Lessons Learned from DIF Algebra Problems

by

Natalia Alexeev

(Under the direction of Jeremy Kilpatrick)

Abstract

In light of the low representation of women in high-level technological jobs, many studies have addressed the issue of gender differences in mathematics. This study investigated those differences from the angle of gender-related differential item functioning (DIF) on algebra and algebra-related items on the Florida Comprehensive Assessment Test (FCAT) and on the 2003 Trends in International Mathematics and Science Study (TIMSS). The goal of the study was to identify the characteristics of items that contribute to DIF, finding differences for female and male students whose total scores on the test matched. More than 300 items from 3 years of the FCAT, Grades 8, 9, and 10, were coded according to the mathematical content, context topic, and other mathematical and nonmathematical characteristics. The Mantel-Haenszel and standardization procedures were used to quantify DIF. Two content categories, geometrical measurement and informal algebra, favored male students. The algebraic manipulations category favored female students. Several context topics were found to contribute to DIF: Recreational topics favored male students, and social studies topics favored female students. Items that required providing an approximate answer were challenging for female students. Items involving converting units favored male students, and items with noncomputed answers favored female students. Characteristics contributing to DIF on the FCAT were compared with characteristics of DIF items from the TIMSS. Data

from the TIMSS were analyzed for U.S. eighth graders. DIF items were identified and characteristics of released items were compared to those on the FCAT. Findings on content categories were confirmed. In addition, items that tested concepts of fractions were common among the DIF items favoring male students. Items with patterns were common among the DIF items favoring female students. The topic of the context did not benefit either gender, although female students had a high proportion of no-context DIF items. The results suggested that there are patterns of differences in mathematics performance for male and female students who presumably have the same ability. The results also indicated that differential functioning should be investigated on the mathematics concept level in addition to the item level to study performance of different demographic or latent groups.

INDEX WORDS: Gender differences in mathematics, Gender-related DIF, TIMSS, Florida state assessment test, Algebra test items

Lessons Learned from DIF Algebra Problems

by

Natalia Alexeev

M.S, Moscow University, Russia, 1986

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2008

Lessons Learned from DIF Algebra Problems

by

Natalia Alexeev

Approved:

Major Professor: Jeremy Kilpatrick

Committee: Martha Carr
Edward Azoff
Sybilla Beckmann

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2008

Dedication

To Boris and Valery

with love and gratitude

Acknowledgments

This study would not be possible without the help of many people.

I would like to thank my adviser, Jeremy Kilpatrick, and my committee members, Martha Carr, Edward Azoff, and Sybilla Beckmann, for their help in conducting and completing my study.

Allan Cohen gave me access to the data for this study and helped me with many aspects of the study. Seock-Ho Kim gave me advice on statistical methods. My friend and fellow graduate student, Hye-Jeong Choi, was always willing to discuss my research and was very supportive. Ted Shifrin arranged a good teaching schedule for me so I was able to pursue research while continuing to teach. I am thankful to them and to everyone who helped me with research and teaching or just being supportive of me during this time.

I am grateful to my parents and my sister Nadezhda. Special thanks go to Valery and Boris, for their enormous moral and technical support around the clock, for their love, and for believing in me.

Thank you!

Table of Contents

List of Figures

List of Tables

CHAPTER 1

ALGEBRA PERFORMANCE AND GENDER EQUITY

"The pursuit of new scientific and engineering knowledge and its use in service to society requires talent, perspectives and insight that can only be assured by increasing diversity in the science, engineering, and technological workforce," writes the National Science Foundation (NSF, 2008) in announcing a new program "ADVANCE: Increasing the Participation and Advancement of Women in Academic Science and Engineering Careers." The goal of this program is "to cultivate a world-class, broadly inclusive science and engineering workforce, and expand the scientific literacy of all citizens." Our society progresses toward technological complexity, and the need for highly educated people increases every year. To meet this growing demand, according to NSF, "it is important that every American has an opportunity to achieve and to contribute in mathematics, engineering, and science." Women are graduating from college in increasing numbers. In 2005, more than half of all undergraduate degrees were awarded to women, but less than a third of the undergraduate degrees in hard science, mathematics, and engineering were awarded to women. Women are underrepresented in science, mathematics, and almost all related fields.

This issue is multifaceted, and it requires a complex approach to understand and solve the problem. One of the approaches is to look at gender differences at the high school level. This study investigated mathematical and nonmathematical characteristics of algebra and algebra-related problems on assessment tests that contribute to gender differences in performance. The study sought to identify those characteristics and elaborate on possible reasons related to teaching and learning algebra, gender equity, and fairness of testing.

Algebra as a Gateway

On March 14, 2008, π day in the mathematical community, major newspapers ran articles on the report by the National Mathematics Advisory Panel (NMAP) released the day before. Although it was not front-page news, it was sufficiently important for major newspapers to report the event. The first paragraph of the *New York Times* article summarized the report before explaining the details:

> American students' math achievement is "at a mediocre level" compared with that of their peers worldwide, according to a new report by a federal panel, which recommended that schools focus on key skills that prepare students to learn algebra. (Lewin, 2008, p. 20)

A *Washington Post* article reached the same conclusion:

> A presidential panel declared math education in the United States "broken" yesterday and called on schools to focus on ensuring that children master fundamental skills that provide the underpinnings for success in higher math and, ultimately, in high-tech jobs. (Glod, 2008, p. A06)

The NMAP was established in April of 2006 by President G. W. Bush via Executive Order 13398 to advise him and Secretary of Education Margaret Spellings on the best use of scientifically based research on the teaching and learning of mathematics with a clear emphasis on the preparation of students for entry into, and success in, algebra. According to the U.S. Department of Education press release on April 18, 2006, the following topics were among those to be addressed by the panel:

- The skills needed for students to learn algebra and be ready for higher levels of mathematics.
- The appropriate design of systems for delivering math instruction that combine elements of learning, curricula, instruction, teacher training, and standards, assessments and accountability.

- Research needs in support of mathematics education.

Secretary Spellings pointed out that the government and citizens were concerned that U.S. students were not performing as well as students from other developed countries on international mathematics and science assessment tests and that the country was losing its edge in global technological competition. She also stressed that all students need solid mathematics skills regardless of their chosen path: college or directly to the workforce.

The panel, consisting of 19 mathematicians, education experts, and psychologists, produced a 120-page report (NMAP, 2008) on the importance of preparing students for algebra, which has always been considered a gateway to later success (Mervis, 2007) and on ways to achieve that goal. The panel expressed concern that American students were not succeeding in mathematics studies:

> This Panel, diverse in experience, expertise, and philosophy, agrees broadly that the delivery system in mathematics education—the system that translates mathematical knowledge into value and ability for the next generation—is broken and must be fixed. This is not a conclusion about any single element of the system. It is about how the many parts do not now work together to achieve a result worthy of this country's values and ambitions. (NMAP, p. xiii)

However, the panel was optimistic: "On the basis of its deliberation and research, the Panel can report that America has genuine opportunities for improvement in mathematics education" (p. xiii).

U.S. students' difficulties in learning mathematics are not restricted to just one strand of mathematics. Nevertheless, performance in algebra has been seen by researchers as a main concern. The courses Algebra I and Algebra II are considered essential to higher level mathematics courses.

The NMAP (2008) reflected on the changes that happened in education in the previous 20 years and weighed in on a long battle in mathematics education on the way mathe-

matics should be taught without taking sides (Mervis, 2008), stressing the importance of understanding and practice in mastering important skills.

Economic and technological changes are transforming the world of work. To succeed in an increasingly competitive economy, all students must learn how to solve complex problems, work with sophisticated representations, and make judgments on accuracy of information (Barley & Orr, 1997; National Research Council [NRC], 2001). The standards-based reform movement in education that started in the early 1990s has had an enormous impact on curriculum, instruction, and assessment (Schmeiser, 2006). Central to this approach are curriculum or content standards that express what students should know and do, as well as alignment among all system components (Webb, 2006). Professional organizations of educators created standards for many school subjects. There are standards in fine arts education and language arts, social sciences and science education, as well as in technology and physical education. In 2000, the National Council of Teachers of Mathematics [NCTM] produced *Principles and Standards for School Mathematics*, its fourth in the series of standards documents for mathematics:

> *Principles and Standards for School Mathematics* is intended to be a resource and guide for all who make decisions that affect the mathematics education of students in prekindergarten through grade 12. The recommendations in it are grounded in the belief that all students should learn important mathematical concepts and processes with understanding. (p. ix)

The NMAP (2008) made a special effort to describe the content and demands of school algebra. It reviewed the algebra topics in current state standards, the algebra objectives in the National Assessment of Educational Progress (NAEP) Grade 12 test, and the algebra standards in Singapore's mathematics curriculum. It compiled a list of major topics of school algebra and recommended that they be used in state curriculum frameworks and state assessment tests. After defining the major topics, the panel addressed the critical skills and essential concepts that prepare students for algebra. It reviewed the critical skills in the curricula of high-performing countries in the Trends in International Mathematics and Science Study

(TIMSS), the National Council of Teachers of Mathematics (NCTM, 2006) *Curriculum Focal Points for Prekindergarten through Grade 8 Mathematics: A Quest for Coherence*, a 2007 American College Testing (ACT) survey, and an NMAP-sponsored survey of 743 teachers of introductory algebra across the country (NMAP, 2008).

On the basis of the review and taking into consideration the structure of mathematics itself, the NMAP (2008) proposed as a critical foundation of algebra, important concepts and skills essential for students to learn thoroughly prior to algebra course work. The panel recommended a coherent, focused curriculum that would include critical foundations with adequate depth and that logically progressed from less-sophisticated to more-sophisticated topics (p. xvii).

The NMAP (2008) concluded that we do not have a full understanding of how students learn algebra and how to better prepare them to enter algebra courses. However, it found indications that too many students in middle or high school algebra classes were seriously unprepared for learning even the basics of algebra:

> The types of errors these students make when attempting to solve algebraic equations reveal they do not have a firm understanding of many basic principles of arithmetic. Many students also have difficulty grasping the syntax or structure of algebraic expressions and do not understand procedures for transforming equations or why transformations are done the way they are. These and other difficulties are compounded as equations become more complex and when students attempt to solve word problems. (p. 32)

Among the recommendations, the panel suggested that research to identify early predictors of success or failure in algebra is needed; the clues for these predictors can probably be determined by closely inspecting the performance of high school students in algebra courses.

Gender Equity

In July 2008, a study on gender differences in mathematics performance (Hyde, Lindberg, Linn, Ellis, & Williams, 2008) became the subject of a news item in the mainstream media. "The Myth of the Math Gender Gap" (Park, 2008) was the title of a *Time* magazine article; "Math Study Finds Girls Are Just as Good as Boys" (Quaid, 2008) reported the Associated Press with articles appearing in different newspapers; "Gender Gap Theory Doesn't Add Up" was the headline on the NBC Nightly News. "Stereotypes are very, very resistant to change, but as a scientist I have to challenge them with data," stated the lead researcher of the study Janet Hyde, professor of psychology at the University of Wisconsin (Fisher, 2008). Hyde et al. studied mathematics scores from 10 state examinations now mandated annually under the No Child Left Behind Act (NCLB) of 2001. They found no meaningful gender difference; the effect sizes were very small, all less than 0.15. They pointed to some evidence of slightly greater male variability in the scores, but they could not explain its causes.

However, it is too early to claim we have reached gender equity in mathematics achievement, according to the *Handbook for Achieving Gender Equity Through Education* (Klein, et al., 2007), the collective work of over 200 gender-equity experts. This book addressed many issues in gender equity in our society and globally, from early childhood education to government policies. This more than 700-page book is a major update and expansion of the previous handbook in 1985. Klein, Kramarae, and Richardson (2007) wrote in the introduction:

> Our empirical research and experience make clear that gender continues to be an important organizing and disempowering principle in the school system. Equity in education is not only a matter of numbers. . . . The *Handbook* includes facts, assumptions, strategies, practices, and content related to curriculum, governance, socialization, psychology, working with diverse populations and multiple educational levels. It is a landmark and definitive piece of work for anyone studying, teaching, or interested in gender equity in education. (p. 1)

According to the handbook, gender differences in state assessment tests and course taking in mathematics are minimal. There are large gender differences, however, in mathematics-related majors and careers. In 2005, more than half (58%) of all undergraduate degrees were awarded to women, but in the hard sciences, mathematics, and engineering, only 26% were awarded to women. In engineering, only 20% of all undergraduate degrees were awarded to women, and about the same percentage (24%) enrolled in graduate school. In mathematics, the situation seems better; 45% of undergraduate degrees were awarded to women (NSF, 2006). Hyde et al. (2008) cited a similar number as proof that girls are not different from boys even in careers in mathematics. However, the number probably includes double majors in mathematics and education. Many states require teachers to have a major in the field they are teaching. This requirement probably explains the much lower percentage (33%) of women enrolled in graduate school in mathematics compared with the percentage of undergraduate degrees in mathematics awarded to women. Although it is a positive sign that more and more mathematics teachers have a major in the field in which they are teaching, women are still underrepresented in mathematics and related fields. Only 27% of the doctoral degrees in mathematics were awarded to women in 2005, although the percentage is higher for master's degrees (44%; NSF, 2006).

To understand why women are underrepresented in the career fields of mathematics, engineering, and science, one needs to examine the issue from many angles: cognitive, social, and psychological perspectives among them. One needs to look for indications of future differences in K–12 education. In 2008, the mean scores on SAT Mathematics (SAT-M) for college-bound seniors were different for male and female students: 533 to 500 (College Board, 2008). The effect size of the difference is not large; however, it should not be ignored. McGraw, Lubienski, and Strutchens (2006) confirmed that male students performed better than female students in mathematics. Although the differences in mathematics achievement on NAEP from 1990 to 2003 were small, they did not diminish over the 13 years. The largest

differences were in the strands of measurement and of number and operations in Grades 8 and 12 and of geometry in Grade 12.

