1-1-1993

An investigation of the effects of conditioning on two ability estimates in DIF analyses when the data are two-dimensional.

Kathleen M. Mazor
University of Massachusetts Amherst

AN INVESTIGATION OF THE EFFECTS OF CONDITIONING ON TWO ABILITY ESTIMATES IN DIF ANALYSES WHEN THE DATA ARE TWO-DIMENSIONAL

A Dissertation

by

KATHLEEN M. MAZOR

Submitted to the Graduate School of the
University of Massachusetts in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

September 1993

School of Education

AN INVESTIGATION OF THE EFFECTS OF CONDITIONING ON TWO ABILITY ESTIMATES IN DIF ANALYSES WHEN THE DATA ARE TWO-DIMENSIONAL

A Dissertation

by

KATHLEEN M. MAZOR

Approved as to style and content by:

Ronald K. Hambleton

Ronald K. Hambleton, Chair

HSwaminathan

H. Swaminathan, Member

Arnold Well, Member

John C Carey

Bailey Jackson, Dean
School of Education

ACKNOWLEDGEMENTS

I have learned a great deal, both academically and otherwise, during my years in this program. I am very grateful to the many people who have helped me over the years, and who have made the experience not only instructive, but enjoyable as well.

First, I would like to thank Ron Hambleton, who directed this dissertation. He has always been immensely generous with his time, his encouragement, and his advice. It was largely his enthusiasm which persuaded me to pursue this area of study, and he has been unwavering in his support in the years since. I truly appreciate the many hours he has devoted to this dissertation and numerous other papers we have worked on together, and I have learned a great deal as a result.

I would like to give a special thanks to Swami. Like Ron, he has been consistently supportive, and generous with his time and his knowledge. Over the years he has worked with, advised, and encouraged me, answering innumerable questions with patience and humor.

I would like to thank Arnie Well for serving on this committee. I greatly appreciate the time he has given, always cheerfully and graciously.

I am also grateful to many fellow students, past and present, who have generously shared their knowledge, resources, and friendship. I leave much richer for having had the opportunity to work with and learn from such an interesting, diverse, and stimulating group of people.

I have been especially lucky to have worked with Brian Clauser on numerous projects. Our collaborations have always helped me to a greater understanding of the problems we have studied, and lead me to new insights and new questions. But what I have valued even more has been his friendship- his support, encouragement, advice, and humor.

Finally I thank my husband, David, and my children, Kerry and Kacie for their understanding and support throughout this process. They have accepted my periodic preoccupations and time away from home with humor and love.

ABSTRACT

AN INVESTIGATION OF THE EFFECTS OF CONDITIONING ON TWO ABILITY ESTIMATES IN DIF ANALYSES WHEN THE DATA ARE TWO-DIMENSIONAL

SEPTEMBER 1993

KATHLEEN M. MAZOR, B.A., UNIVERSITY OF MASSACHUSETTS

M.S., EASTERN WASHINGTON UNIVERSITY

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Ronald K. Hambleton

Differential item functioning is present when examinees of the same ability, but belonging to different groups, have differing probabilities of success on an item. Traditionally, DIF detection procedures have been implemented conditioning on total test score. However, if there are group differences on the abilities underlying test performance, and total score is used as the matching criterion, multidimensional item impact may be incorrectly identified as DIF.

This study sought to confirm earlier research which demonstrated that multidimensional item impact may be identified as DIF, and then to determine whether conditioning on multiple ability estimates would improve item classification accuracy.

Data were generated to simulate responses for 1000 reference group members and 1000 focal group members to two-dimensional tests. The focal group mean on the second ability was one standard deviation less than the reference group mean. The dimensional structure of the tests, the discrimination of the items, and the correlation between the two abilities were varied. Logistic regression and Mantel-Haenszel DIF analyses were conducted using total score as the matching criterion. As anticipated, substantial numbers of items were identified as DIF.

Items were then selected into subtests based on item measurement direction. The logistic regression procedure was re-implemented, with subtest scores substituted for total score. In the majority of the conditions simulated, this change in criterion resulted in substantial reductions in Type I errors. The magnitude of the reductions were related to the dimensional structure of the test, and the discrimination of the items.

Finally, DIF analyses of two real data sets were conducted, using the same procedures. For one of the two tests, substituting subtest scores for total score resulted in a reduction in number of items identified as DIF.

These results suggest that multidimensionality in a data set may have a significant impact on the results of DIF analyses. If total score is used as the matching criterion very high Type I error rates may be expected under some conditions. By conditioning on subtest scores in lieu of total score in logistic regression analyses it may be possible to substantially reduce the number of Type I errors, at least in some circumstances.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS iv

ABSTRACT vi

LIST OF TABLES x

LIST OF FIGURES xi

CHAPTER

I. INTRODUCTION 1

Statement of the Problem 5
Purpose of the Study 13

II. REVIEW OF THE LITERATURE 14

Procedures for Detecting Differential Item Functioning . . 16
The Mantel-Haenszel Procedure 20
The Logistic Regression Procedure 23
Research on the Effectiveness of the MH
and LR procedures 24
Explaining DIF: Examining Item Characteristics 29
A Multidimensional Conceptualization of DIF 37
Support for the Multidimensionality Explanation of DIF . . 48
Summary 57

III. METHODOLOGY 59

Part I 61

Design of the Study 63

Sample Size 63
Test Length 64
Ability Distributions 64
Trait Correlations 65
Dimensional Structure of the Tests and Item
Parameters 65

Data Analysis 69

Part II 69

Subtest Selection 70

Preliminary NOHARM Investigations 70

Data Analysis 72

Part III 73

IV. RESULTS 75

DIF Analyses with Total Score as Matching Criterion . . . 76
Change in Matching Criteria 79

Supplemental Analyses 85
Characteristics of False Positive Items 88

Correspondence Between A Priori Results and NOHARM Results 91
Real Data Results 102

History Test Results 103
Chemistry Test Results 103

V. DISCUSSION AND CONCLUSIONS 106

Discussion 106
Implications 111
Summary 113
Delimitations of the Study 116
Directions for Future Research 118
Conclusions 119

REFERENCES 121

LIST OF TABLES

Table Page

1. Test Dimensional Structure 66

2. Item Parameters 68

3. Total Score Descriptive Statistics 77

4. Number of False Positive Errors with Total Score as Matching Criterion 78

5. A Priori Selected Subtests Descriptive Statistics 80

6. Number of False Positive Errors with Subtests as Matching Criteria 82

7. Summary of False Negative Errors 84

8. Descriptive Statistics for Test 2 Supplemental Conditions . 86

9. DIF Results for Test 2 Supplemental Conditions 87

10. Average Number of False Positive Errors Broken Down by Item Measurement Direction 89

11. Average Number of False Positive Errors Broken Down by Item Discrimination 92

12. Average Number of False Positive Errors Broken Down by Item Difficulty 93

13. Correlations Between True Cosines and Cosines Based on NOHARM Analyses 94

14. Correlations Between Subtest Scores 96

15. Item Classification Accuracy 99

16. Percentage of NOHARM Subtest Items with Measurement Direction Between 25 and 65 Degrees 101

17. Descriptive Statistics for the History and Chemistry Tests . 102

18. Number of Items Identified as DIF 104

LIST OF FIGURES

Figure Page

1. A Comparison of False Positive Error Rates 83

CHAPTER I

INTRODUCTION

Tests have become an integral part of modern society. In the United States test results are used to inform educational decisions regarding placement and advancement from kindergarten through graduate school. Outside of the realm of education, test results are used for selection, advancement, and competency assessment in industry, the military, and in a variety of professions. In addition, test results are often used in program evaluations to help assess the effectiveness of preventative, remedial and other social programs.

Because of the pervasive use of tests, and the importance of the decisions which are made using test results, both tests and the ways in which test results are utilized have come under careful scrutiny. One of the most important and frequently raised questions is whether tests are fair to all examinees. This is a question which has serious social and political ramifications, and which has been the focus of much litigation. Under the general issue of test fairness is the more specific issue of item bias.

The term "bias" has many connotations. In the field of measurement it does not necessarily have the same meaning which a layman might attach to it, and this has sometimes lead to confusion. For instance, one connotation which has drawn considerable attention is apparently biased item content. Minority group advocates and others have found instances of items which portray certain group members in ways that may be considered racist, sexist, stereotypical or demeaning.

While this is certainly undesirable, and such offensive content is best removed, as Scheuneman (1982) notes, such content may not in fact produce differences in performance.

A second apparent unfairness which is sometimes termed "bias" by those unfamiliar with the technical definition of the term is the observation that often there are considerable differences between groups as to how difficult a test or a test item is. Thus, one group may consistently score higher or lower than another on a particular type of test or item. This has been the focus of considerable controversy as some authors have sought to use this as evidence of inherent genetic differences. Such inferences are unfounded. Test scores alone do not provide sufficient information to validate such inferences, especially when there are so many factors known to impact on test performance and which are known to be inequitably distributed in our society.

It is now generally accepted that there may in fact be group differences in performance both at the item and the test level, and that these differences do not necessarily mean that the test is biased or unfair. Instead, such differences may accurately reflect real differences in the skill or ability the test is seeking to measure. Such differences in performance which are due to differences in the underlying ability distributions are typically referred to as "impact" (Holland & Thayer, 1988).

Differential item functioning, in contrast to item impact, refers to differences in performance which are observed _after_ differences in ability are controlled for. Mellenbergh (1989) offers thc following definition: "An item is considered to be biased when it differs in difficulty between subjects of _identical_ ability from _different_ groups."

(p. 128) Thus one group has a relative advantage even when differences in underlying ability distributions are controlled for. By definition any remaining differences in performance are due not to item impact, but to something idiosyncratic about that item, and the interaction between that item and the groups under study. There are a number of statistical techniques used for identifying DIF. The techniques which are currently most preferred are those which are conditional procedures (Mellenbergh, 1989; Hills, 1989; Scheuneman & Bleistein, 1989). Conditional procedures are consistent with the definition of item bias (DIF) which is presented above, as they allow for statistical control of differences in the underlying ability distributions when comparing examinees from different groups.

If an item functions differentially for two groups, it poses a threat to the validity of inferences which are made from the test, as one could argue that that item is measuring something other than what the test purports to measure, or at least something different from what the other items are measuring. As Shepard, (1982) writes, "Item bias methods detect items that are anomalous. Whatever the rest of the items measure, the biased item behaves differently." (p. 24) The question then arises as to what is causing these items to behave differently? Thus the second, and to some authors the more important question (Kok, 1988), becomes one of explanation. Some researchers (e.g. Scheuneman) have sought explanations through careful post hoc analyses of identified items. For the most part, these efforts have not been successful. Shepard et al. (1984) noted that "even minority experts could not predict with greater than chance success what types of items would be difficult for members of a particular group" (p. 95). Nor could such experts explain why certain items were flagged as DIF, while others were not.

The use of expert and/or minority judges for the purpose of identifying biased items is referred to as the judgmental approach. Studies which have compared the results of judgmental with statistical approaches have generally found little convergence between the two methods (Plake, 1980; Engelhard, Hansche, & Rutledge, 1990). What would seem on the surface to be a relatively simple task - looking at items which have been identified using statistical procedures as DIF, and through careful item review determine why those particular items were flagged so that such items could be avoided in the future - has turned out to be a far more difficult task than originally thought.

While the fact that there has been little convergence between the two approaches has been documented in the literature, the question of <u>why</u> this is the case has yet to be answered. Some authors have sought to answer this question by looking even more closely at the characteristics of items flagged statistically (e.g. Scheuneman, 1982, 1987), while others have looked more closely at the statistical procedures which are being used (Shepard, Camilli, & Williams, 1984) to determine whether statistical artifacts may account for the discrepancy. While their research found support for the efficacy of the statistical methods, the use of actual test results means that it is not possible to truly evaluate the power and accuracy of the statistical procedures. Thus, the question remains whether statistical techniques are consistently and correctly identifying true item bias.

Statement of the Problem

The lack of convergence of statistical and judgmental methods for identifying DIF has lead psychometricians to take a closer look at each approach. A careful evaluation of the statistical procedures requires examining whether these procedures are accurate in their classifications. That is, do these procedures consistently identify all of the items which are in fact DIF, without falsely identifying any non-DIF items?

One of the most popular and widely researched statistical procedures for identifying item bias is the Mantel-Haenszel (MH) procedure. This procedure has become known as a kind of industry standard. There are several reasons for the popularity of the MH. One of the most often cited is its theoretical basis, which is consistent with the definition of DIF which stresses that differences in ability distributions should be controlled for. The MH controls for differences in the ability by blocking examinees according to ability. In practice, total test score is most often used, as this is usually assumed to provide the best available estimate of ability. Typically, there is one block for each possible score, resulting in n+1 score categories, where n is the number of test items. If, after the examinees are matched, there is still a significant difference between the two groups in likelihood of success on a given item, the item is considered DIF.

From this brief description it can be seen that central to this procedure is the assumption that the measure which is used as the blocking criterion is a valid estimate of the ability which the test intends to measure. When total test score is used, the assumption is therefore that total test score provides such an estimate. In the case where the test is unidimensional this is a reasonable assumption. Most of the simulation studies which have looked at the performance of the MH have used a unidimensional model to generate the data. These studies have consistently found that the MH accurately identifies most items which are constructed to simulate DIF (Mazor, Clauser & Hambleton, 1992; Clauser, Mazor & Hambleton, 1991; Rogers, 1989). These studies have also found low false positive rates.

While the results of these studies are quite positive, two questions remain. First, are the simulation results generalizable to "real" data sets? Not surprisingly, research using the MH procedure with real data sets has yielded results which are much more difficult to evaluate. The obvious problem is that with real data sets it is not possible to know which items are in fact DIF, so that it is impossible to truly assess the accuracy of classifications. One concern is the instability of the statistic across samples. For example, Hambleton and Rogers (1989) found that when they replicated the MH analyses on two randomly constructed samples (both comparing Anglo-Americans to Native Americans) that the MH was 80% consistent overall, that is 80% of the decisions made on the first analysis were replicated on the cross-validation sample.

The question of whether total test score provides a valid estimate of ability becomes even more difficult with real data sets. There is some research which suggests that changing the criterion will substantially change the classifications of items (e.g. Mazor, Kanjee, & Clauser 1993; Clauser, Mazor, & Hambleton 1991; Ryan, 1991). The question which has yet to be answered is which criterion is the appropriate one, yielding the most accurate item classifications?

Again, with real data sets this question is virtually impossible to answer with certainty.

Another finding which has practitioners concerned is the finding that analyses of the same test items but with different samples may affect the stability of the statistics (Ryan, 1991; Kubiak & Cowell, 1990). Thus, while simulation studies have provided substantial evidence in support of the MH procedure, it is important to remember that these studies have generally used data which was generated by, and therefore fit, a unidimensional model. The results of analyses of real data suggest that in the "real world" the situation may be more complex.

One question which simulation studies have not addressed to date is the question of explanation - that is what makes an item more difficult for one group than another (after conditioning on ability). In general, looking at the characteristics of the items which were flagged has not proven fruitful thus far. However, the answer to this question may lie more in the definition of DIF, than in any particular item characteristics. Typically in simulation studies, DIF items are generated so that there are actual differences in the difficulty parameters between the two groups. This results in differences in the p-values (even after controlling for ability) and the item is flagged as DIF. Thus, simulations build in differences in item difficulties, but generally have not addressed what factors are responsible for such differences.

The question of what factors cause differences between groups in item difficulty (or in other item parameters) is central to an understanding of DIF. To address this question one must return to the definition of DIF. To briefly restate the definition, an item is considered non-DIF if examinees of the same ability have equal probabilities of getting an the item correct, regardless of group membership. However, by this definition, it would seem that if two examinees were indeed of the same ability, and that is the only ability which determines performance on that item, it would be logically impossible for there to be differences in performance (except those due to chance). Therefore, this definition implies that DIF is due to multidimensionality. If the item and the total test score were measuring the same unidimensional ability, or exactly the same weighted composite of abilities, then it would be impossible for differences in performance to exist except due to chance. Thus, if there are significant differences in performance, it must be the case that something other than that estimated by total test score (or whatever matching criterion is used) is influencing performance on that item. Therefore, the test must be multidimensional, and it is this multidimensionality coupled with differences in the underlying multidimensional ability distributions, which explains why an item appears DIF.

This conceptualization of DIF is not in fact new, but has been recognized for some time. The work of Kok (1988), Shealy and Stout (1993) and Ackerman (1992) may be seen as making more explicit the relationship between multidimensionality and DIF, and providing a framework for further work in this area.

An example of how multidimensionality may result in DIF may be useful here. Consider a hypothetical math test, composed of 45 two digit addition and subtraction items, and five problems which also require addition or subtraction of two-digit numbers, but in order to determine which operation to perform the examinee must read three or four sentences which set forth the problem. If the total test score is used as the criterion (and assuming all items are equally discriminating) and the Mantel-Haenszel procedure is run, examinees will be matched on a criterion which is primarily a function of what might be called for simplicity "math ability". (If second run results are used, the total test score is likely to be a purer measure of "math ability" as it is likely that at least some of the word problems would be flagged and therefore eliminated from the criterion.) Thus, matching examinees on total test score will have the effect of matching on math ability. However, five of the items require the second ability, call it reading ability, to be solved. The MH procedure as it is typically implemented (and most other DIF procedures currently in use) match examinees only on the primary ability (or on a weighted composite dependent on the number and type of the items in the test, and the discrimination values of these items). In any case (except when the test and the items are unidimensional) the result is that examinees are matched on some but not all of the relevant abilities, and differences in the underlying conditional ability distributions of the ancillary trait(s) may result in items being flagged as differentially functioning, when in fact differences in performance are due to actual differences in ability. There are a number of studies which provide evidence that this is in fact the case, and these are discussed in the next chapter (see for example, Oshima & Miller, 1990; Ackerman, 1992).

Shepard (1982) notes that the context in which an item is analyzed is extremely important. If a verbal item is embedded in a test comprised otherwise of math problems, then that verbal item is likely to appear biased. If the test is intended to measure only math ability, then the reading ability may in fact be a "nuisance" ability, and it would be desirable to remove items contaminated by that ability. However, it is possible to find examples of tests where it is not desirable to have either the tests or the items be unidimensional. In many situations it is considered preferable to have items be as "realistic" as possible, and in most cases realism means moving away from purely unidimensional items. In some contexts word problems may be considered more "realistic" than pure math problems. If the test is in fact intended to measure this second ability, then flagging such items as DIF is not desirable. But if there are differences in the ability distributions on this second ability, these items may be flagged. In this case test developers would probably not want to remove these items, and such differences would be more correctly labelled item impact than item bias. (Note: Ackerman and others consistently refer to the two abilities as theta and the nuisance ability. But the so-called nuisance ability may be an important ability which the test intends to measure. Therefore, the more neutral terms, first and second ability, or ability A and ability B seem preferable and will be used in this study.)

Thus, this line of reasoning leads to one answer to the question of what causes the differences in performance which are identified as DIF. Namely, that multidimensional items (or combinations of different types of unidimensional items within a single test) are prerequisite to items being identified as DIF. However, it is not just multidimensional items per se which cause the apparent DIF, it is the presence of items sensitive to more than one ability, coupled with between group differences in the multidimensional ability distributions, that results in the potential for bias (Kok, 1988; Shealy & Stout, 1993; Ackerman, 1992). This is because the examinees who are being compared are not in fact comparable. Holland and Thayer (1988) define comparability as "identity in those measured characteristics in which examinees may differ and that are strongly related to performance on the studied item." (p.130) Ackerman (1991) has done some preliminary research which suggests that this is the case.

