as Ackerman proposes or to determine which items form the content appropriate subtest, as Clauser et. al proposed. In a similar vein, it would appear promising to use factor scores formed from the dominant first factor in a factor analysis of the items as a matching criterion. Instead of creating a subtest, factor scores have the advantage that the factor loadings associated with each item give appropriate weight to each item according to its contribution to the construct. In this fashion, subjects with similar factor scores would be more homogenous for the underlying construct than those with the same raw score.

In the aforementioned study of Shepard et al. (1985), the authors examined the utility of factor scores as a matching criterion for the chi-square procedure. The investigation used the responses of 1,000 White examinees and 300 African American examinees on the mathematics test from the High School and Beyond data base. The test is a 32 item, basic skills measure. The authors reported the test contained items that involved "... simple operations, reading graphs, calculating per unit costs, or comparing rates. A few items require basic algebra ... " (p.85). Shepard et al. noted that the verbally-loaded items were most consistently identified as displaying DIF for African American examinees.

In addition to the chi-square DIF procedures, Shepard et. al compared results for the Angoff delta plot, residualized delta plot, pseudo-IRT, and one- and three- parameter IRT. The evaluative criteria for the preceding bias procedures were the unsigned and signed bias indices generated by the three-parameter IRT model. The agreement between the techniques and the IRT models was measured by calculating Spearman rank-order correlations.

To improve on the results of the chi-square procedure, Shepard et al. (1985) investigated the possibility of creating a more unidimensional matching criterion by use of factor scores. The authors factor-analyzed student responses on the mathematics test and used the first-factor score of a principal component solution as the matching criterion for the chi-square method. The authors reported that the correlations with the criterion--the unsigned and signed bias indices generated by the three-parameter IRT model--"... were substantially worse" (p. 92). Shepard et al. did not report correlations between the IRT criteria and the factor-based chi-square procedures, nor did they speculate about the evident inconsistency of the need for unidimensional matching criterion and the poor agreement between the bias indices. The demand in the current literature for a more unidimensional matching criterion, and the incongruent results of the Shepard findings led to the current study.

Research Questions

The validity of all DIF procedures depends critically on the fidelity with which the focal and referent groups are "matched," that is, on the extent to which the groups are "equal" on the ability being measured by the test items under investigation. The validity of the matching criterion comes into question as the data on which the criterion is based depart from unidimensionality. For this reason, the current investigation examined three questions related to the multidimensionality of the matching criterion:

1) *When using total scores and factor scores as matching criteria for empirical data, are the same test items flagged as displaying DIF?*

2) *Do chi-square procedures correctly identify items containing DIF when the total-score matching criterion is composed of item-correct scores from a test in which all items load on the target factor for referent group members and in which a subset of items load on the nuisance factor for focal group members? Is the identification of biased items in such tests improved by the use of factor scores as the matching criterion?*

3) *Do chi-square procedures correctly identify items containing DIF when the total-score matching criterion is composed of item-correct scores from a test in which--for referent and focal group members--a majority of items load on the target factor and in which a subset of items load on the target and nuisance factor? Is the identification of biased items in such tests improved by the use of factor scores as the matching criterion?*

CHAPTER III

METHODOLOGY

The robustness of total observed scores and the efficacy of factor scores as matching criteria were investigated in three studies. The first study involved analysis of empirical data from the administration of two nationally-administered, standardized tests: the *Graduate Management Admission Test* (GMAT) and a high-stakes mathematics achievement test for adults--hereafter referred to as the HSMAT (see Footnote 1). The empirical studies were completed to investigate the first research question: *When using total scores and factor scores as matching criteria for empirical data, are the same test items flagged as displaying DIF?* For the empirical analysis, examinee responses were sampled from a retired form of the *Graduate Management Admission Test* and a retired form of the high-stakes mathematics achievement test. The GMAT analysis was completed on the 1986-1987 (Form 71) administration of the verbal and quantitative subtests. The HSMAT analyses was completed on a 1992 administration of a mathematics subtest. An overview of the two tests and the general procedures for the empirical analyses are described below.

The latter two studies involved the completion of a series of simulations. The first series of simulations were conducted to investigate the second research question: *Do chi-square procedures correctly identify items*

[1] Permission to use the HSMAT data set was granted on the condition that the test be labeled with a generic descriptor.

containing DIF when the total-score matching criterion is composed of item-correct scores from a test in which all items load on the target factor for referent group members and in which a subset of items load on the nuisance factor for focal group members? Is the identification of biased items in such tests improved by the use of factor scores as the matching criterion? For expository purposes, in subsequent discourse this rather lengthy, but necessary, description of the factor structure will be used interchangeably with the label *Factor Structure 1.*

The second series of simulations addressed the third research question: *Do chi-square procedures correctly identify items containing DIF when the total-score matching criterion is composed of item-correct scores from a test in which--for referent and focal group members--a majority of items load on the target factor and in which a subset of items load on the target and nuisance factor? Is the identification of biased items in such tests improved by the use of factor scores as the matching criterion?* In the case of this factor structure, in subsequent discourse the label *Factor Structure 2* will be used interchangeably with the lengthier description.

The programs for use in the analyses are included in Appendix A. Data sets for the simulations study were created using a SAS code developed by Penny (1994); a description of the code is presented below.

Empirical Data Base and Methodology

Examinee responses to forms of the *Graduate Management Admission Test* (GMAT) and a high-stakes mathematics achievement test provided the empirical data for this investigation. The purpose of the *Graduate Management Admission Test* is to assist in the prediction of an examinee's performance in graduate school by measuring the ability of an examinee to "... reason with words, to use mathematical principles, and to work with concepts or abstractions in arriving at solutions to problems" (Graduate Management Admission Council, 1987, p. 4). The complete battery is composed of four verbal sections and four quantitative sections. Two of the sections--one verbal and one quantitative--serve to pilot new items or equate scores; items from these sections do not contribute to an examinee's overall score. In the remaining six operational sections, the 140 items that constitute an examinee's total raw score are presented in a five-option, multiple-choice format. The sections are timed and examinees are penalized for incorrect responses.

The verbal subtest of the GMAT is composed of 75 questions; 50 of the items focus on reading comprehension and 25 items address written expression. The quantitative subtest of the GMAT is composed of 65 questions; the items measure "... basic mathematical skills and understanding of elementary mathematical concepts as well as the ability to reason quantitatively, to solve quantitative problems, and to interpret data given in graphs, charts, or tables" (Graduate Management Admission Council, p. 5). The content of the GMAT quantitative subtest is divided among three areas: arithmetic, algebra, and geometry.

A data set was obtained for a high-stakes mathematics achievement test when preliminary factor analyses of the GMAT verbal and quantitative data revealed the two GMAT subtests to be unidimensional. Plake (personal communication, April 1994) suggested data from the administration of the mathematics test would allow investigation of a test when the first factor accounts for much less of the variance as compared to the GMAT. The HSMAT data set contains examinee classification variables, item responses, and item scores for 7242 examinees. The mathematics subtest contains 50 items which measure five content areas--measurement, algebra, geometry, number relations, and data analysis--at two cognitive levels--set-up answer and solution answer.

For the GMAT and HSMAT, the first step for the data analysis was to sample equal numbers of focal and referent examinees from the complete data sets. Since the number of African American examinees was fewer than white examinees, the number of minority examinees in the data determined the ceiling for the number of examinees sampled. The sample was selected from examinees who indicated their ethnic background as African American or White and had a complete vector of item responses. In the case of the GMAT, a sample of 4,944 examinees was selected from the pool of examinee responses; all 2,444 of the African American examinees were selected and 2,500 white examinees were sampled. For the analysis of the HSMAT data, a sample of 1526 examinees was drawn from the pool of examinee responses. The total population of 763 African American examinees was selected, and 763 of the white examinees were sampled.

Examinee responses for the GMAT verbal and quantitative subtests were scored with a SAS code developed by Harman (1994). The code compared the answer provided by the Educational Testing Service with the response of the examinee and output a vector of correct (1) and incorrect (0) responses for each examinee. The database for the HSMAT contained a vector of correct and incorrect responses for each examinee and no scoring routine was necessary.

The data sets thus created were analyzed using the factor analysis procedure in SAS. The SAS procedure output factor scores which were used as a matching criterion. Factor scores for the empirical study were generated using an orthogonal, unrotated factor solution. The factor scores were standardized to have the same mean and standard deviation as the total score. The scores were then truncated to be integer-level for the Mantel-Haenszel procedure and appended to each examinee's record.

At the culmination of the data manipulation, the GMAT verbal and quantitative and the HSMAT data sets contained a vector of 1s and 0s for each examinee, a total correct score, an unrotated factor score, and a classification variable for ethnicity. The three data sets were submitted to a series of item bias analyses to compare the results achieved when matching on total raw scores and factor scores. Throughout the analyses matching occurred at each observed total score and factor score for the test. More specifically, matching did not involve collapsing of score intervals in this investigation.

Initially, a SAS Macro program written by Harnisch (1991) and a SAS procedure outlined by Camilli and Shepard (1994) were used to complete the DIF analyses. The Mantel-Haenszel alpha was the same for each item in the two printouts, and the Harnisch program was selected for subsequent analyses due to the utility of the printout. In the Harnisch program the summary statistics include α_{MH} (ALPHAMH); α_{MH} transformed to the delta metric (DELTAMH); standardized *p*-difference (DSTD); the chi-square statistic associated with the null hypothesis for Mantel-Haenszel (CHISQMH); the associated probability for the chi-square statistic (PCHIMH); and standardized *p*-difference transformed to the delta metric(DELSTD). In addition, the Harnisch program uses D_{STD} as a flagging criteria. The flagging criteria in the Harnisch program are as follow:

DSTD	Flag
> .10	M++
> .05	M+
< -.05	M-
< -.10	M--

The flagging increases the utility of the DIF information by allowing test practitioners to readily identify potentially biased items for further review while exempting items that do not display DIF from future reviews (Dorans & Holland, 1993).

Harnisch's flagging procedure as described above was used to examine the consistency of an item's DIF status when using total scores and factor scores as matching criteria in the chi-square procedures. Agreement for the two matching criteria was calculated by examining the combined percentage of items consistently flagged as displaying DIF and items consistently not flagged as displaying DIF. Patterns in the shifts of flagging are reported for the HSMAT subtest and the GMAT verbal and quantitative subtests.

Simulation Methodology

In addition to the empirical analysis, two simulation studies were conducted. In the first study the researcher simulated a data structure in which all items loaded on the target factor for referent group members and in which a subset of items loaded on the nuisance factor for focal group members (Factor

Structure 1). This type of factor structure was presented as the first scenario in the *Introduction* where, in the case of rural and urban examinees, the rural examinees (referent group members) had a knowledge base that was absent for urban examinees (focal group members). In essence, the items with vocabulary that was unfamiliar to urban examinees served to create a nuisance factor for this group-- in this case a vocabulary factor.

In the second simulation study the researcher created a data structure in which--for focal and referent group members--a majority of items loaded on the target factor and a subset of items loaded on the target and nuisance factor (Factor Structure 2). In this study the researcher simulated the type of factor structure reflected in the *Literature Review* where several researchers (Camilli & Smith, 1990; Ryan, 1991; Shepard et al., 1985) have reported that math items with a verbal component often display DIF for African American examinees. In the following sections, a description of the general methodology for the creation of the simulated data is followed by a description of the specific methodology associated with the second and third research questions.

Methodology for Simulation of Factor Structures

A SAS program developed by Penny (1994) created the desired factor structures; an example of the program can be found in Appendix A. The general procedure involved the creation of observed scores for the focal and referent examinees. Observed scores were generated for 1,000 focal group members and 1,000 referent group members on 100 items for all the simulations; thus, any effect in the identification of DIF due to sample size or test length were held constant throughout the analyses.

The observed scores were formed by creation of three components: a true score, an error score, and a DIF score. To create the true score for members of each group, a set of item response vectors were generated which correlated with a first factor; in the preliminary steps, the data were continuous. With the exception of the unidimensional data set for the referent group in Factor Structure 1, the process was repeated for each group to create a set of response vectors that correlated with a second factor.

An error component was added to each examinee's true score to reflect a reliability coefficient of .90 for the simulated test. The level of reliability approximated the reliability found in the GMAT and HSMAT in preliminary analyses.

DIF was simulated in items by subtracting a constant from the true score of the focal group; DIF was not simulated for the referent group. An observed score was created by summing the true score, the error score, and the DIF component for each "examinee." The output was standardized to create observed scores with a mean of zero and a standard deviation of one. The standardized data for the two groups were dichotomized by specifying observed item scores that were less than or equal to zero were wrong (0) and observed item scores greater than zero were correct (1).

After the data were dichotomized, the two data sets for the focal and referent groups were concatenated. The simulated data were factor analyzed; the factor scores output; the factor scores standardized (*M* = 50, *SD* = 15) and truncated to integer-level; and the DIF analyses completed to determine if the results were similar to those found with the GMAT and HSMAT data sets.

As was true for the empirical analyses, matching occurred at each observed total score and factor score for the test.

Description of the Data Set for the Second Research Question: Factor Structure 1

In the simulations for the second study, the factor structures for the referent group and focal group were created separately. The correlation of an item with either the first factor or the second factor had a mean of .50 and a standard deviation of 0.1. The correlations were created between the continuous true scores and the factor. When error was introduced into the observed score and the data dichotomized, the correlation between the observed score and the factor loading was suppressed. The item loadings for the observed scores on the factor were generated from a population of items that had a mean loading of 0.30. The mean and standard deviation of the factor loadings were based on preliminary analyses of the empirical data and approximated the loadings found in the HSMAT and GMAT data. The factor structure for the referent group was unidimensional with all items loading on the first factor. A two-factor structure was created for the focal group with the subset of items that loaded highly on the second factor being associated with the nuisance variable. DIF was introduced to only those items that loaded on the second factor.

Methodology for the Second Research Question

The simulation analyses entailed the manipulation of two variables: factor structure and levels of DIF. The factor structure was manipulated to determine the robustness of the total score as a matching criterion in the chi-square procedures as the data set departed from unidimensionality. Three types of factor structures were investigated. For a two factor solution, the first factor accounted for 90% of the common variance, and the second factor accounted for 10% of the common variance (90/10). In the next data set the first factor accounted for 80% of the common variance, and the second factor accounted for 20% of the variance (80/20). In the third data set the first factor accounted for 70% of the common variance, and the second factor accounted for 30% of the variance (70/30). Again, all of the items associated with the second factor had DIF introduced for the focal group members. To account for the effect of the factor structure under various test conditions, the level of DIF was also investigated. Three levels of DIF were defined: -0.5 (high DIF), -0.35 (moderate DIF), and -0.20 (low DIF). These constants were subtracted from the continuous true scores (*M* = 0, *SD* = 1) of focal group members prior to dichotomizing the data.

