# Examining Sources of Gender DIF:

# A Confirmatory Approach

by

SHARON BARNETT

B.Ed., The University of British Columbia, 1986

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

(Department of Educational and Counselling Psychology,
and Special Education)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

December, 2001

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Education and Counselling Psychology, and Special Education

The University of British Columbia
Vancouver, Canada

Date Thursday, December 13, 2001

DE-6 (2/88)

# ABSTRACT

A confirmatory approach based on a multidimensional model (Douglas, Roussos, & Stout, 1996; Shealy & Stout, 1993; Stout & Roussos, 1995) was used to identify sources of differential item functioning (DIF) and differential bundle functioning (DBF) for boys and girls on the British Columbia Principles of Mathematics Exam for grade 12 (PME12). Data consisted of a total of 9404 examinees; 4335 girls and 5069 boys. There were 45 multiple choice items in the exam.

Analyses were completed in two stages. In stage 1, patterns present in the gender DIF research in mathematics were identified. Stage 2 was the statistical confirmation of these patterns. Sources of gender DIF were confirmed for the content areas: polynomial, quadratic relations, logarithms and exponents. Items tapping higher cognitive levels dealing with patterns and relation, word problems, and items containing visuals were also confirmed as a source of DIF. Exploratory analyses indicated that computation items for which no equations are provided may be a source of DIF along with trigonometry items.

This study contributes to an increased understanding of sources of gender DIF that may assist test developers to ensure that mathematics items measure the construct that they are intended to measure and that the test as a whole measures that which it purports to measure. The findings of this research provide an additional source of information about the differential performance of boys and girls that may be used to develop guidelines and test construction principles for reducing gender DIF in mathematics. This research also contributes to a greater understanding of gender differences in mathematics learning and achievement.

# TABLE OF CONTENTS

# LIST OF TABLES

TABLE

# LIST OF FIGURES

FIGURES

# ACKNOWLEDGEMENTS

I would like to extend my sincere gratitude to the members of my research committee, Dr. Bruno Zumbo, Dr. Marion Porath, and in particular, my research supervisor Dr. Kadriye Ercikan for their patience and support in the development of this thesis.

I would like to thank my parents Len and Evelyn Mendes for encouraging me to pursue my goals and for their continuing support.

I would further like to extend my gratitude to my husband Lance Barnett for his continuing encouragement, support, and for his assistance with late night editing. I would like to also acknowledge the skill with which Lance was able to distract our toddler from inspecting and coloring on the thesis preparation materials continually left out.

Finally, I wish to thank Khyber for his patience with Mommy and for sorting my paper clips.

# CHAPTER 1: INTRODUCTION

Studies examining differential item function (DIF) for gender have been conducted for items in a variety of sub-content areas of mathematics. DIF is the psychometric difference in the way a test item functions for two groups. Although gender DIF is often found in items testing proficiency in mathematics little is actually known about the sources of these differences. The fact that gender DIF research has failed to explain the sources of DIF consistently and accurately has direct implications for test developers whose goal is to ensure the internal and external validity of the test. The importance of this study lies in the fact that an increased understanding of sources of gender DIF will assist test developers to ensure that mathematics items measure the construct that they are intended to measure and that the test as a whole measures that which it purports to measure.

On a more local level, this study provides differential performance information specific to the items and subtopic areas of the Principal of Mathematics Exam 12 (PME12) for gender groups. Another important aspect of this study is the methodology. Limited studies to date have examined DIF from a multidimensional perspective. In addition, the confirmatory analysis approach has been used in only a handful of studies. As a result, DIF on individual items on specific tests has been examined but there has yet to be developed a bank of sources of gender DIF. The purpose of this study was to identify DIF patterns and confirm sources of DIF for boys and girls on a test of mathematics. In order to fulfil this purpose, a multidimensional model of DIF and confirmatory analysis were applied to the Principals of Mathematics Exam 12 (PME12) data. In contrast to the traditional approach to examining DIF, the confirmatory approach

requires that hypotheses be derived prior to statistical testing. In this study, hypotheses were formed via an examination of patterns of gender DIF in the research literature. In this study DIF was viewed as due to the multidimensional nature of the construct being assessed. This perspective is notable as one consequence of this perspective is that bundles of items, not simply individual items were tested. It is this multidimensional perspective along with the confirmatory approach that sets this study apart from the traditional gender DIF research.

DIF is closely related to validity, item and test bias, as well as the differential performance of examinee groups. The first part of the literature review discusses research that is related to the distinction between DIF, bias, and performance differences. Validity is also defined. The second part reviews literature on gender DIF in mathematics. Finally, DIF patterns observed in the existing gender DIF research in mathematics that may be used to generate hypotheses about the sources of DIF are presented.

## Statement of Purpose

This study is concerned with identifying sources of gender DIF in tests of mathematics. In this study, a confirmatory approach based on a multidimensional model (Douglas, Roussos, & Stout, 1996; Roussos and Stout, 1996; Shealy & Stout, 1993; Stout & Roussos, 1995) was used to identify sources of DIF for boys and girls on the British Columbia Principles of Mathematics Exam for grade 12 (PME12). Two goals were specified. The first goal was to identify some of the patterns that are present in the gender DIF research in mathematics via an examination of the gender DIF research literature. The second goal was to test the existence of the identified patterns in the British Columbia Ministry of Education's Principals of Mathematics Exam 12.

# CHAPTER 2: REVIEW OF THE LITERATURE

Test results are often the basis for decisions that affect students' educational futures. Typically, if two groups of examinees are matched on the knowledge and skills measured by a given test then those in the matched groups are expected to maintain a similar performance on individual test items. More specifically and in the context of this study, boys and girls, having similar knowledge about mathematics, should perform equally well on items testing mathematics. This is, however, not always the case.

DIF may be defined as " the *unexpected* difference among groups of examinees who are supposed to be comparable with respect to the attribute measured by the item and test on which it appears" (Dorans & Holland, 1993, p. 37). This unexpected difference affects the interpretations based on test results, that is,validity of interpretations. However, researchers do caution us not to interpret DIF necessarily as bias in favor of one group or another. When considering potential sources of DIF, it is helpful to keep in mind the advice of Zieky (1993) who stated, "With all of our work with the index of differential item functioning, as with any other statistic, we must constantly keep in mind the fact that numbers cannot make decisions for us [and] we must take great care not to allow some quantitative system to take the place of informed judgement" (pp. 346 – 347). Ultimately, it is the test developer who must make the decision as to whether an item is appropriate for a test and whether it is biased or not.

Bias may be defined as a source of invalidity. Essentially, it is that which keeps some examinees from demonstrating their knowledge or ability in a given context. It is now widely accepted that DIF does not necessarily mean bias and sources of DIF need to

be examined. Researchers recognize the distinction between DIF and item bias. Camilli (1993) notes the following distinction:

> An item is said to "function differently" for two or more groups if the probability of a correct answer to a test item is associated with group membership for examinees of comparable ability. Statistical indices of DIF are designed to identify such test items. If the degree of DIF is determined to be practically significant for an item and the DIF can be attributed plausibly to a feature of the item that is irrelevant to the test construct, then the presence of this item on the test biases the ability estimates of some individuals. This compound condition when satisfied indicates item bias. (p. 398)

DIF and bias are indeed not synonymous terms. An item statistically flagged with high DIF may not necessarily be an unfair question (Angoff, 1993; Linn, 1993; O'Neill & McPeek, 1993; Schmitt, Holland, & Dorans, 1993; Zieky, 1993). Linn (1993) clearly stated, " DIF does not prove bias. For that matter, the lack of DIF does not prove lack of bias" (p. 352). In order for bias to exist, someone or some group must be placed at an unfair disadvantage or advantage.

Dorans and Holland (1993) highlighted that it is also important to make a distinction between DIF and impact. Impact is defined as a difference in performance between two groups, where differences in item performance reflect differences in overall ability distributions. Dorans and Holland noted that in test and item data, "Individuals

differ with respect to the developed abilities measured by items and tests, and intact groups, such as those defined by ethnicity and gender, differ with respect to the distributions of developed ability among their members" (p. 36). Recall that DIF, in contrast to impact, is an *unexpected difference* between a focal and reference group who *should be* comparable with respect to the attribute being measured by the item and test. Dorans and Holland noted that in test and item data, individuals will differ with respect to the abilities measured by items and tests. Further, groups such as those defined by ethnicity and gender, differ with respect to the distributions of ability among their members.

Angoff (1993) suggested that it is useful to view group differences in performance (both item and test) as consisting, in theory, of two components:

1) the true difference between the groups and

2) an artificial difference brought about by the use of inappropriate and irrelevant content in items.

This second set of differences has been traditionally detected through a differential item functioning analysis. This study, however, utilizes a somewhat different approach to the detection of performance differences between boys and girls on tests of proficiency in mathematics. In this study an examination of the existing DIF literature is used to identify patterns of DIF and formulate hypotheses about sources of gender DIF. These hypotheses are then statistically tested. The other notable difference between this study and that which is traditionally used is that DIF, in this study, is viewed as multidimensional according to the model given by Shealy and Stout (1993). A more detailed exploration of these differences is given below.

## Comparison of Methodology Used in Examining DIF

The confirmatory multidimensional model of DIF (MMD) approach to detecting sources of differential functioning departs from that which is traditionally used for examining differential functioning at the item level. The traditional approach to DIF research and the confirmatory approach used in combination with the multidimensional model of DIF are both presented here for the purpose of comparison and contrast. It is important to recognize the similarities and differences in these two approaches since the first stage of the confirmatory MMD approach used in this study relies on results taken from gender DIF studies that used the traditional DIF approach.

The traditional approach to DIF research and the confirmatory approach used in this study[1] differ in some important ways. First, in this study, DIF was viewed as multidimensional in accordance with a multidimensional model of DIF (MMD) described by Shealy and Stout (1993). MMD is a framework for understanding how DIF occurs based on the assumption that multidimensionality produces DIF. Shealy and Stout's multidimensional model of DIF will therefore be presented. Second, while in the traditional approach items are tested then expert judgement is used to explain the findings, in the confirmatory approach, expert judgement, previous research, or a dimensionality analysis is used to develop DIF hypotheses prior to statistical testing. The role of expert judgement and previous research for both approaches will therefore be explored. Issues surrounding the matching criterion used in the traditional and confirmatory approaches will also be presented as they too differ. While total score is generally used as the matching criterion in the traditional DIF analysis, it will be seen that

[1] "In this study" is used since it is not necessary for the Multidimensional Model of DIF to be used in all confirmatory analyses.

with MMD the preferable matching criterion are the remaining items of a given subtest or the total of scores in a matching subtest. The final product obtained by these approaches also differs. With the traditional approach DIF items and possible sources of DIF may be obtained. However, with the confirmatory analysis used in conjunction with MMD, DIF items and confirmed sources of DIF are obtained.

## *Traditional DIF Approach*

In a traditional DIF statistical analysis, items are statistically tested for DIF one at a time then categorized according to the size and direction of associated DIF. Some of the commonly used statistical procedures for detecting DIF items are the Mantel Haenszel chi-square method, logistic regression, Linn-Harnish method, and SIBTEST. As noted earlier, there are a number of important considerations when conducting a traditional DIF analysis that may be contrasted with the considerations associated with a confirmatory analysis. Considerations in the matching criterion, symmetry, and use of expert judgement (judgmental review) for the traditional DIF analysis are presented here.

### *Matching Criterion*

The matching criterion has been the subject of great discussion in traditional DIF analysis procedures (Clauser, Mazor, & Hambleton, 1991; 1993). Generally, the test score is selected as the matching criterion. This is done for several reasons given by O'Neill and McPeek (1993), "First, the test is a reliable and valid measure administered under standard conditions. Second, a unifactor test measures the same ability as the items in it. Third, test scores are readily accessible for analysis" (p.256). It is evident that the matching criterion must be both valid and reliable, or as Angoff (1993) stated, "It is important that it be so perceived" (p.18). Zieky (1993) noted that no perfect matching

criterion exists. Angoff (1993), in consideration of the Mantel-Haenszel procedure, noted that from a statistical point of view, if the studied item is not included in the matching criterion, then the procedure will not be optimal.