One solution which has been proposed by some authors (e.g. Shealy & Stout 1993, Ackerman, 1992) is that rather than condition on total test score, one should select a valid subtest of items, and that the score on this valid subtest be used as the conditioning criterion. The reasoning here is that if an item is analyzed with the correct criterion, the criterion which accurately estimates the ability (or abilities) which one intends to measure, then the analysis will correctly identify those items which are not measuring this ability or abilities. The decision as to which items to use to construct such a subtest will depend on the intent of the test. Clauser, Mazor, and Hambleton (1991) using a judgmental method of constructing valid subtests found that item classifications did change as a function of the criterion which was used. Ryan (1991) found greater stability across different criteria, but this may be because the criteria used likely did not differ substantially in dimensionality. While the use of valid subtest scores appears to be a reasonable approach, it has yet to be thoroughly investigated.

Swaminathan and Rogers (1990) demonstrated that the MH procedure may be conceptualized as a special case of the logistic regression model. The logistic regression (LR) procedure, like the MH, controls for differences in ability distributions between groups. The LR procedure does this by incorporating an ability estimate (usually, but not necessarily, total test score) into the regression equation. Like the MH, the LR procedure provides a statistical test of whether group membership is significant. Unlike the MH, LR also allows for a test of whether there is an interaction between group membership and ability, which is a test for the presence of non-uniform bias.

Because the LR procedure has been introduced only recently as a procedure for detecting DIF, there is currently much less research available on it than on the MH. However, the research which is available suggests that it performs as well as the MH at identifying uniform DIF, and better at identifying non-uniform DIF. False positive rates were only slighter higher than those associated with the MH, and still quite low (Swaminathan & Rogers, 1990).

The LR model is relevant here not only because it is a promising new technique, but because the regression model lends itself readily to expansion. With respect to the issue of multidimensionality, an estimate of a second ability can easily be incorporated into the LR model. Thus, if much of what is currently being labelled as DIF is due to multidimensionality, then the LR procedure may provide the best model for taking this into account, and thereby could improve item classification accuracy.

If it is possible to model the process which results in items being identified as DIF using currently accepted detection procedures, and then to demonstrate how this apparent DIF essentially "goes away" if the analysis is modified to take into account a second ability, then our understanding of the relationship between DIF detection procedures and multidimensionality will be greatly enhanced. This would have important implications for how multidimensional tests are analyzed. In addition, this could lead to a rethinking of both the judgmental and statistical procedures currently in use, and could well lead to a greater convergence between the two.

Purpose of the Study

The first purpose of the present study was to investigate the conditions which influence whether multidimensional items are identified as DIF. It was demonstrated that multidimensional tests resulted in high false positive error rates when there were between group differences in the underlying multidimensional ability distributions, and examinees were matched on total test score.

The second purpose was to determine whether these high false positive error rates would be reduced by selecting items into relatively more unidimensional subtests, and then conditioning on both subtest scores simultaneously.

Finally, analyses of two real data sets were conducted following the procedures used in the analysis of the simulated data. The purpose of this phase of the study was to assess whether the results obtained in the first part provide a realistic model of what might be encountered "in the real world," and therefore whether the findings from this simulation study were generalizable to real test data.

CHAPTER II

REVIEW OF THE LITERATURE

Studies of DIF may be generally conceptualized as studies of whether different groups show differing responses to test items (Mellenbergh, 1982). In early DIF studies if a particular item was more or less difficult for examinees depending on group membership, the item would be considered DIF. Group differences in ability were not taken into account, which is why approaches using this definition are referred to as unconditional approaches.

While the simplicity of this approach may make it appealing to lay readers, it has lost credibility in the measurement community. It is now widely agreed that differences in performance associated with group membership, while possibly due to DIF, may also be attributable to real differences in ability between the groups under study. For a test to be valid it is desirable that items be sensitive to these differences. When apparent differences in performance can be attributed to differences in the underlying ability, the difference is more appropriately labeled impact rather than DIF (Holland & Thayer, 1988). Because of this, virtually all of the currently accepted definitions of DIF make explicit reference to the need to ensure that underlying differences in ability are taken into account. As Holland and Thayer (1988) write, "Basic to all modern approaches to the study of differential item functioning is the notion of comparing only comparable members of the reference and focal groups." Approaches to DIF which control for underlying between group differences are called conditional approaches.

Shepard et al. (1984) define DIF (item bias) as follows: "For an individual item, bias is defined as the difference in the probability of answering correctly, given equal ability" (p. 101). There are other definitions of DIF in the literature with slight variations in wording, but there is wide if not unanimous agreement as to the two crucial components to this definition: first that there is a difference in performance, and second, that this difference remains after controlling for between group differences in ability.

It would seem a relatively straightforward matter to work from this definition of DIF to develop procedures for identifying and eliminating DIF. As Scheuneman wrote in 1987,

> At one time an orderly progression was envisioned as follows: a) Devise procedures for reliably detecting those items that are performing differently for the groups of interest; b) examine the items and identify causes for the differential performance; c) develop procedures for modifying the items so that the differential performance is reduced or eliminated; and d) develop guidelines for item writers so that future items are free from such biases" (p. 97).

Scheuneman, in retrospect, concluded that the expectation of a straightforward, orderly progression was naive.

The four steps which Scheuneman outlined might be reconceptualized as three: identification, explanation, and elimination. The remainder of this literature review is organized consistent with this framework. First, the most widely accepted procedures for identifying biased items will be presented and discussed. Next, research relevant to the explanation of DIF will be reviewed, with an emphasis on the conceptualization of DIF as multidimensionality. Finally, the implications of a multidimensional explanation of DIF for the third area, the reduction or elimination of DIF will be discussed.

Procedures for Detecting Differential Item Functioning

The definition of DIF presented above is readily translated into item response theory (IRT) terms. In IRT, examinee performance on a test item is modeled as a function of an underlying ability or trait. (Our discussion at this point will focus only on unidimensional IRT models, although multidimensional models are also used, and will be discussed later.) There are several different IRT models currently in use. Logistic models are probably the most popular currently, and may include one, two or three item parameters. One-parameter models model performance as a function of ability and a single item parameter, usually referred to as item difficulty, or b. When there is no guessing (as in the one- or two-parameter models) b is the point where the probability of getting the item correct is 50%. One parameter models are based on the assumptions that items differ only in difficulty, that guessing is minimal, and that all items are equally discriminating. The two parameter model, in addition to the item difficulty parameter, also includes a parameter for item discrimination, referred to as a. The a parameter is proportional to the slope of $P_i(\theta)$ at the inflection point of the curve. The three parameter model includes a third parameter, often referred to as the pseudo-guessing or c parameter, which represents the probability of examinees of extremely low ability answering the item correctly.

Each of the IRT models allows for estimation of an item characteristic curve (ICC). The ICC is a curve which is determined by the specific model chosen, and the item parameters estimated from the data. The ICC specifies the relationship between the probability of success on the item, and the underlying ability or trait.

One of the assumptions of IRT is that the estimates of the item parameters are invariant. This means that these estimates do not depend on idiosyncracies of the sample on which they are estimated, but rather should remain stable across samples. Thus, if a particular item is administered to one group, and the item parameters are estimated, and then the same item is administered to another group, and the item parameters are also estimated, the parameters should be the same (once they have been set to the same scale). If there are differences in the parameters it means that examinees from the two groups are responding differently to the item, which is one way of defining DIF.

This is best illustrated graphically by superimposing the ICC for the second group over that of the first. Then, for any level of ability, it is possible to determine what the probability of success on that item is. If the ICCs are the same, the probability of success will be the same, regardless of group membership. However, if the ICCs differ, the probability of success will also differ for examinees in the range of ability where the curves are divergent. Thus, to define DIF in IRT terms is to say that an item is differentially functioning if the ICCs for that item differ significantly across groups (Hambleton & Swaminathan, 1985).

There have been a number of IRT-based procedures proposed for identifying bias. One of the best known is commonly referred to as Lord's chi-square method (Lord, 1980). In this method, the $\underline{a}$ and $\underline{b}$ parameters are estimated separately for both groups, are transformed onto a common scale, and are then compared simultaneously. The equality of the item parameters is evaluated using a chi-square test.

Another group of IRT based procedures focus on the area between the ICCs. For these procedures the problem is to calculate the area between the curves, and then to determine whether the area reflects a significant difference. Shepard, Camilli and Williams (1984) present formulas for evaluating the area between two curves. However, since that time Raju (1988) has presented formulas for computing the exact area between two ICCs (both signed and unsigned). When these were first presented, no associated test for significance was available. Since that time Raju (1990) has presented procedures for testing the significance of both signed and unsigned areas. If the ICCs do not cross (uniform DIF) the signed and unsigned indices will be the same. However, if they do cross (non-uniform DIF), DIF in one direction in one region may be offset by bias in the other direction in another region, and thus the signed indices may be low.

A third group of IRT-based procedures involves calculating the difference in probabilities of success, and then squaring and summing these. These are aptly referred as the sum of squares (SOS) methods. Shepard, Camilli, and Williams (1984) present formulas for both signed and unsigned SOS indices. They found that of the indices they evaluated (Lord's chi-square, SA, UA, SOS, and USOS) that the sum-of-squares statistics (weighted by the inverse of the variance errors) appeared to be the best.

There is a very clear and direct connection between the generally accepted definition of DIF, and IRT. IRT allows for evaluation of response differences after controlling for or conditioning on ability.

Mellenbergh (1982) argues the reason that IRT methods are to be preferred to other conditional methods, is that IRT methods allow for conditioning on true ability, versus observed score which stands as a proxy for true ability in most other procedures.

While assessing DIF from an IRT perspective has considerable theoretical appeal, there are a number of practical issues which must be considered. One of the most frequently cited is the need for large sample sizes. For procedures which require the use of LOGIST to estimate item parameters (for the three parameter model), a minimum of 1000 examinees per group is recommended. Depending on the testing program, this may or may not be feasible. A second concern is that even if a sufficient number of examinees are available, LOGIST is a difficult and expensive program to run. A third concern is that IRT methods may be conceptually difficult to explain to a naive audience. Fourth, some of the IRT methods do not have associated tests of significance (for instance SOS methods), or the significance test depends on a series of decisions regarding the ability range which is to be considered (i.e. Raju's area method). Fifth, parameters cannot always be equally well estimated for both groups, differences in the ability distributions may be problematic. Sixth, some of the procedures require the practitioner to make decisions which may require some expertise or experience, such as over what range of ability should DIF be evaluated. Finally, the utility of all of the IRT methods is predicated on the fact that the model used must fit the data. Thus, while many authors agree on the theoretical merits of an IRT approach to identifying DIF, (Scheuneman & Bleistein, 1989; Mellenbergh, 1982;

Hills, 1990) these same authors generally acknowledge that practical constraints may preclude the use of such methods in some circumstances.

The practical drawbacks to IRT methods have led practitioners to consider alternate methods. There are procedures which might be considered approximations to IRT techniques, but which overcome some of these practical problems. One of the most popular of these is the Mantel-Haenszel procedure (MH).

The Mantel-Haenszel Procedure

The MH procedure was originally introduced in 1959 by Mantel and Haenszel, who proposed it for use in the retrospective study of disease. Holland and Thayer (1988) introduced the MH procedure to the testing community for the purpose of identifying DIF. They argued that the MH procedure was a natural extension of the chi-square procedures which had been advocated until that time. However, the MH procedure improved on previous approaches by substantially improving the conditioning (going from 5-10 score groups to n+1 score groups where n=number of items).

The MH procedure tests whether the odds of success on a given item are proportional for both groups across all levels of the matching criterion. This is done as follows. First, examinees are sorted into score categories according to their score on the matching criterion. When the total test score is used as criterion, there is one category for each possible score, including zero. It is possible to collapse the data to form fewer score categories if desired. The data are then arranged in a series of 2 X 2 tables, with one table for each score category. The arrangement for the jth matched set of examinees would be as follows:

		Score on the Studied Item		
		1	0	Total
Group	Reference	A_j	B_j	n_{rj}
	Focal	C_j	D_j	n_{fj}
	Total	m_{1j}	m_{0j}	T_j

Where T_j is the total number of focal and reference group members in the jth matched set, n_{rj} is the number of those who are in the Reference group; and, of these, A_j answered the studied item correctly. The other cell frequencies are similarly defined. The null hypothesis of no DIF is that the proportion of examinees passing the item in the reference group equals the proportion for the reference group, for all score group levels.

The Mantel-Haenszel chi-square statistic (MHCHI-SQ) is used to test this hypothesis. This statistic is written as follows:

$$\text{MH Chi-square} = \frac{(|\Sigma_j A_j - \Sigma_j E(A_j)| - \tfrac{1}{2})^2}{\Sigma_j \text{var}(A_j)}.$$

This form includes a continuity correction. The var(A_j) is given by

$$\text{var}(A_j) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{T_j^2 (T_j - 1)}.$$

Under the null hypothesis the MH-CHISQ has an approximate chi-square distribution with one degree of freedom.

Holland and Thayer recommend implementing the MH procedure in a two-step process. In the first step a preliminary analysis is conducted to identify suspect. Next, any items identified as DIF are removed (with the exception of the studied item) and a "purified" total score is calculated, and the MH analysis is repeated with the "purified" total score as the matching criterion. This is so the matching criterion is as "clean" an estimate of ability as possible.

Since it's introduction in 1988, the MH procedure has gained steadily in popularity. In fact, it is has become a kind of industry standard. Both Scheuneman and Bleistein (1989) and Hills (1989) cite the MH as a procedure which is both theoretically sound, practical to implement, and supported by current research findings. There are a number of advantages of the MH which probably contribute to its popularity. First, it allows for matching of examinees at a relatively fine level - that is at every possible score group. While this is not necessarily equivalent to conditioning on true score or true ability, it comes closer than the earlier chi-square approaches and generally satisfies the requirement of ensuring that only comparable members of the reference and focal groups be compared. By conditioning in this way the procedure approximates IRT procedures in one sense, without the need for the often costly and complex computer runs necessary for some of the IRT-based procedures. Writing and running programs to calculate the MH statistics is relatively simple and inexpensive, again as compared to many of the IRT based procedures. In addition, the MH can be used with smaller sample sizes than many of the IRT approaches, although claims that 100 examinees per group are sufficient (Hills, 1990) are probably not warranted (Mazor, Clauser, & Hambleton, 1992).

Another reason for the popularity of the MH which has frequently been cited is that it is conceptually simpler, and therefore more easily explained to many audiences. Finally, and perhaps most importantly, there is considerable research which suggests that MH procedure yields results similar to those obtained with IRT procedures, has very good detection rates, and low false positive rates.

The most frequently cited disadvantage to the MH procedure is that it is relatively insensitive to non-uniform DIF. That is, if an item favors one group at one end of the ability range, and the other group at the other end of the range, that the DIF will essentially cancel itself out, and the item will not be identified. In terms of ICCs, non-uniform bias refers to the case where the difference between the curves is not equal across all ability levels. This reflects an interaction between group and ability level. If the curves cross only at the outer ranges of ability the MH may correctly identify the item as DIF. On the other hand, if the ICCs cross close to the middle of the ability distribution, the MH is not likely to flag the item as the bias will essentially cancel itself out. This can be predicted from a theoretical analysis of the procedure, which does not allow for an interaction of group with ability. This shortcoming of the MH was one of the factors which led to Swaminathan and Rogers (1990) developing a logistic regression procedure as a DIF detection method. While there is some indication that a modification of the MH procedure would increase detection rates for non-uniform DIF (Mazor, Clauser, & Hambleton, 1992) the LR procedure is more statistically sound and can easily be extended to handle multiple conditioning variables.

The Logistic Regression Procedure

The logistic regression model proposed by Swaminathan and Rogers may be written as follows:

$$P(u_{ij}=1)=\exp(\beta_{0j}+\beta_{1j}X_{ij})/[1+\exp(\beta_{0j}+\beta_{1j}X_{ij})]$$

where u_{ij} is the response of person i in group j to the item, $ß_{0j}$ is the intercept parameter, and $ß_{1j}$ is the slope parameter for group j, and X_{ij} is the "ability" of examinee i in group j.

As Swaminathan and Rogers point out, the MH can be conceptualized as a special case of the LR procedure (although it was developed through a different line of reasoning). In addition to allowing for detection of interaction between group and ability, the LR model differs in how the ability variable is treated. The MH procedure treats ability as a discrete variable, and ignores the ordinal nature of the ability scale. In contrast, the LR model makes use of this information. Another advantage of the LR procedure with respect to the current research is that additional variables are readily accommodated by the regression equation, and therefore it can be expanded to allow conditioning on two (or more) variables simultaneously.

Research on the Effectiveness of the MH and LR Procedures

Hambleton and Rogers (1989) conducted a study which compared the results of the MH procedure to an IRT based area method. Given that IRT methods are considered theoretically optimal, correspondence between the IRT methods and the MH would provide evidence of the efficacy of the MH approach. Using data from a statewide high school proficiency exam, Hambleton and Rogers found substantial agreement between the two methods. The items flagged by the MH method were essentially a subset of items flagged by the area method. A close examination of the items consistently missed by the MH but flagged by the area method revealed that the DIF present in four out of the five such items was non-uniform, that is the ICC for the two groups crossed. It was not surprising that the MH did not identify these items, as the MH does not allow for an interaction between group and ability. In fact, as noted above, it is this fact which is often cited as the primary criticism of the MH procedure (e.g., Scheuneman & Bleistein, 1989).

While the Hambleton and Rogers study provides one line of support for the MH procedure, DIF research based on real data is always susceptible to the criticism that one cannot know for certain which items are in fact differentially functioning, and thus detection rates cannot be fairly evaluated. Recent studies using simulated data suggest that the MH has very good detection rates when substantial DIF is present. Mazor, Clauser and Hambleton (1992) simulated several tests using a unidimensional three-parameter logistic model. They introduced DIF items by changing the item parameters for the focal group so that a number of items (approximately 15 percent) were more difficult for the focal group (that is the _b_'s for this group were higher than those for the reference group). They then analyzed these tests with the MH procedure in an effort to identify those items which had been constructed to exhibit DIF. They found very good detection rates with samples of 500, 1000 or 2000 per group. For instance, when comparing groups of equal ability and with 2000 examinees per group, the items which the MH did not flag were those with _p_-differences of .03, a difference of little, if any practical significance. With 200 examinees per group, items with _p_-differences of .17 were missed, and with sample sizes of 100 _p_-differences as large as .23 were missed. The pattern of results for groups of unequal ability (where the mean of the focal group was set to be one standard deviation less than that of the reference group) was similar, but detection rates dropped somewhat. For instance, p- differences of .15, .17 and .23 were missed at sample sizes of 1000, 500 and 200, respectively. When the ability distributions were equal, the percentage of DIF items correctly identified was three to ten percent more than with unequal distributions, depending on the sample size. An examination of the item parameters of the items which were most likely to be flagged revealed these were moderately difficult items, with large b differences. The items most likely to be missed were the most difficult items, items with very small b-differences, and poorly discriminating items. This pattern of results was consistent across all sample sizes. This study also found very low false positive rates for the MH, with only one of the 59 non-biased items being consistently identified at sample sizes of 1000 and 2000.

Rogers (1989) conducted a simulation study which looked at the power of both the MH procedure and the LR procedure. Her results provide support for the viability of both procedures.