The efficacy of factor scores for the varimax rotation and total scores as matching criteria in the chi-square procedures were compared by examining the percentage of spurious flags (false positives) and missed flags (false negatives) across the data sets. The second study involved 18 comparisons: three factor structures, three levels of DIF, and two matching criteria.

Description of the Data Set for the Third Research Question: Factor Structure 2

In the simulations for the third study, the factor structure was the same for members of the focal and referent groups. The data were composed of items that loaded primarily on the first factor and a subset of multidimensional items that loaded secondarily on the second factor. The correlation of the items with the first factor had a mean of .50 and a standard deviation of .1. The correlation of the subset of items with the second factor was determined by manipulating each item's correlation with the first factor. The procedure for creating multidimensional items is described in the following section. DIF was simulated for only those multidimensional items associated with the second factor.

Methodology for the Third Research Question

The simulation analyses for the final research question involved manipulation of three variables: the number of multidimensional items, the level of multidimensionality in items associated with the nuisance factor, and level of DIF. The number of multidimensional items and the levels of multidimensionality in those items were manipulated to examine the robustness of the total score as a matching criterion in the chi-square procedures when the items constructing the data set departed from unidimensionality. Two levels of multidimensionality were introduced; in one variation the loadings of the subset of items associated with the second factor were high (High) as shown in Equation 13:

```
ts[i] = itcor[i]*ts[102] + itcor[i]/1.5*ts[101]                    (13)
        + sqrt(1- (itcor[i]*itcor[i]*(1+1/2.25)))*ts[i];
```

where

ts[i] is the continuous true score of an item

itcor[i] is the correlation of an item with either the primary or secondary factor

ts[102] is the primary factor

ts[101] is the secondary factor

If the items were equally correlated with the two factors, the divisor in the second correlation term (itcor[i]/1.5) would be one. As the divisor increases, the secondary factor will correlate less highly with the items. In another set of analyses the loadings of the items on the second factor were low (Low), as demonstrated in Equation 14:

```
    ts[i] = itcor[i]*ts[102] + itcor[i]/3*ts[101]                    (14)
            + sqrt(1-(itcor[i]*itcor[i]*(1+1/9)))*ts[i];
            where ts[i], itcor[i], ts[101], ts[102] have been previously defined.
```

In addition to the level of multidimensionality, the investigator also manipulated the number of items that were multidimensional and the level of DIF to account for the effect of multidimensional items under various test conditions. In one series of analyses the subset of items that were multidimensional was 10 items (90/10), and another 20 items (80/20), and in a final analysis 30 items (70/30). In addition, three levels of DIF were defined: -.5 (high DIF), -.35 (moderate DIF), and -.20 (low DIF). As stated previously, DIF was introduced for focal group members on all items that comprised the multidimensional subset.

The simulated data were factor analyzed as described in the previous analyses. The efficacy of factor scores for the varimax rotation and total scores as matching criteria in the chi-square procedures was compared by examining the percentage of spurious flags (false positives) and missed flags (false negatives) across the data sets. The investigation involved 36 comparisons: two levels of multidimensionality, three levels of multidimensional items, three levels of DIF, and two matching criteria.

CHAPTER IV

RESULTS

In this chapter the results of the analyses will be presented in the order presented in the methodology section. The results of the use of total scores and factor scores as matching criteria for the two empirical data sets will be presented first, followed by the results for the simulations.

Results for the Empirical Studies

The first research question investigated was: *When using total scores and factor scores as the matching criteria, are the same test items flagged as displaying DIF?* The analysis used data from previous administrations of the GMAT and HSMAT. In the following sections, information is provided about the scores of the focal and referent groups on each test, followed by a description of the factor structure of the data. The consistency of the two matching criteria--total scores and factor scores--is presented next.

Description of GMAT and HSMAT Data Sets

Table 2 gives the mean and standard deviation for each group--African American and White examinees--and the combined groups for the GMAT verbal and quantitative subtests and the HSMAT test. The mean raw score for White examinees for the GMAT verbal subtest was 43.49 (*SD* = 10.60), and for African American examinees the mean raw score was 31.34 (*SD* = 10.43). The combined mean for the two groups was 37.48 (*SD* = 12.14). The reliability (coefficient alpha) of the GMAT verbal subtest based on a sample of 4,944 examinees was 0.90.

For the quantitative subtest of the GMAT the mean raw score for African American examinees was 23.87 (*SD* = 8.66), and the mean raw score for White examinees was 35.86 (*SD* = 9.94). The combined mean for the two groups was 29.93 (*SD* = 11.09). The reliability of the GMAT quantitative subtest based on the sample of 4,944 examinees was α = .91.

Table 2

Mean and Standard Deviation of African American and White Examinees on the GMAT and HSMAT Subtests

	GMAT Verbal		GMAT Quantitative		HSMAT Mathematics	
Group	Mean	SD	Mean	SD	Mean	SD
African American	31.34	10.43	23.87	8.66	23.98	7.82
White	43.49	10.60	35.86	9.94	27.45	9.36
Total	37.48	12.14	29.93	11.09	25.71	8.80

The factor routine in SAS was used to examine the factor structure of the GMAT verbal subtest and the GMAT quantitative subtest; examination of the factor plots (see Figures 1 and 2) and the scree plots (see Appendix B) indicated that each subtest appeared to have a dominant first factor. The first, unrotated factor for the verbal subtest accounted for 83.7% of the common variance in the data; the second factor accounted for an additional 16.3% of the variance.

In a separate factor solution for African American examinees, the first factor for the verbal subtest accounted for 76.7% of the common variance in the data prior to rotation; the second factor accounted for an additional 23.3% of the variance in the verbal data set. For White examinees the first factor of the verbal subtest accounted for 81.1% of the variance in the data prior to rotation; the second factor accounted for an additional 18.9% of the variance. A visual inspection of the factor plots for the two groups (see Appendix B for factor plots) for the verbal subtest revealed similar data structures, a conclusion that was confirmed by the fact that the correlation of the verbal loadings on the first factor for the two groups was .90.

```
                             FACTOR1
                                1

                               .9

                               .8

                               .7

                         W     .6
                     W R  U
                    U S  S QTP .5
                      T      QW V  M   F
                      V        .4   HIEY
                           R    SQLDD Z
                          X    .TGEIZB D
                              XYMKNB B
                               .NCHAG
                               OM                               F
                               U1                               A
                                                                C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0T
                                                                O
                              -.1                               R
                                                                2
                              -.2
```

Figure 1. Plot of Factor Loadings for African American and White Examinees on the GMAT Verbal Test

Prior to rotation, the first factor for the quantitative subtest of the GMAT accounted for 88.2% of the variance in the data, while the second factor accounted for an additional 11.8% of the variance in the data. A separate factor solution for African American examinees revealed the first factor of the quantitative subtest accounted for 83.4% of the variance in the data prior to rotation; the second factor accounted for an additional 16.6% of the variance in the quantitative data set. For White examinees the first factor of the quantitative test accounted for 88.2% of the variance in the data prior to rotation; the second factor accounted for an additional 11.8% of the variance in the quantitative data set. Again, a visual inspection of the factor plots for the two groups (see Appendix B) revealed similar data structures. The correlation of the quantitative loadings on the first factor for the two groups was .67.

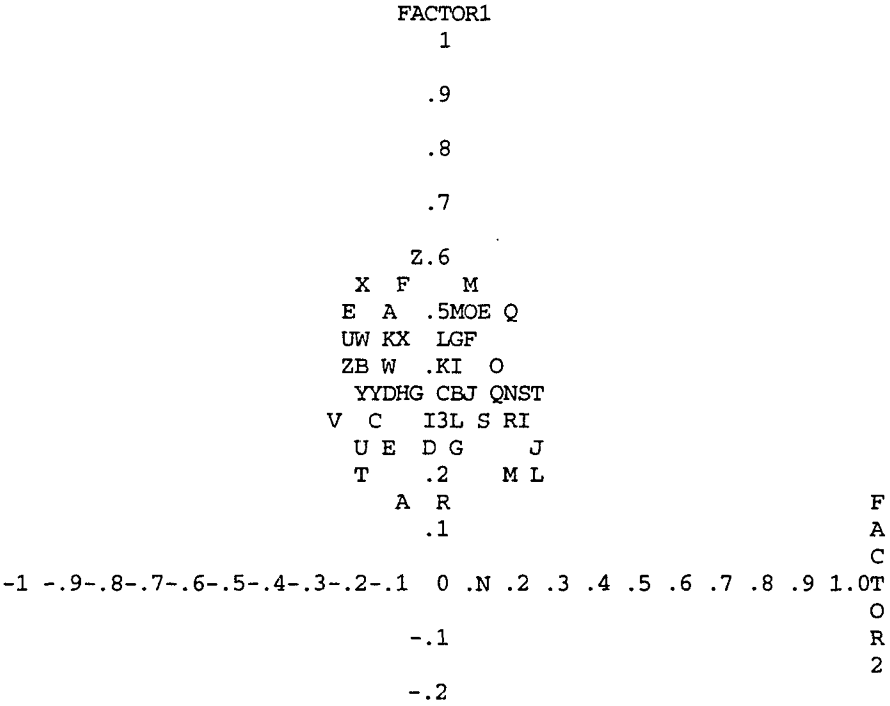

Figure 2. Plot of Factor Loadings for African American and White Examinees on the GMAT Quantitative Test

Inasmuch as the factor scores for GMAT examinees correlated very highly with their raw scores (r = .99 for verbal and r = .995 for quantitative), the use of factor scores as a matching criterion appeared to offer little improvement over total raw score as a matching criterion when measures are unidimensional.

For the mathematics subtest of the HSMAT the mean raw score for White examinees was 27.45 (*SD* = 9.36), and for African American examinees the mean raw score was 23.98 (*SD* = 7.82). The combined mean for the two groups was 25.71 (*SD* = 8.80). The mean score of the African American examinees was approximately half a standard deviation below that of White examinees. The reliability of the HSMAT mathematics subtest based on a sample of 1526 examinees was α = .88.

Analysis of the HSMAT data showed the first, unrotated factor to account for 65.8% of the common variance in the data; the second factor accounted for an additional 34.2% of the variance. Examination of the factor plot (Figure 3) and the scree plot (see Appendix B for scree plot) indicated a two-factor structure.

In a separate factor solution for African American examinees, the first factor of the HSMAT accounted for 86.4% of the variance in the data prior to rotation; the second factor accounted for an additional 13.6% of the variance in the data set. The factor plot is depicted graphically in Figure 4. The HSMAT was clearly unidimensional for African American examinees.

```
                             FACTOR1
                                1

                               .9

                               .8

                               .7
                                 P
                               .6XB
              C             AY SV
                          LH T .DTWN
                         M    F  EJ
                       N    RI .C
                           OIH
                            GU .3
                                         A  E    J
                               .2            O L
                                         DG  W  Z               F
                               .1   BF K     U                  A
                                              M                 C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 Q4SP5 .6 .7 .8 .9 1.0T
                                                                O
                              -.1          V                    R
                                                                2
```

Figure 3. Plot of Factor Loadings for African American and White Examinees on the HSMAT

```
                             FACTOR1
                                1

                               .9

                               .8

                               .7

                               .6
                            X   P
                         J    B.5
                              E  S  T
                       L       A4   Y
                           Z ITDV  C
                         U  J  .3 H GLS
                             W P  IR
                               QF CKF K
                                 BUN    M                       F
                               .1V                              A
                                                                C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1G 0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0T
                                                                O
                              -.1                               R
                                                                2
```

Figure 4. Plot of Factor Loadings for African American Examinees on the High-Stakes Mathematics Test

Somewhat surprisingly, for White examinees the high-stakes mathematics test appeared to be two-dimensional (see Figure 5). The first factor of the mathematics test accounted for 72.8% of the variance in the data prior to rotation; the second factor accounted for an additional 27.2% of the variance. Two replications of this result indicated that the structure was stable across samples. In the first replication the first factor accounted for 70.9% of the common variance, and in the second replication the first factor accounted for 70.5% of the variance. As proved true with the GMAT data, unrotated factor scores for the HSMAT data correlated highly (.95) with raw scores.

```
                             FACTOR1
                                1

                               .9
                    C
                               .8
                                   P
                               .7
                              HY  S B
                            N  .TF VX
                                K   EX W
                               RIR   J  N
                             GOI QC
                               .4

                               .3

                               .2
                                                                F
                               .1          A                    A
                                                                C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .B .3F.4E.5L.J .7 .8 .9 1.0T
                                            G O                 O
                              -.1      D   KW  U  Z             R
                                                                2
                              -.2         Q  M
                                             S P
```

Figure 5. Plot of Factor Loadings for White Examinees on the HSMAT

DIF Results for the GMAT Verbal Items

The consistency with which GMAT verbal items were flagged using total raw scores and unrotated factor scores as matching criteria revealed high levels of agreement. In the following sections the pattern of the changes will be presented relative to the change that occurred when moving from the total-score matching criterion to the factor-score criterion. As shown in Table 3, sixty-three of the 75 items (84%) were consistently identified when using the two criteria. Of the 12 inconsistent items (16%), six items were associated with the referent group. Three of the items changed from intermediate DIF (M+) to no DIF for the referent group, and three items changed from no DIF to intermediate DIF (M+). For the focal group, four items changed from negligible DIF to intermediate DIF (M-), one item changed from intermediate DIF (M-) to negligible DIF, and one item changed from substantial DIF (M--) to intermediate DIF (M-).

Table 3

Pattern of Changes in Flagging Designations for GMAT Verbal Items with Total Score and Factor Score as Matching Criteria

Total Score	Factor Score	Number of items
+ +	+	0
+	No flag	3
No flag	-	4
-	- -	0
- -	-	1
-	No flag	1
No flag	+	3
+	+ +	
No change		63

DIF Results for the GMAT Quantitative Items

For the GMAT quantitative subtest, the percent agreement of items that were flagged for the matching criteria of total score versus factor score resulted in even higher levels of agreement than the GMAT verbal subtest. In Table 4, it can be seen that 60 of the 65 items (92%) were consistently identified when using the two criteria. Of the five inconsistently flagged items, four were associated with the referent group. Three items changed from intermediate DIF (M+) to negligible DIF, and one item changed from negligible DIF to intermediate DIF (M+). For the focal group, one item shifted from intermediate DIF (M-) to no DIF.

Table 4

Pattern of Changes in Flagging Designations for GMAT Quantitative Items for Total Score and Factor Score as Matching Criteria

Total Score	Factor Score	Number of items
+ +	+	0
+	No flag	3
No flag	-	0
-	- -	0
- -	-	0
-	No flag	1
No flag	+	1
+	+ +	0
No change		60

DIF Results for the HSMAT Items

The patterns of differential flagging for the HSMAT data are summarized in Table 5. Thirty-eight of the 50 items (76%) were consistently identified using the two criteria. Of the 12 items (24%) that changed flagging designation, three items changed from substantial DIF (M++) to intermediate

DIF (M+) for the referent groups, and four items changed from intermediate DIF (M+) to no DIF. In the case of focal group members, one item changed from substantial DIF (M--) to intermediate DIF (M-), while four items changed from intermediate DIF (M-) to no DIF.