While Zieky (1993) recognized that test scores are usually the only available practical matching criterion, he discussed the principle that before their performance on test items can be compared, DIF analyses require that examinees be matched in terms of relevant ability, knowledge, and skills. He noted, specifically and in concurrence with Holland and Thayer (1988), that test scores are obviously relevant to whatever knowledge, skill, and ability are being measured by the test. Further, their reliability and validity can be empirically determined, and they are obtained under standardized conditions for all examinees (Zieky, 1993). However, warned Linn (1993), studies of predictive bias require the assumption that the criterion is unbiased.

There is universal recognition of the circularity of the matching criterion question. Angoff (1993) elaborated:

> For if the criterion is itself biased to some degree, then the application of a DIF analysis will certainly be flawed; further, if bias is pervasive in the criterion, than any attempt to identify bias in its component items will inevitably fail. This is the weakness in a circular procedure, in which the "criterion of truth" is not only a test – and therefore already suspect in the minds of some - but is also the test that comprises the very items whose possible bias we seek to determine. On the other hand the circularity cannot be avoided; the test itself is the most relevant and reliable criterion

available to help answer the important questions related to the validity of the test and the inferences one is to draw from it. (p. 17)

Camilli (1993) suggested that the circularity problem could be solved by indicating a true zero point for a set of DIF indices conditioned on total score. Zieky (1993) noted another approach used by Educational Testing Service (ETS) who do a preliminary DIF analysis before an operational use of a set of items as a matching criterion whenever it is technically feasible. Items that have high DIF values are removed from the matching criterion before it is used in an operational DIF analysis. The matching criterion is then inspected to ensure that, before it is used operationally, it is still reasonable in terms of content and statistical aspects of the remaining items. A question of how to judge the matching criterion must then be considered. Linn (1993) proposed three bases for judging the matching criterion:

- *practicality for operational use*. Linn reasoned that since test scores are presently being used, they must, then, be practical.
- *relative superiority among available alternatives*. On this point also, Linn recognized test scores as the best available alternative for matching in most cases. He noted that in order to avoid DIF being associated with valid differences along different dimensions, uni-dimensionality should be added to the list of characteristics for the ideal matching criterion. The same may be said for use of test scores.

*Symmetry*

Researchers must decide whether to examine both negative and positive DIF in order to be fair to both the focal and reference groups, or whether we are going to recognize that DIF research came about as a result of the inequitable assessment of

minority groups and thus consider only negative DIF. Zieky (1993) stated, "The response to the question is based on the meaning that is assigned to the word bias. If bias is thought of as negatively affecting only minority group members or women, then the use of DIF statistics to identify items that are differentially difficult for only those groups is sensible. If, however, bias is seen as an invalid difference between groups then a symmetrical use of DIF is more appropriate" (p. 339).

*Judgmental Review*

Once items have been identified as DIF, in the traditional approach, a judgmental review is conducted where field experts attempt to identify items that may have been statistically flagged. These experts also attempt to explain potential sources of DIF for the statistically flagged moderate and high DIF items. Linn (1993) noted that DIF analyses alone are insufficient and that they need to be used in conjunction with judgmental reviews of test content. Schmitt, Holland, and Dorans (1993) noted that evaluation of items with extreme DIF may provide insight into factors possibly related to DIF. Careful examination of the items with extreme DIF must be undertaken by a variety of experts whose speculation and insight may be used to generate hypotheses. Patterns of gender DIF that may be identified from studies using the traditional approach requires confirmation as attempts at understanding the underlying causes of DIF using substantive analyses of items statistically identified as DI F have been less than successful (Roussos & Stout (1996).

*Confirmatory MMD Approach*

In reaction to the inability of content specialists to identify the sources of DIF consistently and accurately, a confirmatory approach to examining potential sources of

differential functioning of groups of items was suggested (Douglas, Roussos, & Stout, 1996; Roussos & Stout, 1996). Recall that DIF refers to a psychometric difference in the way a test item functions for two groups commonly referred to as the focal (F) and reference (R) groups. The focal group is generally the group of interest. The reference group consists of those with whom the focal group is being compared. It will be seen that in a confirmatory MMD approach items may be viewed as multidimensional and item bundles examined. Hence, in the confirmatory MMD approach to examining potential sources of DIF we refer not to differential item functioning, but differential bundle functioning (DBF). An item bundle may containing two or more items that seem to measure a common secondary dimension in addition to the target ability (Douglas, Roussos, & Stout, 1996). Unlike the traditional DIF approach, in the confirmatory MMD approach potential sources of DIF and DBF are identified a priori. Hypotheses about the size and direction of potential DIF and DBF sources are then formulated. No post hoc judgmental review takes place. If a review of items is completed by field experts it takes place a priori in order to generate the DIF hypotheses. These hypotheses are then tested to either confirm or disconfirm the sources of DIF.

*The Multidimensional Model of DIF (MMD)*

Shealy and Stout (1993) proposed a framework for understanding how DIF occurs based on the assumption that multidimensionality produces DIF. The main construct that the test is intended to measure is called the primary dimension and the addition of dimensions that may produce DIF are referred to as the secondary dimensions. When primary and secondary dimensions characterize item responses the

data are considered multidimensional. According to Shealy and Stout, the *informal* multidimensional model of DIF consists of two conditions (1993):

1) A given item measures a secondary dimension in addition to a primary dimension of the test that the item is intended to measure.
2) A secondary dimension either advantages or disadvantages the focal or reference groups.

When these conditions are presented together, the multidimensional model predicts that the item will manifest DIF.

Douglas, Roussos, and Stout (1996) noted three considerations when conducting the DIF analysis: the purpose of the test, the nature of the secondary dimension, and the examinees of interest. Therefore, judgement as to whether the secondary dimensions are considered benign (caused by an auxiliary dimension) or adverse (caused by a nuisance dimension) in a particular testing situation is key. Stringent requirements were recommended by Douglas, et al. for the assertion of adverse DIF:

1) The matching criterion must result in a construct-valid matching of the examinees on the construct intended to be measured by the test.

2) The secondary dimension of the DIF item must be a true nuisance dimension.

It is noted by Shealy and Stout (1993) that some interpretations in the literature of statistically identified DIF indicate an awareness of the distinction between the auxiliary and nuisance secondary dimensions. However, the informal multidimensional model for DIF, as usually applied in the literature, does not usually distinguish between the two.

Douglas, et al. (1996) suggested the use of dimensional bundles in the analysis of DIF. This "item bundle DIF" is referred to as differential bundle functioning (DBF). They noted that DBF is a natural extension of item DIF.

*The Matching Criterion*

Choosing the matching sub-test is one important decision to be made when completing the statistical confirmatory analysis (Douglas, et al., 1996). Although when testing individual items the most appropriate matching criterion noted was the total test score, more appropriate alternatives are suggested for item bundles. Linn (1993) noted that "to the extent that the content domain of the test is relatively heterogeneous and spans more than a single dimension, there is apt to be a need to use separate sub-scores as the matching criterion for different subsets of items" (p. 350). Linn urged researchers to recognize that DIF techniques, whichever procedure they are based on, the Mantel-Haenszel, the standardization, or an item response theory (IRT) procedure, all suffer from the same serious limitation. They can only detect items that show large differences in one direction or the other relative to the total set of items. It should be noted that if a test contains a combination of areas of knowledge or skill, examinees matched on the total test score may not be matched on each separate area contained in the test. In this study, the most appropriate alternative is the total score of the remaining items in a given subset.

*Symmetry*

In confirmatory analysis the problem of symmetry found in the traditional approach is not evident. Since sources of DIF are of concern, it is imperative to examine DIF in general. While hypotheses state the direction of DIF, these hypotheses are either

confirmed or disconfirmed regardless of whether DIF is identified for the focal versus the reference group.

## Overview of a Confirmatory MMD Analysis

Harris and Carlton (1993) noted that when overall level of mathematics achievement is controlled by matching male and female students on total test score there remain both interest in, and uncertainty about, the various aspects or characteristics of mathematics test items that may be related to male/female differences in performance. Roussos and Stout (1996) suggested that the first step toward understanding these differences is through the specification of the structure of DIF in which items with a common source, or dimension, are placed in a bundle and hypotheses about the potential sources of DBF are determined. These hypotheses would then be tested using statistical procedures. This process has been referred to as confirmatory DIF and/or confirmatory DBF analysis (Douglas, et al., 1996; Roussos & Stout, 1996; Shealy & Stout, 1993; Stout & Roussos, 1995).

Roussos and Stout (1996) recognized that substantive and statistical analyses must be used together such that substantive observations are statistically tested. The first step in a confirmatory analysis is specification of the structure of DIF in which differential functioning hypotheses are generated. Recall that Stout and Roussos (1995) noted that the general cause of DIF is that multidimensionality in items that display DIF are measuring at least one dimension in addition to the primary dimension.

## Previous Use of Confirmatory Analysis

The confirmatory analysis approach to the identification of the sources of differential functioning has been used previously to examine potential sources of gender

DIF in mathematics. Ryan and Fan (1996) completed a confirmatory MMD analysis using SIBTEST. They recognized that girls found problems in some sub-content areas to be differentially more difficult and highlighted the need to examine sub-content areas. Ryan and Fan noted that the notion of DFB is useful for the study of differential items functioning and that the approach added power for detecting patterns of differences. More recently confirmatory analysis has been used to examine sources of translation DIF (Gierl & Khaliq, 2001). Gierl and Khaliq recently addressed the issue of fairness in testing and illustrated the utility of DBF analyses to identify and interpret group differences on achievement tests. They indicated that commonly used statistical methods that are based on a unidimensional concept of ability might lead to erroneous conclusions about DIF.

## Defining Dimensions

Recall that Roussos and Stout (1996) recommended that dimensions, as opposed to items, be analyzed for differential functioning through the analysis of item bundles. Item bundles are groups of items selected so that they measure not only the same primary dimension but also the same secondary dimension. Recall that in the first step of the Roussos–Stout confirmatory approach hypotheses are generated and the structure of potential DIF is determined. The hypothesis specifies whether an item or bundle of two or more items designed to measure the primary dimension also measures an unexpected secondary dimension for examinees in either the reference or focal group.

Four methods that can help test developers specify the structure of DIF are proposed by Roussos and Stout (1996). The four methods are:

1) Use previously published DIF analyses. Primary and secondary dimensions are often outlined by authors who provide interpretations of their DIF results.
2) Use substantive content considerations and judgements to generate DIF hypotheses. Test developers are usually content specialists who can use their substantive knowledge of content area to identify primary and secondary dimensions. They can also help to predict whether one group of examinees will outperform another group of examinees due to a particular examinee group's proficiency on specified dimensions.
3) Analyze archival test data when specific secondary dimensions are prominent. Roussos and Stout (1996) noted that many standardized tests require examinees to solve items in a meaningful context; however, despite the benefits, this might also pose a risk for test equity and fairness if the focal and reference groups are differentially familiar with test contexts. DBF hypotheses can be specified, however, by identifying these contexts and the possible secondary dimensions that are produced by them through the analysis of archival data.
4) Test bundles of items according to some organized principle. Testing the organized bundles rather than individual items can lead to dimensionality-based DIF hypotheses when bundles reflect a secondary dimension (Douglas, et al, 1996).