Rogers varied model-data fit, sample size, test length, shape of the test score distribution, proportion of DIF items, and type and amount of DIF. In this study, the amount of DIF present was operationally defined in terms of the area between the ICCs. Four levels of DIF were simulated, so that the area between the ICCs for the two groups was .2, .4, .6 or .8. Both uniform and non-uniform DIF were simulated. For the items constructed to show uniform DIF, the difference in the b values between the two groups for the four areas were .22, .42, .62 and .82 for the smallest to largest areas, respectively. Each condition was replicated 20 times. Rogers found very good detection rates for both procedures when uniform DIF was present. The performance of the two procedures was very similar in this circumstance, with the MH being slightly better in most conditions. When the ICCs for the two groups differed by more than .2, detection rates were between 65 and 80% across all conditions for both procedures. Detection rates were substantially lower (25-30%) with areas of .2. Rogers noted that detection rates for both procedures improved as sample sizes increased, percent of DIF items decreased, and size of the DIF increased. Overall, Rogers concluded that both the LR and MH procedures were effective in detecting uniform DIF, with MH being slightly better under most conditions.

While the LR and the MH procedure showed very similar performance when the type of DIF was uniform, there were substantial differences between the two procedures when non-uniform DIF was simulated, with the LR procedure producing markedly higher detection rates. While the LR was as effective in identifying non-uniform DIF as it was in identifying uniform DIF, the MH procedure was only about half as effective. The detection rates for the LR procedure with non-uniform DIF were found to be up to 90% higher than the rates for the MH.

In addition to evaluating the performance of these two procedures with respect to identification of DIF items, Rogers also conducted a separate simulation study which evaluated whether the statistics for the procedures met their distributional assumptions. The logistic regression procedure was found to have the expected distributional properties in most conditions. The MH procedure was not distributed as expected in some cases, but there did not appear to be a consistent bias. Rogers concluded that both procedures adequately fulfilled their underlying assumptions. Rogers also looked at the Type 1 error rates for both procedures, and found false positive rates in the expected range. In fact, false positive rates were slightly lower for the MH than for the LR. In conclusion, Rogers recommends the LR procedure over the MH procedure, arguing that LR is a simple, unified, and powerful procedure which enables the detection of both uniform and non-uniform bias. She cites as advantages the fact that it is theoretically defensible, has an associated test of significance, can be used in small samples, and is relatively inexpensive to implement. She notes that LR is more accurate than the MH in detecting non-uniform DIF, and therefore is to be preferred over the MH.

Thus, these simulation studies have provided evidence that the MH procedure is very effective at identifying items in which DIF is known to be present. Such studies have also confirmed that the false positive rates are well within the expected range, and in some cases better than expected. The Hambleton and Rogers study found substantial convergence between the MH and IRT-based area methods with a real data set, further evidence that the results of the MH are valid and accurate.

Thus far we have seen that it is possible to conduct statistical analyses which are consistent with the definition of DIF which we started with, and that it is possible to recover simulated DIF with these techniques. We have argued that the MH procedure provides a good approximation to IRT approaches, as long as the ICCs for the two groups do not cross. If this is the case, that is if non-uniform DIF is present, the LR procedure is more effective in identifying DIF. Thus, in terms of the original tasks as outlined by Scheuneman, either the MH or the LR procedure will provide a reasonably good means of identifying DIF items.

Explaining DIF: Examining Item Characteristics

The next step in the progression outlined by Scheuneman is to determine why certain items are identified as DIF. A number of researchers have approached this problem by examining the items, and looking for item characteristics which would cause members of one group to respond differently than members of the other. While many studies have been reported in this area, perhaps the most consistent finding is that there has been no consistent finding.

Scheuneman has been one of the most prominent researchers in this area. In 1984, she noted that one of the most common hypotheses as to why DIF occurred was that these items had content that was differentially familiar to certain groups. However, as Scheuneman notes, this hypothesis has not been supported. While occasionally such items are identified, Scheuneman notes that "a more common result is that the researcher is unable to interpret his/her results." (Scheuneman, 1984, p. 221). Because of this, Scheuneman argued that the causes of DIF must be pervasive rather than idiosyncratic. That is, while differential familiarity with certain content might affect performance on that particular item, Scheuneman argued that researchers would do better to look for more pervasive sources of DIF, sources that would be likely to influence performance on several items and gave as examples of such influences the adequacy of instructions, reading load and cues to the testwise.

Scheuneman (1987) sought to experimentally induce DIF into a test. Based on previous research and experience, she generated a list of seven general characteristics of test items which she hypothesized could differentially influence performance. These included such things as format, wording, vocabulary, and test-wiseness. Scheuneman then constructed pairs of items so that the target characteristic was present in one item of the pair and absent in the other. These items were then administered as part of a GRE administration, and differences in performance for blacks versus whites were analyzed. While Scheuneman found that manipulating the items in this manner did appear to have differential effects on performance for several of the characteristics investigated, the effects were not always straightforward. The effects of the various item characteristics interacted with other characteristics of the items, suggesting that the manipulated characteristics were not the only characteristics to affect performance. Surprisingly, in some cases the differences between whites on different versions of the questions were greater than the differences between blacks and whites. In conclusion, Scheuneman wrote "What emerges most clearly from this study is how little we know about the mechanisms that produce differential performance between black and white examinees." (p. 117).

Schmitt (1988) looked at items which were identified (using the standardization method) as exhibiting DIF in comparisons between white and Hispanic examinees on the SAT verbal test. She found some evidence that true cognates (words whose stem mean the same in English and Spanish) were somewhat easier for Hispanics as opposed to whites. She also found that items with content of special interest to Hispanics seemed to be a factor in some items (with Hispanics doing better on these items than whites). However, one of the problems with this study, which is in fact common to many studies of this type, is that while a review of the items flagged by the statistical technique may suggest certain hypotheses, these really cannot be confirmed until and unless it is possible to make predictions about the entire set of items - that is if a review of the statistically identified DIF items reveals that these items all contain true cognates, is it also true that the items which were <u>not</u> flagged did not contain true cognates? If the entire set of items is examined (rather than only those items which were flagged) is the presence or true cognates a good predictor of whether a given item will be flagged statistically? Schmitt did conduct a correlational analyses to look at the relationship between DIF statistic values and item characteristics. She reports that the results of this analysis were not conclusive.

Schmitt and Dorans (1990) looked at the characteristics of items found to function differentially for Blacks and Hispanics. They reported some evidence that special interest items and items containing homographs were differentially difficult for certain groups. However, they also noted that there were instances of DIF for which they could find no apparent reason. They concluded by remarking that while their results suggest some of the causes of DIF have been identified, there appear to be other causes which have yet to be identified.

McLarty, Noble, and Huntley (1989) examined the effects of gender related content on DIF. They constructed what they labeled neuter, male and female versions of mathematics and English items. The versions differed in references to male or female names, pronouns, possessives and occupations. The items were then administered to samples of high school students, so that each item was completed by approximately 300 examinees. The data were analyzed using loglinear methods. McLarty et al. tested for two significant interaction effects - first, an interaction between response, item gender, and examinee sex, which would support an hypothesis of differential item difficulty on the basis of sex, and second, an interaction between response, item gender, examinee sex, and examinee ability, which would correspond to a finding of differential discrimination on the basis of sex. In fact, neither of these interactions were significant. McLarty et al. concluded that there was no evidence that there manipulations resulted in sex bias.

Ellis (1989) examined differential item functioning in the context of translated tests. Using an extended process of translations and back translations, she had an American group intelligence test translated into German, and a German group intelligence test translated into English. She then administered both tests (for a total of 251 items) to both American and German examinees. Approximately 200 examinees were in each group. Thus, each group took both tests, but all items were in the examinees' native language. Using Lord's chi-square test, Ellis tested the difference between the item parameters for the two groups. She found ten of the 251 items were identified as differentially functioning using a significance level of .01. She then conducted a content analysis of the items, and found plausible translational or cultural explanations for nine of the ten items identified. For some items the difference in performance appeared attributable to an error or flaw in the translation. For others, differing cultural experiences appeared responsible.

Scheuneman and Gerritz (1990) investigated the relationship between the MH delta statistic and a variety of item characteristics. They classified reading items from the SAT and GRE with respect to content, demand level, propositional analysis, passage structure, and option structure. They then conducted a series of regression analyses with the MH delta statistic as the dependent variable. They found that for the male/female comparisons, the predictor variables they had identified accounted for 25.5% of the variance in the MH delta in the SAT data set, and 44.5% in the GRE data set. For the Black/White comparisons, the percentage of the variance accounted for was 28.4 and 39.7 for the SAT and GRE, respectively. They noted that the effects for passage content were the most marked. In conclusion, Scheuneman and Gerritz remarked that their results suggest that while researchers have often sought a single, identifiable cause of DIF, such a cause may not in fact be present. They suggest that instead DIF may be attributable to an "unfortunate combination" of item features, or the cumulative effect of several small, and singly undetectable, effects. They suggest that this may be the reason that most post hoc analyses of items with extreme DIF values have not generally found explanations.

Thus, from this sample of studies which have looked at the characteristics of items, and sought to explain DIF from this standpoint, it can be seen that the results are inconclusive at best. While a number of studies have identified certain characteristics as associated with differential performance in the context of that study, researchers have consistently found apparent DIF for which they cannot find an explanation. Attempts to predict DIF based on item characteristics have met with limited success. Attempts to elicit DIF based on manipulations of item characteristics have not produced straightforward results. Thus this line of research, focusing on primarily on specific item characteristics has not yet satisfactorily answered the question of why certain items are flagged as DIF.

While the statistical procedures discussed earlier have been demonstrated to be accurate, judgmental procedures for identifying DIF have been much less successful, and generally have much less credibility in the measurement community. This should not be surprising if we return to Scheuneman's sequence of tasks. In this conceptualization, the task of identifying those item characteristics (or other variables) which are responsible for the differences in performance is a necessary first step. Only after the causes of DIF are understood would one approach the task of training judges to identify such items a priori. If we do not know what the judges are to look for, how can they be trained, and how can they be expected to predict which items will be flagged? Plake (1980) argued that demands of the statistical procedures for detecting DIF in terms of requirements of professional expertise, and computer costs and accessibility, made these (statistical procedures) unattractive. It is not entirely clear which statistical techniques Plake is referring to as being prohibitively complex and expensive, but she uses an analysis of variance procedure in her paper. Given recent advances in computer technology, and the widespread acceptance of procedures such as the MH, this argument might not be accepted today. However, at that time she argued that the ready availability of "experts" (with respect to the specific test content), and the fact that expensive computer and statistical consultants could be avoided, made the use of judgmental reviews attractive. Plake acknowledged that any judgmental review was by definition subjective, and thus some assessment of the correspondence between judgmental reviews and the statistical procedures was warranted. Plake conducted such a comparison, and found little relationship between the two procedures. The judges identified twice as many items as did the statistical procedure. It should be noted that the statistical procedure she used (ANOVA) has since been demonstrated to be a less than optimal technique for identifying DIF. However, Plake also noted that the judges often did not agree with each other, and some of their ratings appeared to be determined more by the characteristics of the raters than by characteristics of the items.

More recently, Engelhard, Hansche, & Rutledge (1990) also looked at the convergence of judgmental and statistical procedures. In this study they asked 42 judges to predict which items would function differently for Black and White examinees. They also found very poor convergence between the two techniques, with agreement being in the range which would have been expected by chance. Engelhard et al. did find however that there were some judges whose ratings did show greater convergence.

Shepard, Camilli and Williams (1984) noted the lack of convergence between judgmental and statistical techniques for detecting bias, and wondered whether the statistical techniques might be falsely identifying some items as DIF - that is whether some of the results obtained as a result of statistical analyses might be attributable to statistical artifacts, rather than real DIF. Using responses from thousands of high school students (in the High School and Beyond testing program) they used several IRT-based procedures to look for DIF. These were signed and unsigned area methods, four variations on the SOS methods, and Lord's chi-square. Overall, they looked at pseudo-ethnic comparisons (e.g., white/white comparisons) and contrasted these with true group comparisons. In the pseudo-ethnic comparisons there were few large DIF indices, in contrast to the true ethnic comparisons, where several items were identified. Shepard et al. interpreted this as evidence that the IRT-based procedures were identifying "true" DIF, and not artifacts. They found the weighted SOS statistics to be the best indices for quantifying differences between the ICCs.

Shepard et al. went on to examine the items which were consistently identified as DIF. They found that for the math test results there appeared to be a pattern which suggested a plausible explanation - items which were identified as DIF against Blacks appeared to have a significant verbal component. However, for the math items which were more difficult for whites no explanation was apparent. This was also the case for the vocabulary test, where a review of the items did not suggest any pattern or apparent reason for the difference in performance.

Scheuneman (1982) notes that what item reviewers are most likely to flag as biased are items which are stereotypical or offensive, and while it is important to correct these kinds of items, it is not necessarily these items which produce performance differences. Hills (1989) argues that subjective item reviews (occurring before or instead of statistical analyses) may result in the removal of items which are not actually DIF. Hills implies that this may be detrimental in that "good" items could be removed unnecessarily. Finally it may also be the case that subjective reviews narrow the field in another way - by removing items which are in fact differentially functioning. However, if these items are removed at an initial review stage, it is unlikely they will be administered to examinees, and hence will not be available for statistical analyses. This would thus impact on the degree of apparent convergence.

Thus, neither expert judges nor researchers who conduct extensive post hoc analyses of identified items have been able to satisfactorily predict or explain DIF. This suggests that perhaps a different approach to the problem of explanation is needed. All of the statistical methods discussed above assume unidimensionality. A number of authors have argued that apparent DIF is in fact due to multidimensionality in the data set. The argument and evidence to support this view are presented next.

A Multidimensional Conceptualization of DIF

The argument that differential item functioning is a manifestation of multidimensionality is not a new one. In fact, the definition of DIF directly implies that if DIF is apparent, multidimensionality must be present. Earlier DIF was defined as present if examinees of equal ability, but belonging to different groups, have unequal probabilities of success on an item. One of the most important features of this definition is the concept of comparing only comparable members of the two groups. If one conditions on one ability, the intended to be measured ability, and there are still differences in performance, it therefore follows that the test must therefore be measuring something other than this single ability for at least one of the two groups. Therefore, the test must be multidimensional, with respect to at least one of the two groups. Thus, the apparent DIF must be attributable to this multidimensionality.

Kok (1988) was among the first to explicitly develop this argument. He argued that if it is established that the ICCs for two groups on a given item do in fact differ, that this does not necessarily indicate that the item is unfair. He suggests that "judgments about the possible unfairness of an item requires knowledge of the mechanisms underlying the occurrence of non-coinciding ICCs." (p. 264) To illustrate his point he gives the example of test designed to measure verbal ability, but contains some items which also require some special knowledge, that may not have been covered in all school districts. If examinees from these disadvantaged school districts score lower on these items, and it is because they are actually less able on this special knowledge dimension, "it remains a point of discussion whether the item is unfair" (p. 264). That is, the item or test may be multidimensional. Kok proposes a mathematical model to make explicit the relationship between test multidimensionality and DIF.

Kok (1988) begins by operationalizing the concept of dimensionality. He cites Lord and Novick (1986) who write that "an individual's performance depends on a single underlying trait if, given his value on this trait, nothing further can be learned from him that can contribute to explanation of his performance" (p. 538). In IRT, this is expressed by the concept of local stochastic independence. A test is considered _n_ dimensional in a psychometric sense if stochastic independence between the items is observed only after conditioning on _n_ latent traits. Judgments about the dimensionality of an item are meaningful only with respect to a specific population - a test may be _n_ dimensional in one population, and _n_ + 1 dimensional in another. Thus the dimensionality of a given data set is really a function of both the examinee sample, and the item parameters.

Kok describes three ways that a data set may be unidimensional for a given subpopulation: (1) a single ability is relevant (Kok defines relevance as covariance with the probability of success on an item); (2) other abilities may be relevant in the full population, but in the subpopulation in question these abilities do not covary with the probability of success. This could occur for example if all examinees had the same level of some secondary ability, say reading;
(3) A test may be unidimensional for a given subpopulation even if abilities other than theta are relevant if those abilities affect performance on one item only, analogous to unique factors in factor analysis. Further, Kok writes that "In general, if n abilities are relevant, the test administered in a specific group can still be k dimensional with k<n." (p. 267).

Kok proposes a model which includes a primary ability (theta), and three other abilities to be referred to as n_1, n_2, and n_3. The first of these (n_1), may be conceptualized as a compensatory ability. Kok provides an example of how a compensatory ability might influence test performance as if a test of knowledge of French, and the examinee has no knowledge of French, but with a sufficient knowledge of Spanish could conceivably use his knowledge of Spanish to compensate, at least in part, for his deficit in French. The second of these (n_2) Kok suggests could have to do with the ability of the examinee to understand the test questions. For instance, if a test is written in English, clearly an examinee's ability to understand written English will affect his performance. Finally, Kok postulates that n_3 indicates an examinee's ability to use contextual cues to solve an item, or the testwiseness. Kok proposes that the probability of success on a given item is a function of not just the primary ability, theta, but also, at least potentially, of the three other abilities. Thus, he expresses the item response success probability for item i in group j as:

$$P_j(X_i=1 \mid \xi, n_1, n_2, n_3)=\phi_{3i}(n_3)+[1-\phi_{3i}(n_3)]\phi_{2i}(n_2)\phi_{1i}(\xi+\alpha_{2i}n_1)$$

where ξ, n_1, n_2, n_3 are latent traits, and $\phi_{1i}(\xi, n_1)$, $\phi_{2i}(n_2)$, $\phi_{3i}(n_3)$ are functions which describe the relationship between the separate latent traits, and the response success probability on item j.

Kok develops this model further, and demonstrates that the a necessary condition for the occurrence of DIF is

$$h_1\ (n_1,\ n_2,\ n_3 \mid \xi) \neq h_2\ (n_1,\ n_2,\ n_3 \mid \xi)$$

Thus, Kok proposes that DIF is a possible consequence of between group differences in the conditional distributions of the additional abilities. This could occur for example, if one group were more testwise than another. However, Kok also notes that an item may be multidimensional and <u>not</u> manifest DIF, if the conditional distributions for the two groups are equal. It is also possible that a unidimensional test may be DIF in the case where the test is unidimensional in that while individual items may require more than one ability for solutions, each ancillary ability influences performance on only one item.

Thus Kok's argument is that DIF is a possible consequence of unequal conditional ability distributions. Such differences could result in differences in item parameters if the test data are erroneously assumed to be unidimensional. Thus, items may appear to be differentially functioning (i.e. may exhibit different ICCs) if test developers assume that test results are unidimensional, when this is not the case.

Kok closes stating that this model of DIF has utility in that it posits a common mechanism which can explain DIF in a wide variety of testing situations. He stress however that the model has important implications for DIF research as well. He suggests that rather than simply generating items to have differing unidimensional item parameters, researchers could use multidimensional models and simulate DIF by simulating differences in the underlying ability distributions.

Shealy and Stout (1993), working independently, developed a very similar formulation of DIF. Like Kok, they assert that DIF (and test bias, which can result from the cumulative effects of DIF) can be explained by multidimensionality in the data set.

While Kok posits a primary trait, and three additional traits which are psychologically meaningful, Shealy and Stout refer to a target ability (θ), which is the ability the test is intended to measure, and one or more nuisance determinants, which the test is not intending to measure. Conceivably Kok's testwiseness (n_3) could be considered a nuisance determinant, as could reading ability on a test of American history for example. Shealy and Stout's use of the term "nuisance" implies that one would always wish to measure one and only one trait (in any one given test), and that measurement of any other traits simultaneously is undesirable, and hence a nuisance. Kok's formulation is more neutral with respect to traits other than theta, allowing for the possibility that there may be occasions where measurement of these traits may be desirable.