Female students are underrepresented in extra-curricular mathematics activities. The popular Mathcounts competition for middle school students has about 10% girls at the national-level competition (Lacampagne, Campbell, Damarin, Herzig, & Vogt, 2007). According to the official web site of the Mathematical Association of America (MAA) American Mathematics Competition (AMC), in 2008, about 46% of AMC 10A[1] were female students. Although the participation rate was not low, the average score was quite different from that of male students: 53.4 and 62.6, respectively. On AMC 12A in the same year, the percent of female students went down to 43%. Their average score was 60.8, whereas for male students the average was 69.3 (AMC, 2008). The differences cannot be explained by social factors alone. They may be partially explained, however, by difference in strategies used by female and male students. Gould (as cited in Lacampagne et al., 2007, p. 240) found differences in problem-solving strategies by male and female students of a similar mathematical background when solving novel problems. The female students tended to rely on procedural, rule-bound approaches, and they were less comfortable with the idea of logical equivalence.

One of the findings of research is that male and female students differ in performance on high-level tests such as the SAT-M, ACT Mathematics, or Graduate Record Examination (GRE) Quantitative (Langenfeld, 1997; Snyder, Tan, & Hoffman, 2006). Mathematics fact retrieval is a significant predictor of the accuracy of SAT-M performance. However, fact retrieval is not a significant predictor of the time it takes the examinee to complete a test (Royer & Garofoli, 2005). When does the gender difference in mathematics fact retrieval start to manifest itself? Several researchers have concluded that it happens in elementary school and may be as early as kindergarten. First-grade girls were found to prefer strategies using manipulatives, whereas boys preferred to use retrieval and decomposition in simple addition

[1]There are two exams given, A and B, to accommodate different schedules for spring break.

and subtraction problems (Carr & Jessup, 1997; Carr, Jessup, & Fuller, 1999). Another study (Fennema, Carpenter, Jacobs, Franke, & Levi, 1998) of strategy use in elementary school children from first grade to third grade found a similar pattern: Girls preferred more concrete strategies, whereas boys preferred retrieval and decomposition. In all these studies, the difference in strategy use did not affect performance on the test. Maybe the differences in strategy use in elementary school are manifested as differences in performance years later.

Assessment and Test Fairness

The use of assessment results for evaluative purposes has become legislatively formalized (Schmeiser, 2006). After enactment of the No Child Left Behind Act (NCLB, 2001), assessment tests were used for evaluating not only students' progress but also that of teachers, schools, and districts. Therefore, achievement tests have a big impact on instruction. One task group of the NMAP (2008) worked on assessment. The group addressed five primary questions, two about test content and performance categories and three about item and test design. The focus on item design was on (a) how multiple-choice and various kinds of constructed-response items affect performance, (b) how nonmathematical sources of difficulty or confusion influence performance, and (c) use of calculators on the test. The recommendation of the task group was as follows:

> Much more attention should be paid to the mathematical knowledge being assessed by a particular item and to the extent to which the item addresses that knowledge. (p. 61)

The group said that a better collaboration between mathematicians, mathematics educators, teachers, and psychometricians would help to address this issue. The group also stressed the research needs in test and item design:

> More research is needed on test item design features and how they influence the measurement of the knowledge, skills, and abilities that students use when solving mathematics problems on achievement tests. These design features might have differential impacts across various groups (e.g., gender, race, English language learners). (p. 61)

Test fairness is a very important issue. Researchers should study performance on assessment tests by different demographic groups. Reed et al. (2007) said it best:

> The purpose of gender equity is to protect the rights and privileges of males and females so that both receive equitable and correspondingly fair treatment in the educational system. Testing is an integral part of the educational process for the purposes of institutional accountability, as well as for feedback and monitoring of individual student progress. Although significant strides have been made toward developing professional guidelines to eliminate bias in test instruments and the misuse of test results, it requires continual monitoring to ensure that attention is paid to issues of equity by ethnicity, socialeconomic strata, and gender in the (a) construction of published and teacher-made tests, (b) their administration, scoring, and reporting, and (c) the uses and interpretations of data from the results. (p. 167)

Testing companies and state test development teams employ comprehensive sensitivity reviews and statistical monitoring procedures, such as *differential item functioning* (DIF), to reduce the number of items biased against any demographic group (Zieky, 2006). A test item displays DIF if the measurement properties of the item are different for different subpopulations that are matched with respect to ability or knowledge (Angoff, 1993; Hanson, 1998). DIF studies can help researchers identify what item content, format, or structure factors may be related to the differential performance of a demographic group. DIF statistics in themselves do not address the issue of contributing factors, but further analysis can determine these factors (Reed et al., 2007). These studies are an important part of test fairness reviews and studies of the validity of test results (Mendes-Barnett & Ercikan, 2006; Zieky, 2006). A DIF item is not necessarily a biased item; however, it should be closely examined by specialists.

Research Questions

As an instructor with more than 10 years of experience in teaching introductory mathematics courses in college, I learned firsthand that many of my students lack basic skills in algebra. Even when I think that I know all the mistakes that they can make, they surprise me with one I have never seen before. Most of their mistakes, however, are quite common. Although there are numerous studies on learning algebra, I did not find any that classified these mistakes. I was planning to classify them when my research plans changed because of classes in educational measurement I was taking and a project I did for one of my classes. The project, which was on the growth of strategy use for male and female students in elementary school, made me think about gender differences in mathematics performance. After thinking about how strategy use in elementary school does not appear to affect performance in the early grades but may affect it in high school, I began to wonder whether small differences in performance on mathematics assessments in high school might affect women's representation in mathematics-related careers. It is a really big question to address. I wanted to contribute to addressing it, however. I have combined my interest in basic algebra learning, gender differences, and educational measurement methods in studying gender-related DIF items on algebra assessment tests to try to understand the types and characteristics of items. I was delighted to know that this problem was not important just to me. The conclusions of the NMAP (2008), gender-equity experts, and assessment and evaluation experts gave me this assurance. I agree with Lacampagne et al. (2007) that we need "women to enter the field of mathematics, not only for reasons of fairness and equity, but also because our nation is woefully short of the mathematical talent needed to keep the United States at the forefront of science and technology" (p. 250).

The purpose of this study was to take a closer look at the performance of male and female students on algebra problems on the Florida Comprehensive Assessment Test (FCAT) and in the TIMSS. I addressed the following questions:

1. What are the characteristics of algebra and algebra-related items on which male and female students perform differently on the Florida Comprehensive Assessment Test in 2001, 2002, and 2003, in Grades 8, 9, and 10?

2. How do these characteristics of algebra and algebra-related items change across grades (8–10) and years (2001–2003) ?

3. How do the characteristics of released algebra-related items in Grade 8 TIMSS 2003 on which male and female students perform differently compare with the characteristics identified for the FCAT?

CHAPTER 2

LITERATURE REVIEW

Gender Differences in Mathematics Performance

Gender differences in mathematics performance have been extensively studied in the last 30 years. Two key works in the 1970s sparked the interest in the subject: a study by Lucy Sells (1973) about women at the University of California at Berkeley, and publications on math anxiety by Sheila Tobias (1976, 1978). The critical barrier to women's participation in high-status science and technological fields was seen in the failure to study mathematics. Only 8% of female students entering Berkeley in 1972 had 4 years of high school mathematics, whereas 57% of male students did (Sells, 1973). The report received a lot of attention. Government organizations responded with programs and grants, and many research papers and conferences followed (Chipman, 2005).

Now, female students are taking high school mathematics courses at the same rate as male students. Among high school graduates, more female students have taken Geometry and Algebra II courses than male students (77% vs. 74%, and 64% vs. 60%). The same percentage of students of both genders have taken precalculus (23%) and calculus (11%) (Huang, Taddese, & Walter, 2000; Lacampagne et al., 2007). The only gender difference was in Advanced Placement Calculus exams. In 2008, 49% of students taking exam in Calculus AB were female students, whereas only 41% in Calculus BC were female (College Board, 2008). Regardless of this progress, women are still underrepresented in mathematics and related fields. Government and private organizations continue their work in attracting more women into technological jobs. Many grants and programs are available to help increase women's participation (Lacampagne et al., 2007). Research in gender differences has gained

acceptance and respect in mathematics education research communities (McGraw et al., 2006). What have we learned about gender differences?

The general conclusion from studies of standardized tests in elementary and middle school is that the gender differences in performance are small, and usually there is no advantage for either gender or a very slight advantage for girls (Hyde, Fennema, & Lamon, 1990; Willingham & Cole, 1997). The situation changes in high school. Many studies in the United States report that gender-related differences favoring male students in mathematics test performance tend to increase with age (Gallagher & De Lisi, 1994; Lubienski, McGraw, & Strutchens, 2004).

Between 1990 and 2005, the NAEP tests score increased for male and female students. For eighth graders the small gap in performance, 1 to 4 points, favoring male students persisted over the years (McGraw et al., 2006). The percentage of male students scoring at or above the proficient level almost doubled to 31% and did double for female students to 28%. In 12th grade between 1990 and 2000, male students' scores remained the same, and female students' scores decreased (Lacampagne et al., 2007). McGraw et al. examined gender differences on NAEP by mathematical strands. In the eighth grade, male students performed better in number, data analysis, and measurement. The largest differences were found in Grade 12 in geometry, number, and measurement. The differences tended to be larger at the upper end of the score distribution. McGraw et al. also found that attitudes and self-concepts related to mathematics were more negative for female students than for male students.

At the same time, other researchers found that the magnitude of gender difference on standardized mathematics tests in high school had declined over time (Hyde et al., 1990). According to Hyde et al. (2008), the general population in Grades 2 to 11 after 2000 no longer showed gender differences in math skills.

The one consistent finding of the research on gender in mathematics over the last 30 years has been that male students perform better than female students on standardized college and graduate school admission tests, such as the SAT-M, ACT-M, and GRE Quantitative tests.

Male students outperformed female students on the SAT-M for at least 40 years; in 2005, the gap was about 30 points (Langenfeld, 1997; Snyder et al., 2006). Variability in performance on these tests is much higher for male students than for female students (Willingham & Cole, 1997), and gender differences in test performance are more pronounced in the high range of ability (Benbow & Stanley, 1983).

Although girls lag behind boys on standardized tests, numerous studies have found that they have higher grades (Dwyer & Johnson, 1997; Kimball, 1989). Bridgeman and Lewis (1996), who studied 30,000 students, found that the female students outperformed the male students in undergraduate calculus classes at every level of SAT-M scores. Benbow and Stanley (1982) showed that in a highly select group of boys and girls, the boys performed significantly better on the SAT-M, whereas the girls had significantly better mathematics grades when they took the same demanding mathematics courses as the boys.

Gender differences in mathematics achievement are small compared with the differences within gender (Gallagher & Kaufman, 2005); nevertheless, researchers have found differences between male and female students in multiple areas of mathematics. Female students appear to perform better on algebra problems, whereas male students appear to do better on problem solving. Geometry and measurement items are also easier for male students (Willingham & Cole, 1997; Zenisky, Hambleton, & Robin, 2003b). One of the explanations is that male students have better spatial cognition ability (Battista, 1990). Spatial reasoning can be an important component in solving many types of mathematics problems (Geary, Saults, Liu, & Hoard, 2000; Nuttall, Casey, & Pezaris, 2005). Royer and Garofoli (2005) found that spatial ability is a significant predictor of speed of solution for a set of SAT-M items, although not of accuracy.

Mathematics fact retrieval was a significant predictor of accuracy on the mathematics assessment tests, and male students were better in mathematics fact retrieval (Royer & Garofoli, 2005; Geary et al., 2000). Gender differences in mathematics strategy use for addition and subtraction have been traced to elementary school children (Carr & Jessup, 1997;

Fennema et al., 1998). First grade female students were found to prefer strategies involving manipulatives, whereas male students were more likely to use retrieval (Carr & Jessup, 1997; Carr et al., 1999). Similar results were found in a longitudinal study of young children by Fennema et al. (1998). Carr and Alexeev (2008) found that fluency in mathematics fact retrieval was a significant predictor of the rate of growth in use of cognitive strategies. There are fewer studies of gender differences in strategy use by older students. Gallagher and De Lisi (1994) found that female students tend to use more conventional strategies to solve SAT-M problems, whereas male students are more willing to use unconventional strategies. However, no gender differences were found in the strategies used by male and female students to solve geometry problems (Battista, 1990).

Gender differences in mathematics achievement have been studied from many different perspectives: psychobiosocial (Halpern, Wai, & Saw, 2005), cognitive (Byrnes, 2005; Casey, Nuttall, & Pezaris, 1997; Royer & Garofoli, 2005), cultural (Byrnes, Hong, & Xing, 1997; A. S. Cohen & Ibarra, 2005). Students' self-efficacy, attitude toward mathematics, and self-confidence were found to be related to mathematics performance (Ansell & Doerr, 2000; Lubienski, 2000; Lubienski et al., 2004).

Gender differences continue to be an important topic for research and discussion. Some researchers argue that studying small gender differences might be doing more harm than good by conforming to stereotypes (Boaler, 2003). Others suggest that research on gender differences should pay more attention to earlier grades (Fennema et al., 1998). Studying gender differences in conjunction with socio-economic status and race-ethnicity is another suggested approach (Lacampagne et al., 2007). Regardless of the approach, these studies need to ask the right questions, and gender differences should not be phrased in a putative manner (Lacampagne et al., 2007; Caplan & Caplan, 2005).

Differential Item Functioning

If differential item functioning (DIF) exists across different groups, then research should be conducted to detect and eliminate the aspects of test design and format that may be unfair for participating groups (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 1999, p. 81).

DIF indicates that an item may be measuring something different from or in addition to what it was constructed to measure (Ackerman, 1992). It may be that the item is unfair to one or more subpopulations, although not all items classified as displaying DIF are necessarily unfair (Livingston, 2006; Zieky, 2006). DIF methods measure test invariance: whether the test is performing in the same manner for different groups of examinees (Zumbo, 2007). The most commonly studied cases are gender-related DIF and race-related DIF. Test companies perform these studies to ensure the fairness of the test and validity of results (Mendes-Barnett & Ercikan, 2006).

Many different methods to detect DIF have been developed.[1] According to Zumbo (2007), there are three major frameworks for thinking about DIF: modeling item responses using contingency tables or regression models, item response theory, and multidimensional models. All these frameworks have a working definition of DIF and methods of detecting DIF.