## Gender DIF in Mathematics

Controversy surrounding potential gender bias in mathematics is widespread. Childs (1990) noted the legal case of Sharif versus the New York State Education

Department, in which the plaintiffs charged that by using SAT scores as the sole basis for the award of state merit scholarships, the New York State Education Department was discriminating against girls. It was noted in this case that girls tended to have higher high school grades than the boys with whom they were competing for scholarships. However, since girls also tended to have lower scores on the SAT, they received fewer scholarships.

Examination of differential item functioning for boys and girls stems from a historical recognition of the existence of differential achievement for boys and girls in mathematics. The relationship existing between gender and differential achievement in mathematics has been under examination for many years. Fennema, Carpenter, Jacobs, Franke, and Levi (1998) noted that a powerful scientific discourse involving scholars from many traditions has centered on gender and mathematics. This discourse has had, and is having, an impact beyond the mathematics education community as our understanding of gender and mathematics grows. As early as 1958, differential response patterns between males and females were recognized by Anastasi (1958) in her classical differential psychology test. Research findings since then have been plagued with inconsistencies; however, a few patterns have emerged.

Anastasi (1958) recognized that although differences in numerical aptitude favored boys, the discrepancy in scores did not appear until well into the elementary years. She further noted that if gender differences in computation did appear, they favored girls, whereas boys excelled on tests of reasoning. Research appearing in the 1970s supported these early findings, indicating also that gender differences in mathematics test performance were not consistently observed until early adolescence (Fennema, 1974; Fennema & Sherman, 1977). However, a more recent study examining

academically talented students in grades 2-6 indicated that gender differences for mathematical ability may appear as early as grade 2 (Mills, Ablard, & Stumpf, 1993).

Certainly, many factors affect performance by boys and girls on mathematics items. Haag (2000), in an examination of K-12 Single-Sex Education, noted that girls in single-sex schools may have stronger preferences for subjects such as mathematics and physics than girls educated in schools where boys and girls are mixed. Haag cited Lee and Lockheed's (1990) study of 1,012 students in ninth-grade Nigerian public schools using data drawn from the Second Mathematics and Science Assessment conducted by the International Association for the Evaluation of Educational Achievement (IEA). Mathematics achievement was measured along with stereotypic views of mathematics. Lee and Lockheed found no significant gender gap between mathematics scores of Nigerian boys and girls, once other variables were taken into account. Girls in single-sex schools, however, outperformed other girls in mathematics. In a study by Mallam (1993), it was found that girls in single sex Nigerian schools favored mathematics more than girls in coeducational Nigerian public boarding schools, particularly when mathematics was taught by a female teacher.

## Differential Performance of Boys and Girls in Mathematics

Ascertaining the causes of DIF is of utmost importance (Camilli,1993; O'Neill & McPeek,1993). O'Neill and McPeek (1993) recognized that the major focus of future DIF research should be on trying to uncover testable, verifiable, robust explanations for why DIF occurs when it does. Further, O'Neill and McPeek recognized that a high DIF value is not proof that an item is unfair, and an important consideration is that one must be careful in any interpretation of DIF results. Yet, Roussos and Stout (1996) noted that the

DIF literature makes it evident that attempts at understanding the underlying sources of DIF using substantive analyses of statistically identified DIF items, in general, have been less than successful.

Swaminathan and Rogers (1990) recognized that the detection of DIF in achievement, licensure, and credentialing examinations has become an important issue. The importance of examinations of achievement in mathematics is highlighted here since the presence of bias in this area may prevent girls from entering mathematically oriented professions or pursuing further education in mathematics and related fields. Defining the sources of potential DIF will help to reduce gender bias in tests of mathematics. A challenge in the identification of these sources is the inconsistency of findings, as recognized by Hyde and Jaffee (1998), in gender DIF research.

*Inconsistencies in Research Findings*

In 1974, Maccoby and Jacklin published a compendium of and commentary on studies of sex differences entitled The Psychology of Sex Differences. Shortly after the Maccoby and Jacklin publication, Fennema (1976) conducted a review of the research literature associated with mathematics and noted a possible decline in gender difference from years prior to 1976. Fennema recognized that conclusions regarding male superiority tended to be reached in older studies or from investigations not controlling for the number of mathematics courses taken. A meta-analysis of sex differences in mathematical tasks conducted by Friedman (1989) found only a minute mean effect size ($d$ = -0.024) of the random effects model. Hyde, Fennema, and Lamon (1990) also conducted a meta-analysis involving 100 studies in which they concluded that general statements about gender differences are not warranted since complexities become

obscured. They found that the gender difference was small and that since 1974, the time of Maccoby and Jacklin's publication, the magnitude of the difference had declined. Furthermore, no differences were noted on mathematics concepts regardless of the age level while female superiority in computation declined to zero.

In an examination of sex differences in a causal model of mathematics achievement, Ethington and Wolfle (1984) noted that although studies investigated the reasons for sex differences in mathematics achievement, they failed to be consistent in their conclusions. Studies served to identify a variety of variables that, in one study or another, were shown to have had a significant relationship with mathematics achievement. Gallagher and De Lisi (1994) examined gender differences in mathematics problem solving among high-ability students. They recognized that research examining gender differences in test performance, given specific mathematical content, has found that differences tend to appear on certain problems rather than across all problems and that strategies used to solve problems differed.

*Consistent Patterns in Gender DIF Research*

There has been some notable consistency in the gender DIF research that would be helpful in specifying sources of DIF. Three specific areas that have been consistently examined with regards to achievement in mathematics and gender are: intellectual ability, as defined by high academic performance and precocity (Benbow & Stanley, 1982; Flexer, 1984); grade level, which is generally but not always associated with age; and mathematical content. Although intellectual ability was not a consideration in this analysis, it is recognized that since this exam was designed for students intending to pursue career and further education in mathematics, some prior academic success may be

assumed. As such, studies including high ability samples were not excluded from the literature review and the subsequent substantive analysis. The second variable, grade level, is also a constant in this study. Students taking the exam are all seeking grade 12 graduation. The remaining area, content, will be examined further.

Fennema and Tartre (1985) noted that differences in scores on problem type point to the existence of reasonably consistent sex differences, not always in favor of the boys, on three major groups of problems:

1) measurement and proportion;
2) problems with an obvious spatial component; and
3) abstract problems.

Although labels differ somewhat, the Principles of Mathematics 12 exam (PME12) utilized, generally, these categories. This examination of gender DIF studies deals with four specific areas consistently highlighted in the gender DIF literature and utilized by the PME12: problem solving, patterns and relations, shape and space, and cognitive level tapped by an item. It has also been noted that the context in which an item is presented may contribute to gender DIF (Linn & Hyde, 1989).

*Problem Solving*

It was noted by Linn and Hyde (1989) that the largest gender difference found in quantitative ability was for high school students on problem solving items. However, they also noted that gender differences in problem solving were similar in magnitude to the gender differences in enrollment in advanced courses that emphasized solving word problems. Linn and Hyde used meta-analysis to synthesize results from studies of gender differences in mathematics and science. They found that age trends indicated that

differences favoring boys on problem solving occur in high school. Swafford (1980) examined sex differences in first-year algebra, specifically whether boys and girls with comparable backgrounds in mathematics would have comparable achievement patterns in algebra. Among other things, Swafford was interested in applied problem solving skills that occur during the first year of algebra. He found, in general, that the boys performed better than the girls on the application oriented (consumer) problem solving items. Hyde, Fennema, and Lamon (1990) also found that boys did slightly better than girls on problem solving items when age was not a consideration. An examination of age trend and role of cognitive level indicated that for problem solving items, girls tended to do better in the early years (elementary and middle school) but in high school, boys did better. Berberoglu (1995) completed a DIF analysis of computation, word problem, and geometry questions across gender and SES. In contrast to the studies previously presented, he found that word problem items had the lowest ratio among the groups that indicated DIF against boys and that girls appeared to have an advantage in solving word problem questions.

*Patterns and Relations (Algebra, Arithmetic, and Computation)*

There exists little consistency in research findings reporting DIF in algebra items. In an examination of high ability boys and girls, Flexer (1984) found that despite the absence of a difference in scores by sex on an algebra achievement test, the girls' algebra grades exceeded those of the boys. This relationship, it was noted, remained significant even after the effects of the cognitive predictors were partialled out. On algebra items from data collected by the Second International Mathematics Study (SIMS), Hanna (1986) found the rate of wrong responses was higher for boys than for girls. However, on

SIMS arithmetic and algebra post-test items, Hanna found no statistical difference for correct responses. Harris and Carlton (1993) found, in general, that boys performed relatively better than girls when the primary content of the item was arithmetic as opposed to algebra. Using meta-analysis, Hyde, et al. (1990) found that for arithmetic and algebra items there were no identifiable gender differences when computation was not a consideration.

Inclusion of computation in items was addressed in the research literature although consistency in definitions of computation was shaky at best. In the literature, computation generally referred to the computation associated with solving arithmetic and algebra problems. Of the studies that examined computation in reference to arithmetic and algebra, it was found that DIF results of items tapping this area was mixed, in some research favoring boys (Hanna, 1986; Harris & Carlton, 1993) and in others favoring girls (Flexer, 1984; Hyde, et al., 1990). One possible explanation for this discrepancy might be that arithmetic items tap a lower cognitive level than algebra. When these items were categorized together as computation, the results may have been influenced by what would be a nuisance dimension, in this case cognitive level tapped by the item. When computation was studied in general or compared to problem solving the results were, once again, mixed. For example, Berberoglu's study (1995) found that computation items favored boys, while the results of a meta-analysis conducted by Hyde, et al. (1990) concluded that computation items favored girls.

*Shape and Space (Geometry)*

The greatest consistency in the literature was found in the definition of shape and space. Shape and space refers to geometry and measurement items, sometimes in the

context of spatial ability or spatial visualization (Benbow & Stanley, 1982; Berberoglu, 1995; Doolittle & Cleary, 1987; Hanna,1985; Harris & Carlton,1993; Linn & Hyde,1989; Maccoby & Jacklin,1974; Pattison & Grieve, 1984). Shape and space items on the Principles of Mathematics 12 exam included two subtopic areas; trigonometry and geometry.

In general, it was found that boys tend to do better than girls on shape and space type items, particularly those dealing specifically with geometry. With the exception of Berberoglu (1995) who offered the only contrary findings, it was found that boys did better than girls on shape and space type items.

Berberoglu (1995) noted that in the Mathematics subtest of the University Entrance Examinations in Turkey, girls appeared to have the advantage on geometry items. On geometry and measurement items from SIMS data, Hanna (1985) found that boys did significantly better than girls. Harris and Carlton (1993) found, in general, that boys performed relatively better than girls when the primary content of the item was geometry. In an examination of gender based differential item performance in mathematics achievement, Doolittle and Cleary (1987) also found that girls did more poorly than boys on geometry items. They conjectured that since many of the geometry items in their study contained diagrams, it might support the hypothesis that boys, as a group, tend to develop certain spatial skills to a greater degree than girls, as suggested by Benbow and Stanley (1982) and Maccoby and Jacklin (1974). Looking specifically at whether spatial skills contribute to sex differences on different types of mathematics problems, Pattison and Grieve (1984) recognized that while girls tended to perform better on logical, relatively abstract problems, boys tended to perform better on items dealing

with proportion, scale, and two and three dimensional problems. Pattison and Grieve further noted that sex differences in achievement and participation rates in mathematics are commonly attributed to sex differences in spatial ability. One reason advanced for this hypothesis is that there is a logical connection between mathematical and spatial skills in that that the two skills emerge at about the same time (Benbow & Stanley, 1980; Fennema, 1974). Linn and Hyde (1989) noted that gender differences in spatial ability were declining.