Like Kok, Shealy and Stout postulate that the manifestation of bias may be explained by between group differences in ability on the nuisance determinants, coupled with items sensitive to such abilities. They go on to discuss the implications of this explanation for DIF detection. Their position is that it is impossible to statistically detect bias unless one uses either an external criterion (which must be a valid measure of the target ability), or an internal measure which measures _only_ the target ability. They argue that if the criterion score is influenced by abilities other than the intended to be measured ability, it does not provide an appropriate matching criterion, and may lead to incorrect classifications of items. In response to this dilemma, Shealy and Stout put forward the notion of a valid subtest, which they define as a set of unidimensional items - that is the probability of a correct response to each item in the set depends only on the ability of interest. They note that if every item on a test is contaminated by nuisance determinants that it is not possible to identify a valid subtest, and thus it will be impossible to identify bias (unless a valid external measure is available). They maintain that by matching examinees using this valid subtest score, rather than total test score, (unless the test is unidimensional, in which case they will be the same), group differences in the target ability are appropriately controlled for, and differences due to nuisance determinants can be isolated, and eliminated.

The problem of circularity in using a possibly biased criterion to identify DIF has been noted by other authors as well, and is admittedly a problem with many bias detection procedures. Shepard (1982) noted that DIF procedures which depend on total score for matching, cannot detect pervasive bias in a test. Kok, Mellenbergh and Van Der Flier (1985) also note that using all items (including the DIF one) for computing the total score for estimating the latent trait is a "severe weakness" of the procedures which use this approach. This is the reasoning behind the two-stage implementation of MH procedure, wherein in the first stage potentially biased items are identified, and then the statistic is calculated again, this time conditioning examinees on a total score which does not include the items identified in the first run (with the exception that the biased item is always included).

Ackerman (1992) sought to extend the work of Kok, and of Shealy and Stout, by further elucidating the relationship between multidimensionality and DIF. However, before discussing Ackerman's work, it is necessary to first discuss the multidimensional IRT model which a number of Ackerman's concepts are based on. This model, which is a compensatory multidimensional two-parameter logistic model (M2PL) was developed by Reckase (1985, 1986, 1989). In multidimensional IRT, both compensatory and non-compensatory models are possible. With compensatory models it is possible for high levels of one ability to compensate for low levels of another. Thus, as in Kok's (1988) example above, on a test of French an examinee with a superior knowledge of Spanish could potentially compensate for his lack of knowledge of French. In contrast, noncompensatory models do not allow for such compensation. Thus, in a mathematics test where items depend both on ability to understand written English, and ability to perform certain mathematical operations, an examinee with a superior knowledge of the English language would not be able to use this knowledge to compensate for a lack of knowledge of the mathematical operations required. There is currently some controversy as to which model is more realistic, or better describes actual test performance. At this time there appears to be no definitive answer. It is likely that the answer as to which model is most appropriate depends on the specific testing situation.

In addition to providing necessary background for an understanding of Ackerman's work, Reckase's model is also important in the context of this research as it will provide the model used to generate simulated data. It was chosen because there has already been a significant amount of work done using this model (e.g., Reckase 1985, 1986, 1989; Oshima & Miller, 1990, 1991; Ackerman, 1992), and thus use of this model here will allow for comparisons of results with these studies (which will be discussed below).

Reckase's model may be written as follows:

$$P(\underline{u}_i = 1 | \theta_{ji} a_i,\ d_i) = \frac{e^{\underline{a}\theta_j + \underline{d}_i}}{1 + e^{\underline{a}_i\theta_j + \underline{d}_i}}$$

where $\underline{u}_i$ is the item response

θ_j is a vector of abilities

a_i is a vector of discrimination parameters

and $\underline{d}_i$ is a scalar related to item difficulty.

Reckase sought to find a means of describing multidimensional items in terms which were analogous to the parameters used in unidimensional IRT (UIRT). Thus he developed the concepts of multidimensional item difficulty (MDIFF), multidimensional discrimination (MDISC), and multidimensional information function (MINF).

The concept of item difficulty in a multidimensional space is more complex than in an unidimensional space. In UIRT item difficulty is defined as that point on the ability scale where an item is most discriminating. However, in multidimensional IRT (MIRT) there may be many points where an item is most discriminating. Therefore, Reckase proposed defining MDIFF in terms of both the distance from the origin of the space to the point of maximum change (D) and in terms of the direction specified by the vector of angles, alpha, between the coordinate axes and the line connecting the origin and the point of maximum slope (Reckase, 1989, p. 11). MDISC is defined as the slope of the proportion correct surface at the point of maximum rate of change in the direction, alpha, from the origin. MINF is defined in a manner very similar to the information function in UIRT, but in the MIRT case Reckase notes that the information is indexed by a particular direction. Thus a given item may provide significant information in one direction and not in another. The equations for each of these item features are as follows:

$$\mathrm{MDISC}_i = (\Sigma a^2_{ik})^{1/2}$$

$$\underline{D}_i = \frac{-\underline{d}_i}{\mathrm{MDISC}_i}$$

$$\cos\alpha_{ik} = \frac{\underline{a}_{ik}}{\mathrm{MDISC}_i}$$

and

$$\mathrm{MINF}_{\alpha}(\theta) = P_i(\theta_j)Q_i(\theta_{uj})(\sum_k \underline{a}_{ik} \cos \alpha_{ik})^2$$

where $P_i(\theta_j) = P(\underline{u}_i=1|\theta_j,\ a_i,\ \underline{d}_i)$

and $Q_i(\theta_i) = 1 - P_i(\theta_i)$.

With this brief presentation of Reckase's model, it is now possible to return to a discussion of Ackerman's work.

Central to much of Ackerman's argument is the concept of a reference composite. The reference composite is the score which results when multidimensional items are treated as if they are unidimensional, and a single test score is used to summarize performance. Thus, this score is actually a weighted composite of the underlying multiple dimensions. Ackerman notes that the direction of the reference composite in the latent space is influenced by the characteristics of the underlying multidimensional ability distributions, and the discrimination parameters of the multidimensional items. Because of this, it is possible for the direction of the reference composite to differ for different groups. In this case, the total score would mean different things for the different groups. Thus conditioning on this score in DIF studies is not appropriate, and could yield invalid results.

In order to overcome this problem Ackerman suggests selecting a valid subtest of items, and using this as a criterion score. He notes however that it is probably not realistic to restrict a valid subtest to only those items which measure exactly and only the target ability - in reality most or potentially all items on a test may be influenced to some extent by nuisance determinants. Therefore, Ackerman proposes identifying a validity sector - a group of items which share a similar measurement direction. A validity sector as "a narrow sector (and its mirror image projecting through the origin) constituting the valid subtest items." (1992, p.73) The width of the validity sector is determined by the breadth of the cognitive area being measured (1991).

Items which lie outside the validity sector are considered invalid items - that is they are too heavily influenced by nuisance determinants. Ackerman suggests that these are the items which should be considered biased, and should be deleted from the test. If these items are deleted the total test score is now a valid measure for both groups.

Ackerman (1992) provides didactic examples of how DIF can result from multidimensional items when there are differences in the underlying multidimensional distributions, and the DIF analyses are conducted as if the test were unidimensional. Ackerman notes there are several ways the potential for DIF can occur, and lists four: between group differences in target ability means, between group differences in nuisance ability means, differences in the ratio of the nuisance variance to the target variance, and the correlations between the target and nuisance abilities may differ for the two groups. He then demonstrates how each of these conditions could result in bias.

Ackerman then goes on to provide an empirical example, using simulated data. Using Reckase's model (M2PL), and MIRT parameters estimated from a 25 item math usage test, Ackerman identified a valid subtest of items (items falling within a constructed validity sector). Ackerman then simulated responses for two groups of 1000 examinees, varying both the target and nuisance ability distributions so that there were between group differences in means and standard deviations. He then calculated the reference composites for both groups, and found the direction of the composites to differ substantially. He then analyzed the test using the MH procedure (using MH delta as the test statistic), Stout's simultaneous DIF (SIB) procedure, and an IRT area measure. The SIB procedure identified 6 of the 7 items Ackerman had identified as invalid using the validity sector approach. The MH procedure, using the valid subtest score as the matching criterion, identified 5 of the 7 items. Ackerman re-ran the MH procedure, this time using all the test items, and this time the MH procedure identified 10 additional items - items which Ackerman considered valid. Ackerman suggests this latter result provides an example of how the MH procedure can be misused if it is erroneously assumed that a data set is unidimensional when it is not. The analyses using the IRT area index parallelled the MH results - that is several valid items were identified.

Ackerman's results demonstrate that even if there are no between group differences in the MIRT item parameters, between group differences in the underlying multidimensional ability distributions can result in apparent DIF if the data are analyzed using DIF detection methods that assume unidimensional data. Thus, the multidimensional conceptualization of DIF put forward by Kok and Shealy and Stout is supported. There are also several studies which provide additional direct and indirect support for this viewpoint, and these are discussed below.

Support for the Multidimensionality Explanation of DIF

Oshima and Miller (1991) showed that multidimensional items were identified as DIF when the means of the reference and focal groups on the secondary trait differed. In this study they varied the between group difference on the primary trait means (no difference versus a difference of .5 standard deviations), the between group difference on the secondary trait means (again, no difference versus a difference of .5 standard deviations) and the percentage of items influenced by the secondary trait (5, 10, or 20 percent of the items). Thus they examined 12 conditions, with 10 replications of each condition. The correlation between the two traits was set at zero.

Using Reckase's M2PL model to generate the data, Oshima and Miller simulated responses for two groups of 1000 examinees each to a forty item test. The item parameters for the two groups were the same. The data were analyzed using PCBILOG to obtain unidimensional IRT parameter estimates. The ICCs for the two groups were then compared using signed and unsigned area measures (SA and UA) and signed and unsigned sum of squares (SOS and USOS). Because these measures have no associated significance tests, Oshima and Miller first obtained baseline values and established the criterion that the difference between the ICCs would be considered significant if the value differed from the baseline mean by two or more standard deviations. This is equivalent to identifying an item as biased.

Oshima and Miller found that if there were no differences in the distributions of the secondary traits, multidimensional items were no more likely to be identified as DIF than unidimensional items. This was true regardless of whether or not there were between group differences on the primary trait. If there were differences on the secondary trait, multidimensional items were much more likely to be identified as DIF. Higher detection rates were associated with smaller proportions of multidimensional items. All four indices (SA, UA, SOS and USOS) yielded comparable results. Detection rates (across all indices) ranged from 80-100% in the case where only five percent of the items were multidimensional, from 43-68% where ten percent of the items were multidimensional, and from 24-39% where twenty percent of the items were DIF.

In an earlier study, Oshima and Miller (1990) varied the trait correlations between the reference and focal groups, and again examined the effect this had on the ICCs of the two groups. Using the same M2PL model, they simulated a 40 item test, with two groups of 1000 examinees each for each condition. The correlations between the primary and secondary traits differed for the two groups (except in the baseline condition). In group one the trait correlations were set to be either 0 or .5, while for group two the correlation varied from 0 to 1. A total of nine separate conditions were generated. Oshima and Miller suggest that two of these conditions can be seen as simulating bias. In these two "bias" conditions there is a perfect correlation between the two traits for one group, and of correlation of 0 or .5 for the other. Thus, for one group the test is essentially unidimensional, while for the second group the test is two dimensional.

As in the 1991 study, UIRT estimates were obtained, and the difference between the ICCs for the two groups was evaluated. Again, SA, US, SOS and USOS were used, and again the criteria for significance was that the value exceed the baseline mean by at least two standard deviations. They found that the unsigned indexes (UA and USOS) resulted in a number of items meeting this criterion. For instance, with the correlation between the traits set to 0 for one group and 1.0 for the other, 33 out of the 40 items exceeded the criterion (that is the ICCs were judged to be different) using the unsigned area method. When the trait correlations were set to 0 and .8, 25 items were so identified. The USOS method yielded very similar results. Analyses with the signed indexes yielded results much closer to the baseline results, except in the most extreme conditions.

Oshima and Miller's findings indicate that if a test is unidimensional for one group (the correlation between the two traits equals 1) but not for the other (for instance, the correlation between the two traits equals .5) then it is likely that a number of items will be identified as differentially functioning with the unsigned methods. While the comparison of correlations of 0 versus 1 may be viewed as an extreme case, less extreme between group differences also resulted in a number of items being flagged, even when the correlation was less than one for both groups. Thus, these findings provide further support for the premise that multidimensionality can in fact result in differential item functioning as defined by a lack of invariance across ICCs. Further, it appears that differences in trait correlations between groups did affect the number of items being flagged as such, possibly by influencing the unidimensional *a* estimates. Oshima and Miller noted the need for further research to examine when and under what conditions such differences in correlations occur in practice, and the practical effects of such differences on test scores.

Birenbaum and Tatsuoka (1982) assert that there is always more than one major factor underlying any set of achievement test data. They believe that the dimensionality of such data is related to the number of algorithms which students use to solve test items. They argue that students formulate algorithms (or rules) which they apply, correctly or incorrectly, when responding to test items. Different students use different algorithms, and thus have different response patterns. Birenbaum and Tatsuoka assert that this "adds systematic sources of variation in the data resulting in an increase in the underlying factorial structure of the test" (p. 261). The goal of instruction is to provide students with the correct algorithms. If instruction is successful, this should result in students using fewer algorithms (as they are now using the correct ones) and thus these authors argue that the effect of instruction should be to reduce the dimensionality of the test.

Birenbaum and Tatsuoka simulated data sets wherein they systematically increased the number of algorithms used to generate response patterns. They then conducted a principal components analysis and found that the percentage of the variance explained by the first factor was greater when fewer algorithms were used. As the number of algorithms decreased, coefficient alpha also increased.

Birenbaum and Tatsuoka also examined a real data set. They collected data on 81 seventh grade students prior to and again following instruction in subtraction of signed numbers. Again, they conducted a principal components analysis and calculated coefficient alpha. They report that their results were consistent with their hypothesis that students use fewer algorithms following instruction, and that the analysis revealed increased homogeneity.

These results must be considered weak support at best for their assertion that they were able to reduce the dimensionality of the test, as their measures of dimensionality/homogeneity (terms which they use interchangeably) have not been found to be appropriate measures (Hattie, 1984). However, their hypothesis regarding students use of algorithms in responding to test problems, and the proposition that instruction may serve to reduce the number of algorithms are both worth consideration.

Lautenschlager and Park (1988) also provide evidence that multidimensionality may be evidenced as DIF. While the focus of their study was on the relative merits of two methods of parameter linking, in order to assess the various linking procedures they generated data under several different conditions. Each dataset consisted of items generated using a 3-parameter logistic UIRT model, with identical, normally distributed ability scores for each group. These are the non-DIF items. In addition, a number of DIF items were generated using a two dimensional noncompensatory IRT model. One thousand examinees per group were simulated, with identical normal distributions on the first ability. The number of DIF items were varied so that of the total 54 items either 18, 28 or 46 were DIF. The mean of the distributions of the secondary trait was varied (set at either -.5 or 0), and the correlation between the two traits was also varied. Using Lord's chi-square test (at the .005 level) for the significance of the difference between the unidimensional parameter estimates they found that without parameter linking (the baseline condition) virtually all the non-DIF items were identified as such. In addition, a high percentage of the items constructed to be DIF were identified as such. They note that those items which were missed were those which were only weakly DIF but do not provide further details as to the characteristics of these items. When linking procedures were used the results were less accurate, that is there were a greater number of misclassifications.

The studies discussed above have depended primarily on simulated data to reach their conclusions, and while there are a number of advantages to simulation studies, such studies are often criticized on the grounds that they lack realism - the question often arises as to how generalizable the results of such studies are. However, studies using actual test results can be criticized on the grounds that it is impossible to know absolutely which items are DIF and hence any results must be interpreted with caution. There have been a small number of studies however in which the authors have constructed items with the intent of creating DIF. While one might still question whether it is possible to truly evaluate whether they succeeded, the following two studies are very convincing, and are offered as evidence that it is possible for researchers to construct DIF, that they do so by introducing a items which are sensitive to a second ability, and by using groups who differ in their distributions on that second ability.

The first very clear example of this is presented by Kok, Mellenbergh and Van Der Flier (1985). These authors sought to deliberately construct biased items by writing math items which used Dutch, Spanish or Roman numerals. The examinees (whose native language was Dutch) were randomly assigned to two groups. Both groups got some instruction in Spanish numerals. Then, one group (the Roman group) received instruction in Roman numerals, while the other group (the Spanish group) received additional instruction in Spanish numerals. All 286 examinees were then administered a mathematics test which contained math problems written in Dutch numerals, Spanish numerals, and Roman numerals. Examinees were required to first translate the problem (and write down this translation) and then to write down the correct answer. Thus, there were clearly two abilities required for a correct solution to the problems - first understanding the problem, which for some items required translating the numerals, and then performing the appropriate mathematical operations. In addition, the groups presumably differed substantially in the first ability, as a result of instruction (or lack thereof). Kok et al. checked accuracy of translations to see whether the groups did in fact differ as a function of instruction, and not surprisingly, it was found that Spanish group members were more adept at Spanish numeral translations, and Roman group members were more adept at Roman numeral translations. Kok et al. then analyzed the test results using an iterative logit procedure. Their results suggested that their manipulations did in fact create differentially functioning items, and that the logit procedure did identify many of the items which they had predicted to be differentially difficult for the two groups.

Subkoviak, Mack, Ironson, and Craig (1984) constructed a 50 item vocabulary test consisting of 40 items from a college aptitude test, and 10 items which used black slang vocabulary. They then administered the test to college students, and look for differences in performance between Blacks and Whites (they had over 1000 examinees in each group). Not surprisingly, they found high correlations between the items they had constructed to be differentially functioning (the items requiring knowledge of Black slang) and the items which the DIF detection procedures they were evaluating identified as DIF. In this study knowledge of standard English and knowledge of Black English can be thought of the dimensions or abilities underlying performance. It is also reasonable to presume that there were substantial differences between the two groups in their knowledge of Black slang, and quite possibly in their knowledge of standard English as well.

Thus, both the Kok et al. study and the Subkoviak et al. study provide clear evidence that it is possible to produce items which appear biased by including items which require some skill or knowledge other than what might be considered the primary skill of knowledge, and then administering these items to examinees from groups who have between group differences on this secondary ability.

Mazor, Kanjee, and Clauser (1993) conducted a study with real data which also provides support for a multidimensional conceptualization of DIF. They conducted a series of DIF analyses on responses to two achievement tests. For both tests they made two reference/focal group comparisons. They first compared males and females, and second, they compared examinees who reported English as their best language (EBL) to examinees who reported some other language as their best language(OBL). They began by analyzing the data with both LR and the MH procedure, using total score as the matching criterion. They then repeated the LR analyses, this time expanding the LR equation to include either SAT-V or SAT-M scores in addition to total score. Finally, the MH analyses were repeated, with either the SAT-V or the SAT-M scores substituted for total score as the matching criterion. They found that for the EBL/OBL comparisons including the SAT-V score in the logistic regression equation substantially reduced the number of items identified as DIF. Mazor et al. argued that the SAT-V score provided information on an ability related to facility with written English, an ability which the EBL/OBL groups would be expect to differ on. By including the SAT-V score in the analysis, matching was improved, and thus items which appeared DIF because of this difference in verbal ability (i.e. multidimensionality in the data set) were no longer flagged as DIF. Including the SAT-V scores allowed differences in verbal abilities to be taken into account, with the result that items multidimensional with respect to verbal ability were no longer identified as DIF. Mazor et al. note that external measures such as the SAT-V scores are not always available, and that in some cases internally derived ability estimates may be useful.