Of the five items flagged using total raw score as the matching criterion, three items flagged as substantially biased against the referent group changed to intermediate DIF. One of the two items flagged with substantial DIF for the focal group shifted from substantial DIF (M--) to slight DIF (M-), and one item was consistently flagged as displaying substantial DIF.

Table 5

Pattern of Changes in Flagging Designations of HSMAT Items for Total Score and a Factor Score as Matching Criteria

Total Score	Factor Score	Number of items
+ +	+	3
+	No flag	4
No flag	-	0
-	- -	0
- -	-	1
-	No flag	4
No flag	+	0
+	+ +	0
No change		38

Generally, the DIF analyses of the empirical data demonstrated that the two matching criteria produce consistent results. The studies also indicated a trend from highly consistent identification of DIF for unidimensional data sets such as the GMAT quantitative subtest to less consistency for the multifactorial HSMAT. For the few items in the GMAT subtests which changed flagging designations, there did not appear to be a pattern of changes when moving from the total-score matching criterion to the factor-score matching criterion. The pattern of changes in the flagging designations for the HSMAT data set indicated less bias in items for referent group and focal group members.

Results for the Simulations

As noted earlier, the simulations were conducted to investigate the following two research questions:

Do chi-square procedures correctly identify items containing DIF when the total-score matching criterion is composed of item-correct scores from a test in which all items load on the target factor for referent group members and in which a subset of items load on the nuisance factor for focal group members? Is the identification of biased items in such tests improved by the use of factor scores as the matching criterion?

Do chi-square procedures correctly identify items containing DIF when the total-score matching criterion is composed of item-correct scores from a test in which--for referent and focal group members--a majority of items load on the target factor and in which a subset of items load on the target and nuisance factor? Is the identification of biased items in such tests improved by the use of factor scores as the matching criterion?

Description of Simulated Data Sets

Two types of factor structures were simulated to investigate the above research questions. First, a "test" that was unidimensional for the reference group, but multifactorial for the focal group was simulated (Factor Structure 1). The example cited at the beginning of chapter 1 exemplifies this situation. A verbal analogies test that contains a number of items that would be familiar to examinees raised on a farm but unfamiliar to examinees raised in more urban settings could well be unidimensional for the former examinees, but multifactorial for the latter. That is, the test would be a relatively "pure" measure of verbal analogical reasoning for rural examinees, but would reflect both "vocabulary" and "verbal analogical reasoning ability" for urban examinees. In the case of focal group members, different degrees of factorial complexity were introduced into the "test" by simulating data with 10, 20, and 30 items loading on a second factor. As previously mentioned, these degrees of factorial complexity will be denoted as "90/10, " 80/20," and "70/30," respectively. An example of the simulated 80/20 factor structure is included in Appendix B.

In the second simulation, a "test" that contained inherently multidimensional items was simulated (Factor Structure 2). A quantitative reasoning test with some of the items posed as "word problems" is a case in point. Such a test, although intended to measure quantitative reasoning, necessarily reflects to some extent examinees' verbal ability. To effect the simulation, a subset of 10, 20, and 30 items on a 100 item test were simulated to load on the first and second factor for all examinees (i.e., focal and referent group members). Again, within the context of the discussion, these factor structures will hereafter be denoted as "90/10," 80/20," and "70/30," respectively. The level of multidimensionality was also manipulated; in one simulation the multidimensional items were created to load heavily on both factors (High). Examples of the types of factor structures thus created are included in Appendix B. In the second simulation, the multidimensional items were formed to load primarily on the first factor and secondarily on the second factor (Low). The types of factor structures thus created are also included in Appendix B.

DIF Results for Matching Criteria in Factorially Complex Tests

The total-score and factor-score matching criteria were used with the DIF procedures to investigate their efficacy in the identification of biased items. The factor scores were formed in the preliminary simulation from rotated, orthogonal solutions for all the data sets. The decision to use rotated solutions was based on the high correlation between the total score and factor score seen in the empirical studies. In part, the rotated solution was used in an attempt to "tease" out the multidimensionality of the items.

DIF Results for Factor Structure 1

For data sets where the first factor accounts for approximately 90% of the common variance, total score and a rotated factor score performed almost equally well as matching criteria. As can be seen in Table 6 at the three levels defined for DIF (-.20, -.35, and -.50), when total score is used as the matching criterion in a data set that is primarily unidimensional (90/10), none of the items were spuriously flagged (false positives) at the lowest levels of DIF and only 1% were flagged at the -.50 level. The factor score from a rotated, orthogonal solution did not spuriously flag any items. As seen in Table 7 at the lowest level of DIF (-.20), use of the total-score matching criterion resulted in 10% of the biased items being missed (false negatives), and the use of the factor-score matching criterion resulted in the identification of all biased items.

When the first factor accounted for approximately 80% of the common variance, the number of items that were spuriously flagged increased in tandem with the level of DIF. The total score spuriously flagged from 1% of the items at low levels of DIF to 24% of the items at the highest level of DIF. The factor score did not spuriously flag any items. Total scores as a matching criterion resulted in 25% of the items being missed at low levels of DIF, while the factor score resulted in 10% false negatives.

Table 6

Percentage of Spurious Flags (False Positives) for Factor Structure 1

	10 Items		20 Items		30 Items	
Level of DIF	Total Score	Factor Score	Total Score	Factor Score	Total Score	Factor Score
-.20	0	0	1	0	7	0
-.35	0	0	6	0	27	0
-.50	1	0	24	0	63	0

Table 7

Percentage of Missed Flags (False Negatives) for Factor Structure 1

	10 Items		20 Items		30 Items	
Level of DIF	Total Score	Factor Score	Total Score	Factor Score	Total Score	Factor Score
-.20	10	0	25	10	53	13
-.35	0	0	0	0	0	0
-.50	0	0	0	0	0	0

The trends noted for the previous data structure became more pronounced when the factor structure was such that the first factor accounted for approximately 70% of the common variance and the second factor accounted for the remaining 30% of the common variance. In Table 6, the total score is shown to spuriously flag from 7% of the items at low levels of DIF to 63% of the items at the highest level of DIF. Again, the factor score from a rotated, orthogonal solution did not result in false positives. Total scores as a matching criterion resulted in 53% of the items being missed for low levels of DIF, while the factor score resulted in 13% of the biased items being missed.

DIF Results for Factor Structure 2

In a data set composed of inherently multidimensional items, as the factor structure departed from unidimensionality, the number of spuriously flagged items increased when using the total-score matching criterion. The trend towards improved identification of DIF with the use of the rotated factor score for Factor Structure 1 was reversed when used in Factor Structure 2 to identify items. In Table 8 it can be seen that as the number of multidimensional items that loaded highly on both factors increased, the percentage of false positives went from 0% to 26% when using the factor-score matching criterion. Thus, with the exception of the 70/30 data structure, as compared to the total-score matching criterion the factor-score matching criterion resulted in even a larger number of items being spuriously flagged when data sets contained highly multidimensional items.

The percentage of false negatives increased from 0 to 57% when using total score as the matching criterion (see Table 9). In the case of items with high levels of multidimensionality, the use of factor score as a matching criterion resulted in all simulated biased items being correctly identified.

The use of the factor-score matching criterion resulted in the number of spuriously flagged items increasing from a low of 0% to a high of 73% for items with low levels of multidimensionality (See Table 8). Again the percentage of false positives was higher for the total-score matching criterion than the factor-score matching criterion. Also, with the exception of the 90/10 factor structure, for both matching criteria items with low levels of multidimensionality were incorrectly flagged at a higher rate than items with high levels of multidimensionality. As shown in Table 9, at the lowest level of DIF (-.20), use of the total score as the matching criterion resulted in more false negatives than when the matching criterion was the factor score. For the total-score matching criterion, as the number of multidimensional items increased a concurrent number of items failed to be flagged as displaying DIF.

Table 8

Percentage of Spurious Flags (False Positives) for Factor Structure 2

		10 Items		20 Items		30 Items	
Level of Multidimensionality	Level of DIF	Total Score	Factor Score	Total Score	Factor Score	Total Score	Factor Score
	-.20	0	0	0	0	0	0
High	-.35	0	0	1	5	11	4
	-.50	0	20	4	26	56	1
	-.20	0	0	0	0	0	0
Low	-.35	0	0	0	3	19	19
	-.50	0	17	14	33	69	73

Table 9

Percentage of Missed Flags (False Negatives) for Factor Structure 2

		10 Items		20 Items		30 Items	
Level of Multidimensionality	Level of DIF	Total Score	Factor Score	Total Score	Factor Score	Total Score	Factor Score
	-.20	0	0	35	0	57	0
High	-.35	0	0	0	0	0	0
	-.50	0	0	0	0	0	0
	-.20	20	10	30	0	63	0
Low	-.35	0	0	0	0	3	0
	-.50	0	0	0	0	0	0

Follow-up Simulation Analyses

The above results suggest that for items that are inherently multidimensional, matching on factor scores results in an unusually large number of spuriously flagged items (i.e., false positives). To further investigate this phenomenon, a series of additional simulations were undertaken using two adjusted factor scores. A "purified factor score" was computed as any ordinary factor score, except that biased items were eliminated from the computation. A "factor-based score" was created by simply summing item scores for all unbiased items. (It should be noted that the factor-based score could just as accurately be called a "purified total score." For real data sets, items loading above some pre-specified value on the nuisance factor or factors would be eliminated, and examinees would be matched on their total score on the remaining items.) Note that since the resulting matching criterion is unidimensional, no factor rotation is necessary.

In order to examine how these adjusted factor scores compare to total raw score as matching criteria, a Monte Carlo simulation with a least 100 replications would be desirable. However, a single replication required construction of 27 data sets that were each analyzed separately for DIF. From start to finish, one run took approximately six hours to complete. For this reason, a Monte Carlo was deemed impractical. To glean at least a preliminary notion of the efficacy of the two adjusted factor scores as matching criteria, five replications were conducted. Summary information of the results are shown in Tables 10 through Table 13. Complete results for each replication are included in Appendix C.

Follow-up DIF Results for Factor Structure 1

As previously seen in the first series of simulations, as the data set departed from unidimensionality the number of items that were spuriously flagged increased when using the total-score matching criterion. As shown in Table 10, as the data structure departed from unidimensionality, and as the level of DIF increased, the number of false positives increased. On a test where the first factor accounted for approximately 90% of the variance, the percentage of spuriously flagged items ranged from less than 1% to 2%. At the other extreme, when the first factor accounted for 70% of the common variance, the percentage of spuriously flagged items ranged from 3% for low levels of DIF to 59% for high levels of DIF. Factor scores and factor-based scores spuriously flagged items less than 1 percent of the time across factor structures and DIF levels.

Using total score as the matching criterion, the percentage of false negatives increased as the data set contained more items with DIF (see Table 11). At the lowest level of DIF, the percentage of biased items that failed to be identified with the total-score matching criterion increased from 8% in a

90/10 factor structure to 39% for a 70/30 factor structure. The largest percentage of biased items failed to be flagged at the lowest level of DIF. At a moderate level of DIF (-.35) the total-score matching criterion failed to identify between one and two percent of the biased items for the five replications. When using purified factor scores or factor-based scores as the matching criterion, the average percentage of false negatives ranged between one and two percent.

Table 10

Average Percentage of Spurious Flags (False Positives) for Factor Structure 1

	10 Items			20 Items			30 Items		
Level of DIF	Total Score	Purified Factor Score	Factor-based Score	Total Score	Purified Factor Score	Factor-based Score	Total Score	Purified Factor Score	Factor-based Score
-.20	<1	<1	<1	1	0	0	3	<1	<1
-.35	<1	<1	<1	7	0	0	26	<1	<1
-.50	2	<1	<	20	0	0	59	<1	<1

Table 11

Average Percentage of Missed Flags (False Negatives) for Factor Structure 1

	10 Items			20 Items			30 Items		
Level of DIF	Total Score	Purified Factor Score	Factor-based Score	Total Score	Purified Factor Score	Factor-based Score	Total Score	Purified Factor Score	Factor-based Score
-.20	8	2	2	28	1	2	39	1	1
-.35	2	0	0	0	0	0	1	0	0
-.50	0	0	0	0	0	0	0	0	0

Follow-up DIF Results for Factor Structure 2

Once again, as the data set departed from unidimensionality the number of items that were spuriously flagged increased when using total score as the matching criterion. As shown in Table 12, the trend for increasing number of spurious flags was consistent across data structure and the amount of DIF simulated. In a data set where only 10 of the items displayed high levels of multidimensionality, no items were spuriously flagged for the five replications. At the other extreme, when 30 items with high levels of multidimensionality were included in the data, the percentage of false positives for the total-score matching criterion ranged from 0% for low levels of DIF to 64% for high levels of DIF. At high levels of multidimensionality, the purified factor score and factor-based score were not associated with any false positives.

Similar trends occurred for low-level multidimensional items. In a data set where only 10 of the items displayed low levels of multidimensionality, no items were spuriously flagged for the five replications. With low-level multidimensional items composing a subset of 30 items, the percentage of false positives using the total-score matching criterion ranged from 0% at low levels of DIF to 67% for high levels of DIF. At low levels of multidimensionality, no items were spuriously flagged when purified factor score and factor-based score were used as matching criteria.

For data sets with subsets of high-level multidimensional items, the percentage of biased items that failed to be flagged using the total-score matching criterion increased as the data set contained more items with DIF (see Table 13). For the high-level multidimensional items, at the lowest level of DIF the percentage of false negatives increased from 2% in a 10 item subset to 59% for a 30 item subset. For the low-level multidimensional items, at the lowest level of DIF the percentage of biased items that failed to be identified increased from 6% in a 10 item subset to 55% for a 30 item subset. No false negatives occurred at moderate and high levels of DIF for the five replications when using the total-score matching criterion. Also, no biased items were missed when purified factor score and factor-based score were used as matching criteria in data with low or high levels of multidimensionality.

Table 12

Average Percentage of Spurious Flags (False Positives) for Factor Structure 2

		10 Items			20 Items			30 Items		
Level of Multidimensionality	Level of DIF	Total Score	Purified Factor Score	Factor-based Score	Total Score	Purified Factor Score	Factor-based Score	Total Score	Purified Factor Score	Factor-based Score
	-.20	0	0	0	0	0	0	0	0	0
High	-.35	0	0	0	<1	0	0	13	0	0
	-.50	0	0	0	6	0	0	64	0	0
	-.20	0	0	0	0	0	0	0	0	0
Low	-.35	0	0	0	0	0	0	15	0	0
	-.50	0	0	0	8	0	0	67	0	0

Table 13

Average Percentage of Missed Flags (False Negatives) for Factor Structure 2

		10 Items			20 Items			30 Items		
Level of Multidimensionality	Level of DIF	Total Score	Purified Factor Score	Factor-based Score	Total Score	Purified Factor Score	Factor-based Score	Total Score	Purified Factor Score	Factor-based Score
	-.20	2	0	0	21	0	0	59	0	0
High	-.35	0	0	0	0	0	0	0	0	0
	-.50	0	0	0	0	0	0	0	0	0
	-.20	6	0	0	21	0	0	55	0	0
Low	-.35	0	0	0	0	0	0	0	0	0
	-.50	0	0	0	0	0	0	0	0	0

Follow-up HSMAT Analysis

The contamination of matching criterion with multidimensional items makes the use of total scores problematic. The use of purified factor scores or factor-based scores appears to offer the necessary unidimensional matching criterion. It is unreasonable to assume that test developers will perfectly identify problematic items. To simulate the manner in which a test practitioner might operationalize the purification of the factor score and the factor-based score, the HSMAT data were reanalyzed using purified matching criteria.