*Cognitive Level*

Cognitive level being tapped by an item has been found to be a factor in gender performance on mathematical items. Linn and Hyde (1989) revealed that gender differences arise among high school students on mathematics items dealing with complex applications, sports, or science. Fennema and Tartre (1985) noted that sex-related differences in mathematics favoring boys have been found on higher cognitive level tasks by age 17, citing the work of Armstrong, (1981), Carpenter, Linquist, Matthews, and Silver (1983), and Fennema (1984). Harris and Carlton (1993) coded items according to cognitive complexity and found that there was a consistent and significant shift in relative performance of boys and girls as cognitive complexity increased. Girls were found to do better on items requiring lower level mental processing while boys did better on items requiring higher level mental processing. Kaplan and Plake (1982) investigated the relationship between achievement in mathematics and level of cognitive development for college age males and females. Performance differences were found in mathematics on level of cognitive development for the students in the Kaplan and Plake study. Specifically, differences were found on items identified as abstractions, proportions,

probability, and isolation of variables. Although girls were not found to be different from the boys in terms of level of cognitive development, the girls scored significantly lower than the boys in performance in mathematics. Hyde, Fennema, and Lamon (1990) examined the magnitude of gender differences as a function of the cognitive level of the mathematics test items through meta-analysis. They found the strongest between group effects for cognitive level although no general superiority of boys or girls was reported.

## Summary

Differential Item Functioning (DIF) that is defined as a psychometric difference in the way a test item functions for two groups is used to identify potentially biased test items. A distinction needs to be made between DIF, bias, and impact.

For both the traditional approach to DIF detection and the confirmatory MMD approach, selection of the matching criterion is of great consideration. However, with the confirmatory MMD approach the matching criterion is dictated by the items present in a bundle and the matching subtest. While symmetry is a consideration in the traditional DIF approach, when confirming sources of DIF, symmetry is required. Use of expert opinion also differs in the two approaches. While expert opinion may be used in some forms of substantive analyses to specify the structure of DIF, it is not a requirement since the structure of DIF may also be specified through an examination of the DIF literature for patterns or via a statistical dimensionality analysis. The Multidimensional Model of DIF was presented along with previous DIF studies using this model.

Gender DIF has been of great concern for many years. Although some patterns have emerged regarding the response patterns of boys and girls on mathematical items, the literature is broad and hypotheses regarding observations need to be confirmed

statistically. Consistency in findings exists for problem solving, algebra, geometry, and cognitive level of an item. In general boys were found to do better than girls on problem solving items, geometry, and higher cognitive level items. An examination of the results from DIF studies focussed on algebra/patterns and relations revealed only that findings are inconsistent.

# CHAPTER 3: METHOD

This chapter outlines the method used to confirm sources of gender DIF in mathematics. The source and characteristics of the data will be described along with a description of the two stages of the confirmatory analysis: a substantive examination of previous DIF studies used to specify DIF hypotheses and the consequential statistical confirmation of the hypotheses.

This study will use confirmatory MMD analysis based on a procedure described by Roussos and Stout (1996), and Douglas, Roussos and Stout (1996) to investigate and identify pattern differences in gender responses to multiple choice items on the British Columbia Principles of Mathematics 12 Exam (PME12). Further, this research proposes to identify sources of DIF via an examination of item bundles, thus contributing to a body of confirmed DIF hypotheses that may be used to develop guidelines and test construction principles for reducing gender DIF. As a control for Type I error and in recognition of item dimensionality, the remaining items in the subsets are used as the matching criterion (Douglas, et al, 1996).

## Data

Data from the British Columbia Provincial Principals of Mathematics Examination for grade 12 (PME12) were used to test the existence of the identified patterns of gender DIF specified through an examination of previous gender DIF studies. The purpose of the British Columbia Provincial Examination Program as indicated by the BC Ministry of Education (2001), is to establish performance standards at the provincial level, promote accountability through data obtained from the provincial examinations and

school marks, and ensure equity to students who plan to enter into post-secondary studies. Further, it is indicated that results from the British Columbia Provincial Examinations will enable schools, school districts and the ministry to make informed decisions.

The data consisted of the year 2000, June administration of the PME12 that included 14,309 students. The Principles of Mathematics 12 exam was designed for students who wish to pursue a career in mathematics or a related field such as engineering, business, or economics. There were 45 multiple choice questions worth 66% of the examination. The remaining 34% correspond to written-response items.

## Instrument

The year 2000 PME12 was constructed such that items tapped three cognitive levels based on a modified version of Bloom's (1956) Taxonomy of Educational Objectives. Knowledge items checked student recall of terminology, specific facts, conventions, classifications, and notation. Two levels of understanding and application were used. The middle level items required students to recognize a correct answer after performing a single computation, substituting values into a formula, or identifying specific characteristics of a graph or diagram. The higher cognitive level items, however, required students to form and solve equations, manipulate expressions, produce a graph or diagram, or interpret a graph or diagram. Items tapping the highest mental processes included items that required the student to analyze, synthesize, and evaluate.

The year 2000 administration of the PME12 contained six problem solving items including problem sets and cross-topic problems. Items identified as patterns and relations included sequences and series, polynomial functions, logarithms and exponents,

quadratic relations, and quadratic functions sub-topics. There were 28 patterns and relations items. Included in the shape and space category were trigonometry and geometry items. There were 14 shape and space items.

Procedure: Confirmatory Analysis

Roussos and Stout (1996) proposed a confirmatory analysis in which substantive and statistical differential bundle functioning analyses were combined. In the first step the structure of potential DIF is specified. In this study, the structure of DIF was specified via an examination of previous gender DIF research. Inconsistencies and patterns were noted. Hypotheses were then generated. This was followed by a statistical analysis to confirm or disconfirm the hypothesized sources of DIF. This was followed by a DIF analysis and exploratory analyses.

Roussos and Stout (1996) recognized the possibility of developing DIF free items through the development of hypotheses based on the nature of the underlying dimensionality structure of the test. Recall that this approach was based on the multidimensional conceptualization of DIF in which each potentially DIF item is assumed to be influenced by, at most, two dimensions (primary and secondary). Shealy and Stout (1993) noted that DIF is due to the presence of two factors:

1. The item is sensitive to not only the construct $\theta$ that the item is intended to measure, but also to a secondary construct $\eta$, that may be multidimensional.

2. A difference exists between the two demographic groups of interest in their conditional distributions on $\eta$ given a fixed value of $\theta$, written $\eta|\theta$.

Two functions are essential to the multidimensional model of DIF:

1) the function relating the probability of a correct response to proficiencies $\theta$, and $\eta$, denoted by $P(\theta, \eta)$. The function $P(\theta, \eta)$, does not depend on group membership because once all of the relevant examinee proficiency dimensions for answering the item are known, group membership is, by definition of the completeness of the latent proficiency space, no longer a predictor of performance on the item.
2) the conditional probability density function of the proficiency $\eta$ for examinees with fixed proficiency $\theta$ on the primary dimension, denoted by $f_G(\eta|\theta)$ (where subscript G is a sub-population of the entire exam taking population. G may be either the focal or reference group).

If both of these factors are present, then DIF will occur; otherwise no DIF will occur. In other words, an item will only exhibit DIF if the focal and reference groups that have been equated on a primary dimension differ in distribution on a secondary dimension. Further, the item response function for the item must be sensitive to the secondary dimension. If the conditional distributions on the secondary dimension do not differ on the secondary dimension, even if an item is sensitive to the secondary dimension, no DIF will occur.

*Specification of the Structure of DIF*

Previous gender DIF studies were used to investigate and specify the structure of DIF as proposed by Roussos and Stout (1996). Douglas, et al. (1996) noted that naturally occuring testlets, that may be referred to as bundles, are obvious candidates for hypotheses of collective DIF amplification. The PME12 provided the base structure of

the content bundles. Definitions used in previous studies that were congruent with those used in the PME 12 were used to determine the inclusion of items in the content and cognitive level bundles. The structure of the context bundles however, was determined through an examination of patterns found in the DIF literature. Specifically, common gender DIF context findings and the definitions used to describe context were used to identify categories of context specific gender DIF. Items in the PME12 were then organized into bundles according to these definitions.

*Summary of Hypotheses*

Hypotheses drawn from the substantive review are summarized here. It is hypothesized that patterns identified in the literature will be confirmed through an analysis of dimensional bundles drawn from the PME12, specifically:

*Content*

1. Items and bundles dealing with cross-topic problem solving will exhibit moderate to high level differential functioning in favor of the boys.

Bundles dealing with algebra subtopics identified as patterns and relations will exhibit differential functioning in the following manner:

2. Items and bundles dealing with sequences and series will exhibit moderate differential functioning. No particular direction of differential functioning is hypothesized.
3. Items and bundles dealing with polynomial functions will not exhibit significant differential functioning.
4. Items and bundles dealing with quadratic systems will not exhibit significant differential functioning.

5. Items and bundles dealing with quadratic relations will not exhibit significant differential functioning.
6. Items and bundles dealing with exponential and logarithmic functions will not exhibit significant differential functioning.
7. Items and bundles dealing with geometry will exhibit high level differential functioning in favor of the boys.

*Cognitive Level Tapped by the Item*

1. Items and bundles tapping the highest cognitive level will exhibit moderate to high level differential functioning in favor of the boys.
2. Items and bundles dealing with knowledge only, lower cognitive level items will exhibit moderate to high level differential functioning favoring the girls.

*Context*

1. Items and bundles that include story (word problems) will exhibit high level differential functioning favoring the boys.
2. Items and bundles presented with a visual included in the question will exhibit high level differential functioning favoring the boys.

## Exploratory Analyses

The following exploratory analyses were also completed:

- Items and bundles dealing with trigonometry will be tested. There exist insufficient data for trigonometry items tested in isolation for a reasonable directional hypothesis to be formed.
- Although research on computation items appears in the literature (Berberoglu, 1995; Flexer, 1984; Hyde, Fennema, & Lamon, 1990) the use of equations in the items is

not examined. In this study a distinction is made between the items in which an equation is clearly provided and those for which an equation is given in part, or not at all. Since the presence of an equation may be a visual cue and provides an additional structure from which to correctly respond to the item, it is hypothesized that items and bundles containing and not fully containing an equation in the question may each display differential functioning. Since the level and direction of this hypothesized DIF may not be synthesized from the existing literature an exploratory analysis was completed.

## Statistical Analysis

The second step in the Roussos–Stout confirmatory MMD approach is the statistical testing of the hypothesis for the purpose of confirming the hypothesized sources of differential functioning. The simultaneous item bias statistic (SIB) and bias estimator $\beta$ was calculated by the computer program SIBTEST (Stout & Roussos, 1995). PARDUX was used to test items for DIF. PARDUX is an item response theory (IRT) based scaling and calibration program that also provides DIF analyses based on a procedure given by Linn and Harnish (1981).

### *SIBTEST*

SIBTEST is a computer program developed by Shealy and Stout (1993). Its development was based on a multidimensional item response theory model of DIF and is organized such that bundles or items may be tested for DBF or DIF respectively. With simultaneous item bias, $\beta$ is interpreted as the difference in the expected total score between the focal and reference groups. The SIB statistic $\beta_u$ is computed by first dividing the N-items into an n-item valid subtest and a set of N-n suspect items. The valid

items may be identified as the items that weight most highly on a particular factor. The remaining items are classified as suspect items and can be tested one at a time or collectively. Once the test is split into two categories, the total score on the suspect items (Y) and the valid subtest score (X) respectively given as:

$$Y = \sum_{i=n+1}^{N} U_i \qquad \text{and} \qquad X = \sum_{i=1}^{n} U_i$$

are computed from the dichotomous item scores, $U_i$. The statistics $\bar{Y}_{Rh}$ and $\bar{Y}_{Fh}$ represent the average Y for all examinees that obtain a valid subtest score X=h (h=0,1,2…,n). These statistics are then calculated for both the reference and focal groups. Generally, given the null hypothesis of no DIF, a two-tailed hypothesis test, SIB-p ($z=1.98$, $p<0.05$) is conducted to check for uniform DIF in either group of interest. When a priori hypotheses about the direction of DIF are made then the option to use a one-tailed test is viable. Shealy and Stout (1993) provide a model based parameter measuring the amount of unidirectional DIF present:

$$\hat{\beta}_U = \sum_{h=0}^{n} \hat{p}_h (\bar{Y}^*_{Rh} - \bar{Y}^*_{Fh}),$$

where $\hat{\beta}_U$ is an estimate of $B_U$ and $\bar{Y}_{Rh}$ and $\bar{Y}_{Rh}$ values are used to obtain an adjusted true score theory base regression correction.