Summary

Research on differential item functioning can be seen as facing three major challenges - identification, explanation, and elimination of DIF items. There are currently a number of widely accepted procedures which are used to identify DIF. IRT-based techniques are generally accepted as theoretically preferred, but not always feasible in applied settings. Because of this, techniques such as the MH procedure have been accepted as reasonable approximations. The MH procedure has been shown to be powerful and to have low false positive rates, and therefore it's current popularity and acceptance appear to be well founded. The primary shortcoming of the MH procedure is it's relative insensitivity to non-uniform DIF. The logistic regression procedure presented by Swaminathan and Rogers (1990) is sensitive to non-uniform DIF, and thus may gain in popularity as more researchers become familiar with this procedure, and more results using LR are published. A second advantage of the LR procedure is that the regression equation is easily elaborated to include more terms, and thus in addition to allowing for interactions between group and ability (the term which allows for the assessment of non-uniform bias) it is possible to include a second measure of ability. This becomes especially desirable if DIF is conceptualized in terms of multidimensionality.

The second challenge facing DIF researchers is the challenge of explanation. A number of researchers have attempted to look at the characteristics of specific items identified as DIF, and to find commonalities which suggest possible explanations. The findings of such studies have generally been mixed. In a number of studies positive results have been found, but in many cases the results are not consistent, and generally researchers have not been able to predict which items will be identified as DIF.

A second, more generalized explanation of DIF is that DIF is manifested as a result of multidimensionality in the data set. This approach does not contradict the first approach, but might be view as a more general conceptualization. If two groups differ in their performance on a given item it must be because they are not matched on all the relevant abilities. Thus, items which depend on more than one ability, and where the two groups differ in their distributions on this ability, have the potential to display bias. This explanation of DIF has implications for DIF detection procedures. Some researchers conceptualize abilities other than the primary or target ability as "nuisance" abilities, and imply that items too heavily influenced by such abilities should be removed. These researchers advocate changing the criterion which is used to match examinees by selecting only valid items. The result is presumably a more pure measure of the target ability. However, it may be the case that such items are tapping an important ability, one that test users wish to assess. In this case it may not be desirable to delete items which are multidimensional. However, standard DIF analyses may well identify such items as DIF, as would analyses with a "pure" matching criterion. Thus, there is a need to evaluate procedures which would allow for simultaneous conditioning on more than one ability.

CHAPTER III

METHODOLOGY

The purpose of this study was to investigate how two-dimensional tests and items impact on the results of the MH and LR DIF detection procedures. It was anticipated that multidimensionality in a data set would lead to high false positive error rates and poor accuracy in identifying true DIF items if only one ability was taken into account. If this was found to be true, the second part of the study would focus on whether improving the matching criterion by taking into account the second dimension would decrease false positive errors without increasing false negative errors.

In order to investigate the conditions under which items in a two dimensional test would be falsely classified as DIF using the standard MH and LR procedures a simulation study was conducted. A simulation study was necessary because only by using simulated data was it possible to know whether or not there were between group differences in the item parameters.

Because high false positive error rates were in fact obtained under most of the conditions simulated, part II investigated whether these rates could be reduced. One modification which had been suggested for improving the accuracy of the MH procedure is to select a valid subtest of items, and to use the score on that subtest (rather than total test score) as the matching criterion for the MH. Therefore, in the second part of this study items were selected into subtests, and the subtests were used as the matching criterion. It was expected that if it were possible to correctly identify valid or pure subtests, that many false positive errors would be eliminated. It was further anticipated that this procedure would yield more accurate results with respect to relatively pure or unidimensional items in a multidimensional test, but would not increase the accuracy of classification of items which were multidimensional. Therefore a breakdown of false positive rates by item characteristics was conducted.

Also in this second part of the study the LR procedure developed by Swaminathan and Rogers (1990) was evaluated. Valid subtest scores were included in the logistic regression equation in lieu of total score. In the case where performance on an item depends on two abilities, and the regression model includes only one ability estimate as a predictor, the model is in fact underspecified, and it was anticipated that incorrect classifications would result. That is, group membership may be significant, when in fact it is not, but is functioning as a proxy variable for a secondary ability which is unequally distributed for the groups of interest. It was expected that by taking both abilities into account (by including both subtest scores in a single equation) that false positive error rates would decrease.

All of the above analyses were conducted using simulated data. In order to begin to assess the generalizability of the findings of Parts I and II Part III of this study applied the above procedures to two real data sets. Two achievement tests were first analyzed with both the MH and LR procedures using total score as the matching criterion. Valid subtests of items were selected, subtest scores for each examinee were calculated, and then both the MH and LR procedures were repeated, this time with valid subtest scores as the matching criteria. The results of the total score analyses were compared with the subtest score analyses.

Part I

The purpose of this phase of the study was to assess whether a second dimension influencing a data set would result in misclassification of items as DIF. Prior research had suggested that if the items of a test are sensitive to more than one ability, and if there are between group distributional differences on one of the abilities, multidimensional item impact should be identified as DIF when total score is used as the matching criterion. If these results were confirmed, this would provide further support for a multidimensional explanation of DIF. In addition, by examining how false positive rates vary according to item measurement direction, item discrimination, trait correlations, and the dimensional structure of the test, our understanding of the relationship between multidimensionality and apparent DIF was furthered.

All of the simulated data used in this study was generated using Reckase's two dimensional compensatory model, M2PL (Reckase, 1985, 1986). The computer program MULTISIM (Narayanan, 1992) was used to generate the data.

The degree of relationship between the two (or more) dimensions measured by the test will be likely to impact the identification of items as biased even when the correlations are the same for both groups. In the extreme case, that is when for both groups there is a perfect correlation between the traits, the presence of a second trait will make no difference, as it is redundant with the first, and thus there is only one trait. However, when there is a less than perfect correlation between the two traits, the validity of the total test score as an estimate of examinee ability will become increasingly questionable, and become an increasingly poor conditioning variable as the correlation decreases. Thus the magnitude of the correlation, and any between group differences in correlations, can be expected to impact on the validity of a DIF analysis (Oshima & Miller, 1990).

Another important variable is the extent to which the items and the test are multidimensional. This will be referred to as the dimensional structure of a test. A test can be multidimensional in one of two ways. First, all of the test items may be "pure" items. In this case the items themselves may be unidimensional, but the test may be composed of more than one type of item. Thus, while performance on any given item may depend only on one underlying ability, the test as a whole may have a number of items measuring ability A, and a number of other items measuring ability B. That is, the items themselves are unidimensional, but the test is not. In this case, if there are differences in the underlying ability distributions, the total test score will not provide a valid matching criterion.

The second, and perhaps more realistic way for a test to be multidimensional is that it may be composed in part or entirely of multidimensional items. That is, performance on at least some of the individual items is influenced by more than one underlying ability. For instance, a math item which requires only reading several numerals and an operand may well be unidimensional. However, an item which requires the examinee to read a complex item stem, set up the problem and then decide on and perform a mathematical operation is probably requiring several abilities. There appears to be a move currently away from the former, more "pure" test item, towards the latter, more "realistic" item type, based on the argument that this is the type of problem one is more likely to encounter in the real world. This may be true, but it is also true that the latter type of item is not strictly unidimensional, which will clearly effect any analyses which requires the assumption of unidimensionality.

Finally, the actual parameters of the items are likely to impact on whether or not an item is identified as DIF. Mazor, Clauser and Hambleton (1992) found that in unidimensional data sets the difficulty of the item, the size of the difference in difficulty parameters between the two groups, and the discrimination parameter of the item all influenced whether a biased item was correctly identified.

Design of the Study

Sample size and test length were held constant in all phases of this study. Between group ability distribution differences were also held constant, except for one series of supplemental analyses described below. Trait correlations and the dimensional structure of the tests including item measurement direction (the relative influence of the dimensions on the items) and item discrimination were varied systematically to investigate the influence of each of these variables.

Sample Size. Sample sizes of 1000 examinees per group were used. This may be considered a "best case" scenario, as samples of this size are not routinely available in practice. However, because the focus of the study was not on the impact of sample size per se, it was necessary to chose a sample size that would be adequate to provide a "fair test" of the procedures of interest. Research using data generated using unidimensional IRT models suggests that samples of this size are sufficient to detect most DIF, and items which are missed are those with small differences between groups, differences which would be expected to have virtually no practical impact (Mazor, Clauser, & Hambleton, 1992). This research also suggests that a sample of this size would be expected to yield few false positive errors, yet would be expected to correctly flag virtually all items which were DIF to any meaningful degree. While larger sample sizes may yield even greater accuracy, in practice it is unlikely that practitioners will have access to such large groups for analyses, as the usual group size is generally estimated to be between 200 and 500.

Test Length. A test length of 66 was used. This was considered realistic as most achievement tests range between 35 and 85 items. This number also allowed for 75 percent of the items to be simulated to be predominantly sensitive to ability A, and 25 percent to be predominantly sensitive to ability B, while allowing for six levels of item discrimination to be crossed with eight levels of item difficulty.

Ability Distributions. Ackerman (1992) described four conditions which may result in multidimensional item impact appearing as DIF. First, the groups may differ in their means on the primary ability. Second, the groups may differ in their means on the second ability. Third, the ratio of the variances of the two abilities may differ. Fourth, the correlations between the first and second abilities differ for the groups. The present study focused on the second condition, between group differences in means on the second ability.

For all of the simulated data sets used here there were no between group differences on the first ability (A). A difference in the means of the two groups on the secondary ability was simulated, so that the focal group mean was one standard deviation lower than the reference group mean. After reviewing the results it was decided that two additional supplementary simulations would be conducted for one subset of conditions. In both of these supplementary simulations, the groups differed on the second ability (B) as described above. However, for the first supplementary set of simulations, the reference group mean was set to be one half of a standard deviation greater than the focal group mean on the first ability (consistent distributional differences). For the second supplementary set of simulations, the reference group mean was set to be one half of one standard deviation less than the focal group mean on the first ability (crossed distributional differences).

<u>Trait Correlations</u>. Two different conditions were simulated to investigate the impact of the correlations between the two abilities. In the first condition a correlation of .3 for both groups was simulated. In the second, the correlations were .7 in both groups.

<u>Dimensional Structure of the Tests and Item Parameters</u>. The dimensional structure the item sets was varied to result in three dimensionally different tests. This was done by varying the relative sensitivity of the items to each ability. This is most succinctly expressed as the item measurement direction. The number of items at each measurement direction for each test is presented in Table 1.

Test 1 consisted of 48 items which measured only ability A (items with a measurement direction of 0 degrees), and 16 items which measured only ability B (items with a measurement direction of 90 degrees).

Table 1

Test Dimensional Structure

	Number of Items at Each Measurement Direction (In Degrees)						
Test	0	15	30	45	60	75	90
1	48						16
2		24	24		8	8	
3	12	12	12	12	8	8	

Test 2 consisted entirely of multidimensional items. There were no "pure" A or "pure" B items. Of the 48 items which were more sensitive to dimension A than dimension B, 24 items had a measurement direction of 15 degrees, and 24 had a measurement direction of 30 degrees. Of the 16 items which were more sensitive to dimension B, 8 had a measurement direction of 60 degrees, and the final 8 had a measurement direction of 75 degrees.

Test 3 consisted of 48 predominantly A items (12 items each at 45, 30, 15 and 0 degrees) and 16 predominantly B items (8 items at 75 degrees and 8 at 60 degrees). Thus test 3 had some pure A items, like test 1, but also had some items which were equally sensitive to ability A and B.

From the above it can be seen that the items were essentially grouped into two blocks - the 48 items which were most sensitive to ability A (for simplicity the 12 items in test 3 which are at 45 degrees are referred to as A items, even though they are in fact equally sensitive to both abilities), and the 16 predominantly B items.

Item difficulty was systematically varied across all 64 non-DIF items. Eight levels of difficulty were simulated, with the difficulty parameters set at -1.75, -1.25, -.75, -.25, .25, .75, 1.25, or 1.75. Thus for the 48 A items, there were 6 items at -1.75, 6 items at -1.25, etc. For the 16 B items, there were 2 items at each difficulty level (see Table 2).

For the A items item difficulty was completely crossed with multidimensional item discrimination (MDISC). The discrimination parameters were set to be .2, .4, .6, .8, 1.0 or 1.2.

For the 16 B items the discrimination parameters were not completely crossed within each simulation. Rather, the 16 B items were either low discrimination (.2 or .4) medium discrimination (.6 or .8) or high discrimination (1.0 or 1.2). Thus, while for each simulation the entire range of discrimination values was covered in the A items, the B items had a restricted set of discrimination values, which allowed the impact of discrimination on the B items to be studied separately. Thus, for each test (Test 1, Test 2, and Test 3) and each level of trait correlation (.3 or .7) there were three simulations. The discrimination parameters were always the same for the A items. The discrimination parameters for the B items were either low, medium or high, as described above. These are referred to as the Low, Medium and High discrimination sets below.

Each test had two additional items at 0 degrees, which were true-DIF items. For these two items there was a between group difference in the difficulty parameter of .5. No differences in discrimination

Table 1

Test Dimensional Structure

	Number of Items at Each Measurement Direction (In Degrees)						
Test	0	15	30	45	60	75	90
1	48						16
2		24	24		8	8	
3	12	12	12	12	8	8	

Test 2 consisted entirely of multidimensional items. There were no "pure" A or "pure" B items. Of the 48 items which were more sensitive to dimension A than dimension B, 24 items had a measurement direction of 15 degrees, and 24 had a measurement direction of 30 degrees. Of the 16 items which were more sensitive to dimension B, 8 had a measurement direction of 60 degrees, and the final 8 had a measurement direction of 75 degrees.

Test 3 consisted of 48 predominantly A items (12 items each at 45, 30, 15 and 0 degrees) and 16 predominantly B items (8 items at 75 degrees and 8 at 60 degrees). Thus test 3 had some pure A items, like test 1, but also had some items which were equally sensitive to ability A and B.

From the above it can be seen that the items were essentially grouped into two blocks - the 48 items which were most sensitive to ability A (for simplicity the 12 items in test 3 which are at 45 degrees are referred to as A items, even though they are in fact equally sensitive to both abilities), and the 16 predominantly B items.

Item difficulty was systematically varied across all 64 non-DIF items. Eight levels of difficulty were simulated, with the difficulty parameters set at -1.75, -1.25, -.75, -.25, .25, .75, 1.25, or 1.75. Thus for the 48 A items, there were 6 items at -1.75, 6 items at -1.25, etc. For the 16 B items, there were 2 items at each difficulty level (see Table 2).

For the A items item difficulty was completely crossed with multidimensional item discrimination (MDISC). The discrimination parameters were set to be .2, .4, .6, .8, 1.0 or 1.2.

For the 16 B items the discrimination parameters were not completely crossed within each simulation. Rather, the 16 B items were either low discrimination (.2 or .4) medium discrimination (.6 or .8) or high discrimination (1.0 or 1.2). Thus, while for each simulation the entire range of discrimination values was covered in the A items, the B items had a restricted set of discrimination values, which allowed the impact of discrimination on the B items to be studied separately. Thus, for each test (Test 1, Test 2, and Test 3) and each level of trait correlation (.3 or .7) there were three simulations. The discrimination parameters were always the same for the A items. The discrimination parameters for the B items were either low, medium or high, as described above. These are referred to as the Low, Medium and High discrimination sets below.

Each test had two additional items at 0 degrees, which were true-DIF items. For these two items there was a between group difference in the difficulty parameter of .5. No differences in discrimination

Table 2

Item Parameters

	Predominantly A Items						Predominantly B Items					
	MDISC						Low MDISC		Medium MDISC		High MDISC	
Difficulty	.2	.4	.6	.8	1.0	1.2	.2	.4	.6	.8	1.0	1.2
-1.75	1	2	3	4	5	6	49	57	49	57	49	57
-1.25	7	8	9	10	11	12	50	58	50	58	50	58
- .75	13	14	15	16	17	18	51	59	51	59	51	59
- .25	19	20	21	22	23	24	52	60	52	60	52	60
.25	25	26	27	28	29	30	53	61	53	61	53	61
.75	31	32	33	34	35	36	54	62	54	62	54	62
1.25	37	38	39	40	41	42	55	63	55	63	55	63
1.75	43	44	45	46	47	48	56	64	56	64	56	64
							65	66	65	66	65	66

NOTE: Items 65 and 66 are DIF items, with item difficulties of -.25 and .25 for the reference group, and .25 and .75 for the focal group.

parameters for the two groups was simulated, so that the simulated true-DIF was uniform. The discrimination parameters for these items were the same as the discrimination parameters of the B items for the set.

Thus, altogether 18 conditions were investigated, as there were three test structures, two levels of inter-trait correlations, and three levels of discrimination for the B items. Each condition was replicated ten times.

Data Analysis

For this first phase of the study all tests were analyzed using the MH and LR procedures. The MH analysis was be done using the program written by Rogers and Hambleton (in press), using total score as the matching criterion (MH-T). The logistic regression analyses were conducted using SPSS-X. The logistic regression analysis using total score as criterion is referred to as the LR-T analysis.

The .01 level of significance was used in all analyses. False positive error rates were calculated for analysis for each condition. In addition, false negative error rates were also calculated.

In order to fully understand what types of items were most likely to be incorrectly identified as DIF, and under what conditions, the number of times each item was identified (out of 10 replications) was calculated, items were grouped according to item characteristics and false positive error rates for each group of items were calculated.

Part II

The second part of the study investigated whether changing the conditioning variables used in the Mantel-Haenszel and logistic regression procedures would result in decreased false positive error rates.

Subtest Selection

The use of valid subtest scores was investigated. Valid subtests were constructed in two ways. First, subtest items were selected based on the specifications which were used to generate the data. Items with a measurement direction of 0 to 25 degrees were selected into subtest one. Items with a measurement direction of 65 to 90 degrees were selected into subtest 2. These subtests are referred to as a priori subtests 1 and 2. Thus, this first method of subtest selection allowed assessment of the subtest as criterion analysis under the most favorable conditions possible, that is when it is known, a priori, which items form unidimensional scales.

The second way valid subtests were constructed was based on the results of the NOHARM (Fraser, 1981) procedure. NOHARM was used to perform a nonlinear factor analysis, and items were assigned to subtests based upon these empirical results. Item measurement direction was calculated, and items were selected into subtests in the same way as described for the a priori subtests. These subtests are referred to as NOHARM subtests 1 and 2. The results obtained using the NOHARM subtests procedure were compared to those obtained using the a priori subtests described above to determine to what extent the factor analysis recovered the structure of the tests.

Preliminary NOHARM Investigations. A preliminary question was what set of NOHARM estimates to use to estimate the discrimination parameters, and subsequently the item directions. Discussions with researchers at ACT and Professor Terry Ackerman at the University of

Illinois revealed that they used the latent trait estimates. Pre-preliminary work suggested that of the factor analytic solutions the Promax rotation would be the most appropriate. Therefore a systematic comparison of the latent trait solution and the Promax solution was conducted. Two tests were simulated - much like tests 1 and 2 described above. Test 1 consisted of 50 pure A items and 16 pure B items. Test 2 consisted of 25 items at 15 degrees, 25 items at 30 degrees, eight items at 60 degrees, and eight items at 75 degrees. Responses for 3000 examinees were simulated to provide 2000 reference group examinees and 1000 focal group examinees. Three levels of correlation between the abilities were used: 0, .3 and .7. Each data set was analyzed using NOHARM four separate times - once using 1000 reference group examinees, once using 1000 focal group examinees, once using 2000 reference group examinees, and a final time using 1000 reference group and 1000 focal group examinees.

Latent trait and Promax results were used to calculate item directions (cosines) and items were selected into subtests. Correlations between true cosines and the two estimated cosines were calculated for each sample.