In modeling item responses using contingency tables or regression, examinees are matched on their ability prior to examining group differences. Matching is usually done on the total score on the test. This framework includes two major classes of DIF detection: Mantel-Haenszel (MH) (Holland & Thayer, 1988) and logistic regression approaches (Swaminathan & Rogers, 1990).

In item response theory (IRT), the main focus is on differences in the item characteristic curves (ICC) for each group. The most common methods of detecting DIF are signed area

[1]The method used in this study is discussed in more detail in chapter 3 on page 37.

tests, unsigned area tests, and nested model testing using likelihood ratio tests (Zumbo, 2007).

In the multidimensional models framework, the main assumption is that all tests are multidimensional even though only one primary dimension of the test is of interest. Researchers have not been very successful in identifying sources of DIF by studying individual items with high levels of DIF (Zumbo, 2007). The disjunction between substantive and statistical analysis is not a new issue, and it represents a major deficiency in studying group differences (Gierl, 2005). The basis of the multidimensionality framework is discussed in Ackerman (1992). The most representative method of this framework is the simultaneous item bias test (SIBTEST) (Shealy & Stout, 1993). The approach implemented in SIBTEST allows one to investigate potential sources of multidimensionality that may cause DIF. This method requires studying bundles of items as opposed to individual items. Roussos and Stout (1996) developed a multidimensionality-based DIF analysis paradigm to connect substantive and statistical DIF analysis and link it to SIBTEST (see Gierl, 2005, for a description and a discussion of new developments). During the substantive stage of analysis the dimensional structure of the test is evaluated. Some of the organizing principles are based either on test specifications or content and others on psychological analysis, whereas empirical analysis can also be used. In the second statistical stage of the analysis, SIBTEST is used to test the hypothesis and quantify the size of DIF. The theory-based hypothetic-deductive strategy can be used with any of DIF methods and not only with SIBTEST (Zumbo, 2007).

Even with these well-developed strategies for identifying DIF, the causes of DIF in many comparisons are still elusive. The desire to learn why DIF occurs has led DIF researchers to a new generation of conceptual and methodological development (for a brief review, see Zumbo, 2007).

Gender-related DIF in Mathematics

Most researchers on gender-related DIF in mathematics items have examined these differences as a function of a group membership. A different approach to examining gender-related DIF items is to make the characteristics of the items the central focus. This approach can potentially lead to finding characteristics of the items that may be measuring a dimension or dimensions of performance that were not intended, and in this way, help researchers better understand what may cause the gender DIF. Li, Cohen, and Ibarra (A. S. Cohen & Ibarra, 2005; Ibarra & Cohen, 1997; Li, 2002; Li, Cohen, & Ibarra, 2004) examined DIF as a function of structural characteristics of items. A coding system of item characteristics based on multicontext theory (Ibarra, 1996) was used to predict gender DIF on a college mathematics placement test. According to multicontext theory, people from different cultural backgrounds may have different expectations of the kinds of information to be communicated from their environment. These expectations in turn result in differences in ways of processing information, one result of which is that performance on test items is negatively affected. The coding system developed by Li et al. (2004) consisted of two domains: a social-cultural domain and a mathematics problem domain. Item format was included in the mathematics domain. The structural characteristics considered in the social-cultural component were the following: the nature of the topic, real-world-applicability, and spatial reasoning. In the mathematical component, the following characteristics were tested: algebra or geometry, definition-based question, indefinite answer question, symbol problems, mathematical reasoning, congruity, and connection between answers and solutions. Results suggested that gender DIF was related to certain structural characteristics of the items. The multicontext coding scheme was correct in 76% of the predictions of male DIF, female DIF, or no DIF. The report did not specify which characteristics were better predictors than others.

In a study examining strategy use on multiple-choice (MC) and free-response (FR) test items, Gallagher (1992) found that female students used the same strategies, computational or algorithmic, regardless of the format of the item, whereas male students used different

strategies, algorithmic on FR items and less computational strategies on MC, including working from the given answers. Male students were also more likely than female students to guess the answer on MC items (Gallagher & Kaufman, 2005). Other researchers found that male students do better on MC items except for algebra items, and female students do better on constructed response items (Burton, 1996; Garner & Engelhard, 1999; Henderson, 2001); however, there is no plausible explanation. In the Henderson study, there were no DIF items among the gridded-response items, and the MC items had more DIF items favoring male students than female students.

There are several studies on context and no-context items on a test. Koedinger and his colleagues (Koedinger, Alibali, & Nathan, 2008; Koedinger & Nathan, 2004; Nathan & Koedinger, 2000a, 2000b) found that students do better on items in context. Their conclusion is limited because they used only very basic items concerning retail sales. They could conclude only that students performed better when the context is retail than when the item is without a context. Swafford (1980) found that male students performed better than female students on consumer-oriented word problems in algebra. Several other studies (Hyde et al., 1990; Mendes-Barnett & Ercikan, 2006; Ryan & Chiu, 2001) found that male students did relatively better on word problems, but the topic of the context was not specified. Kaminski, Sloutsky, and Heckler (2008) concluded that learning a concept in a generic instantiation allows for transfer, whereas learning a concept in a context hinders transfer to a different context. They concluded that in assessing mathematics performance, the topic of the context should be more generic in nature. It is possible that male students have an advantage when the topic is more familiar to them than to female students. According to some studies on gender differences (A. S. Cohen & Ibarra, 2005; Gallagher & De Lisi, 1994; Gallagher et al., 2000), there are “male” and “female” topics. Examples often given are race cars as a male topic and dresses on sale as a female topic. Harris and Carlton (1993) mentioned that items with topics such as money, time, fractions, rate, linear and liquid measure, averages, and areas resulted

in mean DIF values favoring male students and only percentages and counting topics favored female students.

Items with noncomputed solutions have been identified as a possible contributor to DIF. Noncomputed solutions have as answers formulas, expressions, and so forth (Harris & Carlton, 1993). According to Harris and Carlton, noncomputed solution items favor female students. Not everyone agrees on that issue, Hyde et al. (1990) found that female students are better at computational tasks.

A. S. Cohen and Ibarra (2005) found that items with graphs or figures are less likely to be biased against female students if these objects are commonly found in the real world or have a practical application. Other researchers (Harris & Carlton, 1993; Mendes-Barnett & Ercikan, 2006; Ryan & Chiu, 2001) concluded that items with visual stimuli such as figures, graphs, or tables favored male students. They concluded that the probable explanation is a gender difference in spatial ability.

Bielinski and Davison (1998) found a gender-by-item-difficulty interaction. Their study on nine forms of a basic skill test in mathematics in Grade 8 confirmed that male students tended to outperform female students on the hardest items, whereas female students outperformed male students on the easiest items.

Many studies have been done on gender differences in mathematics and gender-related DIF. Nevertheless, many questions remain open and are waiting for researchers to address them. With my study I wanted to contribute to understanding why more women do not pursue careers in mathematics-related fields.

CHAPTER 3

STUDY 1: FLORIDA COMPREHENSIVE ASSESSMENT TEST 2001

To address the research questions, I use mixed methods: a quantitative and qualitative design. I used data from the Florida Comprehensive Assessment Test (FCAT): students' responses and test items. The focus of the study was test items, for which I built a classification system that I subsequently tested, redesigned, and tested again. These studies were performed on data from the FCAT, one for each of 3 years. The analysis and results for each year are reported in separate chapters.

Method

The first study was exploratory. I constructed an initial classification based on previous research in mathematics education, assessment, and measurement. I included other attributes that I could think of and classify. Although previous research indicated that some item attributes can contribute to differential item functioning (DIF), I made no hypothesis prior to the statistical analysis.

In this method section, I describe the data and the objectives of the FCAT. In the procedure section, I describe the process of building and revising the item classification and the statistical method that I used to evaluate DIF and to identify categories that contribute to DIF. Finally, I present the results and conclusion.

The Florida Comprehensive Assessment Test

The FCAT is administered annually to all Florida public school students in Grades 3–11. The FCAT is a high-stakes test for students as well as for educators. Achieving a passing score

on the Grade 10 FCAT Reading and the Grade 10 FCAT Mathematics tests is a statewide requirement for graduation. FCAT results serve as a major source of data for determining the grades that the Florida Department of Education (FDOE) assigns to schools and reports annually. According to the FCAT *Handbook* (FDOE, 2005), the FCAT consists of two parts: norm-referenced tests (NRT) in reading and mathematics, which compare the achievement of Florida students that of with students nationwide; and criterion-referenced tests (CRT) in reading, mathematics, science, and writing, which measure student progress toward meeting the Sunshine State Standards (SSS) benchmarks (FDOE, 1996). Both the FCAT SSS and the FCAT NRT are used to measure achievement and guide instruction of individual students. From 2000 to 2004, the test used for the NRT was the Stanford Achievement Test, Ninth Edition (Stanford 9 or SAT 9, published by Harcourt Assessment). The FCAT NRT was not analyzed in this study. All references to the FCAT in this report are references to the criterion-referenced SSS FCAT.

The FCAT is the latest and most comprehensive development in statewide educational assessment, which started more than 30 years ago with the Florida Educational Accountability Act of 1971. The first operational FCAT Mathematics was administered in 1998 in Grades 5, 8, and 10. In 1999, the Florida Legislature expanded the statewide assessment program to include reading and mathematics in Grades 3–10 and required students to pass the Grade 10 FCAT SSS in reading and mathematics in order to graduate from high school. This requirement was first applied to the 2003 graduating class (FDOE, 2005).

Data Sample

Data were analyzed for Grades 8 through 10 from 2001. I used one form for each grade because that was all I had access to. Only students for which no special accommodation was made were retained in the analysis. The distribution of the data sample by gender, grade, and race is provided in Table 1. The students' racial category was not used in the analysis;

Table 1:
Distribution of 2001 Data Sample by Grade, Gender, and Race

	Gender		Race[a]			
Grade	Female	Male	White	Black	Hispanic	Total
8	4,807	4,242	5,199	2,084	1,431	9,049
9	4,982	4,734	5,573	2,291	1,553	9,716
10	4,186	3,699	4,712	1,679	1,199	7,885

[a] Data for other racial groups are omitted.

it is presented only to help describe the population. A description of the item data is given in the next section.

Procedure

There were 150 items on the SSS FCAT Mathematics 2001 for Grades 8, 9, and 10 (50 for each grade). For this study, I included all number items, all measurement items, some very basic geometry items, some items from data analysis and probability, and all algebra items. The rationale for including additional items was based on several arguments. Skills in arithmetic translate to some degree to skills in algebra because one view of algebra is as generalized arithmetic (Usiskin, 1988). At the same time, algebra has roots in geometry (Charbonneau, 1996; Radford, 2001). Knowing the areas of simple figures, proportionality of lengths (Radford, 1996), and proportionality in general (Post, Behr, & Lesh, 1988) is essential to understanding algebra. Most area and perimeter items had been classified in the measurement strand, and items on proportionality could be found in the number, measurement, and geometry strands. The three items from the data analysis and probability strand that were included in this study can easily be categorized as algebra or arithmetic items.

They were probably classified in the data analysis and probability strand because of their context: Two items were about population, and the third one included data on planets, but the item itself was about estimating percentage. Table 2 shows the number of items in each grade that were included in this study and the total number of items in each strand on the FCAT.

Table 2:
Distribution of 2001 FCAT Items by Strand and Grade

	Grade			
Strand	8	9	10	Total
Number	13	8	11	32
Measurement	10	9	8	27
Geometry	1 (6)	5 (12)	3 (10)	9 (28)
Algebra	11	13	13	37
Data analysis and probability	0 (10)	2 (8)	1 (8)	3 (26)
Total	35 (50)	37 (50)	36 (50)	108 (150)

Note. The total number of items is in parentheses when not all items in a strand were used in the analysis.

Classification of Items

Content Domain

The main purpose of the study was to identify the characteristics of items on which male and female students with the same ability perform differently. Previous research on the classification of mathematics items suggested the content domain as the first important characteristic. The content of the item was classified more specifically than just a strand such as algebra or geometry. The starting point of the classification of the content domain for Study 1 was the

classification in a study (Kilpatrick, Mesa, & Sloane, 2007) of U.S. students' performance in algebra in the Trends in International Mathematics and Science Study (TIMSS), which in turn was a modified classification from a special National Assessment of Educational Progress (NAEP) study (Kilpatrick & Gieger, 2000) of eighth graders who were taking or had taken algebra. Because not only algebra items were studied but also related strands, the benchmarks from Sunshine State Standards (FDOE, 1996) were very helpful in classifying algebra-related items and refining the algebra item classification.

The classification that was used for the first stage of analysis had eight categories: number, geometrical measurement, informal algebra, pattern, setting-up/translation, functions, algebraic manipulations, and other. Each category has several subcategories.

- *Number*: This category includes items with basic understanding of numbers, forms of numbers, and properties of numbers. Items in this category are either no context or minimal context. This category is significantly narrower than the number strand in the Sunshine State Standards (FDOE, 1996).
  - Form. This subsection includes scientific notation, position on the number line, comparing fractions, and equivalent forms of number.
  - Properties of numbers. This subsection includes items that ask about subsets of numbers that satisfy certain conditions.
  - Order of operations. In addition to direct questions about order of operations, items that require performing calculations are included (no substitution, just direct calculation).
- *Geometrical measurement*: Items are on basic measurement items such as finding perimeter, area, or volume. Basic formulas are usually provided either in the item or in the table at the beginning of the FCAT test booklet.
  - Length, perimeter, area, volume, and Pythagorean Theorem using basic formulas. (Formulas are provided.) Items require simple substituting and calculating. They

may require applying the Pythagorean Theorem in very simple cases, usually calculating the hypotenuse, but in some cases legs. The theorem is given in the formula table.

  - Items require finding lengths (but not proportion in similar figures and maps) or application of perimeter, area, volume, or the Pythagorean Theorem in a more complex setting such as combining two or more areas, or finding the perimeter of a complicated figure. The items are usually given in a real-world context.
  - Similarity, ratio: Items are on proportion of lengths only in similar geometric figures and map scales. A picture may be or may not be included. Other types of proportion are included in the informal algebra category.
  - Other: Combinations of subcategories or other geometrical measurement items that do not fit in other subcategories are included here.