The HSMAT was selected since the structure of the data reflected the structure of the simulated data for urban and rural examinees (Factor Structure 1). Item loadings for the two groups can be found in Appendix D. Items with high, positive loadings on the second factor were removed from the analysis. The 18 items removed were 1, 2, 4-7, 10-12, 13, 17, 21, 23, 27, 40, 43, 46, and 49. The loadings on the second factor for the items ranged from .21902 to .61162. After removing items that did not load on the same factor for both groups, the combined data set was submitted to a two-factor, non-rotated orthogonal solution.

The removal of the contaminating items resulted in the correlation of item loadings for White and African American examinees changing from .10 for the complete set of 50 items to .18 for the subset of 32 items. For the combined group, the first, unrotated factor accounted for 75.9% of the common variance in the selected HSMAT items; the second factor accounted for an additional 24% of the variance. In the separate factor solution for African American examinees, the first factor for the purified test accounted for 84.4% of the variance in the data prior to rotation, and for White examinees the first factor of the mathematics test accounted for 83.0% of the variance in the data prior to rotation. Even though 18 items were removed from the analysis, the factor plot for the combined group indicated that some items highly related to the second factor were not removed (see Figure 6).

In Table 14 through Table 16 the results of the DIF analyses have been summarized. Note that when using total score and a purified factor score as the matching criteria, twenty-two percent of the items changed flagging designation (see Table 14). When the matching criteria were the fifty-item factor score and a purified factor score, 12% of the items changed flagging designation (see Table 15). Finally, when the matching criteria were a purified factor score and a factor-based score, 10% percent of items were inconsistently identified (see Table 16).

As mentioned previously, the factor-based score could be labeled a purified total score; and, thus, the factor-based and purified factor score comparison is analogous to the total score and factor score comparison made at the beginning of the *Results* section. When the two matching criteria were total score and factor score, 24% of the items were inconsistently identified--

```
                            FACTOR1
                                1

                               .9

                               .8

                               .7
                 A                  F
                               .6
                            X   L H KN
                             B .I  J
                         D       R  QF
                              GCB O  V
                              ZT
                             S .3

                               .2
                                                                F
                               .1            E                  A
                                               M                C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0T
                                             Y                  O
                              -.1             A                 R
                                           E                    2
                              -.2
```

Figure 6. Factor Plot for African American and White Examinees with Purified Item Set

compared to 10% for the purified analogs. It appears that the creation of a more unidimensional matching criterion resulted in more consistent flagging.

Table 14

Pattern of Changes in Flagging Designations of HSMAT Items for Total Score and a Purified Factor Score as Matching Criterion

Total Score	Purified Factor Score	Number of items
+ +	+	3
+	No flag	4
No flag	-	0
-	- -	2
- -	-	1
-	No flag	1
No flag	+	0
+	+ +	0
No change		39

Table 15

Pattern of Changes in Flagging Designations of HSMAT Items for Factor Score and a Purified Factor Score as Matching Criterion

Factor Score	Purified Factor Score	Number of items
+ +	+	0
+	No flag	1
No flag	-	2
-	- -	2
- -	-	0
-	No flag	0
No flag	+	1
+	+ +	0
No change		44

Table 16

Pattern of Changes in Flagging Designations of HSMAT Items for a Purified Factor Score and a Factor-based Score as Matching Criterion

Purified Factor Score	Factor-Based Score	Number of items
+ +	+	0
+	No flag	0
No flag	-	1
-	- -	2
- -	-	1
-	No flag	0
No flag	+	1
+	+ +	0
No change		45

CHAPTER V

DISCUSSION

In first section of this chapter a summary of the results will be presented. Following the summary, the implications for future research will be explored.

Summary of Results for Empirical Studies

The results of the empirical studies indicated the consistency of flagging for the total-score and factor-score matching criterion was affected by the dimensionality of the data. The total-score and factor-score matching criteria flagged items differently as the data set became increasingly multidimensional; however, in the case of tests that are primarily unidimensional, the two matching criterion resulted in essentially the same flagged items.

The GMAT verbal and quantitative subtests appeared to be primarily unidimensional and displayed similar factor structures for African American and White examinees. In the case of the GMAT quantitative subtest, the first factor accounted for 88.2% of the common variance; and 92% of the items received the same flag designations across the two matching criterion. For the GMAT verbal subtest, the first factor accounted for 83.7% of the common variance, and 84% of the items received the same flag designation.

The factor structure for the high-stakes mathematics test (HSMAT) appeared to be multifactorial, and separate factor plots for African American and White examinees revealed different factor structures for the two groups. Somewhat surprisingly, the factor structure for White examinees appeared to have two factors, while the test appeared to be primarily unidimensional for

African American examinees. The HSMAT data set had a first factor that accounted for 65.8% of the common variance. For this apparently two-factor data set, 76% of the items were consistently flagged. For the three empirical tests there appeared to be a trend toward lower levels of consistency for the two matching criteria as the data departed from unidimensionality.

Summary of Results for Simulated Studies

The inconsistency of the two matching criteria indicated a need to explore the efficacy of total-score and factor-score matching criteria as a test departed from unidimensionality. In a series of simulations, tests were created to model two different factor structures. In the first factor structure (Factor Structure 1), the items for the referent group members were formed to load solely on the first factor. In the case of focal group members, different degrees of multidimensionality were introduced into the "test" by correlating the factor loadings of 10, 20, or 30 items with a second factor.

In a second series of simulations, the majority of the 100 items for the referent and focal group members were formed to load on the first factor. A subset of 10, 20, or 30 multidimensional items were created to load on the first and second factor for all examinees--focal and referent group members (Factor Structure 2). The level of multidimensionality was also manipulated. In one series of analyses, the multidimensional items were created to load on both factors, and in the second series of simulations the multidimensional items were created to load primarily on the first factor with a minor loading on the second factor. In both factor structures, differential item functioning was simulated for focal group members in the subset of items associated with the nuisance factor. The two types of factor structures were retained in the preliminary investigation and the final series of simulations.

Summary of Results for Factor Structure 1

For tests composed of items loading on one of two factors (Factor Structure 1), the factor score from a rotated, orthogonal solution as a matching criterion resulted in no spurious flagging of any items for the 90/10, 80/20, and 70/30 tests. When the simulated test was primarily unidimensional (90/10), total score and the rotated factor score performed almost equally well as matching criteria. None of the items were spuriously flagged at the lowest levels of DIF and only 1% were flagged at the -.50 level for the total-score matching criterion. As the simulated test departed from unidimensionality, use of the total score as a matching criterion resulted in increasingly larger numbers of items being spuriously flagged.

In the preliminary investigation, for tests that were primarily unidimensional (90/10), the two matching criteria resulted in biased items being missed at only the lowest level of DIF (-.20). The use of the total score as the matching criterion resulted in greater numbers of false negatives as the test became more factorially complex. The use of the factor score as the matching criterion also resulted in larger numbers of biased items being missed as the test became more factorially complex; however, the percentage of false negatives was lower for factor scores. Thus, compared to the total-score matching criterion, the factor score appeared to improve the identification of biased items.

Summary of Results for Factor Structure 2

The preliminary results for the factor structure containing inherently multidimensional items (Factor Structure 2) called into question the efficacy of the unrotated factor score as a matching criterion. For primarily unidimensional tests (90/10) in which 10% of the items were constructed to have high levels of multidimensionality, none of the items were spuriously flagged with the total-score matching criterion; however, the factor-score matching criterion spuriously flagged 20% of the items at the -.50 level. As the tests increased in factorial complexity, the number of false positives increased for the total-score matching criterion. The factor score, however, spuriously flagged even more items as the data set became more factorially complex.

For the data set composed of items with high levels of multidimensionality, the use of the two matching criteria resulted in biased items being missed at only the lowest level of DIF (-.20). As the test became increasingly more complex, use of the total-score matching criterion resulted in more false negatives. No items were missed when the factor score was used as the matching criterion for tests composed of items with high levels of multidimensionality.

Tests with low-level multidimensional items followed the same pattern of more false positives as the factor structure departed from unidimensionality. As was the trend for the tests with high-level multidimensional items, for the low-level items the factor score was associated with higher levels of spurious flagging than the total score. In more factorially complex tests, use of the total score as a matching criterion resulted in increasing numbers of false negative errors, that is, biased items that were not flagged. By contrast, there were few false negatives when factor scores were used as the matching criterion.

While use of the factor score improved the identification of items at low levels of DIF, the factor-score matching criterion was associated with a higher level of spurious flagging than the total-score matching criterion. Neither criteria--total score or factor score--appeared to function appropriately as the factor structure departed from unidimensionality.

The preliminary simulation results summarized above prompted a set of additional analyses designed to "purify" the matching criterion. The results of these analyses will be briefly summarized.

Summary of Follow-up Simulation Analyses

To create a "purified factor score" the factor score was computed as any ordinary factor score, except that biased items were eliminated from the computation. A "factor-based score" was created by summing item scores for all unbiased items. (The factor-based score could just as accurately be called a "purified total score.") True to the pattern detected in the preliminary analyses, as the data structure departed from unidimensionality, and as the level of DIF increased, the number of spurious flags proliferated. This trend was seen in Factor Structure 1 and Factor Structure 2.

Summary of Follow-up Results for Factor Structure 1

In the case of Factor Structure 1, the percentage of spuriously flagged items for the total-score matching criterion ranged from less than 1% to 2% for the 90/10 factor structure to as high as 59% for the 70/30 factor structure.

Purified factor scores and factor-based scores spuriously flagged items less than 1 percent of the time across factor structures and DIF levels.

Using total score as the matching criterion, the percentage of biased items that failed to be flagged increased as the data set contained more items with DIF. When using purified factor scores or factor-based scores as the matching criterion, the average percentage of false positives ranged between one and two percent.

Summary of Follow-up Results for Factor Structure 2

The trends for detection of biased items were the same for tests composed of either high-level or low-level multidimensional items (Factor Structure 2). In a data set with only 10 multidimensional items--high or low levels of multidimensionality--less than 2% of the items were spuriously flagged using the total-score matching criterion. As has been previously seen, increasing factorial complexity was accompanied by greater numbers of false positives when matching on total-score. At high and low levels of multidimensionality, no items were spuriously flagged when purified factor score and factor-based score were used as matching criteria.

The percentage of biased items that failed to be flagged using the total-score matching criterion increased as the data set contained more multidimensional items. No false negatives occurred at moderate and high levels of DIF for the five replications when using total score as the matching criterion. In addition, the use of purified factor score and factor-based scores as matching criteria resulted in the identification of all biased items.

Summary of Follow-up HSMAT Analysis

The elimination of contaminated items was applied to a test-developer's situation by application of the information gained in the previous studies to the HSMAT data set. The consistency of the flagging of items was examined for the total score, the factor score, the purified factor score, and the factor-based score. Twenty-two percent of the HSMAT items changed flagging designation when using the total-score and purified factor-score matching criteria. When the matching criteria were the 50-item factor score and a purified factor score, twelve percent of the items changed flagging designation. Finally, when the matching criteria were purified factor scores and a factor-based scores the percentage of items switching flagging designation was 10%.

As mentioned previously, the factor-based score could be labeled a purified total score; and, thus, the factor-based and purified factor score comparison is analogous to the total score and factor score comparison made at the beginning of the *Results* section. When the two matching criteria were total score and factor score, the flags changed for 24% of the items--compared to 10% for the purified analogs. It appears that the creation of a more unidimensional matching criterion resulted in more consistent flagging.

Implications for Future Studies

Virtually all text books on tests and measurement emphasize the importance of investigating the factor structure of newly-developed tests to ensure that the internal test structure is consistent with the developer's theory and intended use. The results of this investigation suggest that the factor structure of tests should be examined not only for the entire sample of test takers, but, where sample sizes allow, for relevant subpopulations of examinees as well. This is advisable not only to ensure that the test is measuring the same underlying factors across subpopulations of test takers, but also to ensure that a subsequent DIF analysis equates the groups under study using the appropriate criterion. If a sizable percentage of the test items load on a factor other than that intended by the developer, or if the items are inherently multidimensional (e.g., word problems), then the results of this investigation suggest that the usual practice of matching groups on the total score is probably incorrect.

The design of the present study was developed from the multidimensionality model of DIF advanced by Shealy and Stout (1993) and Ackerman (1992). In the model, if DIF exists then the item must be multidimensional--that is it must measure a target ability and a nuisance ability. For DIF to exist, the focal group and the referent group must differ on the nuisance ability. With real tests, it will usually be the case that if DIF is present (that is, if the test contains both a target and a nuisance factor, and subgroups of the population differ on the nuisance factor), then it is likely that items will load differentially on the nuisance factor. That is, it is likely that item responses will depend to varying degrees on the nuisance factor. In the models created for this study, all DIF items were modeled to load equally on the nuisance factor. Thus in a data set with 70 items loading primarily on the first factor (the target ability) and 30 items loading on the second factor (the nuisance ability), when DIF was simulated at a low level (-.20), it was so defined for all 30 items. Additional research should expand the current investigation to simulate items that load differentially on the nuisance parameter. This would more nearly simulate real tests and may provide more explicit guidance to practitioners.

The inability of the factor-score to identify biased items in data sets composed of inherently multidimensional items (Factor Structure 2) appears to support the earlier work of Shepard et al. (1985). The model used in this study to simulate data for Factor Structure 2 was based on the findings of Camilli and Smith (1990), Ryan (1991), and Shepard et al. (1985) that DIF was often displayed in verbally-loaded mathematics items that required the "nuisance" ability of reading. Shepard et al. reported that chi-square measures of DIF which used a factor-score matching criterion correlated poorly with IRT-based measures of DIF. The inability of the factor-score in the current study to identify biased items appears to support Shepard's conclusions. Since the total-score matching criterion also performed poorly in the identification of biased items, investigation needs to continue in the creation of an unidimensional matching criterion for data sets composed of multidimensional items.