The Promax rotation yielded better results than the LT parametrization in terms of both the correlations between the cosines and in terms of the number of items correctly classified with some exceptions. Therefore, it was decided that the loadings obtained with the Promax rotation would be used in the investigation.

The analyses using 1000 reference group examinees and 1000 focal group examinees had results that were generally as good or very close to as good as the analyses using 2000 reference group examinees and generally better than either of the 1000 examinee analyses. Therefore, the combined reference and focal groups were used with the NOHARM procedure.

Data Analysis

The MH and LR procedures were repeated, this time conditioning on subtest scores rather than total scores. First, the MH analysis was implemented using a purified total score as criterion. The purified total score was based only on those items which were not identified on the first MH analyses (but always including the studied item). This is referred to as the MH-P analysis.

All items were reanalyzed with the LR procedure, expanded to include subtest scores in the regression equation. First both a priori subtest scores were substituted for total score (LR-A), then both NOHARM subtest scores were substituted for total score (LR-N). Because the LR model allowed both subtest scores to be incorporated into the model simultaneously, it was hypothesized that substantial reductions in false positive error rates would be obtained in this condition. The characteristics of the false positive items were also investigated as described under Part I.

The correspondence between the results obtained using the a priori subtests and the NOHARM subtests was examined, both by a comparison of false positive error rates for the LR-A and LR-N analyses, and by examining the item classifications and correlations among the scores.

It was expected that the MH-T and LR-T analyses would both result in relatively high false positive error rates. Incorporating the two subtest scores into the LR equation was expected to improve matching and thus reduce false positive errors. It was further expected that the a priori subtests would result in more accurate matching than the NOHARM subtests, and thus the LR-A analyses would yield lower false positive error rates than the LR-N analyses.

Part III

Data from the College Board achievement tests for Chemistry and History were analyzed to determine whether real test data would yield results at all similar to those obtained in the simulated conditions described above.

Each test was first shortened to 66 items (using random item selection) to make test length equal to that used in the simulations. For the Chemistry test, one thousand white and one thousand Asian American examinees were randomly selected from the item response data which were available for use in the study. For the History test, one thousand male and one thousand female examinees were randomly selected.

First, the MH and LR procedures were implemented as described in Part I, conditioning was on total test score only. Next, a nonlinear factor analysis was conducted using NOHARM, to assess whether the data were multidimensional and whether meaningful valid subtests could be constructed. The NOHARM results suggested that the History test data were adequately fit by two dimensions, while the Chemistry test data were better fit by three dimensions.

Subtests for the History test were constructed following the same procedure as was used for the simulated data. The MH procedure was then repeated, using purified total score as criterion. The logistic regression procedure was also repeated, with both subtest scores included in the equation in lieu of total score.

Because three dimensions were identified for the Chemistry test, the NOHARM estimates were submitted to a cluster analysis (using SPSS-X). The cosine distance was used. Based on a three cluster solution, items were sorted into three subtests. The MH and LR procedures were repeated as described for the History test, expect that for the Chemistry test, three rather than two subtest scores were used. As in Part II, the results of the successive analyses were compared, with respect to the different criterion scores used, and with respect to differences in the MH and LR results.

CHAPTER IV

RESULTS

Six conditions were simulated for each of three different test structures. The dimensionality of the tests was varied by varying the measurement direction of the items. Test 1 consisted of 48 pure A items (with a measurement direction of 0 degrees) and 16 pure B items (with a measurement direction of 90 degrees). Thus no item was multidimensional, but the two types of items, taken together, resulted in a multidimensional test. Test 2 consisted of 48 items which were more sensitive to dimension A, and 16 items which were more sensitive to dimension B, but all items were influenced to some extent by both dimensions. Finally, test 3 consisted of 12 pure A items, 24 items which were more sensitive to A than B, 16 items which were more sensitive to B than A, and 12 items which were equally sensitive to both conditions. Thus, the difference between tests 2 and 3 was that for test 2 the first 48 items had measurement directions of 15 or 30 degrees, while the measurement directions for the first 48 items in test 3 ranged from 0 to 45 (at 15 degree intervals). However, the parameters for the last 16 items were the same for tests 2 and 3. All three tests contained two DIF items, which were pure A or predominantly A items. For all of the above simulations there was no difference in the underlying ability distributions for the two groups on ability A, while the reference group mean was one standard deviation greater than the focal group mean on ability B.

For each test, two levels of correlation between the two underlying abilities were simulated, with this correlation specified as either .3 or .7.

Item difficulty was systematically varied from -1.75 to 1.75 at intervals of .5 for all items. For the first 48 items MDISC was systematically varied from .2 to 1.2, at intervals of .2. The discrimination of the last 16 items (and the 2 DIF items) was varied to create three different discrimination conditions for each test (and each level of correlation). Thus, in the low discrimination condition the discrimination of these items was either .2 or .4, in the medium discrimination condition the discrimination of these items was either .6 or .8 and in the high discrimination condition the discrimination of these items was either 1.0 or 1.2. Thus the discrimination of the first 48 items (generally the pure A or predominantly A items) did not change across discrimination conditions, but the discrimination of the last 16 and the two DIF items did change.

Descriptive information for the three tests under the six studied conditions is presented in Table 3. For all tests, as the discrimination of the B items increased, the standard deviation of the samples increased, as did the between group difference in means.

DIF Analyses with Total Score as Matching Criterion

DIF analyses using total score as the matching criterion were the first analyses to be carried out. Both logistic regression (LR-T) and Mantel-Haenszel (MH-T) procedures were implemented for all data sets. The results of these analyses are presented in Table 4. The LR-T and

Table 3

Total Score Descriptive Statistics[1]

Test	Statistic	Low MDISC		Medium MDISC		High MDISC	
		r=.3	r=.7	r=.3	r=.7	r=.3	r=.7
1	$\bar{X}$	32.7	32.6	32.1	32.1	31.6	31.7
	SD	7.3	7.7	7.9	8.5	8.4	9.3
	$\bar{X}_R - \bar{X}_F$	.6	.7	1.4	1.9	2.4	3.2
2	$\bar{X}$	31.5	31.5	31.1	31.1	30.6	30.7
	SD	8.3	9.0	9.4	10.3	10.4	11.3
	$\bar{X}_R - \bar{X}_F$	2.8	2.9	1.7	2.8	5.0	4.7
3	$\bar{X}$	31.6	31.6	30.9	31.1	30.6	30.6
	SD	8.0	8.7	9.2	10.1	10.2	11.2
	$\bar{X}_R - \bar{X}_F$	3.0	2.7	3.8	3.8	4.7	4.3

[1] Statistics represent averages across ten replications.

MH-T procedures yielded very similar results, with the MH-T being slightly more conservative in most conditions.

From Table 4, it can be seen that substantial numbers of false positives were obtained in several analyses. The highest numbers of false positives were obtained in the high discrimination conditions for all three tests. For test 1 close to fifty percent of the items were identified as DIF in the high discrimination conditions. Fewer items were identified as DIF in the medium discrimination conditions for all tests, although rates were still high, ranging from 19 to 41 percent depending on the test and condition. The fewest number of false

Table 4

Number of False Positive Errors with Total Score As Matching Criterion[1]

Condition		Test 1		Test 2		Test 3	
MDISC	Correlation						
		LR-T	MH-T	LR-T	MH-T	LR-T	MH-T
MDISC	Correlation						
Low	.3	8.2	8.2	4.9	3.7	11.6	10.8
	.7	8.9	8.5	4.1	3.9	9.8	9.5
Medium	.3	25.7	24.6	14.1	13.3	19.4	18.6
	.7	24.1	23.5	11.9	11.9	20.1	18.6
High	.3	32.8	32.3	20.1	19.7	24.2	23.1
	.7	33.2	31.0	19.1	18.6	23.3	22.4

[1] Averaged across ten replications.

positives were obtained in the low discrimination conditions, with 6 to 18 percent of the items being identified as DIF.

There were considerable differences in false positive rates across the three test structures. Test 1 showed the most marked increase in rates of false positives across the three discrimination conditions. While test 3 had similar rates at the lowest discrimination condition, rates in the medium and high discrimination conditions were not as high. Test 2 had the lowest false positive rates of all three tests in all conditions. There were minimal differences in the number of false positives across trait correlation levels. The size of the correlation between the traits appeared to have little influence on the DIF results from the three tests. Rates for failures to identify the two DIF items were extremely low for both the LR-T and MH-T analyses. False negative error rates ranged from 0 to .8 percent across all conditions.

Change in Matching Criteria

Three additional DIF analyses were conducted to determine whether a change in the matching criteria would result in improved accuracy, i.e. lower false positive rates. First the MH procedure was re-implemented, this time using a purified total score as the matching criterion (MH-P). The purified total score was calculated by removing all the items which were identified as DIF in the MH-T analysis (except the studied item) and using that score as the matching criterion.

The logistic regression procedure was then implemented two additional times, this time with subtest scores substituted for total score in the logistic regression equation. Subtest scores were calculated by selecting the "most pure" items and using only those items to calculate subtest scores. Thus, items which measured primarily dimension A (that is, had a measurement direction of 25 degrees or less) were selected into subtest 1, and those which measured dimension B (that is, had a measurement direction of 65 degrees or more) were selected into subtest 2. Two sets of subtests were formed, referred to as the a priori subtests and the NOHARM subtests. The a priori subtests were selected based on the true item cosines (those used to generate the data) and the NOHARM subtests were selected based on the cosines calculated from the NOHARM estimates of the item discriminations. Descriptive information for the a priori subtests is presented in Table 5.

Table 5

A Priori Selected Subtests Descriptive Statistics

Test	Subtest	Statistics	Low MDISC		Medium MDISC		High MDISC	
			r=.3	r=.7	r=.3	r=.7	r=.3	r=.7
1	1	$\bar{X}$	25.1	25.1	25.1	25.1	25.1	25.1
		SD	6.8	6.9	7.0	6.9	7.0	7.0
		$\bar{X}_R - \bar{X}_F$	.3	.2	-.6	.1	.2	.1
1	2	$\bar{X}$	7.5	7.5	7.0	7.0	6.5	6.5
		SD	2.0	2.0	2.8	2.8	3.6	3.6
		$\bar{X}_R - \bar{X}_F$	.9	.9	1.9	2.1	2.8	2.9
2	1	$\bar{X}$	12.1	12.1	12.1	12.1	12.0	12.0
		SD	3.9	4.0	3.9	4.1	4.0	4.2
		$\bar{X}_R - \bar{X}_F$	.6	.7	.6	.7	.7	.7
2	2	$\bar{X}$	3.4	3.4	3.2	3.2	3.0	3.0
		SD	1.3	1.3	1.6	1.7	2.0	2.0
		$\bar{X}_R - \bar{X}_F$	.4	.4	.9	.9	1.4	1.4
3	1	$\bar{X}$	12.0	12.0	11.9	11.9	11.9	11.9
		SD	3.7	3.9	3.8	3.9	3.9	4.0
		$\bar{X}_R - \bar{X}_F$	.3	.3	.3	.3	.4	.2
3	2	$\bar{X}$	3.4	3.4	3.2	3.2	3.0	3.0
		SD	1.3	1.3	1.6	1.7	2.0	2.0
		$\bar{X}_R - \bar{X}_F$	.4	.4	.9	.9	1.3	1.2

The results of these three additional DIF analyses are presented in Table 6. Figure 1 allows for comparison across all analyses. The MH-P analyses resulted in minimal or no changes as compared to the MH-T analyses for false positive rates for all the low discrimination conditions of all three tests, and for the medium and high discrimination conditions of tests 2 and 3. The greatest changes in false positive rates for the MH-T analyses to the MH-P analyses were obtained in the medium and high discrimination conditions of test 1. In these conditions a substantial decrease in false positive rates was obtained when the purified total score was used in lieu of total score.

The logistic regression analyses resulted in dramatic reductions in false positive rates in several conditions. The most marked change was on test 1, where both the LR-A and the LR-N analyses resulted in substantial reductions in all conditions. These analyses resulted in false positive rates 50 percent to 98 percent lower than the rates obtained when total score was used as criterion.

For tests 2 and 3 substantial reductions were obtained in the medium and high discrimination conditions, but not in the low discrimination conditions. In the low discrimination conditions of tests 2 and 3 the LR-A and LR-N procedures resulted in increases rather than decreases in false positive rates, with one exception (the LR-N analysis for the low MDISC, r=.3 condition).

The pattern of results for the LR-A and LR-N analyses were similar, but the actual numbers of false positives obtained differed in the various conditions. On test 1 the LR-A analyses tended to flag fewer false positives than the LR-N analyses. However, on tests 2 and 3

Table 6

Number of False Positive Errors with Subtests as Matching Criteria

		Test 1			Test 2			Test 3		
MDISC	Correlation	MH-P	LR-A	LR-N	MH-P	LR-A	LR-N	MH-P	LR-A	LR-N
Low	.3	7.5	1.2	1.4	3.7	10.3	6.6	10.5	16.7	6.6
	.7	8.9	1.1	4.5	4.1	9.3	6.0	8.2	16.7	11.9
Medium	.3	16.5	.5	.5	12.8	6.4	3.6	18.9	11.8	3.2
	.7	16.7	1.0	1.8	12.5	7.8	6.5	18.3	13.1	6.2
High	.3	16.9	.7	.7	18.2	6.6	2.1	24.4	8.1	2.6
	.7	17.2	1.5	1.8	17.8	6.0	7.0	22.9	8.4	5.5

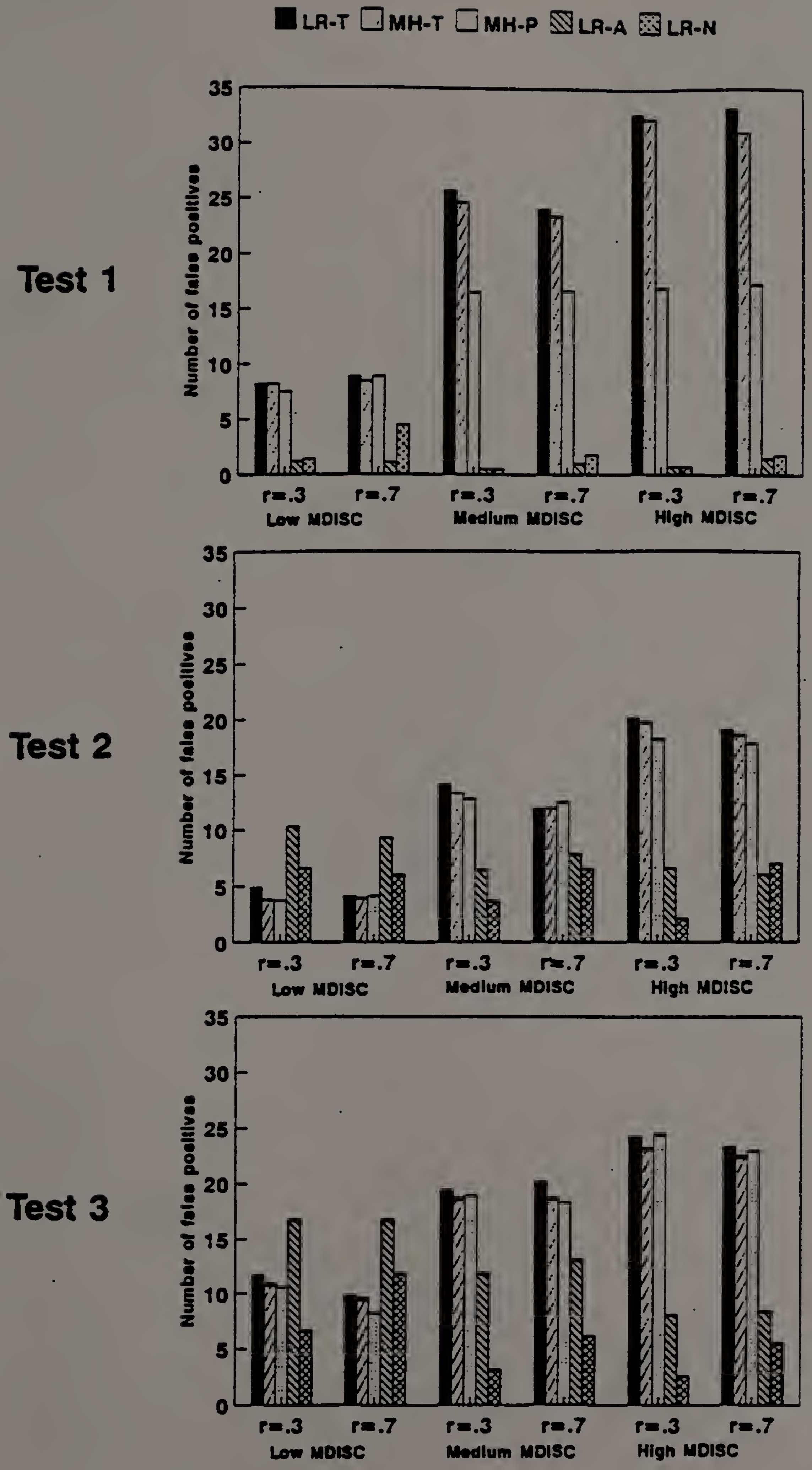

Figure 1. A Comparison of False Positive Error Rates

the LR-N analyses yielded the lowest false positive rates. This is an unexpected result, and will be discussed in some detail later.

When total score was used as the matching criterion, the number of false positives increased as the discrimination of the B item set increased. This trend was reversed for the LR-A and LR-N analyses, where there was a tendency for false positive rates to be lower (or unchanged) in the higher discrimination conditions.

The correlation between the underlying traits had virtually no impact on false positive rates in the LR-A analyses. However, in the LR-N analyses a lower false positive rates were associated with the lower correlation conditions.

False negative error rates associated with all analyses are reported in Table 7. Reported percentages are based on two items occurring in six conditions and ten replications for each test. Thus, for each test there were 120 opportunities for false negative errors. False negative error rates for the MH-P, LR-A, and LR-T analyses were higher than false negative error rates for the MH-T and LR-T analyses.

Table 7

Summary of False Negative Errors

	Analysis				
Test	LR-T	MH-T	MH-P	LR-A	LR-N
1	0	0	0	2.5	2.5
2	.8	.8	4.2	4.2	3.0
3	0	.8	1.7	10.0	8.0

NOTE: Error rates were calculated as the percentage of times DIF items were missed of a possible 120 opportunities: 2 DIF items, in six conditions per test, and ten replications for each condition.

Supplemental Analyses

Because of the unexpected results obtained for the low discrimination conditions of tests 2 and 3, namely that more false positives rather than fewer were observed when subtest scores were used as criteria, two further conditions were investigated. These were variations on the low discrimination condition of test 2. In both conditions the item parameters were the same as those described above for the low discrimination condition of test 2, but the ability distributions were changed. In both cases the reference group mean on ability B was one standard deviation higher than the focal group mean on ability B, as it was in all other simulations. The changes were in the distributions on ability A. In the first case, the reference group mean on ability A was set to be .5 greater than the focal group mean on ability A, thus the reference group was more able than the focal group on both dimensions (consistent difference). In the second case the reference group mean was set to be .5 less than the focal group mean on ability A, thus the reference group was less able than the focal group on one dimension, but more able on the second (crossed difference). This was done to assess whether the direction of the ability distribution differences influenced false positive error rates. Both correlation levels were simulated.