- *Informal algebra*: This category includes a wide variety of items with nongeometrical proportions and rate. They are arithmetic items that do not involve variables but may involve modeling of arithmetic expressions. The items are given in a real-world context.
  - The first subcategory includes items with percentage, nongeometrical proportion, or ratio, but not the items that include rate.
  - Rate: This subcategory includes items with speed, acceleration, cost per hour and other rates. The student is not necessarily required to find a rate, but rate is involved in them.

- *Pattern*: This category includes identification of a rule, finding one or more terms, or matching given general rules.
  - Number sequence: This subcategory is specifically reserved for number sequences. If two number sequences are given with relations between them (as in a table), then it is a function.

  - Pictorial: It includes items with tiles, where one is to find a specific design for the tile or expanded figures.
  - Function: The function can be given as a table of values, a graph, or a verbal description. In a verbal description, it is usually described as a rate of change.
  - Other relations: The item includes combinations of the above subcategories: for example, a pictorial representation and a function given by a table.

- *Setting-up, translation*: Items involve translating real-world situations into equations, inequalities, or functions. The item usually requires matching because items are multiple choice (or gridded response, which allows only a numerical answer). Therefore, items ask specifically about an expression. This category is different from pattern/functions in the sense that the rule for setting up a function is given. The situation should be just translated into a function, whereas in patterns, the rule should be determined. This category is subdivided into three subcategories:
  - Equation
  - Inequality
  - Function

- *Functions*: Items involve understanding functions; finding one variable given the other. This category is different from patterns in the way that the function is given: by either an equation or a graph.
  - Finding a dependent or independent variable. The task involves either finding the value of the variable on a given graph or substituting for a given variable in the equation and calculating another variable.
  - Interpreting or finding parameters such as slope.
  - Determining the impact of changing parameters.
  - Matching a function and its graph.

- *Algebraic manipulations*: Given an equation, expression, or inequality, items ask students to solve or simplify. The item can be given in context or not. All types of equations are in this category because there are not too many of them in the FCAT. For other types of tests, this category may be extended. The category has three subcategories:
  - Simplifying expression
  - Solving equation
  - Solving inequality
- *Other* includes either combinations of the previous eight categories or categories that are not listed above. The purpose of this category is to group items that do not fit into a single subcategory described above so that the category can be refined later.

The distribution of items by category and original strand on the FCAT is shown in Table 3. Two graduate students in mathematics and a lecturer in mathematics helped in refining the classification. They classified the items according to my preliminary classification, and then we discussed all items that we disagreed on. I did not calculate interrater agreement. After I decided that the content classification was reasonable, two graduate students in mathematics education coded selected items: 5 items at Grade 8 and additional items from Grade 9 in the order presented on the test. The first rater coded 27 items and the second rater coded 23 items (25% and 21% of the sample, respectively). The interrater reliability for the first rater and me was moderate (Cohen's kappa $= .71$) and for the second rater and me was high (Cohen's kappa $= .81$). In a three-way discussion afterward, the two raters and I came to an agreement on every coded item. In some cases, the descriptions of the subcategories were not very clear. Discussion of the remaining disagreements led to refinement of some categories. The raters pointed out to me that some items can easily belong to different categories depending on the way the item is solved. An example is Item 21 from Grade 9 (this item was released by FDOE):

> A circle that has a radius of 5 inches has an area of 25π square inches. If the radius is doubled, what is the area of the new circle?

Table 3:
Distribution of 2001 FCAT Items by Strand and Category

	Strand					
Category	Number	Measurement	Geometry	Algebra	Data analysis & probability	Total
Number	11					11
Geometrical measurement	2	18	5	1		26
Informal algebra	18	6		1	1	26
Pattern		1	1	14	1	17
Setting-up/ translation		1		4		5
Functions			2	9	1	12
Algebraic manipulations	1			3		4
Other		1	1	5		7
Total	32	27	9	37	3	108

A. 10π square inches B. 50π square inches C. 100π square inches D. 200π square inches (Source: FDOE[1] (FCAT, 2006b, p. 21))

From my point of view, the clear intention of the item designers was to test students' understanding of the impact of changing parameters: If the radius is doubled, then the area will quadruple. The item would be classified in the subcategory functions/determine the impact

[1]The Florida Comprehensive Assessment Test (FCAT) Mathematics items appear by permission of the Florida Department of Education, Office of Assessment, Tallahassee, Florida 32399-0400.

of changing parameters. However, the item can be done in a different way. If a student finds the radius $r = 5 \times 2$ and substitutes it into the formula for the area of a circle, then the item would be classified in the subcategory of geometrical measurement/basic area, perimeter.

After the discussion, I went over the classification again with the clear intention to find items that could be classified differently depending on the solution method and checking my coding again for all items. After this recoding, I moved six items into the category of other.

Cognitive Complexity

Another common dimension in classification is cognitive complexity or cognitive demand. Many classifications of cognitive complexity have three levels: low, moderate, and high. In the current study, only multiple-choice (MC) and gridded-response (GR) items were considered; short response and extended response items were not included. I had access only to these types of items. In literature, GR items are called constructed-response items. Therefore, very few if any items were from the highest level of complexity. The difference between low and moderate levels is almost negligible. None of my attempts to find or build a complexity classification more suitable to the current analysis was successful. As a result, I did not use the cognitive complexity dimension in classifying the test items.

Form of the Item

Another common dimension for classifying items is the representations used, which usually include the statement of the item and the form of the answers. In Kilpatrick et al. (2007), numerical, verbal, graphical, symbolic, and pictorial categories were used. I decided to use a different approach and test which characteristics of the form, apparent and subtle, contribute to DIF. Below I explain the characteristics I tested in this study.

Type of Response

For the GR items a numerical answer was required, and several possible forms of that answer were accepted as correct. The data available to me did not contain actual responses for the GR items but only whether the response was correct or incorrect. The data included students' responses for all the MC items. In the present study, the distribution of items for each grade was approximately one third GR and two thirds MC items.

Topic of the Context

Items were classified as no-context or context items. After studying the FCAT items, I decided to introduce seven topics for the first study:

- *No context*
- *Sport/recreation/transportation*: Athletic races; recreation activities, such as boating or bicycling; and traveling by car, ship or airplane; recreational and athletic facilities without measuring dimensions or business
- *Physical sciences*: Physics, chemistry, biology (other than population of bacteria)
- *Population*: Growth, density
- *Retail/currency/business*: Items with cost, earnings, or currency
- *Social sciences*: Geography (maps), history
- *Measuring*: Finding dimensions or area, perimeter, and volume for real world objects, such as perimeter of a picture frame or area of a runway
- *Other*: Topics not mentioned above

These topics cover well all of the items from the FCAT 2001. However, new topics may and probably will be introduced in other years. Interrater agreement for me and

Table 4:
Distribution of 2001 FCAT Items by Topic and Grade

	Grade			
Topic	8	9	10	Total
No context	3	7	2	12
Sports/recreation/transportation	6	5	8	19
Physical sciences	5	4	7	16
Population	3	4	1	8
Retail/currency/business	9	4	6	19
Social studies	2	2	3	7
Measuring	7	8	4	19
Other		3	5	8
Total	35	37	36	108

and each rater was high. The first rater coded 27 items out of 108 (25%), and Cohen's kappa was .91; the second rater coded 23 out of 108 items (21%), and Cohen's kappa was .85. After discussion, agreement was reached for all items, and only a clarification for the sport/recreation/transportation topic was necessary. The distribution of items according to topic and grade is presented in Table 4.

Other Characteristics

In addition to domain and context, I considered 14 other item characteristics. Some properties of the items are known to be important in contributing to DIF. Some were included because they were known to make items more difficult for students, although there is no evidence

that they contribute to DIF. Other characteristics were included because they have easily manifested features such as type of number or a inclusion of a table.

- *Type of solution*: A solution might be a computed or noncomputed. When a solution is computed, the answer is a number or numbers. However, some numbers are not computed: for example, items on understanding of scientific notation. Noncomputed solutions have as answers formulas, expressions, and so forth (Harris & Carlton, 1993). Items that test understanding of scientific notation also belong to noncomputed solutions in this classification. According to Harris and Carlton, noncomputed solution items favor female students. Computed solutions are divided into two groups: one-step solution and more-than-one-step solution. The criterion for one step is two numbers given in the item that are to be combined by an algebraic operation. If counting units are involved or if the student has to calculate one of the numbers, then it is not a one-step item. Interrater agreement between me and one of the raters, a graduate student in mathematics education, was high (Cohen's kappa = .91) and moderate with the other (Cohen's kappa = .65). After discussion, we reached agreement on all coded items, and a better explanation of the system was developed.

- *Type of numbers*: Numbers in an item were divided into only integers and not only integers. All numbers in the statement of the item and in the choices given in MC items were considered. For GR items, answers were not taken into consideration.

- *Figures*: Figures were divided into picture only, picture with information, or geometric figure. Classification by figure is not complicated. Some items simply have a picture that does not carry any relevant information. Others have a picture with some dimensions shown or that specifies the length the examinee has to find. The third type is a geometry figure without any real-life elements.

- *Tables*: Tables were divided into two types: for pattern and other. Like figures, this characteristic is apparent.

- *Graphs*: Graphs are classified as function graphs, the *xy*-grid, and the number line. A. S. Cohen and Ibarra (2005) found that items with graphs or figures are less likely to be biased against female students. Other researchers (Harris & Carlton, 1993; Mendes-Barnett & Ercikan, 2006; Ryan & Chiu, 2001) concluded that items with visual stimuli such as figures, graphs, or tables favored male students. There are pie graphs on the FCAT 2001 test, but of the few data analysis items included in this analysis, none has a pie graph.

- *Estimated answers*: Usually the words *approximately* or *estimate* are included in this type of item; however, rounding is not included. Items with undetermined answers such as "none of the above" or that ask the student to estimate the answer usually favor male students (A. S. Cohen & Ibarra, 2005). The guidelines for developing tasks for FCAT tests (FDOE, 2001) specifically prohibit the use of undetermined answers:

  > Responses such as "None of the Above", "All of the Above," and "Not Here" should NOT be used. Responses such as "Not Enough Information" or "Cannot Be Determined" should NOT be used unless they are a part of the benchmark being assessed. They should not be used as distracters for the sake of convenience. (pp. 9–10)

- *Answer is part of the item*: These items have an answer that is sometimes called a forced answer, and in some sense all MC items can be considered forced-answer items. In this study, if the answer cannot be found or calculated before the student looks at the choices, then the answer is part of the item. The number of these items is limited.[2] The following item is an example:

  > The distance from Tom's house to his school is 2 km to the nearest kilometer. Which of these could be the actual distance?
  >
  > a. 3 km b. 2.9 km c. 2.6 km d. 1.6 km

[2]Most items of this type are in the NRT part of the FCAT, which was not included in the current study.

- *Percentage*: Specifically the item asks the student to find a percent, or the percent is given in the statement. Harris and Carlton (1993) found that items with percents favored male students.

- *Fractions*: Decimal fractions are not included here. Fractions and especially mixed numbers are difficult to use with a calculator. There are limited number of items with fractions on the FCAT because the guideline for developing tasks for the FCAT (FDOE, 2001, p. 8) require the use of decimal notation for numbers with metric units.

- *Distracting information*. Some of the items have extraneous information. For example, in one item the maximum speed and average speed of a plane were given, whereas only average speed was required in the calculation. The item is one step, but it has distracting information.

- *Rate*: Items include quantity per unit time or per another unit. The rate is usually mph or $/hour, but is not limited to those. This characteristic is different from the content category of informal algebra/rate; the content area category can include functions, setting-up/translations, algebraic manipulations, or patterns.

- *Ratio*: Ratio items belong to at least two possible subcategories: similarity and ratio in the geometrical measurement category and proportion and ratio in informal algebra. To test whether ratio contributes to DIF, I decided to make it a separate characteristic.

- *Converting units*. Converting any units can be included. In this study it involves mostly converting nonmetric linear units; in the FCAT, a conversion table is provided to students.

- *Items with vehicles in context.* Vehicles include but are not limited to cars, planes, ships, boats, and bicycles. Items can be of different topics, such as sports/recreation/transportation, retail/currency/business, or measuring.

Table 5:
The mth Slice of a 2×2 *Contingency Table for Item Score by Group*

	Item score		
Group	Right (R)	Wrong (W)	Total (N)
Focal (f)	R_{fm}	W_{fm}	N_{fm}
Reference (r)	R_{rm}	W_{rm}	N_{rm}
Total (t)	R_{tm}	W_{tm}	N_{tm}

DIF Evaluation

Mantel-Haenszel and Standardization Procedures

Two DIF assessment procedures have been used by Educational Testing Service (ETS) since the mid 1980s: Mantel-Haenszel (Holland & Thayer, 1988) and standardization (Dorans & Kulick, 1983, 1986). According to Dorans and Holland (1993), these procedures are related and complement each other well. Both methods use total score as a measure of compatibility, but they also are flexible enough to use different ways of matching groups.

For each item, the data used in the Mantel-Haenszel (MH) method are in the form of a table, where m is a score level. For every mth slice, the contingency table shown in Table 5 has information on the number of correct and incorrect item scores for each group and totals. The group of interest is called the focal group (f), and the group used for comparison is the reference group (r).

An estimate of the constant odds ratio was given by Mantel and Haenszel (1959):

$$\alpha_{MH} = \left[\sum_i R_{rm} W_{fm} / N_{tm}\right] / \left[\sum_i R_{fm} W_{rm} / N_{tm}\right]$$

Holland and Thayer (1988) converted α_{MH} into a difference expressed in the so-called delta metric. This metric is used by ETS for expressing DIF.

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

Positive values of *MH D-DIF* favor the focal group, and negative values favor the reference group.