One line of investigation might be to gauge the effect of the removal of multidimensional items from a data set on the consistency of the flagging designation. The results would provide further evidence about the need and utility of a purified criterion. The high consistency of flagging designations, and the apparently unidimensional factor structures of such instruments as the GMAT verbal and quantitative subtests, indicate such tests offer little opportunity for testing the purification process; however, tests with more multidimensionality would be candidates for inspection.

The present investigation is a beginning, but by no means exhausts the analytical possibilities in investigations of alternative matching criteria. Assuming more streamlined computer simulations can be devised, it would be informative to undertake many more replications than was practically feasible here in order to gain a better understanding of the distribution of false negative and false positive flags as a function of test factor structure. It would also appear possible to equate focal and reference group members *simultaneously on all factors that underlie a given test.* (Note that this is not the same as equating on total test score.) In this way aberrant items that measure trivial factors or draw upon very specialized knowledge could be identified.

The results of the current study are relevant to the two-stage DIF analysis outlined by Dorans and Holland (1993). The authors referred to the first stage of the two-step procedure as a "...*criterion refinement* or *purification* step" (p. 60). In essence, in the first stage *all* items are used to compose the matching criterion and are submitted to a DIF analysis with this criterion. In the criterion refinement step, test items flagged as displaying sizable DIF are eliminated in the subsequent formation of a "purified" matching criterion. In the second step, the DIF analysis is then repeated with the purified matching criterion.

The current findings indicate that even as a test becomes more factorially complex the criterion refinement step will identify items with moderate to high levels of DIF. Interestingly, the pervasive presence of low amounts of DIF, as seen in complex factor structures of this study, failed to be detected with the total-score matching criterion. In addition, the presence of multidimensionality resulted in spurious flagging. Depending on Dorans and Holland's definition of *sizable*, the information provided by the criterion refinement study may be resulting in false positives being appropriately included or inappropriately excluded in the creation of the purified matching criterion. Further lines of investigation might be directed at the impact of low-levels of DIF and the ability to use information about the factor structure to improve the purification of the matching criterion.

A consistent, if somewhat baffling, finding in DIF research is that test developers and subject matter experts have been frustrated in their attempts to find any substantive or experiential thread running through items that have been flagged. Moreover, flagged items appear to be indistinguishable from other, non-flagged items on the test. Green (1991) has suggested that part of the problem stems from the use of focal groups and referent groups that are defined sociologically or biologically (race, gender, etc.), rather than on some more substantive basis. As a result, he argues, the groups are too heterogeneous. More homogeneous groups (e.g., individuals who have taken the same courses) may result in flagged items that cohere in some theoretically or practically logical fashion.

The results of the present study suggest an alternative possibility. A principal finding of this investigation was that matching on total score for factorially complex tests results in a substantial spurious flagging of items as biased. Inasmuch as the overwhelming majority of DIF studies in the past used either the Mantel-Haenszel, the standardization procedure, or a variant of the chi-square approach (all of which use total score as the matching criterion), it is possible that the lack of substantive coherence among flagged items was a simple consequence of the fact that the items were spuriously flagged. A matching on factor scores may well result in the identification of items that do in fact allow explanations that are traceable to the experiential backgrounds of African American examinees, women, and other groups of examinees that have been the focus of bias studies in the past. At the very least, before researchers abandon traditionally disenfranchised groups as appropriate foci for investigation of bias, it would be wise to ensure that their analytical procedures were appropriate.

BIBLIOGRAPHY

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Angoff, W. (1972, Sept.). *A technique for the investigation of cultural differences.* Paper presented at the annual meeting of American Psychological Association, Honolulu.

Baghi, H., & Ferrara, S. (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.

Baghi, H., & Ferrara, S. (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. (1981). A criticism of Scheuneman's item bias technique. *Journal of Educational Measurement, 18,* 59-62.

Baker, F. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement, 11,* 111-141.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, *Statistical theories of mental test scores* (chapters 17-20). Reading, MA: Addison-Wesley.

Bleisten, C. A., & Schmitt, A. (1987). *The effect of criterion selection on the evaluation of differential item functioning on SAT analogy items between black and white examinees.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, DC.

Block, N., & Dworkin, G. (Eds.). (1976). *The IQ Controversy.* New York: Pantheon.

Bond, L. (1981). Bias in testing. In B. F. Green (Ed.) *New directions for testing and measurement.: Issues in testing--Coaching, ethnic bias, and disclosure, (8),* 55-77.

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

Camilli, G., & Smith, J. (1990). Comparison of the Mantel-Haenszel test with a randomized and a jackknife test for detecting biased items. *Journal of Educational Statistics, 15,* 53-67.

Clauser, B., Mazor, K., and Hambleton, R. (1994). The effects of score group width on the Mantel-Haenszel procedure. *Journal of Educational Measurement, 31,* 67-78.

Clauser, B., Mazor, K., and Hambleton, R. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6,* 269-279.

Clauser, B., Mazor, K., and Hambleton, R. (1991). The influence of the criterion variable on the identification of differentially functioning test items using the Mantel-Haenszel statistic. *Applied Psychological Measurement, 15,* 353-359.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5,* 115-124.

Cole, N. (1973). Bias in selection. *Journal of Educational Measurement, 10,* 237-255.

Cunningham, G. (1986). *Educational and psychological measurement.* New York: Macmillan Publishing Company.

Darlington, R. (1971). Another look at "cultural fairness." *Journal of Educational Measurement, 8,* 71-82.

Debra P. v. *Turlington.* (1981). 644 F. 2d 397, 5th cir.

Debra P. v. *Turlington.* (1983). 564 F. Supp. 177 (M.D. Fla.).

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

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

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

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

Educational Testing Service. (1987). *Graduate Management Admission Test: Test Code 71.* Princeton: Author.

Eells, K., Davis A., Havighurst, R., Herrick, V., & Tyler, R. (1951). *Intelligence and cultural differences.* Chicago: University of Chicago Press.

Einhorn, H., & Bass, A. (1971). Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin, 75,* 261-269.

Engelhard, G., Anderson, D., & Gabrielson, S. (1990). An empirical comparison of Mantel-Haenszel and Rash procedures for studying differential item functioning on teacher certification tests. *Journal of Research and Development in Education, 23,* 172-179.

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

Golden Rule Insurance Company et al. v. Washburn et al. 419-76 (stipulation for dismissal and order dismissing case, filed in the Circuit Court of the Seventh Judicial Circuit, Sangamon County, IL (1984).

Gould, S. (1981). *The Mismeasure of Man.* New York: W. W. Norton & Company.

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.

Griggs v. Duke Power Company. 401 U.S. 424 (1971).

Gross, A., & Su, W. (1975) Defining a "fair" or "unbiased" selection model: A question of utilities. *Journal of Applied Psychology, 60,* 345-351.

Hambleton, R., Clauser, B., Mazor, K., & Jones, R. (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., & Jones, R. (1993). *Comparison of empirical and judgmental methods of detecting differential item functioning.* (Laboratory of Psychometric and Evaluative Research Report No. 231). Amherst, MA: University of Massachusetts, School of Education.

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

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Publications.

Harman, A. (1994). SAS code for scoring GMAT examinee responses. Greensboro, NC: Author.

Harnisch, D.L. (1991). *Techniques for assessing differential item performance on achievement tests.* Proceedings of the Sixteenth Annual SAS Users Group International Conference. Cary, NC: SAS Institute Inc., 1503-1508.

Hills, J. (1989). Screening for potentially biased items in testing programs. *Educational Measurement: Issues and Practice, 8* (4), 5-11.

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.

Hoover, H., & Kolen, M. (1984). The reliability of six item bias indices, *Applied Psychological Measurement, 8,* 173-181.

Ironson, G., & Subkoviak, M. (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement, 16,* 209-225.

Kim, J., & Mueller, C. (1978). Factor analysis: Statistical methods and practical issues. Sage University Paper series on Quantitative Applications in the Social Sciences, 14. Beverly Hills, CA: Sage Publications.

Kubiak, A., & Colwell, W. (1990, April). *Using multiple statistics with the same items appearing in different test forms.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston.

Lai, M., & Saka, T. (April, 1993). *Using differential item functioning procedures to improve interpretation of and performance on the verbal subtest of the SAT.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta.

Larry P. v. Riles. 343 F. Supp. 130b (N.D. Cal. 1972). (preliminary injunction), affirmed, 502 F. 2d 963 (9th Cer 1974), opinion issued No. D-71-2270 RFP (N.D. Cal. October 16, 1979).

Linn, R. (1973). Fair test use in selection. *Review of Educational Research, 43,* 139-161.

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

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

Mazor, K., Clauser, B., & Hambleton, R. (1991). *The effect of sample size on the functioning of the Mantel-Haenszel statistic* . Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

McPeek, W., & Wild, C. (1986, April). Performance of the Mantel-Haenszel statistic in a variety of situations. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Parents in action on special education (PASE) v. Hannon. No. 74-C-3586. (N.D. IL. July 16, 1980).

Penny, J. (1994). SAS code for simulating two-factor structure in a data set. Greensboro, NC: Author.

Petersen, N., & Novick, M. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13,* 3-29.

Plake, B. (1980). A comparison of a statistical and subjective procedure to ascertain item validity. *Educational and Psychological Measurement, 40,* 397-404.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

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

Resnick, D. (1982). History of educational testing. In A. K. Wigdor & W. R. Garner (Eds.) Ability Testing: Uses, Consequences, and Controversies, Volume II. Washington, DC: National Academy Press.

Rudner, L. (1978). Using standard tests with the hearing impaired. *Volta-Review, 80,* 31-40.

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

SAS Institute. (1985). *Statistics user guide: Version 5.* Cary, NC: Author.

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

Scheuneman, J., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27,* 109-131.

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

Shealy, R., & Stout, W. (1989, April). *A procedure to detect test bias present simultaneously in several items.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Shealy, R., & Stout, W. (1993). An item response theory model for test bias and differential test functioning. In P. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 367-388). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Spray, J., & Miller, T. (1992). *Performance of the Mantel-Haenszel statistic and the standardized difference in proportions correct when population ability distributions are incongruent* (ACT- RR-92-1). Iowa City, Iowa: American College Testing Program.

Thorndike, R. (1971). Concepts of culture fairness. *Journal of Educational Measurement, 8,* 63-70.

Tittle, C. (1982). Use of judgmental methods in item bias studies. In R. Berk (Ed.), *Handbook of Methods for Detecting Test Bias* (pp. 31-63). Baltimore: John Hopkins University Press.

Wigdor, A. K. (1982). Psychological testing and the law of employment discrimination. In A. K. Wigdor Y W. R. Garner (Eds.) *Ability Testing: Uses, consequences, and controversies, Part II: Documentation Section.* Washington DC: National Academy Press.

Wright, D. (1986). *An empirical comparison of the Mantel-Haenszel and standardization methods for detecting differential item performance* (Statistical Report No. SR-86-99). Princeton, NJ: Educational Testing Service.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26,* 55-66.

APPENDIX A

SAS Code for Analyses

SAS Code to Create Samples

```
Options LS=80 PS=59;
  Filename New 'Scratch:[Scratch.Johnsong]GMAT_5000.Dat';
Data Score;
Infile GMAT4234 Missover;
 Input State $ 67-68 YOB 101-102 Gender $ 103 Admin 104-107 Degobj
121
        Citizn 124-126 Umajor 132-133 Gradtn 134-137 Pace 138 Area 139-
140
        Race 141 Edexp 143 Form 202-203 @204 (Ans1-Ans25) (1.)
       @244 (Ans26-Ans45) (1.) @284 (Ans46-Ans70) (1.) @324 (Ans71-
Ans95) (1.)
        @364 (Ans96-Ans115) (1.) @404 (Ans116-Ans140) (1.0) Vright 449-451
        Qright 461-463 Tright 473-475 Fs_v 458-460 Fs_q 470-472 Fs_t 482-
484;
  If Gender NE ' ' and (Race=2 Or Race=3) and Form=71 and Fs_q Gt '0' and
    Fs_v GT '0';
Proc Sort Data=Score;
 By Race;
Data Bycount;
 Set Score;
 By Race;
  N+1;
  If Last.Race;
  Output;
  N=0;
 Keep Race N;
Data Sample;
 Merge Score Bycount;
 By Race;
  If First.Race Then K=2500;
    Retain K;
  If Uniform(0)<K/N Then Do;
   Output;
   K=K-1;
  End;
   N=N-1;
```

```
Proc Sort Data=Sample;
 By Race;
Data All;
 Set Sample;
File New;
 By Race;
Put Race 1 @3 (Ans1 - Ans140) (F1.) Vright 144-146 Qright 148-150
        Tright 152-154 Fs_v 155-156 Fs_q 158-159 Fs_t 161-162;
```

SAS Code to Score Student Responses for the GMAT

```
Options Ls=80 Ps=59;
  Filename New 'Scratch:[Scratch.Johnsong]MH71_All.Dat';
  Filename One 'Scratch:[Scratch.Johnsong]GMAT_5000.Dat';
Data First;
Infile One Missover;
Length Q1-Q140 $ 2;
Retain Key1-Key140;
Array Key{*} Key1-Key140;
Array Ans{*} Ans1-Ans140; *Ans Are Student Answers;
Array Q{*} $ Q1-Q140; *Character Vars Used With Proc Tabulate;
Array S{*} S1-S140; *The S's Are Scored Responses;
If _N_=1 Then Do;
  Input @12 (Key1-Key140) (1.);
  Delete;
  End;
Else Do;
  Input Race 1 @3 (Ans1 - Ans140) (F1.) Vright 144-146 Qright 148-150
   Tright 152-154 Fs_v 155-156 Fs_q 158-159 Fs_t 161-162;
  Do I=1 To 140;
    If Key{I}=Ans{I} Then Do;
      S{I}=1;
      End;
    Else S{I}=0;
    End;
  Drop I Key1-Key140;
  End;
Data All;
 Set First;
 Keep Race S1 -- S140;
File New;
 Put @1 (S1 -- S140) (1.) Race 142 Fs_v 144-145 Vright 147-148 Fs_q 150-151
   Qright 153-154;
```

SAS Code to Create Factor Scores for Use as Matching Criteria

```
Filename Rfac 'Rfactor.Dat';

Data New;
Infile Rfac Lrecl=80 Recfm=V Missover;
Input  @1 (Item1 - Item4) (F1.) @5 (Item11 - Item56) (F1.)
       @53 (Factor1) (F2.) @58 (Rawscr) (F2.) Ethnic 64;

Proc Factor Data=New Scree Priors=Smc
   N=2 Reorder Plot Out=F1scores;
 Var Item1 -- Item4 Item11--Item56;
Title 'Factor Analysis For Complete Data Set';

Proc Means Data=New;
Var Rawscr;

Proc Standard Data=F1scores Out=Zquant Mean=25.7634804 Std=8.8491931;
Var Factor1 Factor2;

Filename Facfile 'Twofac.Dat';
Data _Null_;
Set Work.Zquant;
File Facfile Lrecl=80 Recfm=V;
Put @1 (Item1 -- Item4) (F1.) @5 (Item11 -- Item56) (F1.)
       @53 (Factor1) (F2.) @59 (Factor2) (F2.)
       @65 (Rawscr) (F2.) Ethnic 71;
```