Descriptive statistics for the supplemental test and subtests are presented in Table 8. The same sequence of DIF analyses were performed for these conditions as was performed for the previous conditions. Results of these analyses are presented in Table 9. The consistent distributional difference conditions resulted in relatively few false positives when total score was used as the matching criterion. In

Table 8

Descriptive Statistics for Test 2
Supplemental Conditions

Conditioning Variable	Descriptive Statistic	Consistent Distributional Differences		Crossed Distributional Differences	
		r=.3	r=.7	r=.3	r=.7
Total Score					
	$\overline{X}$	33.0	33.0	30.2	30.2
	SD	8.6	9.6	8.0	8.7
	$\overline{X}_R - \overline{X}_F$	5.9	5.6	.1	.2
A Priori					
Subtest1	$\overline{X}$	12.8	12.8	11.4	11.4
	SD	4.0	4.1	3.8	4.0
	$X_{\overline{R}} - X_{\overline{F}}$	2.2	2.0	.7	.8
Subtest2	$\overline{X}$	3.4	3.4	3.4	3.4
	SD	1.3	1.4	1.3	1.3
	$X_{\overline{R}} - X_{\overline{F}}$	.5	.5	.4	.4

contrast, in the crossed distributional difference condition a substantial number of false positive errors were made in both the LR-T and MH-T analyses. There were no false negative errors associated with the LR-T or MH-T analyses for either condition.

Table 9

DIF Results for Test 2 Supplemental Conditions[1]

	Analysis				
Distributional Difference	LR-T	MH-T	MH-P	LR-A	LR-N
Consistent					
r=.3	2.7	2.3	2.2	13.7	22.5
r=.7	1.0	1.0	.7	11.3	13.2
Crossed					
r=.3	11.9	10.3	10.9	5.1	4.9
r=.7	10.3	9.6	10.2	6.4	8.1

[1] Number of false positives per test (of a possible 64) averaged across ten replications.

When the purified total score was used as the matching criterion there were minimal changes in false positive rates for both conditions. Again, no false negative errors were observed.

Use of subtest scores in the LR analyses resulted in substantial changes. In the consistent distributional difference condition false positive rates increased dramatically in both the LR-A and LR-N analyses. Changes in the crossed difference condition were in the opposite direction, with the LR-A and LR-N analyses yielding fewer false positives than the LR-T and MH-T analyses. The LR-A analyses missed 2 (of 40) DIF items in the consistent difference condition, and 1 (of 40) in the crossed difference condition. The LR-N missed 4 (of 40) and 2 (of 40) in the consistent and crossed difference conditions.

Characteristics of False Positive Items

In order to further understand the results of the series of DIF analyses, and the changes in classification with the changes in criterion, additional descriptive analyses were conducted. Item false positive identification rates were broken down by item measurement direction, item discrimination, and item difficulty. Item false positive identification rates were calculated by grouping items according to the variable of interest (e.g., item discrimination), calculating the number of replications on which each item was identified as DIF, and then averaging across the items in the group. There were 10 replications, so that an identification rate of 10 would mean that all items of that type were identified on all replications. The two DIF items were not included in this series of descriptive analyses, as the goal was to identify what item characteristics were associated with false positive identifications.

False positive rates broken down by item measurement direction are presented in Table 10. The numbers in the table reflect the average number of times items of a given measurement direction were incorrectly identified as DIF, out of a possible ten replication. The rates for the various analyses of test 1 reveal that when total score is used as the matching criterion, it is the items at 90 degrees (the pure B items) which are likely to be identified. The pure B items are in the minority, and clearly the items at 0 degrees (pure A items) have a greater influence on total score. In the MH analyses, when items identified as DIF are removed and total score recalculated for the MH-P analysis, the purified score is now influenced even more by pure A items and rates for items at 0 degrees approach 0, while rates for the pure B

Table 10

Average Number of False Positive Errors
Broken Down by Item Measurement Direction

Test	Analysis	Item Measurement Direction in Degrees						
		0	15	30	45	60	75	90
1	LR-T	1.9	-	-	-	-	-	8.2
	MH-T	1.8	-	-	-	-	-	8.1
	MH-P	.1	-	-	-	-	-	8.2
	LR-A	.1	-	-	-	-	-	.2
	LR-N	.2	-	-	-	-	-	.5
2	LR-T	-	1.4	.2	-	4.7	6.4	-
	MH-T	-	1.3	.1	-	4.2	6.2	-
	MH-P	-	.8	.1	-	5.3	6.3	-
	LR-A	-	.1	1.2	-	5.3	.4	-
	LR-N	-	.4	.8	-	1.7	1.4	-
3	LR-T	4.0	1.7	.3	2.0	4.5	6.1	-
	MH-T	3.9	1.6	.2	2.0	4.1	5.9	-
	MH-P	3.7	1.1	.2	2.3	4.6	6.0	-
	LR-A	.2	.5	1.4	4.5	5.5	.3	-
	LR-N	.6	.7	1.3	2.4	.9	1.1	-

items remain high. In contrast, when the two subtest scores are incorporated into the LR equations, rates for both the pure A and pure B items drop to almost 0.

Tests 2 and 3 differ from test 1 in that most of the items are not pure with respect to either dimension, and there is greater variability in the measurement directions of the items. Rather than two levels of measurement direction as on test 1, test 2 has 4 levels, and test 3 has 6. There are corresponding differences in false positive identification rates as a result.

When total score is used as the matching criterion for test 2, it is the items at 60 and 75 degrees which are most likely to be identified. This is also true for test 3, and in fact the actual rates for items at these directions are very similar. However, for test 3, the next highest false positive rates are found for the items at 0 degrees. The lowest rates for both tests are for the items at 30 degrees.

Changing the criterion from total score to purified total score for the MH resulted in slight reductions in false positive rates for items at 15 degrees, and no change for items at 30 degrees for both tests. Small increases in rates for items at 45, 60 and 75 degrees were noted.

When the subtest scores were substituted for total score in the LR analyses, the pattern of results changed. While the LR-T analyses tended to identify the most discrepant items (items at 75 or 60 degrees, and then those at 0 degrees) the LR-A was more likely to identify items at 60 or 45 degrees. In contrast, the item false positive rates for LR-N analyses tended to show much less variability across item measurement direction levels.

False positive identification rates broken down by item discrimination are presented in Table 11. This table reveals that for the LR-T, MH-T and MH-P analyses false positive rates increased as item discrimination increased, without exception. This trend was not as clear for the LR-A and LR-T analyses. For test 1, the highest false positive rates were associated with the lower item discriminations. The LR-A analyses tended to identify the higher discrimination items at higher rates, but there were some exceptions to this for test 3. The LR-N analyses tended to have low rates overall, but there was not a clear relationship between false positive rates and item discrimination for tests 2 and 3.

False positive rates for items broken down by item difficulty are presented in Table 12. From this it can be seen that there was a tendency for items of moderately difficulty to have higher false positive rates as compared to the relatively more easy or more difficult items. This tendency was consistent across tests and across analyses.

Correspondence Between A Priori Results and NOHARM Results

It was apparent from several of the results presented above that the results obtained with the LR-A analyses often differed from the results obtained with the LR-N analyses. Therefore, a more detailed assessment of the correspondence between the a priori selected subtests and the NOHARM selected subtests was conducted.

Correlations between the true cosines and the cosines calculated using NOHARM estimates are presented in Table 13. The relationship between the true and NOHARM estimated cosines is important as items were

Table 11

Average Number of False Positive Errors
Broken Down by Item Discrimination

Test	Analysis	Item Discrimination					
		.2	.4	.6	.8	1.0	1.2
1	LR-T	.7	2.3	3.3	4.0	4.7	5.5
	MH-T	.6	2.3	3.3	3.8	4.6	5.4
	MH-P	.6	1.9	2.5	2.6	2.6	2.7
	LR-A	.3	.3	.1	.1	.1	.1
	LR-N	.5	.6	.2	.1	.1	.1
2	LR-T	.3	1.0	1.7	2.6	2.8	3.6
	MH-T	.3	.8	1.5	2.5	2.7	3.4
	MH-P	.2	.7	1.6	2.4	2.8	3.1
	LR-A	.3	.9	1.0	1.5	1.6	2.0
	LR-N	.8	1.0	.4	.9	.6	1.2
3	LR-T	.4	.8	2.8	3.2	4.9	5.0
	MH-T	.3	.7	2.5	3.0	4.7	4.8
	MH-P	.3	.7	2.6	3.0	4.8	4.8
	LR-A	.3	1.4	1.3	3.1	2.0	3.8
	LR-N	.8	1.0	.4	.9	.6	1.2

Table 12

Average Number of False Positive Errors
Broken Down by Item Difficulty

		Item Difficulty			
Test	Analysis	-1.75 to -1.25	-.75 to -.25	.25 to .75	1.25 to 1.75
1	LR-T	3.2	3.7	3.8	3.1
	MH-T	3.1	3.6	3.7	3.0
	MH-P	2.1	2.2	2.3	2.3
	LR-A	.1	.1	.2	.1
	LR-N	.3	.3	.3	.2
2	LR-T	1.7	2.1	2.3	1.8
	MH-T	1.6	2.0	2.2	1.7
	MH-P	1.6	1.9	2.0	1.7
	LR-A	1.1	1.3	1.3	1.1
	LR-N	.8	.9	.9	.7
3	LR-T	2.1	3.1	3.3	2.8
	MH-T	2.1	2.9	3.1	2.7
	MH-P	2.1	3.0	2.9	2.9
	LR-A	1.4	2.1	2.3	1.9
	LR-N	.7	1.0	1.0	.9

Table 13

Correlations Between True Cosines and Cosines Based on NOHARM Analyses[1]

Condition	Test 1	Test 2	Test 3
low MDISC			
r=.3	.91	.48	.72
r=.7	.68	.25	.52
medium MDISC			
r=.3	.98	.86	.87
r=.7	.95	.67	.75
high MDISC			
r=.3	.99	.91	.89
r=.7	.97	.84	.81

[1] Correlations were calculated for each replication separately, and then averaged.

selected into subtests based on cosines. Differences in cosines could lead to different items being selected into subtests. In general, the cosine correlations were higher when the correlation between the underlying traits was .3 than when it was .7. Cosine correlations also tended to be higher as the discrimination of the B items increased. Finally, cosine correlations were higher for test 1 (where all items were sensitive to only one dimension) as compared to tests 2 and 3 where most items were sensitive to both dimensions.

Correlations between the a priori selected subtests and NOHARM selected subtests are presented in Table 14. While the pattern of relationship among the subtest scores are not as clear as for the cosines, the same general trends are present. That is, higher a priori/NOHARM subtest correlations are associated with the lower correlation between the underlying traits, more highly discriminating B items, and with test 1 as opposed to tests 2 and 3. In addition it is noteworthy that for most of the conditions (13 of 18) the correlations between the two NOHARM subtests was within .1 of the correlations of the two a priori subtests with each other.

Information on item classification accuracy is presented in Table 15. In this table the percentage of items missed refers to the percentage of test items whose true cosines were within the specified limits for one of the subtests, but which were not assigned to that subtest. Thus these items should have been included but were not. Items which were cross-classified were items which should have been included on one subtest, but were incorrectly included on the other subtest. Items which were correctly classified were those which included in the correct subtest. Again, higher correct classification rates are associated with the lower trait correlation, higher discrimination of the B items, and with test one as compared to tests 2 and 3. Cross-classifications and missed classifications also tend follow this pattern.

In addition to the relatively infrequent cross-classifications which were noted in the NOHARM subtests, the NOHARM assignements tended to include items which were not included in either of the a priori subtests - that is items with a true measurement direction greater than

Table 14

Correlations Between Subtest Scores

Test	Conditions	A priori 1 with NOHARM 1	A priori 2 with NOHARM 2	A priori 1 with A priori 2	NOHARM 1 with NOHARM 2
1					
	low MDISC				
	r=.3	.98	.94	.13	.12
	r=.7	.75	.78	.29	.20
	medium MDISC				
	r=.3	.99	1.0	.18	.18
	r=.7	.99	.99	.45	.44
	high MDISC				
	r=.3	.91	.92	.20	.20
	r=.7	.99	.99	.50	.49

Continued on the next page.

Table 14, continued:

Test	Conditions	A priori 1 with NOHARM 1	A priori 2 with NOHARM 2	A priori 1 with A priori 2	NOHARM 1 with NOHARM 2
2					
	low MDISC				
	r=.3	.80	.54	.21	.27
	r=.7	.73	.49	.30	.30
	medium MDISC				
	r=.3	.93	.85	.39	.54
	r=.7	.85	.81	.52	.53
	high MDISC				
	r=.3	.94	.91	.47	.63
	r=.7	.94	.92	.61	.69

Continued on the next page.

Table 14, continued:

Test	Conditions	A priori 1 with NOHARM 1	A priori 2 with NOHARM 2	A priori 1 with A priori 2	NOHARM 1 with NOHARM 2
3					
	low MDISC				
	r=.3	.90	.61	.19	.40
	r=.7	.80	.59	.28	.31
	medium MDISC				
	r=.3	.96	.81	.34	.50
	r=.7	.88	.82	.49	.52
	high MDISC				
	r=.3	.96	.88	.41	.55
	r=.7	.94	.90	.58	.66

Note: Correlations were calculated for each replication separately, and then averaged.

Table 15

Item Classification Accuracy

		Mean Percentage of Items[1]		
Test	Condition	Missed	Cross-Classified	Correctly Classified
1				
	low MDISC			
	r=.3	14	1	86
	r=.7	36	9	56
	medium MDISC			
	r=.3	3	0	97
	r=.7	6	0	94
	high MDISC			
	r=.3	3	0	97
	r=.7	4	1	95
2				
	low MDISC			
	r=.3	36	12	52
	r=.7	50	19	31
	medium MDISC			
	r=.3	9	1	90
	r=.7	30	6	64
	high MDISC			
	r=.3	3	1	96
	r=.7	9	3	88

Continued on the next page.

Table 15, continued:

		Mean Percentage of Items[1]		
Test	Condition	Missed	Cross-Classified	Correctly Classified
3				
	low MDISC			
	r=.3	24	3	73
	r=.7	41	15	45
	medium MDISC			
	r=.3	6	0	94
	r=.7	31	0	69
	high MDISC			
	r=.3	3	1	96
	r=.7	16	1	83

[1]All cell percentage are averages across ten replications.

25 and less than 65 degrees. This means that the items which were included on the NOHARM subtests are more varied with respect to measurement direction than those included on the a priori subtests. Table 16 presents the percentage of NOHARM subtest items which fall into this category. For tests 2 and 3 a substantial number of items with measurement directions within this range were included in a NOHARM subtest. This is consistent with all of the above. In summary, the primary finding of the comparisons between the NOHARM and the a priori subtests was that the NOHARM results were generally consistent with the a priori results. There were some differences, as not unexpectedly the

Table 16

Percentage of NOHARM Subtest Items with Measurement Direction Between 25 and 65 Degrees

Condition	Test 2	Test 3
low MDISC		
r=.3	43	41
r=.7	47	41
medium MDISC		
r=.3	38	36
r=.7	40	44
high MDISC		
r=.3	44	35
r=.7	38	36

NOTE: Percentages are averaged across replications.

Percentages are not reported for Test 1 as there were no items at these measurement directions for Test 1.

correspondence between the two was not perfect. Differences were greatest when the items were less discriminating, and when the correlation between the underlying abilities was greater. Interestingly, the fact that the NOHARM classifications resulted in subtests that were more varied than the a priori subtests in terms of item measurement direction appeared to improve (reduce) false positive error rates when subtest scores were used as criterion.

Real Data Results

In the final phase of this investigation the procedures discussed above were applied to two real data sets. These tests were both achievement tests, one in the area of history, the other in the area of chemistry. Both tests were shortened to 66 items (using random item deletion). The reference group for the history test was males, and the focal group was females. The reference group for the chemistry test was whites, and the focal group was Asian Americans. Descriptive statistics for these two data sets are presented in Table 17.

Table 17

Descriptive Statistics for the History and Chemistry Tests

			History			
	Total		Subtest 1		Subtest 2	
Group	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Males	40.4	10.6	12.3	3.2	13.0	4.6
Females	37.5	10.6	11.5	3.4	11.5	4.4
Combined	38.9	10.7	11.9	3.3	12.3	4.5

				Chemistry				
	Total		Subtest 1		Subtest 2		Subtest 3	
Group	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Whites	32.8	12.5	16.6	6.0	9.9	4.9	6.0	2.9
Asian Americans	32.8	11.1	17.4	5.3	9.3	4.4	5.8	2.9
Combined	32.8	11.8	17.0	5.6	9.6	4.6	5.9	2.9

History Test Results

The results of the NOHARM analysis of the history data set suggest that the two-dimensional model (with c=0) provided an adequate fit to the data. The root mean square of the residual matrix was .004499 well below .08944, which is suggested by the author of NOHARM as a rough guideline for assessing goodness of fit. Recent research has suggested that converting the residuals to z-scores and then evaluating the percentage of z-scores greater than 1.96 provides additional information as to the goodness of fit. In this case the percent of z-scores less than -1.96 and greater than 1.96 was 4.4289, again, further evidence of an adequate fit.

Based on the NOHARM results, two subsets of items were formed, and two subtest scores were calculated based on these items. Then the same series of DIF analyses were conducted as were conducted for the simulated data sets. The results of the DIF analyses of the history data set are presented in Table 18. The LR-T and the MH-T analyses yielded very similar results, with the LR-T procedure identifying one more item. Changing to the purified total score for the MH procedure did not result in any changes in item classifications. Similarly, substituting the subtest scores for total score in the LR procedure also did not result in any changes in classifications.

Chemistry Test Results

The two-dimensional NOHARM solution (with c=0) was determined not to provide an adequate fit to the data. While the root mean square was .00562, still well below the recommended .08944, the percent of z-scores less than -1.96 or greater than 1.96 was 8.4, higher than desirable.

Table 18

Number of Items Identified as DIF

	Analysis			
Test	LR-T	MH-T	MH-P	LR-N
History	16	15	15	16
Chemistry	16	13	14	8

Therefore, a second NOHARM analysis was conducted, this time with a three-dimensional solution requested. The root mean square reduced slightly to .00507, and the percent of z-scores outside the acceptable range reduced to 5.4 . Thus, it was judged that the three-dimensional model provided an adequate fit to the data, and that therefore it would be appropriate to form three subtests. Cluster analysis was used to sort the items into subtests. The items were clustered based on cosine distances between the NOHARM-estimated discrimination parameters. A three-cluster solution was used. Three subtest scores were then calculated for each examinee.

The results of the DIF analyses for the chemistry data set are also presented in Table 18. For this data set the LR-T analysis identified 3 more items than the MH-T analysis, and the MH-P analysis identified one item more than the MH-T analysis. However, the LR-N procedure (where all three subtest scores were included in the LR equation) resulted in the fewest number of DIF items being identified of any of the analyses. In fact, the LR-N procedure resulted in fifty percent fewer items being flagged as DIF as compared to the LR-T procedure.

CHAPTER V

DISCUSSION AND CONCLUSIONS

Discussion

The results presented in Chapter IV confirm earlier research that multidimensional item impact may be identified as DIF when there are underlying distributional differences between the two groups, and total test score is used as the matching criterion. This investigation found relatively high false positive error rates when total test score was used as the matching criterion when in fact there were no between group differences in the multidimensional item parameters. This was true for both the Mantel-Haenszel and the logistic regression procedures which produced very similar results. The extent to which this was true was influenced by the dimensionality of the test, and the discrimination parameters of the items in the test. Both of these factors influenced the relative impact each dimension has on total score. When total test score was more influenced by items of one dimension, using total score as the matching criterion was more likely to identify items which were most heavily influenced by another dimension. As total score is more evenly influenced by both dimensions, it is the more extreme items (of both dimensions) that are more likely to be identified.