The standardization's item discrepancy indices are used to flag items for further visual inspection with help of graphs of empirical item response functions or differences between empirical item response functions for different groups. These indices include *STAND P-DIF* and its delta metric version *STAND D-DIF*. The former index is used more often, although the latter one has a smaller variance and correlates higher with *MH D-DIF*. More information about the MH method and the *STAND D-DIF* index can be found in Appendix A.

ETS has classification rules for the *MH D-DIF* and *STAND P-DIF* indices. An item is classified as negligible DIF (Category A) if *MH D-DIF* is not statistically different from 0 or if the absolute value of *MH D-DIF* is less than 1.0. An item is classified as large DIF (Category C) if *MH D-DIF* is significantly greater than 1.0 in absolute value, and the absolute value of *MH D-DIF* exceeds 1.5. Other items are classified as moderate DIF, Category B. The range of values of *STAND P-DIF* is between -1.0 and 1.0. Absolute values of the index less than .05 constitute negligible DIF; absolute values between .05 and .1 are indications that further inspection of the item is needed; absolute values greater than .1 are rare and require careful inspection of the items (Dorans & Holland, 1993).

I chose MH and *STAND P-DIF* for the analysis for two reasons. First, these two procedures are used by ETS, and ETS developed generally accepted guidelines for classifying the magnitude of DIF. Second, the procedures are easy to implement in programming, eliminating the need to buy expensive software. Although the results of the two procedures are highly correlated, I decided to use both of them and report both indices in most cases. *MH D-DIF* can be misleading in large samples, but it is easier to interpret in the classification of DIF. *STAND P-DIF* does not depend on the size of the sample.

The analyses reported for these two statistics were performed using a Perl computer program written by Boris Alexeev, a graduate student in mathematics, for the purposes of this study. A description of the procedure used in the program can be found in Appendix B.

Identification of Characteristics That Contribute to DIF

For the indices *STAND P-DIF* and *MH D-DIF*, I used a one-way analysis of variance (ANOVA), or a two-tailed t test in the case of only two groups, to identify categories and characteristics of items that contributed to DIF. I compared the means of *STAND P-DIF* and *MH D-DIF* for all characteristics of items that I coded. In some cases, I report only the *STAND P-DIF* results when the results were very similar; however, in other cases I used the *MH D-DIF* results or both to allow comparison to other years. A post hoc Tukey multiple comparison procedure was used to identify the direction of significant effects. Harris and Carlton (1993) did a similar analysis for one-stage *MH D-DIF*, although on a different type of test and with different characteristics tested. I used two-way ANOVA to identify interactions between item characteristics.

Results

FCAT SSS Strands

An analysis of all 149 FCAT items[3] for Grades 8 through 10 in 2001 with respect to the official classification by strand showed that the mean *STAND P-DIF* values and the mean *MH D-DIF* values were significantly different ($F(4, 144) = 3.695$, $p < .01$ and $F(4, 144) = 3.576$, $p < .01$, respectively). A post hoc Tukey test showed that across the three grades male students performed relatively better on items in the measurement strand than female students matched on the total score on the test, whereas the reverse was true for the algebra strand and the data analysis and probability strand. The *STAND P-DIF* and *MH D-DIF* means are shown in Table 6. These results are partially supported by the findings of other

[3]The total number of items on the FCAT was 150. One item had corrupted data.

Table 6:
Mean DIF Values for Strands in Study 1

Strand	Number of items	*STAND P-DIF* (SD)	*MH D-DIF* (SD)
Number	32	−.004 (.032)	−0.064 (0.492)
Measurement	27	−.021 (.039)$^{a\,b}$	−0.315 (0.561)$^{a\,b}$
Geometry	28	−.003 (.025)	−0.050 (0.402)
Algebra	36	.004 (.036)^{a}	0.056 (0.505)^{a}
Probability and data analysis	26	.010 (.027)^{b}	0.138 (0.373)^{b}
Total	149	−.003 (.033)	−0.043 (0.491)

Note. Two entries with the same superscript in the same column are significantly different ($p < .05$) according to the Tukey post hoc test.

researchers (Ansell & Doerr, 2000; Lubienski et al., 2004; McGraw et al., 2006), who concluded that male students usually perform better than female students on items in the measurement strand. However, the results are different for the data analysis and probability strand, which showed that the female students performed relatively better on this strand. Of course, one has to be careful when comparing absolute difference in performance to performance matched on the total score. If students are matched by the total score, in some sense by ability, one would expect that their performance would be the same on items with a particular characteristic, therefore the difference in *STAND P-DIF* and *MH D-DIF* means is unexpected and requires one to study the items. When only 107 algebra and algebra-related items were tested by the strand, there were no statistically significant differences in *STAND P-DIF* or *MH D-DIF* means for the different strands, although all measurement

and algebra items were retained for the analysis. The means of *STAND P-DIF* for the 9 retained geometry items and the 3 data analysis and probability items went down to −.005, (SD = .031) and −.028, (SD = .021), respectively.[4] The overall *STAND P-DIF* mean for 107 items changed to −.007 (SD = .036), down from −.003 (SD = .033) for all 149 items. This change shows that the retained items in both strands are more tilted toward male students than the unretained items from those strands.

Content Classification and Characteristics

Content Categories

The content classification had eight categories. The *STAND P-DIF* and *MH D-DIF* means for all categories are shown in Table 7. Although a one-way ANOVA showed that the means were different across all categories ($F(7, 99) = 2.747$ and $F(7, 99) = 2.605$, $p < .05$, respectively), a post hoc Tukey test showed no differences in the means in multiple comparisons. As Table 7 shows, the algebraic manipulations category has higher mean indices than the other categories. However, the size of the category was not large enough to provide a definite answer. Some categories had very few items.

Topic of the Context

The results for *STAND P-DIF* and *MH D-DIF* means for different topics are provided in Table 8. A Levene statistics test showed that the topic subgroups barely passed a test of homogeneity of variance for *STAND P-DIF* with p = .051, whereas the variances of the topic subgroups were the same for *MH D-DIF*, p = .092. The ANOVA for topic showed that the means across the topics were significantly different for *STAND P-DIF* and *MH D-DIF* ($F(7, 99 = 5.688$ and $F(7, 99) = 6.176$, $p < .001$). In addition to the Tukey post hoc test for *MH D-DIF*, a Tamhane post hoc analysis was performed for *STAND P-DIF* that took into consideration that the variances may be different.

[4]The *MH D-DIF* means went down, too.

Table 7:
Mean DIF Values for Content Categories in Study 1

Content category	Number of items	*STAND P-DIF* (SD)	*MH D-DIF* (SD)
Number	11	.013 (.030)	0.178 (0.399)
Geometrical measurement	26	−.018 (.037)	−0.226 (0.547)
Informal algebra	26	−.016 (.030)	−0.258 (0.486)
Pattern	16	.005 (.039)	0.050 (0.535)
Setting up/translating	5	.008 (.022)	0.118 (0.295)
Functions	12	−.001 (.028)	−0.011 (0.436)
Algebraic manipulations	4	.030 (.020)	0.488 (0.294)
Other	7	−.028 (.046)	−0.374 (0.641)
Total	107	−.007 (.036)	−0.094 (0.523)

An interesting detail is that the results of the post hoc analyses for *STAND P-DIF* and *MH D-DIF* means are different. The lowest *STAND P-DIF* mean is for the population topic, whereas the lowest *MH D-DIF* mean is for the sports/recreation/transportation topic. Both topics, along with measuring, heavily favored male students but to a much lesser extent. On the other hand, topics such as social studies and physical sciences favored female students, but to an even lesser extent than the measuring topic favored male students. None of the topics that favored female students had mean significantly different from the topics in the retail/currency/business, no-context, and other. That is, the social studies topics favored female students with respect to the population and the sport/recreation/transportation topics but not with respect to the retail/currency/business topic.

Table 8:
Mean DIF Values for Topic in Study 1

Topic	Number of items	*STAND P-DIF* (SD)	*MH D-DIF* (SD)
No context	11	.007 (.025)^{a}	0.102 (0.358)^{a}
Sports/recreation/transportation	17	−.032 (.046)^{bc}	−0.520 (0.626)^{abcde}
Physical sciences	16	.016 (.027)^{bde}	0.209 (0.374)^{bfg}
Population	8	−.033 (.016)^{adfg}	−0.450 (0.260)^{fh}
Retail/currency/business	19	−.000 (.023)^{f}	0.013 (0.373)^{c}
Social studies	9	.015 (.022)^{cg}	0.261 (0.322)^{dhi}
Measuring	20	−.022 (.034)^{e}	−0.303 (0.504)^{gi}
Other	7	.011 (.034)	0.196 (0.560)^{e}
Total	107	−.007 (.036)	−0.094 (0.523)

Note. Two entries with the same superscript in the same column are significantly different, $p < .05$ according to a Tamhane post hoc test for *STAND P-DIF* and a Tukey post hoc test for *MH D-DIF*.

As I previously reported (page 41), the result for content categories was not conclusive. Although an ANOVA showed that the *STAND P-DIF* and *MH D-DIF* means for categories were different, a post hoc analysis did not confirm that result. At the same time there were significant differences among the means for context topics. I performed a two-way ANOVA on the *STAND P-DIF* and *MH D-DIF* means to check on interactions of topic and category. I report only the *MH D-DIF* results here. The analysis of variance indicated a significant main effect of item topic on the value of *MH D-DIF*, $F(7, 67) = 4.294$, $p = .001$. There was no main effect of item category on *MH D-DIF*, $F(7, 67) = 1.241$, $p = .293$. There was a significant interaction between topic and category, $F(7, 67) = 2.476$, $p = .002$. Although

I did not go deeper into investigating the results for two-way ANOVA, it is clear that the possible difference in the means for content categories may be attributed to the item context. Partial eta squared for the interaction term was .48, that is, 48% of the variance in the *MH D-DIF* mean is uniquely attributable to this interaction term. This interaction effect is easy to see from the example of the observed *MH D-DIF* means for cells with more than one item shown in Table 9.

Other Characteristics

I performed two-tailed t tests on characteristics with only two levels. Some of the characteristics showed differences in *STAND P-DIF* and *MH D-DIF* means, and others did not. The means for characteristics that did not show significant differences in means are shown in Table 10, and the results for characteristics that demonstrated significant *STAND P-DIF* mean differences are shown in Table 11.

There was no significant difference in the *STAND P-DIF* means for multiple-choice (MC) and gridded-response (GR) items, although the *STAND P-DIF* mean was slightly higher for GR items. It is also interesting that all four category B DIF items favoring male students were MC, whereas a category B DIF item favoring female students was a GR. Because only a small number of items were DIF items, it is not surprising that in the present study the means were not different for MC and GR items.

Problems where answers were part of the item (so-called forced answers) and items with distracting information were not any different in means from the items that did not have these characteristics. In addition to different topics, items were tested as having context or not. The difference was not significant.

Visual stimuli in the item did not demonstrate any differences in *STAND P-DIF* means. A two-tailed t test was performed on all visual stimuli together and by type (table, figure, graph) as well as ANOVA by different types of figure (picture, picture with information,

Table 9:
MH D-DIF Means for Category-by-Topic Interactions in Study 1

Category Topic	Number of items	Mean	SD
Geometrical measurement			
Sports/recreation/transportation	3	−0.817	0.278
Social studies	4	0.088	0.316
Measuring	18	−0.258	0.506
Informal algebra			
Sports/recreation/transportation	8	−0.803	0.334
Physical sciences	4	−0.042	0.131
Population	3	−0.238	0.179
Retail/currency/business	8	−0.114	0.223
Other	2	0.607	0.369
Functions			
No context	2	0.120	0.038
Sports/recreation/transportation	2	0.147	0.239
Population	3	−0.562	0.283
Retail/currency/business	3	0.040	0.316
Other			
Sports/recreation/transportation	2	−0.840	0.132
Retail/currency/business	2	0.235	0.135

geometric figure), table (pattern or other), and graph (function, *xy* grid). All analyses showed no difference in the means.

Table 10:
STAND P-DIF Means for Item Characteristics That Did Not Contribute to DIF in Study 1

Item	Yes		No	
characteristic	*n*	Mean (SD)	*n*	Mean (SD)
Multiple choice	69	−.007 (.035)	38	−.006 (.036)
Integers only	61	−.009 (.036)	46	−.004 (.036)
Figure	25	−.014 (.036)	82	−.004 (.036)
Table	15	−.003 (.043)	92	−.007 (.034)
Graph	11	−.005 (.031)	96	−.007 (.036)
Visual stimuli	50	−.010 (.037)	57	−.004 (.034)
Forced answer	12	.002 (.031)	95	−.008 (.036)
Percents	11	−.016 (.031)	96	−.006 (.035)
Fractions	12	−.002 (.035)	95	−.007 (.036)
Distracting information	16	−.007 (.031)	91	−.006 (.037)
Ratio	16	−.007 (.044)	91	−.007 (.034)
Context	96	−.008 (.036)	11	.007 (.025)
One step[a]	12	−.009 (.022)	66	−.011 (.040)

[a] Test performed on items with computed solutions.

The difference in *STAND P-DIF* means for the type of solution was significant (Table 11). Noncomputed solutions favored female students, whereas computed solutions favored male students. But it is interesting to note that there were no DIF items among the noncomputed solution items. In the computed solution items, *STAND P-DIF* means for one-step items and more-than-one-step items were not significantly different (Table 10).

Table 11:
STAND P-DIF Means for Item Characteristics That Contributed to DIF in Study 1

Item characteristic	Number of items	Mean (SD)	t (p value)	Effect size
Type of solution				
Noncomputed	29	.005 (.026)	2.45	
Computed	78	−.011 (.038)	(.017)	0.45
Estimated answers				
No	91	−.001 (.033)	4.22	
Yes	16	−.039 (.035)	(.000)	1.14
Converting units				
No	102	−.005 (.035)	2.57	
Yes	5	−.046 (.038)	(.011)	1.17
Vehicles in context				
No	90	−.000 (.032)	4.48	
Yes	17	−.039 (.036)	(.000)	1.19
Rate				
No	81	−.002 (.035)	2.21	
Yes	26	−.020 (.036)	(.029)	0.51

No differences were found in type of number. Neither fractions nor decimals contributed to DIF in the FCAT 2001. The result was the same for items involving percent or ratio

(Table 10). However, items with rate had a significant difference in *STAND P-DIF* mean, although the effect size, Cohen's *d*, was not large[5] (Table 11).