SAS Macro for Analyzing Differential Item Functioning

Harnisch, D.L. (1991) MHPROG SAS Macro Listing

Source: Harnisch, D.L. (1991). Techniques for Assessing Differential Item Performance on Achievement Tests. Proceedings of the Sixteenth Annual SAS Users Group International Conference. Cary, NC: SAS Institute Inc., 1503-1508.

```
/*This macro computes DIF indices based on the Mantel-Haenszel procedure.
Users must specify the test questions using variable names of Q1-Qn where n
represents the number of items on the test. These variables must be scored as
1 for correct and 0 for incorrect. The variable TOTAL must be created which
represents the performance of the students on the criterion measure of
interest. The SAS data set name which includes the binary coded test
questions, total score, and the discrimating variable name (coded 1 for focal
and 0 for reference) are used as arguments on the SAS macro MHPROG. For
example, a SAS data set by the name of GR11MATH.DIF. containing 50 items
binary coded and summed to create a variable TOTAL along with sex coded 1
for males and 0 for females would be written: MHPROG
(GR11MATH.DIF,SEX,50);        */

%MACRO MHPROG(Data,Comvar,Nitem);
Proc Summary Data = &Data Nway;
    Class Total &Comvar;
    Var Q1 - Q&Nitem;
    Output Out = Mhisum N = N1 - N&Nitem Sum = R1 - R&Nitem;

Proc Sort Data=Mhisum; By Total ;

Data Mhsums ;
    Set Mhisum ;
    By Total;
     Array Ns{&Nitem} N1 - N&Nitem;
     Array Rs{&Nitem} R1 - R&Nitem;
    If First.Total And Last.Total Then Do;
      Output ;
       Do I = 1 To &Nitem;
        Ns(I) = 0.0;
        Rs(I) = 0;
      End;
```

```
       If &Comvar = 0 Then &Comvar = 1;Else &Comvar = 0;
      Output;
      End;
    Else Output;
 Proc Sort ; By Total &Comvar;
 Data Mhbase;
      Set Mhsums (Keep=N1-N&Nitem R1-R&Nitem);
     If Mod(_n_,2)=1;
 Data Mhfocal;
      Set Mhsums (Keep=N1-N&Nitem R1-R&Nitem);
     If Mod(_n_,2)=0;

 Data Dtots; Set Mhsums; If Mod(_n_,2)=0; Keep Total;

Proc Iml;
   Eps = 0.0000001 ; /* 1.0e-7 */

   Use Mhbase;
   Read All Into Prebs;
    Nlevel=Nrow(Prebs);
     Rbs=Prebs( | 1:Nlevel,&Nitem+1:&Nitem*2 | );
    Nbs=Prebs( | 1:Nlevel,1:&Nitem | );
   Free Prebs;

   Use Mhfocal;
   Read All Into Prefs;
    Nlevel=Nrow(Prefs);
     Rfs=Prefs( | 1:Nlevel,&Nitem+1:&Nitem*2 | );
    Nfs=Prefs( | 1:Nlevel,1:&Nitem | );
   Free Prefs;

 Wbs = Nbs - Rbs;
 Wfs = Nfs - Rfs;


 /* There Are Now Four Matricies Which Constitute The Table Cells */
 /* Each Column Of These Matricies Corresponds To A Table Cell */
 /* The Columns Correspond To The Different Items   */

 Alphas = (Rbs # Wfs ) / ( Rfs # Wbs );

 Ms = ( Wbs # Rfs ) / ( Nfs + Nbs );
 Alpha_ms = (Rbs # Wfs) / (Nfs + Nbs);
```

```
Alphamh = Alpha_ms (| +, |)  / Ms(| +, |);


Create Summary1   From  Alphas ;
Append  From  Alphas ;

Free Ms Alpha_ms  Alphas;

Mus = Nbs # (Rbs + Rfs ) / (Nfs + Nbs);
Sigmas = ( Nbs # Nfs # ( Rbs + Rfs ) # ( Wbs + Wfs ) ) /
        ((Nfs + Nbs ) # ( Nfs + Nbs ) # ( Nfs + Nbs -1 )) ;
Term1 = Rbs(| +, |) ;
Term2 = Mus(| +, |) ;
Term3 = Sigmas(| +, |) <> (Eps # J(1, Ncol(Sigmas), 1.0) ) ;
Term4 = Abs ( Term1 - Term2 ) - 0.5 ;

Free Wfs Wbs Term1 Term2 ;

Chisqmh = ( Term4 # Term4 ) / Term3 ;

Free  Term3 Term4 ;

Pchimh = J(Nrow(Chisqmh),Ncol(Chisqmh),1.0) - Probchi(Chisqmh , 1.0) ;
Pfs = Rfs /  Nfs ;
Pbs = Rbs /  Nbs ;
Ds = Pfs - Pbs ;

*Creating Sas Data Sets For Plots;
Use Dtots; Read All Into Dtotsm;
Nrtot = Nrow(Dtotsm);
Drefm = Repeat(0,Nrtot,1);
Dfocm = Repeat(1,Nrtot,1);
Diffs = Ds | | Dtotsm;
Refpc = Pbs | | Dtotsm | | Drefm;
Focpc = Pfs | | Dtotsm | | Dfocm;
Allpc = Refpc//Focpc;
*Create Sds.Ddiffs From Diffs; *Append From Diffs;
Create Sds.Dallpc From Allpc; Append From Allpc;

Create Summary2   From Ds  ;
 Append   From Ds  ;
Create Summary3   From  Pfs  ;
 Append   From  Pfs  ;
```

```
  Create Summary4  From  Pbs ;
   Append  From  Pbs ;

  Free Rbs Rfs ;

  Efs = Nfs # Pfs ;
  Ebs = Nfs # Pbs ;

  Free Pfs Pbs ;

  Psubf = ( Efs (| +, |) / Nfs(| +, |) );
  Phatf = ( Ebs (| +, |) / Nfs(| +, |) );
  Dstd = Psubf - Phatf ;

  Free Nfs Nbs ;

  /* Bound The Probabilities Away From The Ends */

  Psubf = (Eps # J(Nrow(Psubf),Ncol(Psubf),1.0)) <> Psubf ;
  Psubf = ((1.0 - Eps) # J(Nrow(Psubf),Ncol(Psubf),1.0)) >< Psubf ;
  Phatf = (Eps # J(Nrow(Phatf),Ncol(Phatf),1.0)) <> Phatf ;
  Phatf = ((1.0 - Eps) # J(Nrow(Phatf),Ncol(Phatf),1.0)) >< Phatf ;

  Cnams = { 'psubf ', 'phatf'};
 Outps = (Psubf `) | | (Phatf`) ;
 Create Pests From Outps (| Colname = Cnams |);
  Append From Outps;
 Free Cnams Outps ;

 Deltaps = J(Nrow(Psubf), Ncol(Psubf), 13.0) - 4.0 # Probit( Psubf ) ;
 Deltaphs = J(Nrow(Phatf), Ncol(Phatf), 13.0) - 4.0 # Probit( Phatf ) ;
 * Print Deltaps Deltaps ;

  Alphamh = (Eps # J(Nrow(Alphamh),Ncol(Alphamh),1.0)) <> Alphamh ;

 Deltamh = -2.35 # Log ( Alphamh ) ;
 Deltastd = -2.35 # Log ((Phatf # (1 - Psubf))/(Psubf # (1 - Phatf)));

 * Print Deltamh Deltastd ;

* Print Alphamh Deltamh Dstd Chisqmh Pchimh Deltastd Deltaps Deltaphs;
*Transposing The Matrix To Yield Item By Index Table;
Talphamh=Alphamh`; Tdeltamh=Deltamh`; Tdstd=Dstd`;
```

```
Tchisqmh=Chisqmh`;  Tpchimh=Pchimh`;   Tdelstd=Deltastd`;

 *Print Talphamh Tdeltamh Tdstd Tchisqmh Tpchimh Tdelstd;
 Outs = Talphamh | | Tdeltamh | | Tdstd | | Tchisqmh | | Tpchimh | | Tdelstd;
 Cname={'alphamh','deltamh', 'dstd', 'chisqmh', 'pchimh','delstd'};

 Create Results From Outs (| Colname = Cname |);
  Append From Outs;

Data Results;Set Results; If Alphamh=0 Then Do ;
    Deltamh=.; Dstd=.;Chisqmh=.;Pchimh=.;Delstd=.;End;
    If Dstd>.10 Then Flag='m++';
    Else If Dstd>.05 Then Flag='m+ ';
    Else If Dstd<-.10 Then Flag='m--';
    Else If Dstd<-.05 Then Flag='m- ';
    Else Flag='  ';
Proc Print Data=Results;
    Title2 "Mantel-Haenszel Statistics: By &Comvar";
    Title3 'flag Column Indicates Level Of Dif For Gender Group';
*Proc Means Data=Results;
*    Var Alphamh Deltamh Dstd Delstd;
*    Title2 "Descriptive Statistics Of Mh-Parameters: By &Comvar";
*Proc Corr Data=Results;
*    Var Alphamh Deltamh Dstd Delstd;
*    Title2 "Correlations Among Mh-Parameters: By &Comvar";
*If Ttest Wanted;
/*
Proc Ttest Data=&Data;
    Class &Comvar;
    Var Total Q1-Q&Nitem;
    Title2 "Ttest Of Performance On Items: By &Comvar";
*/
%Mend;
```

SAS Code to Simulate Factor Structure 1

```
OPTIONS LS=80 PS=59;
FILENAME NEW 'SCRATCH:[SCRATCH.JOHNSONG]Monte70_50.DAT';
/*
This data step creates 1 trait, followed
by the 100 items that tap that trait.

The trait is in ts101, while ts1-ts100 are
the true scores on items that tap that trait.

*/
data rcont;

  array seed[101] seed1-seed101; /* seeds for true scores and traits */
  array ersed[100] ersed1-ersed100; /* seeds for error components */
  array ts[101] ts1-ts101; /* true scores and latent traits */
  array errs[100] errs1-errs100; /* error components */
  array itcor[100] itcor1-itcor100; /* correlation of true score with
                              latent trait */

  keep ts1-ts100
     errs1-errs100;

  /* Initialize seeds first for the true scores to be kept in "ts,"
     then for the error components kept in "errs." */
  tempseed = 123456789;
  do i = 1 to 101 by 1;
    call ranuni(tempseed,seed[i]);
   seed[i] = int(seed[i]*tempseed);
   end;
  do i = 1 to 100 by 1;
    call ranuni(tempseed,ersed[i]);
   ersed[i] = int(ersed[i]*tempseed);
   end;

  /* Create 100 item correlations (correlations with latent traits). */
  tempseed = 987654321;
  m = 0.50;
  dispers = 0.1;
  do i = 1 to 100 by 1;
    call rannor(tempseed,itcor[i]);
   itcor[i] = m + dispers*itcor[i];
```

```
      end;

    /* Create data for referent examinees */
    do j = 1 to 1000 by 1;

      /*
         Create 101 random normal deviates. Remember that
         1-100 are the true scores on the items and 101
         is the latent trait
      */
      do i = 1 to 101 by 1;
        call rannor(seed[i],ts[i]);
      end;

      /* Correlate 100 true scores with trait (kept in 101)
         and normalize these later.
      */
      do i = 1 to 100 by 1;
        ts[i] = itcor[i]*ts[101] + sqrt(1-(itcor[i]*itcor[i]))*ts[i];
      end;

      /* Create error components.
         These will need to be normalized later.
      */
      do i = 1 to 100 by 1;
        call rannor(ersed[i],errs[i]);
      end;

      output; /* response vector in continuous form */
    end;

run;

/* Standardize true scores to be Z's. */
proc standard data=rcont out=rcont mean=0 std=1;
  var ts1-ts100;
run;

/* Standardize errors. */
proc standard data=rcont out=rcont mean=0 std=.31;
  var errs1-errs100;
run;
```

```
data fcont;

 array seed[102] seed1-seed102; /* seeds for true scores and traits */
 array ersed[100] ersed1-ersed100; /* seeds for error components */
 array ts[102] ts1-ts102; /* true scores and latent traits */
 array errs[100] errs1-errs100; /* error components */
 array itcor[100] itcor1-itcor100; /* correlation of true score with
                          latent trait */

 keep ts1-ts100
    errs1-errs100;

 /* Initialize seeds first for the true scores to be kept in "ts"
     and then for the error components kept in "errs." */
 tempseed = 123456789;
 do i = 1 to 102 by 1;
   call ranuni(tempseed,seed[i]);
   seed[i] = int(seed[i]*tempseed);
   end;
 do i = 1 to 100 by 1;
   call ranuni(tempseed,ersed[i]);
   ersed[i] = int(ersed[i]*tempseed);
   end;

 /* Create 100 item correlations (correlations with latent traits). */
 tempseed = 987654321;
 m = 0.50;
 dispers = 0.1;
 do i = 1 to 100 by 1;
   call rannor(tempseed,itcor[i]);
   itcor[i] = m + dispers*itcor[i];
   end;

 /* Create data for subjects. */
 do j = 1 to 1000 by 1;

   /*
      Create 102 random normal deviates.  Remember that
     1-100 are the true scores on the items and 101-102
     are the latent traits.
   */
   do i = 1 to 102 by 1;
     call rannor(seed[i],ts[i]);
     end;

     /* Correlate 100 true scores with traits (kept in 101-102).
        Normalize these later
    */
    do i = 1 to 30 by 1;
      ts[i] = itcor[i]*ts[101] + sqrt(1-(itcor[i]*itcor[i]))*ts[i];
    end;
    do i = 31 to 100 by 1;
      ts[i] = itcor[i]*ts[102] + sqrt(1-(itcor[i]*itcor[i]))*ts[i];
    end;

     /* Create error components.
        These will need to be normalized later
    */
    do i = 1 to 100 by 1;
      call rannor(ersed[i],errs[i]);
    end;

     output; /* response vector in continuous form */
  end;

run;

/* Standardize true scores to be Z's. */
proc standard data=fcont out=fcont mean=0 std=1;
  var ts1-ts100;
run;

/* Standardize errors.  */
proc standard data=fcont out=fcont mean=0 std=.31;
  var errs1-errs100;
run;

/* Create reference group. */
data ref;
  set rcont;
  array dif[100] dif1-dif100;

  /* All DIF is nil for the reference group. */

  do i = 1 to 100 by 1;
    dif[i] = 0;
  end;
  group = 1; /* Define reference group id here. */
  drop i;
run;

/* Create focal group. */
data focal;
  set fcont;
  array dif[100] dif1-dif100;

  /* Here is where DIF is defined.*/
   /* First initialize all the values, then set the particular items. */
  do i = 1 to 100 by 1;
    dif[i] = 0;
  end;
  do i = 1 to 30 by 1;
    dif[i] = -.50;
  end;
  group = 0; /* Define focal group id here. */
  drop i;
run;


/* Now concatenate and then sort the data sets. */
data total;
 set ref focal;
run;
proc sort data=total;
  by group;
run;


/* Now create continuous observed scores using DIF. */
data total;
  set total;
  array ts[100] ts1-ts100;
  array errs[100] errs1-errs100;
  array dif[100] dif1-dif100;
  array obs[100] obs1-obs100;

  keep ts1-ts100
      errs1-errs100
      dif1-dif100
      obs1-obs100
      group;

   do i = 1 to 100 by 1;
     obs[i] = ts[i] + errs[i] + dif[i];
   end;
run;

/* Standardize continuous observed scores. */
proc standard data=total mean=0 std=1;
   var obs1-obs100;


/* Create dichotomous response vectors
    and compute raw score. */
data total;
   set total;

   array ts[100] ts1-ts100;
   array obs[100] obs1-obs100;
   array errs[100] errs1-errs100;
   array dif[100] dif1-dif100;
   array score[100] score1-score100;

   drop i;

   /* Dichotomize observed score. */
   do i = 1 to 100 by 1;
     if obs[i] le 0 then score[i] =  0;
                 else score[i] = 1;
   end;
   total= sum(of score1-score100);
run;


data focal;
set total;
if group=0;

/*
    Factor analyze dichotomous response vectors,
     save results, and then score.  When finished, data set FOCAL will
     include the two factor scores in addition to the other variables.
*/

proc factor data=focal priors=smc scree n=2 rotate=varimax plot
       out=fscores;
  var score1-score100;
  title1 'Factors found in focal data set';


/* Look at the correlations. */
proc corr data=fscores;
  var total factor1 factor2;
  title1 'Correlations of Focal Group';
  title2 'Raw and Factor Scores';
run;

data refer;
set total;
if group=1;
/*
    Factor analyze dichotomous response vectors,
     save results, and then score.  When finished, data set REFER will
     include the two factor scores in addition to the other variables.
*/
proc factor data=refer priors=smc scree n=2
       plot out=rscores;
  var score1-score100;
  title1 'Factors found in referent data set';


/* Look at the correlations. */
proc corr data=rscores;
  var total factor1 factor2;
  title1 'Correlations of Referent Group';
  title2 'Raw and Factor Scores';
run;

data comb;
set fscores rscores;
/*
  This step is to transform the factor scores to a more
  reasonable scale.
*/
```

```
proc standard data=comb  out=total m=50  std=15;
  var factor1 factor2;

proc factor data=comb priors=smc scree n=2 rotate=varimax plot
            var score1-score100;
   title1 'Factors found in combined data set';


/* Now truncate factor scores so they will suit m-h analysis. */
data total;
  set total;
  factor1 = int(factor1);
  factor2 = int(factor2);
run;


DATA ALL;
 SET total;
 KEEP group score1-score100 total FACTOR1 factor2;
FILE NEW;
 PUT @1 (score1 -- score100) (F1.) total 102-104 @107 (factor1) (f2.)
    @110 (factor2) (f2.) group 113 ;
```