Much of the applied DIF literature uses the classification system set by ETS whereby items are put into one of three categories, "A", "B", or "C". A level items may be considered to be free of DIF. B level DIF indicates moderate DIF and C level items are those items with the most severe or large DIF (Zwick & Ercikan, 1989).

Linn (1993) stated that he would prefer to see a graphical presentation of the DIF results but noted a concurrence with Zieky (1993) that information overload is a real concern. Further, Linn concurred with the use of A, B, C category labels with an indication of direction, noting that it is a reasonable attempt to get the critical information to test developers without overwhelming them.

Douglas, et al. (1996) used a $p$ value of 0.05 as the cut off in determining DIF while Gierl, Bisanz, Boughton, and Khaliq (2001) used an adaptation of the ETS item classification guidelines for single items with DIF classified as A, negligible DIF when $|\beta| < 0.059$. B level, or moderate DIF, is indicated when $0.059 \leq |\beta| < 0.088$. C level DIF, or severe DIF, is indicated when $|\beta| \geq 0.088$. This study therefore takes both the p value and $|\beta|$ thus adding the following $p$ value classification for B and C level DIF. B level DIF is additionally indicated in this study when $|\beta| \leq 0.059$ and $p < 0.05$ and C level DIF is indicated when $0.059 \leq |\beta| < 0.088$ and $p < 0.05$. The addition of the p value consideration allows items to be classified in a manner similar to that used in PARDUX.

*PARDUX*

For DIF detection, PARDUX utilizes the Linn-Harnish method (Linn & Harnish, 1981). Using this method, estimates of the three parameter logistic model given the item parameters: $a$, item discrimination, $b$, the item difficulty, and $c$, the lower asymptote were obtained based upon all available cases in the sample. This provides an estimate of each person's location along the latent trait. $P_{ij}$, the estimated probability that a person $j$ would answer item $i$ correctly, were obtained from these estimates using the three parameter logistic model. The $a$, $b$, and, $c$ estimates are based on the model averaged both over

members of a particular target group and for members of subgroups of that target group defined by regions on the theta scale. The estimates are then compared to the observed proportion correct for the target group and for each of the subgroups. Linn and Harnish noted that if person j is a member of a subgroup g then the proportion of people in subgroup g expected to get item i correct according to the model is:

$$P_{ij} = \frac{1}{n_g} \sum_{j \in g} P_{ij}$$

such that $n_g$ is the number of persons in subgroup g. The proportion of people in the complete target group, $P_i$, would then be:

$$P_i = \sum_g n_g P_{ig} / \sum_g n_g$$

The observed proportion correct on item i for subgroup g, $O_{ig}$, is, then, the number of people in subgroup g who answer item i correctly divided by $n_g$. For the complete focal group:

$$O_i = \sum_g n_g O_{ig} / \sum_g n_g \ .$$

The difference, given as $D_i = O_i - P_i$, is an index of the degree to which members of a given group perform better or worse than expected on a given item. In the Linn-Harnish (1981) method, differences in subgroups on the theta scale provide an index for each region on the theta scale. The Linn-Harnish procedure computes the observed and expected mean responses and the observed minus predicted (difference between observed and expected) for each item by deciles of: a) the specified subgroup; b) the subgroup as a whole. Parameter estimates obtained from the entire sample are used to compute the expected values and ability estimates ($\theta$) are computed for members of the specified subgroup. The direction of DIF is indicated by the sign associated with the magnitude

(obtained minus expected mean) of difference. Three levels of DIF indices are used. Level 1 DIF indicates negligible DIF. This includes items with $|Z| < 2.58$. Level 2 DIF indicates moderate DIF. This includes items with $|Z| > 2.58$ and the expected mean difference is $>0.10$. Level 3 DIF indicated high level, or severe DIF. This includes items with $|Z| \geq 2.58$ and the expected mean difference is $>0.10$.

A distractor analysis was also used for items or bundles flagged with DIF or DFB. Schmitt, Holland and Dorans (1993) noted that differential distractor information could assist in gaining additional insight into the causes of DIF. Further, researchers may gain a greater understanding of the respondents' cognitive processes with additional knowledge about which distractors differentially attracted a specific subgroup.

# CHAPTER 4: RESULTS

This chapter summarizes results of four sets of analyses:

(1) Analysis of test results using descriptive statistics separately for boys and girls,

(2) An analysis of gender DIF research literature leading to a summary of gender DIF patterns and the subsequent definition of dimensions used in the creation of hypotheses,

(3) Confirmatory DIF analyses for testing the hypotheses, and

(4) Exploratory DIF analyses.

## Descriptive Statistics

There were 45 multiple choice items in the PME12. There were 4335 girls in the focal group and 5069 boys in the reference group for a total of 9404 examinees were included in the analyses. In order to avoid language effects, only examinees indicating English as their first language were included in the statistical analyses. The mean score for the boys was 30.20 (SD= 7.47) and 30.07 (SD=7.12) for girls. Reliability was calculated using KR-20. The coefficient for the boys was KR-20=0.8647and for the girls KR-20=0.8537. See Table 1 for a summary of the descriptive statistics.

Table 1

Summary of the Descriptive Statistics

| | Mean | SD | Reliability |
|---|---|---|---|
| Boys N=5069 | 30.20 | 7.47 | 0.87 |
| Girls N=4335 | 30.07 | 7.12 | 0.85 |

## Defining the Dimensions

The dimensions were defined according to a procedure suggested by Roussos and Stout (1996). Roussos and Stout recommended that primary and secondary dimensions identified in previous DIF research may be used to define dimensions to be studied. Recall that the primary dimension is the intended construct to be measured by a set of items. The secondary dimension is the additional, possibly DIF inducing, dimension. The secondary dimension is potentially composed of two parts: 1) a benign auxiliary dimension, and 2) an adverse nuisance dimension. Multidimensionality in the test structure that is not due to gender differences is accounted for by the auxiliary dimension. The differential existence of the nuisance dimension for the gender groups is the focus of the hypothesis testing.

Roussos and Stout (1996) further recommended that bundles be tested according to some organized principal. The existing organization of the PME12 served to provide structure to the examination of the DIF literature. Specifically, the gender DIF literature was examined using the structure provided by the PME12. The PME12 was organized according to content and cognitive level areas. The examination of the gender DIF literature therefore was focused on the specific content and cognitive level areas identified in the PME12. The content areas were: problem solving, patterns and relations, and shape and space. Within these content areas were more specific subtopics: problem sets, polynomial, quadratic relations, quadratic functions, logarithms and exponents, sequence and series, geometry, and trigonometry. These more specific content areas were used to define the bundles since each represented a very specific construct being

measured. Three cognitive levels were tapped by the items: level 1–knowledge, level 2–understanding and application, and level 3–a higher level of understanding and application. The cognitive levels defined the bundles. Since cognitive level bundles were drawn from a general subtopic the matching criterion was the total score of the remaining items from the total test.

The gender DIF literature was examined and studies with subtopics that were similar to those in PME12 were identified. The level of DIF, the direction of DIF, and other potentially relevant information (such as the additional examination of context or specified purpose for the study that may suggest the need for an exploratory analysis) were used in the definition of the subtopics. The DIF literature provided enough relevant information to formulate hypotheses for all of the subtopics except trigonometry. The information relevant to trigonometry was generally too closely tied to geometry or tested within a context that may have influenced the interpretation. Therefore, an exploratory, rather than confirmatory, analysis was conducted for trigonometry. The formation of the hypotheses are presented in detail here. A summary of the general subtopics and the specific subtopics used to identify bundles for the confirmatory analysis is provided in Table 2.

1) Content areas tested, cognitive level tapped by the item, and the definitions used in the PME12 were compared to those used in the DIF literature. Studies with content area and cognitive level definitions congruent to those found in the PME12 were used to formulate a hypothesis about the level and the direction of DIF.

As dictated by the PME12, content bundles included subcategories of problem solving, patterns and relations, and space and shape. Cognitive level bundles were created

based on whether they tapped the lowest (knowledge) or highest (higher level understanding and application: analysis, synthesis, and evaluation) cognitive levels in the Principles of Mathematics 12 exam and,

2) Patterns and consistencies in definition in the gender DIF literature were identified for item context. Items fitting the definitions generally used in the literature were placed in bundles. In other words, context bundles were formed in consideration of gender DIF patterns related to context identified in the literature. An exploration of the gender DIF literature revealed that the inclusion of an equation on computation items and use of visuals in a question might differentially influence responses of boys and girls. Further, items placed in a word problem context were also consistently found to contain DIF.

As dictated by gender DIF research literature, four context bundles were created. Directional hypotheses were formed for word problems and items containing a visual cue. An exploratory analysis was completed for the two computation bundles since no particular direction may be hypothesized using patterns appearing in the gender DIF literature.

Specific definitions are given for each content, cognitive level, and context bundle. Table 2 indicates the area, general subtopics, and specific subtopics used to identify bundles for the confirmatory analysis. Recall that the remaining items in the general subtopic served as the matching criterion. When the number of items remaining in the general subtopic was equal to or less than the number of items in the bundle, the remaining items in the total test score was used as the matching criterion. In order to identify the presence of a secondary dimension, using SIBTEST, bundles were analyzed

for differential bundle functioning. The number of items in the bundles ranged from 2 to 13.

Table 2

Area, General Subtopics, and Specific Subtopics Used to Identify Bundles for the Confirmatory Analysis

| Area | General Subtopic | Specific (Bundle) Subtopic | Number of Items in Each Bundle |
|---|---|---|---|
| Content | Problem Solving | General Problem Solving | 6 |
| | Patterns & Relations | Polynomials | 5 |
| | Patterns & Relations | Quadratic Relations | 5 |
| | Patterns & Relations | Quadratic Functions | 3 |
| | Patterns & Relations | Logarithms & Exponents | 8 |
| | Patterns & Relations | Seq. & Series | 8 |
| | Shape & Space | Geometry | 4 |
| Cognitive Level | Total PME12 | High Cognitive Level | 13 |
| | Total PME12 | Low Cognitive Level | 6 |
| Context | Problem Solving* | Word Problems | 2 |
| | Shape & Space* | Visuals | 10 |

*The matching criterion used was the score on the total test with the items in the bundle not included in the analysis. The number of items in the general subtopic was too small to produce reliable results.

*Content Bundle Definitions*

Definitions used in the creation of content bundles was dictated primarily by the PME12. Bundles were created for specific subtopics within each of the general subtopics: problem solving, patterns and relations, and shape and space items.

*Problem Solving*

The Principles of Mathematics 12 Exam categorized items with cross-topic problems including word problems along with unique problem sets as problem solving. In the literature, problem solving generally refers to problem solving in context (word problems), applied problem solving, and/or reasoning ability in problem solving. (Benbow & Stanley, 1982; Berberoglu,1995; Hyde, Fennema, & Lamon,1990; Linn & Hyde,1989; Pattison & Grieve, 1984). An example of a problem solving item is shown in Figure 1.

*Figure 1.* Example of a Problem Solving Item

| |
|---|
| 43. Determine the product of all of the positive divisors of 72. |
| A. $72^{5}$ |
| B. $72^{6}$ |
| C. $72^{10}$ |
| D. $72^{12}$ |

Recall that a separate bundle for word problems was also created and categorized as context in order to be consistent with the literature. The general problem solving bundle contained six items identified by the PME12 as either cross topic problem solving, or unique problem solving items, one of which was a word problem. It was hypothesized that bundles containing problem solving items would exhibit DBF in favor of the boys.