Two characteristics, items with estimated answers and items with vehicles in context, had significant differences in *STAND P-DIF* means and large effect size, Cohen's *d* equals to 1.14 and 1.19, respectively (Table 11). These results were not unexpected, although the large effect sizes of the differences were surprising. Another characteristic with large effect size was converting units. All five items that involved converting units highly benefited male students.

Overall, the analysis of FCAT 2001 items produced interesting results. At the same time, it is premature to make any definite conclusions.

DIF Items

In 2001, there were four category B DIF items (1 in Grade 8, 1 in Grade 9, and 2 in Grade 10) favoring male students and one B-DIF item favoring female students (Grade 10). Only two of these items have been released by the FDOE, one favoring male students and one favoring female students. All released items can be found on the FCAT home page (FDOE, 2008). Figure 1 shows the empirical response functions (IRF) for female and male students for the following released B-DIF Grade 10 item favoring male students:

> Tanisha and some friends from her bicycle club went on a training ride from West Palm Beach to Miami. They planned to ride 45 miles from West Palm Beach to Fort Lauderdale, another 10 miles to Hollywood, and finally 15 miles to Miami. Tanisha's bicycle got a flat tire north of Miami, and she was unable to complete the training ride. Her odometer showed she had traveled 60 miles. Approximately what percent of the training ride did Tanisha complete?
>
> (Source: FDOE[6] (FCAT, 2005, p. 6))

[5]J. Cohen (1988) defined effect sizes as "small, $d = 0.2$," "medium, $d = 0.5$," and "large, $d = 0.8$."

[6]The Florida Comprehensive Assessment Test (FCAT) Mathematics items appear by permission of the Florida Department of Education, Office of Assessment, Tallahassee, Florida 32399-0400.

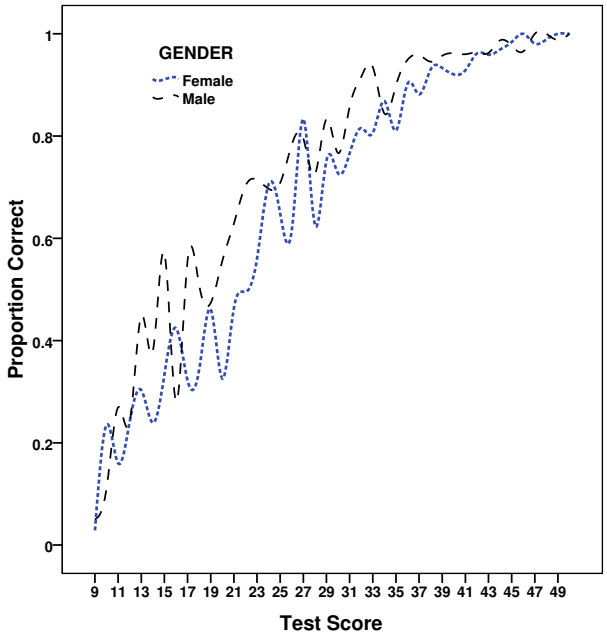


Figure 1: Empirical IRF for a B-DIF item favoring male students.

The p value, the proportion correct, is high for this item. By the FCAT classification it is categorized as an easy item. Overall, 80% of the students answered this item correctly: 75% of the female students, and 85% of the male students. Mathematically this item is straightforward and requires the student to add three numbers and divide the fourth number by the sum. However, the wording of the item is complicated. The item is quite long and has the names of four Florida cities. It also includes the word *odometer*, whose meaning more male students may know. The question has the word *approximately*. The estimated answers characteristic includes items with the word *approximately*. I found that this characteristic benefits male students.

The item above had the highest male DIF, but the magnitude of the *STAND P-DIF* was not the highest. Figure 2 shows the empirical item response functions for male and female students for an unreleased B-DIF item benefiting male students that had the highest *STAND P-DIF* by magnitude (-0.1001). The item was classified in the informal algebra/rate

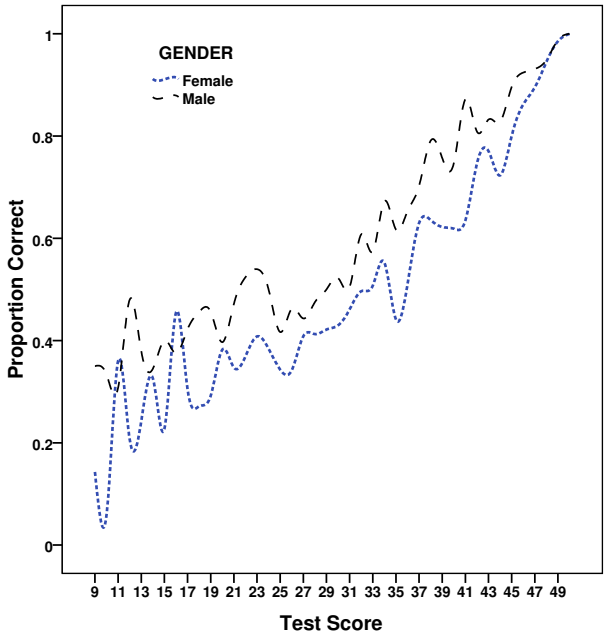


Figure 2: Empirical IRF for an item with highest magnitude of *STAND P-DIF*.

subcategory. Mathematically, this item is more complex than the previous one. It involves understanding of relation between distance, rate, and time for two runners and one bicyclist. This item was average in difficulty, and it was answered correctly by 60% overall: 53% of the female students and 67% of the male students. The item involves a bike race and the names of little-known Florida towns. The statement of the item is short and concise, and the numbers are presented in a table. Like the previous item, the question contains the word *approximately*.

The two previous items were from Grade 10. The B DIF item from Grade 8 involves bicycles too. This item was classified in the finding lengths subcategory of geometrical measurement. The item was solved correctly by 60% of the female students and 70% of the male students. The item has just two numbers given: the distance in feet and the number of loops. It asks the student to find a total distance in miles. The most popular distracter, 29% for female students and 18% for male students, was the quotient of the larger number

and the smaller one, which suggests that these students did not read the item carefully. I think it is very unlikely that so many students did not understand the difference between the total distance and the distance for just one loop. Two other distracters took into account a possible confusion between feet and yards. This item has the word *approximately* too.

The fourth category B DIF item was from Grade 9. This item was not about bicycles; it involved estimating a length for which four approximate measurements were given in a table. I classified this item in the length, area, volume subcategory of geometrical measurement that requires using basic formulas in a more complex setting than just applying a basic formula. The topic was measuring real-life objects. The item asks about a *reasonable estimate* and requires the student to convert inches into feet. The item was solved correctly by 41% of the female students and 52% of the male students. The most popular distracter was the sum of the four measurements in the table.

All four DIF items favoring male students required estimated answers. Three of the four were in the sports/recreation/transportation topic; all of them involved a bicycle race or ride, and the fourth involved measuring. All items were MC and required an answer to be computed with a more-than-one-step solution. Two items required converting linear units and were from the subcategory of geometrical measurement that involves finding length, perimeter, area, or volume in a more complex setting than just applying a basic formula. Two items were from the informal algebra category, one with rate and another without rate. Two items had nonpattern type of table, and two other items did not have any visuals. One item required calculating percentage.

The empirical item response functions for the lone B-DIF item favoring female students are shown in Figure 3:

> Krista has decided to enter a local marathon. As part of her training, she is going to increase the number of miles she runs every week by 3 miles. If Krista runs 12 miles in

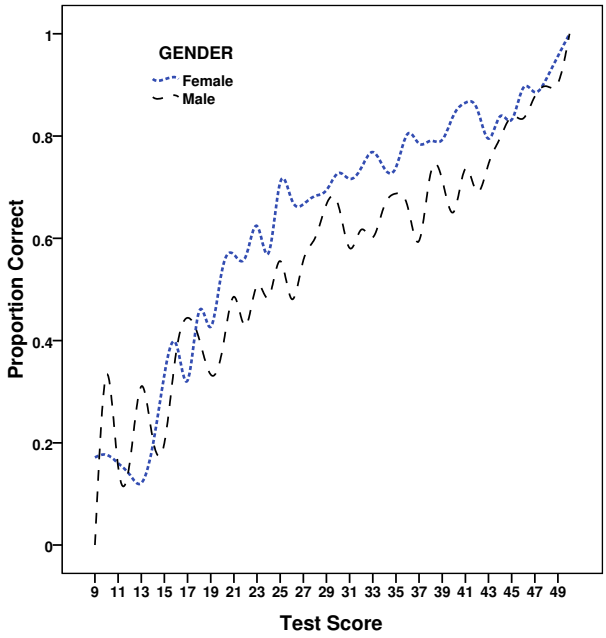


Figure 3: Empirical IRF for a B-DIF item favoring female students.

the first week, how many miles will she run during the ninth week?

(Source: FDOE[7] (FCAT, 2005, p. 8))

The topic of the item, sports/recreation/transportation, usually favors male students according to the analysis; however, the item does not include any vehicles, and the subcategory, pattern/function, is mostly neutral with slight tilt toward female students. The item was GR with a more-than-one-step computed solution. The problem was solved correctly by 70% of the female students and 65% of the male students.

Conclusion

According to the first study, the male DIF items clearly fit a profile that is more likely to benefit male students: the informal algebra or geometrical measurement category, the

[7]The Florida Comprehensive Assessment Test (FCAT) Mathematics items appear by permission of the Florida Department of Education, Office of Assessment, Tallahassee, Florida 32399-0400.

sports/recreation/transportation or measuring topic, an estimated answer, converting units, and a computed answer. The female DIF item is not so clearcut. Although the category is not one that benefits male students, the topic is. One possible reason that the female students did relatively better on the item may have been that they explicitly wrote the pattern table without trying to figure out the expression, whereas male students may have tended to write an expression and may have used 9 for the number of weeks instead of 8. I could not check my hypothesis because I did not have access to actual answers on the gridded items, but only whether the response was correct or incorrect.

I conclude that one candidate for a challenging category for female students is geometrical measurement, and informal algebra is a close second. This finding confirmed the conclusion of several studies (Gallagher & De Lisi, 1994; Harris & Carlton, 1993; McGraw et al., 2006; Mendes-Barnett & Ercikan, 2006; Willingham & Cole, 1997).

The most challenging topics for female students are sports/recreation/transportation, population, and measuring. Not many studies have addressed the issue of topic. Many studies have found that male students do relatively better on word problems, but I am not aware of studies that considered different topics. I found only a brief remark in Harris and Carlton (1993) that items with topics such as money, time, fractions, rate, linear and liquid measure, averages, and areas resulted in mean DIF values favoring male students and only percentages and counting topics favored female students.

Among other characteristics, items involving estimated answers, vehicles, converting units, and rate contribute heavily to DIF favoring male students. A. S. Cohen and Ibarra (2005) cited the first two characteristics, one as items with indefinite answers and another as a male topic, in their study as favoring male students.

Although all DIF items had computed solutions, the noncomputed answer characteristic has been found to contribute to female DIF. This result confirmed the conclusions of the Harris and Carlton (1993) study, although the definition of noncomputed solution was slightly

different. I included scientific notation in noncomputed solutions because it includes no computation but understanding of the concept.

After analyzing the results of the first study, I realized that the classification of categories should be changed. The number of subcategories needed to be reduced. Otherwise, it would be impossible to run an ANOVA because so many subcategories would have only a few items.

The hypotheses that I wanted to test in the second study were the following:

- The majority of items with *STAND P-DIF* less than $-.05$ are from the geometrical measurement and informal algebra categories: items that require the students to find lengths, perimeter, or area of non-basic shapes, and items that have rate of change in the form of linear speed.
- The majority of the items with *STAND P-DIF* less than $-.05$ are on the topics recreation, measuring, or population.
- Estimated answers and converting units are factors contributing to male DIF.

CHAPTER 4

STUDY 2: FLORIDA COMPREHENSIVE ASSESSMENT TEST 2002

Method

The Florida Comprehensive Assessment Test

There were no major changes in the FCAT from 2001 to 2002. The distribution of items by official strand was similar to 2001 and is shown in Table 12. Three measurement items did not fit my classification; therefore, they were not retained for the analysis. These items were geometric and involved the measurement of angles. Two more geometry problems were included in the analysis than in Study 1. The total number of items is roughly the same in both years.[1] The FCAT reuses some of the problems from the previous year for linkage purposes; there were 43 items retained from 2001.

Sample

The distribution of data by demographic groups for 2002 is given in Table 13. The sample size was smaller than in 2001. There were 30 forms in 2002, whereas only 15 forms were used in 2001. In each case, I used data from one form. Only students that did not require special accommodation were included in the analysis.

[1] In 2001, one retained item had corrupted data; therefore, the analysis was performed on 107 items, the same as in 2002.

Table 12:
Distribution of 2002 FCAT Items by Strand and Grade

	Grade			
Strand	8	9	10	Total
Number	11	9	10	30
Measurement	12	7 (9)	8 (9)	27 (30)
Geometry	1 (6)	7 (12)	3 (8)	11 (26)
Algebra	11	11	14	36
Data analysis and probability	0 (9)	2 (9)	1 (8)	3 (26)
Total	35 (49)	36 (50)	36 (49)	107 (148)

Note. The total number of items on 2002 FCAT is in parentheses when not all items in a strand were used.

Classification

Changes to Content Category Classification

After analyzing the results for 2001, I decided to reduce the number of subcategories by aggregating some of them. The following is the modified content category classification:

- *Number*: Basic understanding of numbers, forms of the numbers and properties. Problems in this category are either no context or minimal context. I decided to eliminate subcategories.