SAS Code to Simulate Factor Structure 2

```
OPTIONS LS=80 PS=59;
FILENAME NEW 'SCRATCH:[SCRATCH.JOHNSONG]M701_50.DAT';
/*
This data step creates the 2 traits, followed
by the 100 items that tap those traits.

The traits are in ts101 and ts102, while ts1-ts100 are
the true scores on items that tap those traits.

*/
data cont;

  array seed[102] seed1-seed102; /* seeds for true scores and traits */
  array ersed[100] ersed1-ersed100; /* seeds for error components */
  array ts[102] ts1-ts102; /* true scores and latent traits */
  array errs[100] errs1-errs100; /* error components */
  array itcor[100] itcor1-itcor100; /* correlation of true score with
                               latent trait */

  keep ts1-ts100
     errs1-errs100;

  /* Initialize seeds first for the true scores to be kept in "ts,"
      and then for the error components kept in "errs." */
  tempseed = 123456789;
  do i = 1 to 102 by 1;
    call ranuni(tempseed,seed[i]);
   seed[i] = int(seed[i]*tempseed);
   end;
  do i = 1 to 100 by 1;
    call ranuni(tempseed,ersed[i]);
   ersed[i] = int(ersed[i]*tempseed);
   end;

  /* Create 100 item correlations (correlations with latent traits). */
  tempseed = 987654321;
  m = 0.50;
  dispers = 0.1;
  do i = 1 to 100 by 1;
    call rannor(tempseed,itcor[i]);
   itcor[i] = m + dispers*itcor[i];
```

```
    end;

  /* Create data for subjects. */
  do j = 1 to 1000 by 1;

    /*
       Create 102 random normal deviates.  Remember that
      1-100 are the true scores on the items and 101-102
      are the latent traits
    */
    do i = 1 to 102 by 1;
      call rannor(seed[i],ts[i]);
    end;

    /* Correlate 100 true scores with traits (kept in 101-102).
      Normalize these later.
    */
    do i = 1 to 30 by 1;
      ts[i] = itcor[i]*ts[102] + itcor[i]/1.5*ts[101] +
           sqrt(1-(itcor[i]*itcor[i]*(1+1/2.25)))*ts[i];
    end;
    do i = 31 to 100 by 1;
      ts[i] = itcor[i]*ts[102] + sqrt(1-(itcor[i]*itcor[i]))*ts[i];
    end;

    /* Create error components.
      These will need to be normalized later.
    */
    do i = 1 to 100 by 1;
      call rannor(ersed[i],errs[i]);
    end;

    output; /* response vector in continuous form */
  end;

run;

/* Standardize true scores to be Zs. */
proc standard data=cont out=cont mean=0 std=1;
  var ts1-ts100;
run;

/* Standardize errors.*/
```

```
proc standard data=cont out=cont mean=0 std=.31;
  var errs1-errs100;
run;

/* Create reference group. */
data ref;
  set cont;
  array dif[100] dif1-dif100;

  /* All DIF is nil for the reference group. */

  do i = 1 to 100 by 1;
    dif[i] = 0;
  end;
  group = 1; /* Define reference group id. */
  drop i;
run;

/* Create focal group. */
data focal;
  set cont;
  array dif[100] dif1-dif100;

  /* Define DIF. */
  /* First initialize all the values, then set the particular items. */
  do i = 1 to 100 by 1;
    dif[i] = 0;
  end;
  do i = 1 to 30 by 1;
    dif[i] = -.50;
  end;
  group = 0; /* Define focal group id.  */
  drop i;
run;


/* Concatenate and then sort the data sets. */
data total;
 set ref focal;
run;
proc sort data=total;
  by group;
run;
```

```
/* Create continuous observed scores using DIF. */
data total;
  set total;
  array ts[100] ts1-ts100;
  array errs[100] errs1-errs100;
  array dif[100] dif1-dif100;
  array obs[100] obs1-obs100;

  keep ts1-ts100
      errs1-errs100
      dif1-dif100
      obs1-obs100
      group;

  do i = 1 to 100 by 1;
    obs[i] = ts[i] + errs[i] + dif[i];
  end;
run;

/* Standardize continuous observed scores. */
proc standard data=total mean=0 std=1;
  var obs1-obs100;


/* Create dichotomous response vectors
     and compute raw score */
data total;
  set total;

  array ts[100] ts1-ts100;
  array obs[100] obs1-obs100;
  array errs[100] errs1-errs100;
  array dif[100] dif1-dif100;
  array score[100] score1-score100;

  drop i;

  /* Dichotomize observed score. */
  do i = 1 to 100 by 1;
    if obs[i] le 0 then score[i] =  0;
              else score[i] = +1;
  end;
  total= sum(of score1-score100);
run;


/* Factor analyze dichotomous response vectors,                */
/* save results, and then score.                        */
/* When finished data set TOTAL will include the two factor scores */
/* in addition to the other variables.                    */
proc factor data=total priors=smc scree n=2 rotate=varimax plot
        out=fscores;
  var score1-score100;
  title1 'Factors found in combined data sets';


/* Look at the correlations. */
proc corr data=fscores;
  var total factor1 factor2;
  title1 'Correlations of Raw and Factor Scores';
run;

/*
  This step is to transform the factor scores to a more
  reasonable scale.
*/
proc standard data=fscores out=total m=50 std=15;
  var factor1 factor2;

/* Truncate factor scores so they will suit m-h analysis. */
data total;
  set total;
  factor1 = int(factor1);
  factor2 = int(factor2);
run;

DATA ALL;
 SET total;
 KEEP group score1-score100 total FACTOR1 factor2;
FILE NEW;
 PUT @1 (score1 -- score100) (F1.) total 102-104 @107 (factor1) (f2.)
   @110 (factor2) (f2.) group 113 ;
```

APPENDIX B

Scree Plots and Factor Plots

```
     |
     |
     |
  10 +
     |
     |     1
     |
     |
     |
     |
   8 +
     |
     |
     |
     |
     |
     |
   6 +
E    |
i    |
g    |
e    |
n    |
v    |
a  4 +
l    |
u    |
e    |
s    |
     |
     |
   2 +
     |      2
     |
     |
     |      34
     |        5
     |         6790124
   0 +                56790124567901245679012456790124567901245679012
     |                                                      34567901245
     |
     |
     |
     |
     |
  -2 +
     |
     |
     -----+-------+-------+-------+-------+-------+-------+-------+-------+------
          0      10      20      30      40      50      60      70      80
                                        Number
```

Figure B1. Scree Plot of Eigenvalues for the GMAT Verbal Test

```
                             FACTOR1
                                1

                               .9

                               .8

                               .7

                               .6
                         W
                     WSR VU    .5
                    U    S TPP
                       T   Q  W.V
                     V    XR        M I  F
                               .SRLF LJEY C
                              X MGJIDAZ  D
                               YT EB Z B
                               OMCKAG                           F
                              U.PK                              A
                                J                               C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0T
                                                                O
                              -.1                               R
                                                                2
                              -.2
```

Figure B2. Plot of Factor Loadings for African American Examinees on the GMAT Verbal Test

```
                             FACTOR1
                                1

                               .9

                               .8

                               .7

                           W   .6
                           U
                               .5
                      R   VS T P
                  U  T         .4V M
                     V       R  WL  LE
                               .3SGDFEY
                          X    XYIEJZB
                               .ANIA
                                OC                              F
                               MU                               A
                                                                C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0T
                                                                O
                              -.1                               R
                                                                2
                              -.2
```

Figure B3. Plot of Factor Loadings for White Examinees on the GMAT Verbal Test

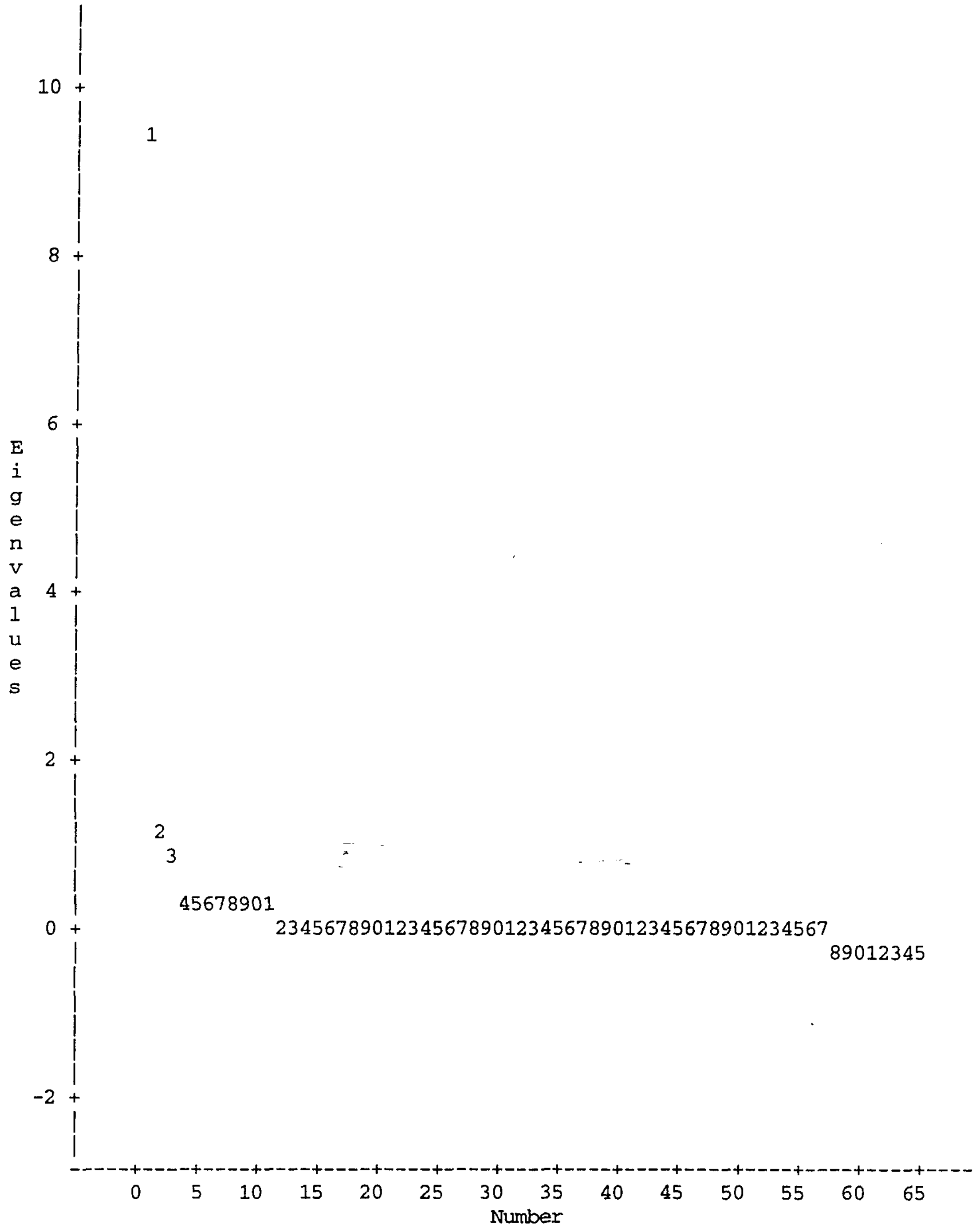

Figure B4. Scree Plot of Eigenvalues for the GMAT Quantitative Subtest

```
                             FACTOR1
                                1

                               .9

                               .8

                               .7

                               .6
                            X
                              F.5 Z
                           E  A     M
                         ZB  K .4 M O
                           Y D  KG FE  Q
                        V Y CH GLI B  H
                         U C EC JDJ O
                          T   HDBLS SNP
                                 G  PRIT                        F
                             A .1   M L                         A
                                                                C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0T
                                 N                              O
                              -.1                               R
                                                                2
                              -.2
```