*Patterns and Relations*

In the Principles of Mathematics 12 Exam, the patterns and relations subtopic included items involving polynomials, exponential and logarithmic functions, quadratic relations, quadratic systems, and sequences and series. An example of a polynomial item

and a quadratic systems item are given in Figures 2 and 3, respectively. Patterns and relations was the area with the least consistent definitions found in the literature. It was hypothesized that bundles dealing with patterns and relations defined as those dealing with polynomials, exponential and logarithmic functions, quadratic relations, quadratic systems, and sequences and series would not exhibit DBF.

*Figure 2.* Example of a Polynomial Item

4. Solve the following inequality for $x$, given that $a$, $b$ and $c$ are constants such that $a < b < c$.

$$(x-a)^3 (x-b)^2 (x-c) > 0$$

A. $x > c$

B. $x < a$ or $x > c$

C. $x < c,\ x \neq a,\ x \neq b$

D. $a < x < c,\ x \neq b$

*Figure 3.* Example of a Patterns and Relations Quadratic Relations Item

6. Which equation represents a rectangular hyperbola?

$x^2 \cdot 3y^2 = 9$

$x - y = 9$

$x^2 - y = 9$

$x^2 \cdot y^2 = 9$

*Shape and Space*

Shape and space is defined by the PME12 as those items dealing with geometry and trigonometry. It was hypothesized that items and bundles dealing with geometry would exhibit DBF in favor of the boys. A bundle was also created with trigonometry as

the primary dimension for the purpose of comparison and exploration since studies of shape and space or measurement items may also include trigonometry items.

*Figure 4.* Example of a Trigonometry Item

35. Solve:

$\tan\theta - \cos^2\theta = 1/2, 0 \le 0 < 2\pi$

A. 0.36, 3.50

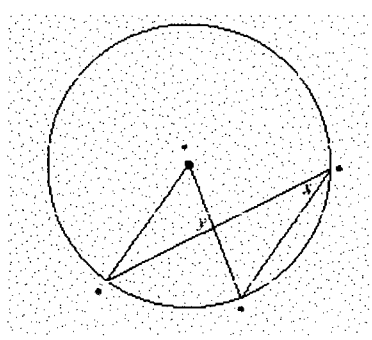
B. 0.79, 3.93

C. 0.86, 2.74

D. 0.88, 2.94, 3.26, 3.74

*Figure 5.* Example of a Geometry Item

41. In the diagram below, O is the center of the circle and OR$\|$PQ. Determine an expression for $y$ in terms of $x$.



A. 2x

B. 3x

C. 180° - 2x

D. 180° - 3x

The two item types are analyzed separately in this study. Figures 4 and 5 provide examples of a trigonometry item and a geometry item, respectively.

*Cognitive Bundle Definitions*

The PME12 tapped three cognitive levels. Cognitive level 1 included only knowledge items, cognitive level 2, included items where understanding and application referred to single computation items, and cognitive level 3 where understanding and application referred to analysis, synthesis, and evaluation. Two of these cognitive levels were tested: level 1, tapping the lower level cognitive level- knowledge, and level 3, the higher cognitive level- understanding and application: analysis, synthesis, and evaluation.

Item examples for the two cognitive level poles are shown in Figures 6 and 7.

*Figure 6.* Example of a Level 1 –Knowledge Only Cognitive Item

21. Which of the following is an arithmetic sequence?

A. 1, 3, 5, 7

B. 1, 2, 4, 8

C. 1, 3, 5, 8

D. 1, $\sqrt{2}$, 2, $\sqrt{8}$

*Figure 7.* Example of a Level 3- High Cognitive Level Item

28. Evaluate $\sum_{k=2}^{4} \log_6 k$

A. 0.60

B. 1.23

C. 1.77

D. 4.00

In the gender DIF research literature, items identified as tapping higher cognitive levels were found to favor boys while those tapping lower cognitive levels were found to

favor girls (Armstrong, 1981; Carpenter, Linquist, Mattews, & Silver, 1983; Fennema, 1984; Fennema & Tartre, 1985; Harris & Carlton, 1993; Kaplan & Plake, 1982). It was therefore hypothesized that items and bundles tapping the highest cognitive level would exhibit moderate to high level DBF in favor of the boys. It was also hypothesized that items and bundles dealing with knowledge only (lower cognitive level items) would exhibit moderate to high level DBF favoring the girls.

*Context Bundle Definitions*

Harris and Carlton (1993), after examining the Doolittle and Cleary study (1987), recognized a need to study a broader range of item characteristics in order to identify relative strengths and weaknesses for male and female students. They suggested that the format in which the test items are presented be examined along with, when applicable, the subject matter in which test items are embedded. Analyses were completed for item context. Item bundles were created for context in the following manner:

1) When problem solving was presented in the context of a story (word) problem it was hypothesized that DBF would be detected in favor of the boys (Hyde, Fennema, & Lamon, 1990; Linn & Hyde, 1989; Swafford, 1980). There were only two items presented in this context. Figure 8 provides an example of a problem solving item presented in a story (word) problem context.

2) It was hypothesized that items incorporating a visual would display a high level of DIF in favor of the boys. A consideration in this hypothesis was the association between use of a visual and the association with spatial skills items. Boys, as noted by Benbow and Stanley (1982), tend to develop certain spatial skills to a greater degree than girls. Figure

9 is an example of an item with a visual included. The geometry item presented in Figure 5 provides a second and different example of an item in which a visual is included.

*Figure 8.* Example of a Story (Word) Problem

45. A farmer has 40 m of fencing to enclose a rectangular pasture. One side of the pasture must include part of one side of his building. If the side of the building is 60 m, determine the maximum area the farmer can enclose.

A. 100 $m^2$

B. 200 $m^2$

C. 400 $m^2$

D. 625 $m^2$

*Figure 9.* Example of an Item with a Visual Included

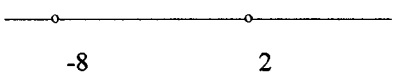
8. Solve $|x+5| > 3$

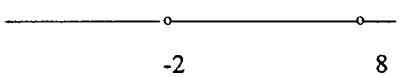
A.



B.



C.



D.



## Statistical Analyses

PARDUX was used for analyzing differential functioning at the item level and SIBTEST was used to analyze bundles and test hypotheses. Confirmatory analyses were completed, followed by the exploratory analyses.

## *Confirmatory Analyses*

Confirmatory analyses were completed for subtopic bundles of content areas, cognitive level tapped by the item, and context. The computer program SIBTEST was used for both the confirmatory and exploratory analyses of the bundles. Results of the statistical test of the hypotheses associated with the bundles are provided here.

### *Problem Solving*

It was hypothesized that bundles dealing with general problem solving items would exhibit differential functioning in favor of boys. Indeed, C level DBF in favor of boys was detected ($\beta$= 0.098, p=0.000).

### *Patterns and Relations*

Results from the patterns and relations bundles, consisting of algebra subtopics, varied by subtopic. The specific patterns and relations subtopic bundles tested were polynomial, quadratic relations, quadratic systems, logarithms and exponents, and sequence and series. It was hypothesized that items and bundles dealing with polynomial functions would not exhibit significant DBF. The data did not support this hypothesis. Moderate, or B level, DBF favoring girls was detected for the polynomial bundle ($\beta$= -0.059, p=0.000). The data did, however, support the hypothesis of no DBF for the quadratic systems bundle. It was hypothesized that the bundle for quadratic relations items would not exhibit significant DBF yet moderate DBF was detected ($\beta$=-0.052, p=0.021). Contrary to the hypotheses, logarithms and exponents displayed B level DBF ($\beta$=-0.086, p=0.001). It was hypothesized that the bundle dealing with sequence and series would exhibit moderate DBF. This hypothesis was not supported by the statistics. The sequence and series bundle revealed no DBF.

*Shape and Space.*

One confirmatory analysis was completed for shape and space. A bundle was tested for geometry. It was hypothesized that DBF would be found in favor of the boys for geometry. Analysis of the geometry bundle produced no DBF. Results of the bundle analysis for content are displayed in Table 3. Some items appear in more than one bundle as the item contains aspects of more than one content subtopic.

Table 3

Summary Table of Content Bundles

| Reference/Description | Bundle | $\beta$ | Level | p value | Favors |
|---|---|---|---|---|---|
| Problem Solving | 3,5,35,43,44,45 | 0.098 | C | 0.000 | BOYS |
| Polynomials | 1,2,3,4,5 | -0.059 | B | 0.000 | GIRLS |
| Quadratic Relations | 6,7,11,12,13 | -0.052 | B | 0.021 | GIRLS |
| Quadratic Systems | 8,9,10 | -0.001 | A | 0.931 | |
| Logarithms & Exponents | 14,15,16,17,18,19, 20,28 | 0.086 | B | 0.001 | BOYS |
| Sequence & Series | 21,22,23,24,25,26, 27,28 | -0.002 | A | 0.920 | |
| Geometry | 39,40,41,42 | 0.008 | A | 0.736 | |

*Cognitive Level*

It was hypothesized that DBF favoring girls would be found on the lower cognitive level items, and DBF favoring boys on the higher cognitive level items. Analysis of the lower cognitive level, knowledge only, item bundle revealed no DBF ($\beta$ = 0.0, p = 0.999). High cognitive level items produced C level DIF ($\beta$ = 0.353, p =

0.000) in favor of the boys as was hypothesized. The results of the cognitive level analyses are summarized in Table 4.

Table 4

Summary Table of Cognitive Bundles

| Reference/Description | Bundle | $\beta$ | Level | p value | Favors |
|---|---|---|---|---|---|
| Level 1-Knowledge | 2,6,14,21,29,30 | 0.0 | A | 0.999 | |
| Level 3-Higher | 4,5,12,13,19,20, 27,28,36,37,38, 42,45 | 0.353 | B | 0.000 | BOYS |

*Context*

Two context effects were hypothesized and tested - word problems and inclusion of visuals. For word problems it was hypothesized that DBF would be detected in favor of the boys. The word problem bundle did indeed produce C level DBF ( $\beta$= 0.0999, p= 0.000) favoring boys. It was hypothesized that primary dimension bundles including items presented with a visual would exhibit high level DBF favoring the boys. The DBF hypothesis was confirmed for visuals ( $\beta$= 0.141, p=0.000). The results of the context analyses are summarized in Table 5.

Table 5

Summary Table of Context Bundles

| Reference/Description | Bundle | $\beta$ | Level | p value | Favors |
|---|---|---|---|---|---|
| Word Problems | 37,45 | 0.099 | C | 0.000 | BOYS |
| Visual Included | 2,8,29,30,31,39,40, 41,42 | 0.141 | C | 0.000 | BOYS |

## *Exploratory Analyses*

Exploratory analyses were completed initially for trigonometry, and computation with and without an equation clearly provided in the question. Post-hoc exploratory analyses were then conducted for higher cognitive level to see if the C level DBF detected was attributable to one or both of the content areas that contributed items to the higher cognitive level bundle. A summary of the results from exploratory analyses is shown in Table 6.

### *Trigonometry*

An exploratory analysis was completed for trigonometry items. Moderate DBF in favor of boys on shape and space trigonometry items was identified ($\beta$=0.076, p=0.024).

### *Computation*

Two exploratory analyses were conducted for computation. Bundles were created using the expertise of two teachers of mathematics who were consulted in the selection of the items. Aware of the possibility that boys and girls may be cued differentially by the presence or lack of an equation in an item, two bundles were created: a) computation items with equations clearly present in the question and b) computation items where equations are not clearly presented in the question. Included in this bundle were the items for which the equation was provided in the exam appendix since it was necessary for the student to select the correct equation from multiple options. C level DBF ($\beta$= -0.23, p= 0.000) in favor of girls was detected for the bundle including computation items with no equation provided yet the bundle containing computation items with equations provided was found to display no DBF.