- *Geometrical measurement*: Basic measurement items that involve distance, length, perimeter, area, or volume. Basic formulas are usually provided either in the item or in

Table 13:
Distribution of 2002 Data Sample by Grade, Gender, and Race

	Gender		Race[a]			
Grade	Female	Male	White	Black	Hispanic	Total
8	2,661	2,289	2,700	1,167	886	4,950
9	2,750	2,653	2,949	1,265	987	5,403
10	2,203	1,845	2,335	880	660	4,048

[a] Data for other racial groups are omitted.

the table at the beginning of the test booklet. I reduced the number of subcategories to two:

  - Formula/scale: Using a single basic formula, the Pythagorean Theorem, or a basic ratio or scale for similar geometric figures and maps.
  - Modified formula: Using more than one basic formula or a modified basic formula. This is the main difference from the first subcategory. The items are usually given in a real-world context.

- *Informal algebra*: Word problems not involving variables, but may involve modeling of arithmetic expressions. The category also includes items with nongeometrical proportions, percents, or rate. I reduced the number of subcategories categories to two:
  - Proportionality: Proportionality, percent, ratio, and other arithmetic items without rate or geometrical measurement.
  - Rate: Items with rate, but not necessarily asking the student to find one.

- *Pattern*: Identification of a rule, finding one or more terms that are not distant, or matching given general rules. I reduced the number of categories to two by aggregating subcategories:
  - Function: Number sequence and function. The function can be given as a table of values, a graph, or a verbal description. In a verbal description, it is usually described as a rate of change.
  - Pictorial: Items with tiles, where one is to find a specific design for the tile, or items with expanded figures, and mixed patterns such as pictorial and function at the same time.
- *Setting-up/translation*: Translating real-world situations into equations, inequalities or functions. Items usually require matching because items are MC or GR. Therefore, items ask specifically about an expression. This category is different from pattern/functions in the sense that the rule in setting up a function is given. It should be just translated into a function, whereas in patterns, the rule should be determined. I removed all subcategories. Although inequality requires a different type of thinking than equation or a functions, it is not a common item on assessment tests.
- *Functions*: Understanding functions; finding one variable given another. This category is different from patterns in the way that the function is given: by either an equation or a graph. The number of categories was reduced to two:
  - Variable: Finding a dependent or independent variable. The task involves either finding the value of the variable on a given graph or substituting for a given variable in the equation and calculating for another variable.
  - Interpretation: Interpreting, finding parameters such as slope, or determining the impact of changing parameters; matching a function and its graph.

- *Algebraic manipulations*: Given an equation, expression, or inequality, solve or simplify. The item can be given in context or without. All types of equations and inequalities are in this one category because there are so few of them in the FCAT. For other types of tests, this category may be extended. There are no subcategories.

- *Other*: Includes either combinations of the previous eight categories or categories that are not listed above.

The number of main categories remained the same. I reduced the number of subcategories from 25 to 12, and for simplicity I will call them all *categories*. I have 12 categories, some of them sometimes grouped, such as informal algebra/proportionality and informal algebra/rate into informal algebra when both categories exhibit the same behavior or feature. The distribution of items by category and original strand on the FCAT is shown in Table 14.

Changes and Additional Characteristics

Initially, I added one new topic: computers. However, I had only three items on this topic; therefore, I decided not to use it. The *MH D-DIF* and *STAND P-DIF* values on these three items were neutral; its elimination did not change the outcome of the analysis of topics. I added two item characteristics. One characteristic was linear speed, not necessary finding the speed. Although this was a subset of the items concerning rate, my observations of the previous year's data suggested that it is more difficult for students than any other rate. The other rates characteristic is also tested. Another characteristic is nonmetric units such as feet, miles, ounces, or gallons. I noticed that converting nonmetric linear units is a factor contributing to gender-related DIF. Checking all nonmetric units is a step that I overlooked in the previous year's analysis.

Table 14: Distribution of 2002 FCAT Items by Strand and Category

	Strand					
Category	Number	Measurement	Geometry	Algebra	Data analysis & probability	Total
Number	9			1		10
Geometrical measurement	0/1	10/5	6/1	1/0		17/7
Informal algebra	13/2	4/4		3/0	1/0	21/6
Pattern	1/0			7/2	1/0	9/2
Setting-up/ translation				9		9
Functions	1/0		0/3	2/2	0/1	3/6
Algebraic manipulations	3			5		8
Other		4	1	4		9
Total	30	27	11	36	3	107

Note. For subdivided categories, the first number is the first category, and the other is the second, see pages 56–59.

Results

FCAT SSS Strands

As with Study 1, I started with analyzing the official classification according by strand. An ANOVA for all 148 FCAT 2002 items showed no significant differences in the mean *STAND P-DIF* values and the mean *MH D-DIF* values ($F(4, 143) = 2.134$, $p = .08$, and

$F(4, 143) = 2.285$, $p = .06$, respectively). For the retained 107 problems, the differences were not significant either ($F(4, 102) = 1.145$, $p = .34$, and $F(4, 102) = 1.446$, $p = .22$, respectively). In the previous study, the algebra and data analysis *STAND P-DIF* and *MH D-DIF* mean values were significantly different from those of the measurement strand when all 149 items were analyzed. Male students did better on items from the measurement strand than the female students with the same total score on the test.

Content, Context, and Other Characteristics

Content Categories

A one-way ANOVA on 12 categories showed a significant difference in *STAND P-DIF* and *MH D-DIF* values ($F(11, 95) = 2.360$, $p = .013$, and $F(11, 95) = 2.278$, $p < .01$, respectively). The results from the content category analysis are shown in Table 15.

The analysis of categories showed that not all types of geometrical measurement are challenging to female students. The formula/scale of geometrical measurement mean for both indices was not significantly different from the means of other categories. The formula/scale items involve using only one basic formula or a well-defined procedure for ratio of lengths. The modified formula category of geometric measurement is not very procedural. Although some items are easy, they still require either altering an existing basic formula or using several formulas. The algebraic manipulations category is very procedural. Therefore, it is not surprising that the mean of the formula/scale of geometrical measurement category is not different from the mean of the algebraic manipulations category.

Overall, the means are not really different between categories with the exception of algebraic manipulations. And even this category mean is different from only four other category means. There are no indications that male students performed significantly better on these four subcategories than matched female students. In Study 2, there were twice as many items in the algebraic manipulations category than in Study 1. The *STAND P-DIF* mean is high,

Table 15:
Mean DIF Values for Content Category in Study 2

Content category	Number of items	*STAND P-DIF* (SD)	*MH D-DIF* (SD)
Number	10	.000 (.031)	−0.021 (0.379)
Geometrical measurement			
Formula/scale	17	.002 (.038)	0.055 (0.534)
Modified formulas	7	−.020 (.031)^{a}	−0.360 (0.520)^{a}
Informal algebra			
Proportionality	21	−.008 (.037)^{b}	−0.097 (0.539)^{b}
Rate	6	−.029 (.036)^{c}	−0.416 (0.417)^{c}
Patterns			
Function	9	.009 (.042)	0.176 (0.616)
Pictorial	2	−.026 (.057)	−0.342 (0.763)
Setting up/translating	9	−.005 (.016)	−0.057 (0.196)
Functions			
Variable	3	−.018 (.047)	−0.280 (0.680)
Interpretation	6	.001 (.031)	0.047 (0.451)
Algebraic manipulations	8	.048 (.022)^{abcd}	0.744 (0.380)^{abcd}
Other	9	−.015 (.045)^{d}	−0.197 (0.607)^{d}
Total	107	−.002 (.040)	−0.022 (0.547)

Note. Two entries with the same superscript in the same column are significantly different according to the Tukey post hoc test, $p < .05$.

it is close to the threshold to be flagged by the procedure. This shows that almost all items in this category were favoring female students.

Topic of the Context

As in the previous study, an ANOVA with respect to topic showed significant differences in *STAND P-DIF* and *MH D-DIF* mean values ($F(7, 99 = 6.021$, and $F(7, 99) = 5.428$, $p <$ .001, respectively). The results of the analysis are shown in Table 16. In this study, the means for measuring were not significantly different from those for any other topic. To summarize the results from the table, two topics were particularly troubling for female students: population and sports/recreation/transportation. On those topics, male students performed significantly better than female students matched on the total score. The means for other topics were not significantly different, although some topics were slightly favorable to male students and some to female students.

Considering that the *STAND P-DIF* and *MH D-DIF* means differed for both the content and context categories, the next question is whether their interaction was significant. I report the results for *MH D-DIF* only. A two-way ANOVA indicated a significant main effect of item topic on *MH D-DIF*, $F(7, 67) = 3.345$, $p = .005$. There was no main effect of item category on *MH D-DIF*, $F(7, 67) = 1.682$, $p = .100$, and no significant interaction between category and topic, $F(7, 67) = 1.132$, $p = .336$. This result is different from that of Study 1, where the interaction was significant. The main effect of item category is not significant when topic is considered. This result was expected because only one category was significantly different from the other categories.

Other Characteristics

I performed two-tailed t tests on characteristics with only two levels. Some of the characteristics showed differences in the *STAND P-DIF* and *MH D-DIF* means, whereas others did not. The means of characteristics that did not show a difference are reported in Table 17, and the means for characteristics that demonstrated mean differences in *STAND P-DIF* are reported in Table 18.

Table 16:
Mean DIF Values for Topic in Study 2

Topic	Number of items	*STAND P-DIF* (SD)	*MH D-DIF* (SD)
No context	14	.019 (.031)^{af}	0.231 (0.372)^{ae}
Sports/recreation/transportation	13	−.043 (.043)^{abcde}	−0.595 (0.579)^{abcd}
Physical sciences	21	.009 (.028)^{bg}	0.129 (0.404)^{bf}
Population	5	−.048 (.035)^{fghi}	−0.616 (0.413)^{efh}
Retail/currency/business	23	.004 (.030)^{ch}	0.076 (0.475)^{c}
Social studies	7	.015 (.032)^{di}	0.342 (0.621)^{dg}
Measuring	16	−.010 (.036)	−0.129 (0.562)
Other	8	.005 (.027)^{e}	0.059 (0.401)
Total	107	−.002 (.038)	−0.022 (0.547)

Note. Two entries with the same superscript in the same column are significantly different according to the Tukey post hoc test, $p < .05$.

In this study, although the difference in the *STAND P-DIF* means for MC and GR items was not significant, it was very close to significance ($p = .053$).[2] It was a somewhat unexpected result because unlike Study 1, there were three GR items out of five B DIF items favoring male students and one MC item out of two B DIF items favoring female students.

Another change from the previous study is that the *STAND P-DIF* means for non-computed solutions were not significantly different from the means for computed solutions. Converting units was not a contributing factor to DIF in this study. However, context and no-context items had significantly different *STAND P-DIF* mean values. The mean of items

[2]But not for the *MH D-DIF* means ($p = .073$).

Table 17:
STAND P-DIF Means for Item Characteristics That Did Not Contribute to DIF in Study 2

Item	Yes		No	
characteristics	n	Mean (SD)	n	Mean (SD)
Multiple choice[a]	67	−.008 (.035)	40	.007 (.041)
Integers only	61	−.006 (.038)	46	.002 (.037)
Figure	26	−.008 (.037)	81	−.001 (.038)
Table	10	−.010 (.046)	97	−.002 (.037)
Graph	13	−.011 (.037)	94	−.001 (.038)
Visual stimuli	49	−.010 (.037)	58	.004 (.038)
Forced answer	11	−.003 (.022)	96	−.002 (.040)
Percents	12	−.015 (.042)	95	−.001 (.037)
Fractions	11	−.009 (.030)	96	−.002 (.039)
Distracting information	10	−.011 (.039)	97	−.002 (.038)
Other rates[b]	20	−.012 (.032)	82	.003(.037)
Ratio	21	.001 (.040)	86	−.003 (.038)
Converting units	8	−.019 (.043)	99	−.001 (.037)
One step[c]	10	.011 (.044)	71	−.005 (.040)
Noncomputed solution	26	−.001 (.029)	81	−.004 (.040)
Non-metric units	42	−.011 (.036)	65	.003 (.038)

[a] Difference in means is close to significant ($p = .053$).
[b] Items with linear speed are excluded.
[c] Test performed on items with computed solutions.

with nonmetric units was not different from the mean of other items. There was no significant difference in the means between items with metric units and items with nonmetric units.

Table 18:
STAND P-DIF Means for Item Characteristics That Contributed to DIF in Study 2

Item characteristic	Number of items	Mean (SD)	t (p value)	Effect size
Context				
No	19	.019 (.033)	2.81	
Yes	88	−.007 (.037)	(.006)	0.72
Estimated answers				
No	86	.003 (.037)	3.23	
Yes	21	−.025 (.036)	(.002)	0.76
Vehicles in context				
No	96	.002 (.034)	4.23	
Yes	11	−.045 (.041)	(.000)	1.35
Linear Speed				
No	101	.000 (.036)	3.00	
Yes	6	−.046 (.036)	(.003)	1.28

Rate was still a contributing factor; however, items with rate but not linear speed were not different in the means of nonrate items. Only six items involved linear speed, and the *STAND P-DIF* mean was significantly lower than the mean of the rate items that did not involve linear speed. These items were not a subset of items with vehicles; only three of them were common for both groups. Because the group was so small, I cannot claim that linear speed is a contributing factor to DIF; however, this characteristic is a candidate for further testing.

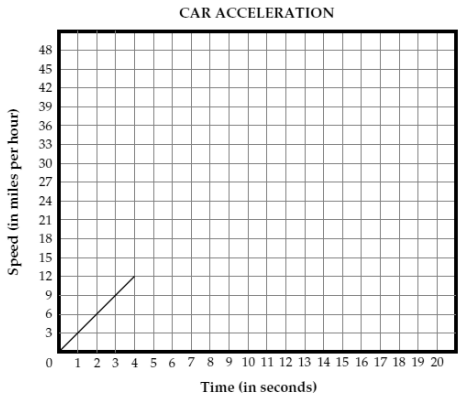
DIF Items

There were five Category B DIF items favoring male students and three items favoring female students. Only three items were released by the FDOE, and they can be found on the FCAT home page (FDOE, 2008). All the released items favored male students. First, I describe items favoring male students and compare them with the items in Study 1. Then I describe items favoring female students.

All five DIF items in 2002 were on the FCAT in 2001. One of the items was an unreleased Category B DIF. In 2001, this item had the highest magnitude of *STAND P-DIF*. This item is about two runners and a bicyclist. I described it on page 49. The other four items were not categorized as Category B DIF items, although three of them were flagged by having *STAND P-DIF* $< -.05$. One item was not flagged by *STAND P-DIF* in Study 1.