At the same time, items which were most discriminating were most likely to be identified as DIF. While such items would be expected to have a greater influence on total score, and thus pull the score more towards the direction of these items, highly discriminating items are also more readily identified by DIF procedures, and are more likely to be identified. When the discrimination parameter was increased for items sensitive to the minor dimension (B), it was more likely that items most sensitive to either dimension would be identified.

For the low discrimination sets the average number of false positive errors ranged from approximately 4 to 11. For the medium discrimination sets this range was approximately 12 to 24, and for the high discrimination sets it was approximately 18 to 32. Thus, in a test with 64 non-DIF items, 6 to 50 percent of the items were identified as DIF depending on the condition. While the conditions simulated here were chosen for illustrative purposes, and may be more extreme in terms of dimensionality and discrimination than those found in practice, the very high false positive error rates found in some conditions suggest that multidimensionality in a data set cannot be ignored, and may have a major impact on the results of DIF analyses.

Part II of this study addressed whether using ability estimates (based on subsets of relatively pure items) in lieu of total score would impact the results of the matching criterion. The answer is clearly yes. In almost all cases changing the criterion resulted in changes in the number of false positive errors. In most of the conditions the changes were dramatic, and were in the desired and predicted direction. However, the impact of change of criterion must be evaluated in terms of the analysis (LR versus MH), the dimensionality of the test, item discrimination and item difficulty.

The logistic regression procedure might be considered the procedure of choice when multiple ability estimates are used, as the logistic regression equation readily accommodates multiple ability estimates and thus allows for simultaneous conditioning on all relevant abilities. The most dramatic difference in identification rates in the LR analyses were observed on test 1, where 75 percent of the items were sensitive to one ability, and 25 percent were sensitive to the second ability. When a priori knowledge of item parameters was used to construct subtests, and both subtest scores were included in the logistic regression equation instead of the single total score, substantial reductions in false positive rates were obtained. In fact, in one condition (high MDISC, r=.7) the change in the percentage of items identified dropped from 50 percent, to only 2 percent. The most extreme reductions were for medium and high discrimination sets, as these were the sets with the highest false positive error rates when total test score was used.

When all or almost all of the items were multidimensional (as in tests 2 and 3), the changes in rates were not as dramatic. For test 2 fewer false positive errors were obtained when the total score was used as criterion, and thus there was relatively less room for improvement. For both tests 2 and 3 the lowest identification rates were still higher than the lowest rates obtained for test 1. However, for the medium and high discrimination sets, substituting the subtest scores for total score did result in a substantial reduction in the number of items identified. For instance, for the high MDISC, r=.7 condition of test 2, 30 percent of the items were identified as DIF when total score was used as criterion, which was reduced to only 9 percent of the items when subtest scores were used. For the same conditions of test 3 the reduction was from 36 percent to 13 percent.

For the low discrimination conditions of tests 2 and 3, substituting subtest scores for total score in the logistic regression equation actually resulted in an increase in the number of items identified. The results of the two further simulations of the low discrimination condition of test 2, highlight the importance of the direction of the differences in the underlying multidimensional ability distributions. When there were differences between the two groups on both abilities, and the differences were in the same direction, even fewer items were flagged as DIF with total score as criterion than in the same condition, but with the groups differing only on one ability. In this case matching on the subtest scores resulted in an increase in the number of false positive errors. This suggests that, in this circumstance, matching on total score provides more accurate matching than matching on subtest scores. However when the distributional differences crossed, so that the focal group mean was greater than the reference group mean on the first dimension, but the reverse was true on the second dimension, matching on total score alone resulted in a more items flagged as DIF. In this case, matching on both subtest scores resulted in a substantial reduction in false positive error rates.

The analysis of item identification rates by item direction for the three tests suggests that the effect of incorporating both subtest scores in the LR analysis of test one is to reduce false positive error rates for items with a dimensionality or measurement direction similar to the items used to construct the subtests. This is the most likely explanation for the differences between the analyses which used the a priori subtests and those which used the NOHARM subtests. The NOHARM subtests contained more items covering a greater range of measurement directions. This resulted in fewer false positive errors for the LR-N analyses of tests 2 and 3, which had items spread across a wider range of measurement directions.

The correspondence between the results obtained using NOHARM and those obtained using the a priori selected subtests suggest that it is possible to use NOHARM to group items into subtests in the way that was done here with reasonable accuracy. As noted above, in many conditions the analyses using the NOHARM selected subtests resulted in lower false positive rates than were obtained in the corresponding analyses using a priori subtests. In other conditions there was very little difference in rates.

The correspondence between the NOHARM selected subtests and the a priori selected subtests was best when items were more discriminating, and when there was less of a correlation between the underlying abilities. In addition, correspondence was also better when the test was composed of items which measure one or the other trait (test 1), rather than each item being multidimensional (as on tests 2 and 3).

One finding that was not expected was the relatively small impact of the size of the correlation of the underlying abilities. A substantial change in the magnitude of this correlation (from .3 to .7) resulted in relatively minor changes in item classifications. The impact of the two levels of correlation was probably most apparent in the NOHARM analyses, and those based on the NOHARM subtests. In general, the higher correlation between the underlying abilities was associated with less accurate NOHARM results.

Because there were only two true DIF items included in each test, results regarding false negative errors must be considered suggestive and not conclusive. However, both the LR-T and the MH-T missed only a single item on a single replication, of the 360 possible identifications. The LR-A and LR-N analyses did have higher false negative error rates, with the LR-N having the lower rate of the two.

The results of the real data analyses are encouraging, particularly the results of the analyses of the Chemistry test. For this test, the substantial reductions in the number of items identified as DIF in the LR-N analysis as compared to the LR-T analysis were similar to the reductions obtained in test 1 and in the higher discrimination conditions of tests 2 and 3. The fact that no such reduction in rates was obtained with the History test suggests that the impact of changing the matching criteria depends on the specific test and sample used. This is consistent with the results obtained with the simulated data sets.

Because these two tests are real, the true or correct item classifications are not known. However, it can be argued that the lower number of DIF items is more accurate. This would be expected both on logical grounds, and based on the relatively low false positive error rates obtained in the simulated data analyses.

Implications

This study has several implications for practice. First, practitioners should be aware that multidimensionality in a data set can result in apparent DIF when there are underlying distributional differences and total test score is used as the matching criterion. Further, the number of false positive errors may be alarmingly high. For instance, in one condition simulated here, a full 50 percent of the items were flagged as DIF. While the decision as to whether multidimensional items should be removed from a test is a judgmental one, and must be made in the context of the purpose of testing, practitioners should be aware that the results of DIF analyses where total score is used as the matching criterion depend on the dimensionality of the test as a whole, and the discrimination of the items. It is noteworthy that it is not always the most discrepant items which are identified as DIF. In some circumstances use of total test score may result in the most multidimensional items being the ones which are least likely to be flagged. The discrimination of items may be expected to influence false positive rates both by impacting total test score, and because more discriminating items are more likely to be identified. The results presented above also suggest that items of medium difficulty are most likely to be flagged (at least when the underlying ability distributions are similar to those simulated here).

This research demonstrates that by conditioning on more than one ability estimate it is possible to substantially reduce the number of false positive errors obtained in a multidimensional data set. Further, the NOHARM program yielded discrimination parameter estimates which could be used to select subtests with a reasonably high correspondence to the a priori selected subtests. In fact, in a number of conditions the analyses based on the NOHARM selected subtests yielded lower false positive rates than the corresponding a priori analyses.

While using the subtest scores in lieu of total scores resulted in substantial improvement in the accuracy of the DIF analyses in almost all of the conditions simulated here, the reduction in false positive errors was well above the expected levels of 1 or 5%. Thus, practitioners need to be aware that they may be eliminating items which show differential functioning as a result of multidimensional impact rather than DIF.

One finding which may be of concern to practitioners is the increase in false negative errors associated with the change in criteria to subtest scores. As noted above, it is difficult to evaluate this change due to the small number of DIF items included in this study. However, even if increases are close to the magnitude found here, the cost of these errors must be weighed against the very high false positive error rates associated with the total score as criterion. In some circumstances false positive error rates were close to fifty percent, and clearly practitioners cannot afford to remove fifty percent of the items on a test.

While the focus of this study was not to investigate the correspondence of the M-H and LR procedures, the results do provide evidence that when total test score is used as the matching criterion these two procedure yield very similar results. This is important, as the LR regression procedure has only recently been applied to DIF analyses, and thus there is not an abundance of research on this procedure.

Summary

There were two primary purposes of this study. First, to confirm earlier research which demonstrated that multidimensional item impact may be identified as DIF. The second purpose was to determine whether conditioning on multiple internal ability estimates would reduce the number of false positive errors.

In order to address these two purposes a simulation study was conducted first. Examinee responses were simulated to three different tests. Each data set was two-dimensional, but the dimensional structure was varied across tests. Each test contained 66 items. The first 48 items were most sensitive to the first ability, the next 16 to the second ability, and the last two were true DIF items, with a between group difference of .5 in the difficulty parameter. The MDISC values for the first 48 items were systematically varied between .2 and 1.2 in each test. However, the MDISC values for the 16 items which were more sensitive to the second dimension were either low (.2 or .4) medium (.6 or .8) or high (1.0 or 1.2) in each test. A sample size of 1000 was used for each reference and focal group. The ability distributions were simulated so that the reference group mean on the second dimension was one standard deviation greater than the focal group mean. Correlations between the two dimensions were the same for both groups, set to be either .3 or .7. Ten replications were conducted for each condition.

The first sets of analyses used total score as the matching criterion. As anticipated, high numbers of non-DIF items were flagged as DIF in several of the conditions with both the LR and MH procedures. The factors which seem to contribute most to high false positive rates were the dimensional structure of the test and the measurement direction and discrimination of the items. Items most likely to be identified were high discrimination, moderate difficulty items with measurement directions most discrepant from the direction of the majority of items on the test. The correlation between the underlying abilities had little impact on false positive error rates.

The second part of this study investigated whether a change in matching criterion resulted in a change in false positive error rates. Subtests were selected in two ways - first based on the parameters used to generate the data (a priori subtests) with only the items which had a measurement direction within 25 degrees of a given factor being selected into the subtest for that factor. Thus the most multidimensional items were not included in either subtest. Each data set was also analyzed using NOHARM, and the same subtest item selection procedure was carried out using the NOHARM a-parameter estimates rather than the generating parameters.

The results of this phase of the study provided evidence that the change in criterion from total score to subtest score(s) resulted in substantial changes in false positive rates. First each data set was analyzed again using LR, this time with subtest scores used in lieu of total score. In most (but not all) conditions this change in criteria resulted in substantial reductions in false positive rates. The magnitude of the reductions appeared to be strongly related to the dimensional structure of the test, and the discrimination of the items.

The correspondence between the a priori selected subtests and the NOHARM selected subtests varied as a function of the dimensional structure of the test, the correlation between the two underlying abilities, and the discrimination of the items. In several conditions the NOHARM selected subtests resulted in even greater reductions in false positive errors than the a priori selected subtests, without an increase in false negative errors. Thus it appears that the NOHARM procedure does provide estimates of discrimination parameters which are adequate for selecting items into subtests, at least under conditions similar to those studied here.

In general, as the discrimination of the items loading primarily on the second factor increased, the number of false positive errors obtained using total score increased. At the same time, the number of false positive errors obtained using subtest scores decreased.

In Part III of this study the procedures described above were applied to two real data sets. The results of the real data analyses were consistent with the simulated data analyses, and suggest that the procedures investigated here are feasible for application to real test data. For the Chemistry test substituting subtest scores for total score as the matching criterion resulted in substantial reductions in the number of items identified as DIF.

Delimitations of the Study

While the results presented above are very encouraging, there are several limitations which must be noted. First, it was not possible to investigate fully all of the variables which might be expected to influence how a change from total score to subtest score might influence the results of DIF analyses. For instance, sample size has been shown to influence detection rates in studies of DIF using unidimensional data sets, and thus would be expected to have an impact in multidimensional data sets as well. In fact, sample size may be even more important with multidimensional data, as sample size could well influence the stability of the parameter estimates obtained with programs such as NOHARM. Test length, and in this case subtest length as well, are variables which would also be expected to influence DIF analysis results. The subtests used in the present study were sometimes very short, and thus matching on the second dimension was sometimes done with a score of modest reliability. Longer subtests might have yielded even greater reductions in false positive errors in some conditions.

A second set of limitations has to do with the fact that while several important variables were investigated, it was not possible to investigate each variable exhaustively. For instance, the dimensional structure of the test as a whole appears to be an important factor in multidimensional DIF studies. The present investigation looked at three different tests, chosen to represent two extreme cases, and one mixed case. However, there are limitless other combinations of item parameters which could be used to generate two-dimensional data sets, and other combinations may yield other results.

As with any simulation study, the question of generalizability of results is an important one. The item parameters used in the simulation part of this study were chosen to be within the boundaries of what might be expected to be found in practice, but it is not argued that they are typical or representative. Also, the simulated tests were limited to two-dimensions, which may not be typical of what may be found in practice. The analyses of the real data sets suggests that some tests may have dimensionality greater than two. The simulation phase of the study also assumed that the dimensionality of the data set was known (that is a two-dimensional solution was requested with NOHARM), rather than checking the fit of successive solutions, which would be necessary with real data sets.

In the present study only two DIF items were included in each test, because the primary research questions had to do with false positive rather than false negative error rates. Because there were so few DIF items, comparisons between the various analyses with respect to false negative error rates must be considered tentative. In addition, the between group difference in item difficulty on these items was substantial, and thus the relative sensitivity of the various procedures to different amounts of DIF is not known.

Items were selected into subtests based on only one decision rule, with an arbitrary cutoff. Different decision rules would be expected to result in different items being selected, and thus would be likely to impact on false positive rates.

Directions for Future Research

Several of the limitations noted above suggest directions for future research. First, studies similar to this but which investigate other test lengths, other sample sizes, different item parameter combinations, and different dimensional structures, including tests with three and four dimensions, would be valuable. There are several potentially fruitful areas of research related to determining the dimensionality of both tests and items. Further research is needed to provide guidelines to practitioners on how to determine the number of dimensions required to fit a given data set. In addition, further research on the factors which influence the accuracy of the NOHARM parameter estimates is needed, as the correspondence between the NOHARM estimates and the true parameters is not perfect, and seems to be related to several variables.

Alternatives to NOHARM analyses could also be investigated. It may be that simpler, widely available factor analysis techniques would provide factor loadings which would allow the items to be sorted into subtests as accurately as the NOHARM procedure allows.

Given that IRT-based procedures are generally considered theoretically preferable to procedures such as the MH and LR with unidimensional data, one might argue that IRT-based procedures are the procedures of choice in multidimensional DIF analyses as well. Multidimensional DIF analyses using an IRT model would involve estimating item parameters for the reference and focal groups separately, and then comparing the estimates. Future research might compare the results of such a DIF analysis with the type of subtest-based analyses investigated here.

Conclusions

This study confirmed that multidimensional item impact may be identified as DIF when there are between group differences in the underlying ability distributions, and total score is used as the matching criterion in LR or MH analyses. Under some circumstances the false positive error rates were alarmingly high. When subtests composed of items selected to be relatively more "pure" with respect to each dimension were used in lieu of total score and the logistic regression procedure was repeated, the number of false positive errors was reduced substantially in most conditions studied. This was also found to be true with one of the two real data sets studied.

This study is important because it is one of the first to investigate possible solutions to the problem of differentiating multidimensional item impact from DIF. While simulated data were used extensively, considerable care was taken to evaluate the procedures under conditions where the true item parameters were not known. This, in concert with the results of the real data set analyses suggest that not only does LR offer a potential solution to this dilemma, but that implementation of this procedure is feasible for the practitioner.

REFERENCES

Ackerman, T. A. (1991, November). Measurement direction in a multidimensional latent space and the role it plays in bias detection. Paper presented at the International Symposium on Modern Theories in Measurement: Problems and Issues. Montebello, Quebec.

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

Birenbaum, M., & Tatsuoka, K. K. (1982). On the dimensionality of achievement test data. Journal of Educational Measurement, 19, 259-266.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1991). Examination of various influences on the Mantel-Haenszel statistic. Paper presented at the meeting of the American Educational Research Association, Chicago.

Ellis, B. (1989). Differential item functioning: Implications for test translations. Journal of Applied Psychology, 74, 912-921.

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

Fraser, C. (1981). NOHARM: A FORTRAN program for non-analysis by a robust method for estimating the parameters of 1-, 2-, and 3-parameter latent trait models. Armidale, Australia: University of New England, Centre for Behaviourial Studies in Education.

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

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer Academic Publishers.

Hattie, J. A. (1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research, 19, 49-78.

Hills, J. R. (1989). Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, 8, 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.

Ironson, G. H., Homan, S., Willis, R. & Signer, B. (1984). The validity of item bias techniques with math word problems. Applied Psychological Measurement, 8, 391-396.

Kok, F. (1988). Item bias and multidimensionality. In R. Langeheine & J. Rost (Eds.), Latent trait and latent class models (pp.263-275). New York: Plenum Press.

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

Kubiak, A. T., & Cowell, W. R. (1990, April). Using multiple DIF 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, MA.

Lautenschlager, G. J., & Park, D. G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. Applied Psychological Measurement, 12, 365-376.

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

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. Educational and Psychological Measurement, 52, 443-452.

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1993, April). Using logistic regression with multiple ability estimates to detect differential item functioning. Paper presented at the meeting of the National Council on Measurement in Education, Atlanta.

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

Mellenbergh, G. J. (1989). Item bias and item response theory. International Journal of Educational Research, 13, 127-143.

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

Narayanan, P. (1992). MULTISIM: A Fortran V program for generating two-dimensional data. Amherst, MA: School of Education, University of Massachusetts.

Oshima, T. C., & Miller, M. D. (1990). Multidimensionality and IRT-based item invariance indexes: The effect of between group variation in trait correlation. Journal of Educational Measurement, 27, 273-283.

Oshima, T. C., & Miller, M. D. (1991, April). Multidimensionality and item bias in item response theory. Paper presented at the meeting of the American Educational Research Association, Chicago.

Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. Educational and Psychological Measurement, 40, 397-404.

Reckase, M. D. (1985, April). The difficulty of items that measure more than one ability. Paper presented at the meeting of the American Educational Research Association, Chicago.

Reckase, M. D. (1986, April). The discriminating power of items that measure more than one ability. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Reckase, M. D. (1989). The interpretation and application of multidimensional Item Response Theory models and computerized testing in the educational environment. Office of Naval Research Technical Report (N00014-85-C-0241), Arlington, VA.

Rogers, H. J. (1989). A logistic regression procedure for detecting item bias. Unpublished Doctoral Dissertation, University of Massachusetts, Amherst.

Rogers, H. J., & Hambleton, R. K. (in press). MH: A Fortran 77 program to compute the Mantel-Haenszel statistic for detecign differential item functioning. Educational and Psychological Measurement.

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

Scheuneman, J. D. (1982). A posteriori analyses of biased test items. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 180-198). Baltimore, MD: Johns Hopkins University Press.

Scheuneman, J. D. (1984). A theoretical framework for the exploration of causes and effects of bias in testing. Educational Psychologist, 19, 219-225.

Scheuneman, J. D. (1987). An experimental exploratory study of causes of bias in test items. Journal of Educational Measurement, 24, 97-118.

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

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

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

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

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential test functioning. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 197-239). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Subkoviak, M. J., Mack, J. S., Ironson, G. H., & Craig, R. D. (1984). Empirical comparison of selected item bias procedures with bias manipulation. Journal of Educational Measurement, 25, 301-319.

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