Table 6

Summary Table of Bundles in the Exploratory Analysis

| Reference/Description | Bundle | $\beta$ | Level | p value | Favors |
|---|---|---|---|---|---|
| Trigonometry | 29,30,31,32,33,34, 35,36,37,38 | 0.092 | C | 0.001 | BOYS |
| Computation (No Equation) | 1,5,10,11,12,13,15 24,28 | -0.077 | B | 0.011 | GIRLS |
| Computation (Equation Provided) | 3,4,9,17,18,22 | -0.011 | A | 0.614 | |
| High Cognitive Level (Shape and Space) | 36,37,38,42 | 0.030 | A | 0.183 | |
| High Cognitive Level (Patterns and Relations) | 4,5,12,13,19,20,27,28 | 0.151 | C | 0.000 | BOYS |

*High Cognitive Level by Content*

An examination of the high cognitive level items revealed that they were drawn primarily from two content areas: patterns and relations, and shape and space. The items identified as tapping the high cognitive level were divided into two separate bundles: high cognitive level patterns and relations, and high cognitive level shape and space. Exploratory post-hoc analyses were then completed for these bundles. No DBF was found for high cognitive level shape and space bundles in the exploratory analysis. In contrast, C level DBF ($\beta$=0.151, p = 0.000) in favor of the boys was detected for high cognitive level patterns and relations items. In order to ensure that the DBF was not a result of a particular subtopic of patterns and relations, the item composition of the high cognitive level bundle was examined by patterns and relations subtopic areas. The

individual items did not belong to any one particular patterns and relations subtopic area. Recall that the highest level of DBF detected for patterns and relations subtopics was level B.

## *DIF*

A DIF analysis was completed using PARDUX. Table 7 presents the results of the DIF analysis and the primary dimension they correspond to in the confirmatory analysis.

Since the dimensional bundle should indicate the source of potential DIF, it follows that if individual items are identified as DIF, they will share the features of the primary dimension that was confirmed as a source of DIF. Individual items were therefore tested using the Linn-Harnish method to check for congruence with primary dimension areas confirmed as DIF and as an exploratory measure.

Although a distractor analysis was conducted there were no notable findings nor was non-uniform DIF detected via an examination of the item characteristic curves using PARDUX.

### *Problem Solving*

Items 5 and 45 were flagged as having moderate DIF. Item 5 was a cross-topic, polynomial, problem-solving item. Item 45 was a word problem.

### *Patterns and Relations*

Two polynomial items were identified as DIF, one favoring the girls, the other, the boys. Two quadratic relations items were also identified as containing moderate DIF. The direction of DIF for these items was also mixed. Three logarithm and exponents items were identified with moderate DIF. Two items favored the boys and one item favored the girls. Recall that B level DBF was found for polynomials, quadratic relations and logarithms and

exponents. All but one of the sequence and series items identified as DIF in this area indicated moderate DIF in favor of the boys. No DBF was found for sequence and series.

Table 7

DIF Items Identified by PARDUX and the Corresponding Subtopic Bundle from the DBF Analysis

| Item | Level of DIF | Favors | Subtopic Bundle |
|---|---|---|---|
| 1 | 2 | girls | polynomials |
| 5 | 2 | boys | polynomials |
| 7 | 2 | girls | quadratic relations |
| 13 | 2 | boys | quadratic relations |
| 14 | 2 | girls | logarithms and exponents |
| 15 | 2 | boys | logarithms and exponents |
| 20 | 2 | boys | logarithms and exponents |
| 21 | 2 | girls | sequence and series |
| 23 | 2 | girls | sequence and series |
| 25 | 2 | girls | sequence and series |
| 27 | 2 | boys | sequence and series |
| 29 | 2 | boys | trigonometry |
| 31 | 2 | boys | trigonometry |
| 39 | 2 | girls | geometry |
| 41 | 2 | boys | geometry |
| 42 | 2 | boys | geometry |
| 45 | 2 | boys | problem solving |

*Shape and Space*

Two of the three geometry items flagged with DIF favored the boys while only one favored the girls. Two trigonometry items were flagged with DIF. Both of the trigonometry items identified as DIF indicated moderate DIF in favor of the boys. Recall the no DBF was found for geometry but high DBF was found for trigonometry. All of the shape and space items (geometry and trigonometry) flagged with DIF also included a visual. Five out of the seventeen items (29%) identified with DIF contained visuals.

Table 8

Percentage of Flagged Items Represented by each Subtopic Bundle in the DBF Analysis

| Subtopic | % of Flagged Items* | Favors Girls | Favors Boys |
|---|---|---|---|
| problem solving | 6% | 0 | 1 |
| polynomials | 12% | 1 | 1 |
| quadratic relations | 12% | 1 | 1 |
| logarithms & exponents | 18% | 1 | 2 |
| sequence and series | 23.5% | 3 | 1 |
| trigonometry | 12% | 0 | 2 |
| geometry | 18% | 1 | 2 |
| lower cognitive level | 12% | 2 | 0 |
| high cognitive level | 35% | 0 | 6 |
| visuals included | 29% | 1 | 4 |
| computation (no equat.) | 23.5% | 1 | 3 |

* Percentage does not add to 100 since item categories overlap.

*Cognitive Level of the DIF Items*

Of the seventeen items flagged with DIF, six are items tapping the higher cognitive level. All of the higher cognitive level items flagged indicated moderate DIF in favor of the boys. Two flagged items were identified as knowledge only, both in favor of the girls in congruence with the DIF literature.

*Computation*

Of the items identified as having DIF four were computation items with no equation included. This represents 24% of the items identified with DIF. None of the computation items with equation provided were identified with DIF. Recall that B level DBF was detected for the bundle containing computation items without equations, and no DBF was detected for computation items with an equation provided.

# CHAPTER 5: DISCUSSION

In this study a two stage confirmatory MMD approach was applied to the Principles of Mathematics 12 exam. The findings of this research will add to the powerful scientific discourse centered on gender and mathematics, and provide an additional source of information about the differential performance of boys and girls. These confirmed sources of differential item functioning may be used to develop guidelines and test construction principles for reducing gender DIF in mathematics. This research also contributes to a greater understanding of gender differences in mathematics learning and achievement. Two goals were inherent in this confirmatory method of analysis. The first goal was to identify some of the patterns that are present in the gender DIF research in mathematics. This was accomplished through a review of the gender DIF literature in which patterns of differential functioning of items for boys and girls on tests of mathematics were noted. The second goal was to test the existence of these patterns in the British Columbia Ministry of Education's Principals of Mathematics Exam 12. The confirmatory MMD approach used in this study was of particular importance. While information obtained from the DIF research provides insight into the patterns present for boys and girls in mathematics, the confirmation of these patterns serves to highlight the sources of DIF and allow test developers to make more informed decisions.

Lack of attention to the dimensionality of items has, in part, resulted in DIF reports using mixed content. Hyde, Fenema, and Lamon (1990) noted that their analysis of mathematics content was less than successful because a high number of studies used either a mixture of mathematics content or simply did not report the content. The mixture

of mathematical content in the literature was seen to the greatest degree in pattern and relations subtopics and in the inclusion of a combination of trigonometry and geometry in analyses.

This chapter summarizes results by each source of differential functioning considered in the study; compares findings to previous research; and identifies issues that need to be addressed in future research.

## Problem Solving

In the gender DIF literature, problem solving generally referred to word or story problems (Berberoglu, 1995;Doolittle & Cleary, 1987; Harris & Carlton, 1993; Lane, Wang, & Magone, 1995; Swafford, 1980). Recall that two problem solving hypotheses were tested, one for general problem solving and one for problem solving in context, and that previous studies indicted that problem solving items favored boys. The mixture of applications for the term problem solving was noted in the analysis of the gender DIF literature. Problem solving was also used to describe quantitative problem solving (Pallas & Alexander, 1983), problem solving strategies, problem type (Gallagher & De Lisi,1994), and problem solving as a process (Flexer,1984). Problem solving was also studied in association with cognitive level (Hyde, Fennema & Lamon, 1990). In this study, problem solving was confirmed as a source of gender DIF when the item is presented in the form of a story problem or when the problems are non-context specific. Even when the definitions of problem solving as a source of differential functioning were not identical in previous research, the findings were the same. Whether problem solving items are studied according to type, strategy or content, the conclusion is that, on problem

solving items, boys tend to perform better than girls who are at the same ability level. Results of this study lead to the same conclusion.

*Patterns and Relations*

The primary dimensions identified in patterns and relations were often placed together under the heading of algebra or computation, and no differential function was expected to be found since the results of previous studies were quite mixed. However, when algebra was broken down into specialized skill areas it became clear that differential functioning was present in some but not all of the specialized skill (subtopic) areas studied. Algebra/patterns and relations subtopics produced a mixture of confirmed sources of DIF with some subtopics favoring the girls, some favoring the boys. In this study, moderate differential functioning was found only for polynomials and quadratic relations favoring girls and in logarithms and exponents favoring boys. Some subtopic areas such as quadratic systems and sequence and series produced no differential functioning. The mixture of subtopics found to favor the boys and girls in this area might account for the mixed findings of previous studies since the findings of these studies would be influenced by the specialized skill areas tapped by the items, and subsumed under the category algebra or patterns and relations.

*Space and Shape*

In the gender DIF literature space and shape items were generally defined as geometry and/or measurement with DIF items in these areas consistently favoring the boys (Berberoglu, 1995; Doolittle & Cleary, 1987; Hanna, 1985; Harris & Carlton, 1993; Linn & Hyde, 1989; Maccoby & Jacklin ,1974; Pattison & Grieve, 1984). Given this consistent finding in the literature, the results for the space and shape bundles containing

either geometry or trigonometry items was surprising. Although differential functioning was confirmed for individual geometry *items* it was not confirmed for geometry. Moderate differential functioning, however, was found for trigonometry. An examination of the composition of the geometry bundle showed that despite the fact that the geometry items showed no differential functioning the individual items corresponded to higher cognitive level items and utilized a high number of visuals both of which contained high DBF (see table 5). Perhaps the DIF findings for space and shape in current research reflect DIF in higher cognitive level items, or are a factor of use of visuals in the items. In contrast to what might be expected given the findings of previous gender DIF research, the present study did not find geometry to be a source of gender DIF. Items containing visuals often associated with geometry was, however, confirmed as a source of gender DIF.

Another possible explanation for this discrepancy in performance of boys and girls on shape and space items may be taken from Casey, Nuttall, and Pezaris (2001) who examined spatial-mechanical reasoning skills versus mathematics self-confidence as mediators of gender differences on mathematics subtests. One conclusion that may be drawn from their study and used to illuminate some of the findings in this study is that choice of strategy may account for differences in scores on spatial skills items. An investigation into the choice of strategy used by boys and girls on these items may serve to explain the differential functioning of items in this area.

## Cognitive Level

The two cognitive levels tested were the lower level cognitive level- knowledge, and the higher cognitive level- understanding and application: analysis, synthesis, and

evaluation. It was hypothesized that the higher cognitive level bundle would be found to favor boys while those tapping lower cognitive levels would favor girls (Armstrong, 1981; Carpenter, Linquist, Mattews, & Silver, 1983; Fennema, 1984; Fennema & Tartre, 1985; Harris & Carlton, 1993; Kaplan & Plake, 1982). Contrary to expectations, no differential functioning was found for the lower level, knowledge bundle. This finding should be encouraging for teachers and administrators as it indicates that for this exam, there was no difference in the ability of boys and girls to answer a question for which knowledge of the principles of mathematics was the only consideration. A high level of DBF was confirmed for the high cognitive level bundle in the hypothesized direction. All of the higher cognitive level items favored boys. This finding was in congruence with the DIF literature. Perhaps, as was suggested by Casey, Nuttall, and Pezaris (2001), greater exposure to and/or greater interest in areas of mathematics might serve to explain this finding. This finding merits further investigation into why boys do better on higher cognitive level items.