Figure B5. Plot of Factor Loadings for African American Examinees on the GMAT Quantitative Test

```
                             FACTOR1
                                1

                               .9

                               .8

                               .7

                               .6
                              F
                           XZ FMH
                            M  O     Q
                         XAK L JFO    T
                          ED  GUKPBBI   I
                          WZIW RGHBJC L N
                           ECYY VP    S  L
                             V .2H  M
                            UC  T                               F
                              A.1                               A
                                  N                             C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0T
                                                                O
                              -.1                               R
                                                                2
                              -.2
```

Figure B6. Plot of Factor Loadings for White Examinees on the GMAT Quantitative Test

```
     |
   8 +
     |
     |
     |  1
     |
   7 +
     |
     |
     |
     |
   6 +
     |
     |
     |
     |
   5 +
     |
E    |
i    |
g    |
e  4 +
n    |    2
v    |
a    |
l    |
u  3 +
e    |
s    |
     |
     |
   2 +
     |
     |
     |
     |
   1 +
     |
     |     3 4
     |        56
     |           78 901 23 45
   0 +                       6 78 901 23 456 78 901 23 45
     |                                                   6 78 901 23 456 78 90
     |
     |
     |
  -1 +
     |
     +-+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+--
       0      5     10     15     20     25     30     35     40     45     50
                                        Number
```

Figure B7. Scree Plot of Eigenvalues for the High-Stakes Mathematics Achievement Test

```
                             FACTOR1
                                1

                               .9

                               .8

                               .7

                               .6
                               X
                              XVT
                              NRYQ
                              NBHE
                             XEWUL
                             CZYA
                               ACY
                               QW
                               H                                F
                               .1                               A
                                          R  F                  C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 OQ .TBIE .J .6 .7 .8 .9 1.0T
                                        D NC H A                O
                              -.1                               R
                                                                2
                              -.2
```

Figure B8. Plot of Factor Loadings for Focal Group Members for Factor Structure 1 (80/20)

```
                             FACTOR1
                                1

                               .9

                               .8

                               .7

                               .6
                               J X
                         Z U T Y5V A
                           VNEGHBBXQC
                            DLWEENBF
                           R S IBTUG   K
                              AYCMX
                           C  OD ZP
                            Q  .2
                           O    W  H                            F
                               .1                               A
                                                                C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0T
                                                                O
                              -.1                               R
                                                                2
                              -.2
```

Figure B9. Plot of Factor Loadings for Referent Group Members for Factor Structure 1 (80/20)

```
                             FACTOR1
                                1

                               .9

                               .8

                               .7

                               .6

                               .5
                                   L SWXI
                               .4   XFK Y
                                 C PZYEV
                               .3 UJATC
                                  GHRVXB    BST
                               .2  HFEO     CMJD
                                  N        APIG H               F
                               .1       Q                       A
                                                                C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0T
                                                                O
                              -.1                               R
                                                                2
```

Figure B10. Plot of Factor Loadings for Focal Group Members for Factor Structure 2 with High-level Multidimensional Items

```
                             FACTOR1
                                1

                               .9

                               .8

                               .7

                               .6

                               .5
                                    L W I
                               .4   XZFXY
                                 C  ZYEM
                               .3 UUAVC
                                  TGYVXT   R MS N
                               .2  HFEO   PCDBHJ
                                  N    Q   EALF                 F
                               .1                               A
                                                                C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0T
                                                                O
                              -.1                               R
                                                                2
```

Figure B11. Plot of Factor Loadings for Referent Group Members for Factor Structure 2 with High-level Multidimensional Items

```
                             FACTOR1
                                1

                               .9

                               .8

                               .7

                               .6

                               .5
                                    L T
                               .4    ZF W QI
                                   WCABKMXY
                               .3 RURLZ MB S
                                   TUVAXBFONT
                               .2 NIGQS KCC D
                                     HYO AE  H                  F
                               .1      F    I                   A
                                             L                  C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0T
                                                                O
                              -.1                               R
                                                                2
```

Figure B12. Plot of Factor Loadings for Focal Group Members for Factor Structure 2 with Low-level Multidimensional Items

```
                             FACTOR1
                                1

                               .9

                               .8

                               .7

                               .6

                               .5

                               .4   L MJYTI
                                    CDXISN
                               .3   XTCAOFBW
                                 F MJSAE BEB
                               .2   FUHOKAN L
                                     IEQH   C                   F
                               .1   N HQT  W                    A
                                          R                     C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0T
                                                                O
                              -.1                               R
                                                                2
```

Figure B13. Plot of Factor Loadings for Referent Group Members for Factor Structure 2 with Low-level Multidimensional Items

Appendix C

Results of Replications for Simulation Studies

Table C1

Number of Spurious Flags (False Positives) and Missed Flags (False Negatives) for Factor Structure 1 (90/10)

	DIF Level: -.20						DIF Level: -.35						DIF Level: -.50					
	Total Score		Purified Factor Score		Factor Based Score		Total Score Score		Purified Factor Score		Factor Based Score		Total Score Score		Purified Factor Score		Factor Based Score	
R	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF
1	2	2	1	0	1	0	3	1	1	0	1	0	4	0	1	0	1	0
2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
5	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Note.
R = Replication
SF = Spurious Flags
MF = Missed Flags

Table C2.

Number of Spurious Flags (False Positives) and Missed Flags (False Negatives) for Factor Structure 1 (80/20)

	DIF Level: -.20						DIF Level: -.35						DIF Level: -.50					
	Total Score Score		Purified Factor Score		Factor Based Score		Total Score		Purified Factor Score		Factor Based Score		Total Score		Purified Factor Score		Factor Based Score	
R	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF
1	4	6	0	0	0	0	9	0	0	0	0	0	16	0	0	0	0	0
2	0	6	0	0	0	0	2	0	0	0	0	0	17	0	0	0	0	0
3	0	5	0	0	0	1	5	0	0	0	0	0	12	0	0	0	0	0
4	0	3	0	0	0	0	5	0	0	0	0	0	18	0	0	0	0	0
5	0	8	0	1	0	1	6	0	0	0	0	0	16	0	0	0	0	0

Note.
R = Replication
SF = Spurious Flags
MF = Missed Flags

Table C3.

Number of Spurious Flags (False Positives) and Missed Flags (False Negatives)for Factor Structure 1 (70/30)

	DIF Level: -.20						DIF Level: -.35						DIF Level: -.50					
	Total Score		Purified Factor Score		Factor Based Score		Total Score		Purified Factor Score		Factor Based Score		Total Score		Purified Factor Score		Factor Based Score	
R	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF
1	6	11	1	0	1	0	19	0	1	0	1	0	43	0	1	0	1	0
2	1	13	0	1	0	0	18	0	0	0	0	0	40	0	0	0	0	0
3	0	12	0	0	0	0	17	1	0	0	0	0	36	0	0	0	0	0
4	2	9	0	0	0	0	20	0	0	0	0	0	45	0	0	0	0	0
5	2	13	0	1	0	2	16	0	0	0	0	0	43	0	0	0	0	0

Note.
R = Replication
SF = Spurious Flags
MF = Missed Flags

Table C4.

Number of Spurious Flags (False Positives) and Missed Flags (False Negatives) for Factor Structure 2 (90/10) with High Levels of Multidimensionality

	DIF Level: -.20						DIF Level: -.35						DIF Level: -.50					
	Total Score		Purified Factor Score		Factor Based Score		Total Score		Purified Factor Score		Factor Based Score		Total Score		Purified Factor Score		Factor Based Score	
R	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Note.
R = Replication
SF = Spurious Flags
MF = Missed Flags

Table C5.

Number of Spurious Flags (False Positives) and Missed Flags (False Negatives) for Factor Structure 2 (80/20) with High Levels of Multidimensionality

	DIF Level: -.20						DIF Level: -.35						DIF Level: -.50					
	Total Score		Purified Factor Score		Factor Based Score		Total Score		Purified Factor Score		Factor Based Score		Total Score		Purified Factor Score		Factor Based Score	
R	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF
1	0	3	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0
2	0	6	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0
3	0	5	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
4	0	4	0	0	0	0	1	0	0	0	0	0	6	0	0	0	0	0
5	0	3	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0

Note.
R = Replication
SF = Spurious Flags
MF = Missed Flags

Table C6.

Number of Spurious Flags (False Positives) and Missed Flags (False Negatives) for Factor Structure 2 (70/30) with High Levels of Multidimensionality

	DIF Level: -.20						DIF Level: -.35						DIF Level: -.50					
	Total Score		Purified Factor Score		Factor Based Score		Total Score		Purified Factor Score		Factor Based Score		Total Score		Purified Factor Score		Factor Based Score	
R	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF
1	0	14	0	0	0	0	5	0	0	0	0	0	52	0	0	0	0	0
2	0	22	0	0	0	0	5	0	0	0	0	0	42	0	0	0	0	0
3	0	22	0	0	0	0	17	0	0	0	0	0	45	0	0	0	0	0
4	0	16	0	0	0	0	10	0	0	0	0	0	42	0	0	0	0	0
5	0	15	0	0	0	0	10	0	0	0	0	0	44	0	0	0	0	0

Note.
R = Replication
SF = Spurious Flags
MF = Missed Flags

Table C7.

Number of Spurious Flags (False Positives) and Missed Flags (False Negatives) for Factor Structure 2 (90/10) with Low Levels of Multidimensionality

	DIF Level: -.20						DIF Level: -.35						DIF Level: -.50					
	Total Score		Purified Factor Score		Factor Based Score		Total Score		Purified Factor Score		Factor Based Score		Total Score		Purified Factor Score		Factor Based Score	
R	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF
1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Note.
R = Replication
SF = Spurious Flags
MF = Missed Flags

Table C8.

Number of Spurious Flags (False Positives) and Missed Flags (False Negatives) for Factor Structure 2 (80/20) with Low Levels of Multidimensionality

	DIF Level: -.20						DIF Level: -.35						DIF Level: -.50					
	Total Score		Purified Factor Score		Factor Based Score		Total Score		Purified Factor Score		Factor Based Score		Total Score		Purified Factor Score		Factor Based Score	
R	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF	SF	MF
1	0	5	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
3	0	6	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
4	0	4	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0
5	0	5	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0

Note.
R = Replication
SF = Spurious Flags
MF = Missed Flags

Table C9.

Number of Spurious Flags (False Positives) and Missed Flags (False Negatives) for Factor Structure 2 (70/30) with Low Levels of Multidimensionality

	DIF Level: -.20						DIF Level: -.35						DIF Level: -.50					
	Total Score		Purified Factor Score		Factor Based Score		Total Score		Purified Factor Score		Factor Based Score		Total Score		Purified Factor Score		Factor Based Score Score	
R	**SF**	**MF**	**SF**	**MF**	**SF**	**MF**	**SF**	**MF**	**SF**	**MF**	**SF**	**MF**	**SF**	**MF**	**SF**	**MF**	**SF**	**MF**
1	0	11	0	0	0	0	6	0	0	0	0	0	52	0	0	0	0	0
2	0	22	0	0	0	0	11	0	0	0	0	0	50	0	0	0	0	0
3	0	20	0	0	0	0	13	0	0	0	0	0	46	0	0	0	0	0
4	0	15	0	0	0	0	12	0	0	0	0	0	44	0	0	0	0	0
5	0	14	0	0	0	0	11	0	0	0	0	0	41	0	0	0	0	0

Note.
R = Replication
SF = Spurious Flags
MF = Missed Flags

Appendix D

Item Loadings on the HSMAT

	Loadings on First Factor		Loadings on Second Factor	
	African-American	White	African-American	White
ITEM1	0.41791	0.083	-0.00398	0.37669
ITEM2	0.16386	-0.01313	0.0595	0.21902
ITEM3	0.18758	0.8379	0.09369	-0.37105
ITEM4	0.33193	-0.09029	-0.00602	0.25429
ITEM5	0.44493	0.0167	-0.04964	0.45093
ITEM6	0.19353	-0.00352	0.15886	0.33719
ITEM7	0.3026	-0.05553	0.14354	0.40817
ITEM8	0.28438	0.63695	0.06556	-0.05498
ITEM9	0.33574	0.49339	-0.08904	0.01254
ITEM10	0.49557	0.00451	-0.22754	0.61162
ITEM11	0.19244	-0.09409	0.21333	0.38442
ITEM12	0.41957	-0.0187	-0.2867	0.55477
ITEM13	0.14273	0.64525	0.29632	-0.06293
ITEM14	0.1621	0.58181	0.10137	-0.12786
ITEM15	0.36351	-0.03511	-0.07501	0.488
ITEM16	0.53321	0.72993	0.01534	0.121
ITEM17	0.19173	-0.19261	-0.00926	0.35987
ITEM18	0.23524	0.52512	0.13575	-0.02501
ITEM19	0.43259	0.62927	0.05109	0.08198
ITEM20	0.36459	0.62328	-0.04908	0.02766
ITEM21	0.30571	-0.09983	-0.23407	0.5025
ITEM22	0.36877	0.6155	0.01457	0.13124
ITEM23	0.40496	-0.11276	-0.01929	0.40733
ITEM24	0.53515	0.6215	-0.13286	0.14562
ITEM25	0.39618	0.671	0.14022	-0.01017
ITEM26	0.3352	-0.1079	-0.16615	0.63102
ITEM27	0.40415	0.67427	0.14342	-0.01043
ITEM28	0.49166	0.63749	-0.04894	0.16339
ITEM29	0.34421	0.45097	0.12787	0.08247
ITEM30	0.33719	0.60829	-0.14693	0.12939
ITEM31	0.34166	0.53034	-0.0864	0.1359
ITEM32	0.22512	0.59871	0.0203	0.05546
ITEM33	0.01261	0.46789	-0.04759	-0.08676
ITEM34	0.18862	0.4845	0.09754	0.02331

Item Loadings on the HSMAT (continued)

	Loadings on First Factor		Loadings on Second Factor	
	African-American	White	African-American	White
ITEM35	0.25045	0.46292	0.09747	-0.01476
ITEM36	0.27667	0.52479	-0.12367	0.17545
ITEM37	0.18899	0.52858	0.13744	0.01791
ITEM38	0.3158	0.66996	0.18688	-0.01617
ITEM39	0.30147	-0.18909	-0.12163	0.46671
ITEM40	0.44992	0.49804	-0.04204	0.27651
ITEM41	0.21095	0.4551	0.21189	-0.03349
ITEM42	0.27531	-0.26865	-0.0299	0.5225
ITEM43	0.34713	0.46405	-0.05537	0.05812
ITEM44	0.21234	0.50507	0.14607	0.06052
ITEM45	0.29518	-0.26558	0.2179	0.45108
ITEM46	0.43524	0.55273	0.16458	0.16789
ITEM47	0.1287	0.46327	0.07078	0.08231
ITEM48	0.12171	-0.32589	0.05985	0.38111
ITEM49	0.27561	0.56608	-0.09234	0.25716
ITEM50	0.35917	0.5285	-0.04809	0.17937