A post-hoc analysis was completed to determine if differential functioning was present for higher cognitive level in both patterns and relations and shape and space. It was indicated that the higher cognitive level bundle with the shape and space items exhibited no differential functioning. The higher cognitive level patterns and relations item bundle, however, contained high level differential functioning. This finding was unexpected since high differential functioning was found for one of the two space and shape subtopics while only moderate differential functioning was found in some subtopics of patterns and relations. Further research into the cognitive processes used by

students when confronted by these types of items might serve to shed light on these findings.

## Computation

An exploratory analysis was conducted for computation. The research literature suggested that computation items in general might exhibit differential functioning: however, no direction was hypothesized. These findings were generally associated with algebra content. High levels of differential functioning were detected in favor of girls on the bundle for computation with no equation provided in the question. Once again, greater information about the cognitive processes of boys and girls could serve to illuminate these findings since only negligible differential functioning was found for items for which an equation was provided in the question.

The high level of differential functioning identified in the visuals bundle is left unexplained. Further information is required about the way in which boys and girls use the visuals and think about the items for which visuals are provided.

## General Observations

It is interesting to note that confirmed sources of differential functioning that were found to favor the boys were found in bundles containing higher cognitive level items, computation items without equations provided, and problem solving. More research into the cognitive processes involved in solving these item types might reveal a connection between the way in which students solve problems, determine solutions to computation items when no equation is provided, and resolve high cognitive level items, since, at least on the surface, it would appear that logical reasoning may be inherent in all three of these areas.

## Differential Item Functioning Analysis

The analysis of the individual items in the patterns and relations content areas was interesting. Recall that the patterns found in the differential item functioning research was mixed for the patterns and relations items such that it was hypothesized that, when the items were placed in subtopic bundles, no differential item functioning would be detected. Findings from the differential item functioning analysis were consistent with those in the previous research literature. The results of the differential bundle functioning analyses, when compared to the results of the differential item functioning analysis, appear to be incongruent. A close inspection of the results reveals that, similar to what was reported in the literature, these analyses yielded mixed results. Two of the polynomial items were identified as functioning differentially, one favoring the girls, the other, the boys; yet the polynomial bundle displayed moderate differential functioning in favor of the girls. The increased power to detect differential functioning when items are placed in bundles that was highlighted by Douglas, Roussos, and Stout (1996) may also be at play here.

Two quadratic relations items were identified as containing moderate differential item functioning. The direction of differential functioning for these items was also mixed. Three logarithm and exponents items were identified as displaying moderate differential item functioning. Two items favored the boys and one item favored the girls. The lack of congruency may have been a result of information lost when items were placed in bundles since the size and direction of differential functioning in the bundles may serve to cancel out item level differential functioning that may have otherwise been noted.

## Summary and Directions for Future Research

Results from the patterns and relations bundles consisting of algebra subtopics varied by subtopic. This finding supports the decision to test algebra according to its subtopics. Further, the content areas of general problem solving, quadratic relations, polynomials, logarithms and exponents, and trigonometry were confirmed as sources of differential functioning for gender groups mirroring the mixed results found in the literature. It was also confirmed that high cognitive level items are a source of differential functioning and that context such as visuals in an item, presentation of an equation in items requiring computation and word problems should be a consideration in test construction.

Findings in this study indicate that more information about the cognitive processes used by boys and girls when solving word problems and computation items is needed in order to explain differential functioning in tests of mathematics for boys and girls. An increased understanding of the way in which boys and girls use visual content in a problem, solve word problems, use cues to select and use appropriate equations, and generally utilize cues and clues in content and context would be beneficial in gaining an understanding of the sources of gender differences in performance patterns.

There are some limitations to this study that might affect the generalizability of results. Only students who indicated English as their first language were included in the analysis. Test constructors must keep in mind that patterns might be quite different for English as a second language (ESL) examinees.

The PME12 was designed for students in grade 12 who wished to pursue careers or further education in mathematics. As such, prior course taking was not considered in

this analysis. It should be noted that studies traditionally recognized that performance in mathematics for boys and girls has been found to be generally equivalent in the early years with a discrepancy in scores by gender appearing later (Hyde, Fennema, & Lamon, 1990; Kaplan & Plake, 1982; Pallas & Alexander, 1983). It should be recognized that the hypotheses and statistical results in this study may have been quite different for a younger population although more recent high ability studies have indicated that gender differences in mathematics may begin as early as grade 2 (Mills, Ablard, & Stumpf, 1993).

Future research may wish to focus on; use of visuals, word problems, and computation with no equation is present, with an English as a second language (ESL) or high ability population. More confirmatory research using the multidimensional model would serve to build a bank of confirmed hypotheses about the sources of differential functioning for gender in mathematics and in other areas. This would help educators to make decisions about instruction and policy. Results of this study will also assist test developers to ensure validity in tests of mathematics for boys and girls.

# REFERENCES

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

Anastasi, A. (1958). Differential psychology (3rd ed.). New York: Macmillan.

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

Armstrong J. (1981). Achievement and participation of women in mathematics: Results of two national surveys. Journal for Research in Mathematics Education, 12, 356-372.

Benbow, C. P., & Stanley, J. C. (1982) Consequences in high school and college of sex differences in mathematical reasoning ability: a longitudinal perspective. American Educational Research Journal, 19, 598-622.

Berberoglu, G. (1995). Differential item functioning (DIF) analysis of computation, word problem and geometry questions across gender and SES groups. Studies in Educational Evaluation, 21, 439-456.

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

Carpenter, T, Linquist, M, Matthews, W, and Silver, E. (1983) Results of the Third NAEP Mathematics Assessment: Secondary School. Mathematics Teacher, 76, 652-659.

Casey, M. B., Nuttall, R., Pezaris, E.(2001). Spatial-mechanical reasoning skills versus mathematics self-confidence as mediators of gender differences on mathematics subtests using cross-national gender-based items. Journal of Research in Mathematics Education, 32, 28-57.

Childs, R.(2000). Gender bias and fairness. In, ERIC Clearinghouse on Tests, Measurement, and Evaluation, ED328610.

Clauser, B., Mazor, K & Hambleton, K. (1991). Influence of criterion variable on identification of differential item functioning test items using the Mantel-Haenszel statistic. Applied Psychological Measurement, 14, 353 – 359.

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

Doolittle, A., & Cleary, T. (1987). Gender-based differential item performance in mathematics achievement items. Journal of Educational Measurement, 24, 157-166.

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

Douglas, J., Roussos, L., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. Journal of Educational Measurement, 33, 465-484.

Ethington, C. & Wolfle, L. (1984). Sex differences in a casual model of mathematics achievement. Journal for Research in Mathematics Education, 15, 361- 377.

Fennema, E. (1974). Mathematics, learning and the sexes: A review. Journal of Research in Mathematics Education, 5, 126-139.

Fennema, E., Carpenter, T., Jacobs, V., Franke, M., & Levi, L. (1998). A longitudinal study of gender differences in young children's mathematical thinking. Educational Researcher, 27, 6-11.

Fennema, E. & Sherman, J. (1977). Sex related differences in mathematics achievement, spatial visualization, and affective factors. American Educational Research Journal, 14, 51-71.

Fennema, E. & Tartre, L. (1985). The use of spatial visualization in mathematics by girls and boys. Journal for Research in Mathematics Education, 16, 184- 206.

Flexer, B. (1984). Predicting eighth grade algebra achievement. Journal for Research in Mathematics Education, 15, 352 – 360.

Friedman, L. (1989). Mathematics and the gender gap: a meta-analysis of recent studies on sex differences in mathematical tasks. Review of Educational Research, 59, 185-213.

Gallagher, A. & De Lisi, R. (1994). Gender differences in scholastic aptitude test-mathematics problem solving among high ability students. Journal of Educational Psychology, 86, 204-211.

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

Gierl, M. & Khaliq, S. (2001). Identifying Sources of Differential Item and Bundle Functioning on Translated Achievement Tests: A Confirmatory Analysis. Journal of Educational Maeasurement, 38, 164-187.

Haag, P. (2000) K-12 Single-Sex Education: What Does the Research Say? In ERIC Clearinghouse on Elementary and Early Childhood Education, Champaign IL ED444758 .

Hanna, G. (1986). Sex differences in a casual model of mathematics achievement, Journal for Research in Mathematics Education, 15, 361- 377.

Harris, A., & Carlton, S. (1993). Patterns of gender differences on mathematics items on the scholastic aptitude test. Applied Measurement in Education, 6, 137-151.

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

Hyde, J., Fennema, E., & Lamon, S. (1990). Gender differences in mathematics performance: a meta-analysis. Psychological Bulletin, 107, 139-155.

Hyde, J., & Jaffee, C. (1998). Perspectives from social and feminist psychology. Educational Researcher, 27, 14-16.

Kaplan, B., & Plake, B. (1982). Sex differences in mathematics: differences in basic logical skills? Educational Studies, 8 , 31- 36.

Lane, S., Wang, N., & Magone, M. (1996). Gender related differential item functioning on a middle school mathematics performance assessment. Educational Measurement: Issues and Practice, 15, 21-27, 31.

Lee, V. E., & Lockheed, M. M. (1990). The effects of single-sex schooling on achievement and attitudes in Nigeria. Comparative Educational Review, 34, 209-231. EJ 412 239.

Linn, R.(1993). The use of differential item functioning statistics: A discusssion of current practice and future implications. In P. Holland, & H. Wainer (Eds.), Differential item functioning (pp.337-347). Hillsdale, NJ: Lawrence Erlbaum Associates.

Linn, R., & Harnish D. (1981). Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 18, 109-118.

Linn, M., & Hyde, J. (1989). Gender, mathematics, and science. Educational Researcher, 18, 17 - 19, 22- 27.

Maccoby, E., & Jacklin, C. (1974). Psychology of Sex Differences. Palo Alto, CA: Stanford University Press.

Mallam, W.A.(1993). Impact of school-type and sex of the teacher on female students' attitudes toward mathematics in Nigerian secondary schools. Educational Studies in Mathematics, 24, 223-229. EJ 476 667.

Mills, C.J., Ablard, K.E., & Stumpf, H. (1993). Gender differences in academically talented young students' mathematical reasoning: Patterns across age and subskills. Journal of Educational Psychology, 85, 340-346.

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

Pallas, A., & Alexander, K. (1983). Sex differences in quantitative SAT performance: New evidence on the differential coursework hypothesis. American Educational Research Journal, 20, 165-182.

Pattison, P. & Grieve, N. (1984). Do spatial skills contribute to sex differences in different types of mathematical problems? Journal of Educational Psychology, 76, 678-689.

Roussos L. & Stout W. (1996). A multidimensionality based DIF analysis paradigm. Applied Psychological Measurement, 20, 355-371.

Ryan K., & Fan, M. (1996). Examining gender DIF on a multiple-choice test of mathematics: A confirmatory approach. Educational Measurement: Issues and Practice, 15, 15-20, 38.

Schmitt, A., Holland, P., & Dorans, N. (1993). Evaluating hypotheses about differential item functioning. In P. Holland, & H. Wainer (Eds.), Differential item functioning (pp.281-315 ). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Stout, W., & Roussos, L. (1995). SIBTEST Manual. Urbana, IL: University of Illinois, Department of Statistics, Statistical Laboratory for Educational and Psychological Measurement.

Swafford, J. (1980) Sex differences in first-year algebra. Journal for Research in Mathematics Education, 11, 335- 346.

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

Wang, N., & Lane, S. (1996). Detection of gender-related differential item functioning in a mathematics performance assessment. Applied Measurement in Education, 9, 175-199

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

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