The following item had the second highest *MH D-DIF* and *STAND P-DIF* by magnitude among those items favoring male students:

> An automobile testing organization is verifying the acceleration characteristics of a car. The car will accelerate at a rate of 3 miles per hour per second from 0 miles per hour (mph) to 45 mph. The graph below shows the beginning of the ideal acceleration plot.



> If the rate of acceleration remains constant, how many seconds will it take the car to reach its final test speed?
>
> (Source: FDOE[3] (FCAT, 2006a, p. 25))

This item is from Grade 10 and was classified by the FCAT as a geometry item. The benchmark listed for this item is the following:

> Using a rectangular coordinate system (graph), applies and algebraically verifies properties of two-and three-dimensional figures, including distance, midpoint, slope, parallelism, and perpendicularity. (FDOE, 2001, p. 36)

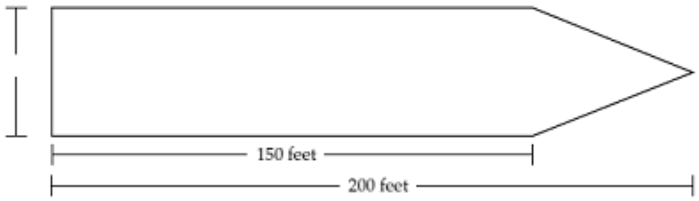
The item also assesses an algebraic thinking benchmark on representing "real-world problem situations using finite graphs" (p. 37).

This item is very confusing. It is not clear whether acceleration is discussed in Florida mathematics classes or how many students had taken a physics course by the time of the test. Male students might have an advantage because they usually have a greater interest in cars and the topic of a car's acceleration is probably more familiar to them. Without the graph, this item might be easier, although still more difficult for female students than male students. It is basically a one-step item if the graph is not used. Another way of solving the item is to use the graph and extend the line. The third way is to look for a pattern in the graph and find a not-so-distant term by understanding the pattern or simply by repeated addition. I classified this item in the content category other because it can be solved in many different ways. Mathematically, this item is very similar to the lone female DIF problem in 2001 about a marathon runner shown on page 51. I can only guess the reason that the female students did so poorly compared to the male students matched on the total score. One reason could be that the female students were turned off by the topic of cars and acceleration. Another possible reason is that they did not understand what to do with the graph because the sentence related to the graph is not very clear.

[3]The Florida Comprehensive Assessment Test (FCAT) Mathematics items appear by permission of the Florida Department of Education, Office of Assessment, Tallahassee, Florida 32399-0400.

Another question is why this item did not have a significant DIF in 2001. One plausible explanation is that in 2002 the item was close to the end of the test, whereas in 2001 it was in the middle of the test. The number correct went down for female students from 63% to 58% and slightly for male students from 72% to 71%. A rush to finish the test on time may have forced some female students to give up on this potentially confusing item.

Another Grade 10 DIF item experienced a big drop from 2001 to 2002 in number correct for both genders but especially for female students, from 36% to 28%, compared with male students, from 45% to 40%. Although DIF is not the same as the difference in percentage correct between male and female students, they are correlated. If the drop was more for female students, it is likely that the item would move to the significant DIF category. The item was not even flagged by *STAND P-DIF* in 2001, and the next year it was a DIF item. The change could not be explained by a change position on the test. Although in 2001 it was at the beginning of the test section, in 2002 it was in the middle of the section. This item is in the modified formula of geometrical measurement, the one that requires using more than one formula or altering the existing one:

> Jackie wants to determine the number of gallons of paint needed to paint the entire deck of a cargo ship. A sketch of the deck is shown below.
>
> 
>
>
> How many square feet will be painted?[4]
>
> (Source: FDOE[5] (FCAT, 2005, p. 17))

I do not have a plausible explanation of why Grade 10 students did so poorly on this item. The formulas for area of a rectangle and area of a triangle were provided in the formula sheet.

[4]One measurement in the figure was missing on the released item, but not on the test. The missing number was 40 feet.

[5]The Florida Comprehensive Assessment Test (FCAT) Mathematics items appear by permission of the Florida Department of Education, Office of Assessment, Tallahassee, Florida 32399-0400.

The solution is straightforward: Add the area of the rectangle to the area of the triangle. One shortcut, which might not save any time, would be to calculate the area of the triangle as half the area of the small rectangle. One possible explanation for the poor performance is that first sentence "Jackie wants to determine the number of gallons of paint" is about gallons, whereas the question is about square feet. Probably, for some students there was no clear connection. Students have about one and a half minutes for each GR item. They may have never painted something and never thought about this relation. This hypothesis may also explain why the male students did better on this item than the female students: They are probably more likely to help parents with painting jobs. Another feature that female students probably did not like was the cargo ship.

Another Grade 10 DIF item had a nautical topic, a boat race. The item has not been released, so I can only describe it. This item had the highest magnitude of DIF. I classified this item as proportionality in informal algebra. The topic was sports/recreation/transportation. The item involved percents and converting days and hours into hours. The word *approximately* was in the question. In 2001, this item was one of the last; 49% of the female students and 62% of the male students solved this item correctly. In 2002, this item was in the middle of the test, and more students solved it: 57% of the female students and 71% of the male students.

According to many studies (e.g., Willingham & Cole, 1997; Zenisky, Hambleton, & Robin, 2003a), female students perform better on algebra items. However, the item below is from the algebra strand, and it was a DIF item favoring male students in Grade 9 in 2002:

> The population of a town is 13,000 and is increasing by about 250 people per year. This information can be represented by the following equation, where y represents the number of years and p represents the population.
>
> $$p = 13,000 + 250y$$

> According to the equation above, in how many years will the population of the town be 14,500?
>
> (Source: FDOE[6] (FCAT, 2006b, p. 10))

In the present study, the female students performed really well on algebraic manipulations. The item was not classified as algebraic manipulations but rather as function/variable that asks the student to find the dependent or independent variable given the other. After substitution for p, the simple linear equation should be solved. I think that something went wrong with the substitution because the female students performed well on algebraic manipulations items. This item was on the FCAT in 2001 in Grades 8 and 9. It was flagged by *STAND P-DIF* at Grade 9 in 2001 but not at Grade 8. My analysis found that the topic of population favors male students, although with so few items on that topic, it is still a hypothesis. There were no other characteristics that favored male students.

The item with the highest magnitude of *STAND P-DIF* ($-.0961$) in 2002 did not make the list of DIF items (*MH D-DIF* $= -0.979$). It was Grade 8 item, and the topic was population. The FDOE did not release the item. The population of a town is given, and students have to find the approximate population some time back when it was approximately $a\%$ smaller. The item had the word *approximately* twice.

None of the three DIF items favoring female students was released. One Grade 10 item was about solving a system of linear equations given in a social science context. The item was solved by only 25% of the female students and 23% of the male students. Although the item is DIF, it was barely flagged by *STAND P-DIF*, which can be explained by the low number of correct responses. Another algebraic manipulation item was a DIF in Grade 8. The item is a word problem in retail/currency /business. However, an equation was set up, and the student could solve the equation without actually reading the entire problem.

[6]The Florida Comprehensive Assessment Test (FCAT) Mathematics items appear by permission of the Florida Department of Education, Office of Assessment, Tallahassee, Florida 32399-0400.

The second Grade 10 item was a pattern problem. Mathematically, it was equivalent to a 2001 DIF item favoring female students about a marathon runner (see page 51). The topic was retail/currency/business, which is usually neutral.

Two DIF items favoring male students in 2001 appeared on the FCAT in 2002. One of them was a DIF item again. The other just barely missed the cutoff. The lone DIF item favoring female students in 2001 was on the test in 2002. This time it was not a DIF item, but was flagged by *STAND P-DIF*.

Conclusion

Study 2 confirmed some of the findings from Study 1. Two context topics, population and sports/recreation/transportation, were prevalent in items favoring male students, as were estimated answers and items with vehicles. Almost all the characteristics in Study 1 that did not favor either male or female students did not favor either gender in Study 2. Only noncomputed solutions moved from significant to nonsignificant, whereas items with context or no-context moved in a different direction. In Study 2, the measuring topic was more neutral than in Study 1. One category, algebraic manipulations, clearly favored female students in Study 2. It was not a change from Study 1, because so few items were in that category in Study 1.

The main conclusion from Studies 1 and 2 is that a foolproof prediction of DIF is not possible; however, it is possible to identify characteristics that are likely contributors to gender-related DIF. Even if one cannot predict DIF in all cases, one can learn more about possible underlying reasons for DIF.

There are at least two implications from knowing contributing factors. One is the possibility of reducing assessment bias by carefully choosing topics. One cannot assess students' mathematical achievement if one measures something nonmathematical. I understand, for example, that topics such as population should be used; it is an important phenomenon to understand. At the same time, however, I do not understand why there were so many bicycle

race items on the FCAT. The topic of measuring is very important, as well, but I do not see why students need to find the area of the deck of a cargo ship instead of, for example, a fancy banner for mathematical competition. Although my complains may seem trivial, my analysis did show that some topics are contributing to DIF and need to be taken seriously.

By knowing mathematical contributors to DIF, teachers can alter classroom instruction to help students with content that contributes to DIF. However, more research should be done, and this study can help identify content categories that create trouble for male students or female students.

For Study 3, I decided not to change categories, topics, or characteristics. The main goal for Study 3 was to confirm the possible contribution of the item characteristics to DIF. The primary hypothesis was the same as for Study 2.

CHAPTER 5

STUDY 3: FLORIDA COMPREHENSIVE ASSESSMENT TEST 2003

Method

The Florida Comprehensive Assessment Test

For those Florida students planning to graduate from high school in 2003, passing the FCAT Grade 10 in mathematics became a requirement. According to the *Miami Herald*, thousands of people protested against the FCAT in May 2003 because 12,500 high school seniors were denied a diploma for not passing the test (Pinzur, 2003). One week earlier, the same newspaper had reported that the average 2003 FCAT scores were the highest in 5 years in all subjects and at all grades (Pinzur & Ovalle, 2003). I am not aware whether the new requirement affected the construction of the test; however, the 2003 test in mathematics does appear easier than in 2001 and 2002.

The distribution of items by official SSS strand was similar to that of 2001 and 2002; it is shown in Table 19. In 2003 the FCAT used 6 items from 2001, 20 items from 2002, and 19 items from both years. In all, 45 items had been analyzed in Studies 1 and 2.

Sample and Classification

The demographic breakdown of the students' data for 2003 is given in Table 20 . The sample size was the largest for all 3 years, only 10 forms were used in that year compared with 15 and 30 in 2001 and 2002, respectively.

The classification of the items did not change from Study 2. I did not add any new characteristics or topics. The distribution of items by category and original strand on the FCAT in 2003 is shown in Table 21.

Table 19:
Distribution of 2003 FCAT Items by Strand and Grade

	Grade			
Strand	8	9	10	Total
Number	11	9	9	29
Measurement	11	9	8 (9)	28 (29)
Geometry	3 (6)	4 (11)	5 (9)	12 (26)
Algebra	12	12	13	37
Data analysis and probability	0 (9)	0 (9)	2 (9)	2 (27)
Total	37 (49)	34 (50)	37 (49)	108 (148)

Note. The total number of items on 2003 FCAT is in parentheses when not all items in a strand were used.

Results

FCAT SSS Strands

An ANOVA of all 148 FCAT items from 2003 with respect to SSS strands showed that the *STAND P-DIF* means were significantly different ($F(4, 143) = 2.521$, $p = .04$), whereas the *MH D-DIF* means were not significantly different ($F(4, 143) = 2.209$, $p = .07$). According to a Tukey post hoc test, *STAND P-DIF* means for measurement and algebra strands were significantly different ($M = -.009$, $SD = .031$ and $M = .017$, $SD = .039$, respectively). For the 108 items used in the analysis, the *STAND P-DIF* and *MH D-DIF* means were significantly different across strands ($F(4, 143) = 4.487$, $p = .002$ and $F(4, 143) = 3.539$, $p = .01$). The *STAND P-DIF* means for the measurement and algebra

Table 20:
Distribution of 2003 Data Sample by Grade, Gender, and Race

	Gender		Race[a]			
Grade	Female	Male	White	Black	Hispanic	Total
8	7,944	7,231	8,401	3,507	2,688	15,175
9	8,861	8,188	8,982	4,051	3,353	17,049
10	7,321	6,389	7,734	2,940	2,493	13,710

[a] Data for other racial groups are omitted.

strands were still significantly different. In addition, for both indices, *STAND P-DIF* and *MH D-DIF*, the means for the algebra and the data analysis and probability strands were significantly different. However, because I retained only two data analysis and probability items, I do not think this result is reliable. This result is different from that of the previous studies. In 2002, there were no significant differences in means for strands, whereas in 2001 the mean for measurement was significantly different from the means for algebra and for data analysis and probability when all items were considered, and there were no significant differences when only items used in the analysis were tested.

Content, Context, and Other Characteristics

Content Categories

An ANOVA showed that the *STAND P-DIF* and *MH D-DIF* means for the categories were significantly different ($F(11, 96) = 3.093$, $p = .001$ and $F(11, 96) = 3.050$, $p = .002$, respectively). The means for both indices are reported in Table 22. A Tukey post hoc test indicated that the *STAND P-DIF* and *MH D-DIF* means for algebraic manipulations were different from the means for number, both categories for geometrical measurement and informal

Table 21: Distribution of 2003 FCAT Items by Strand and Category

	Strand					
Category	Number	Measurement	Geometry	Algebra	Data analysis & probability	Total
Number	12					12
Geometrical measurement	0/1	8/6	4/2	1/0		13/9
Informal algebra	12/1	3/5		2/1	1/0	18/7
Pattern			0/2	10/4	0/0	10/6
Setting-up/ translation		2	1	7		10
Functions	0/1		0/2	2/3	0/1	2/7
Algebraic manipulations				3		2
Other	2	4	1	4		11
Total	29	28	12	37	2	108

Note. For subdivided categories, the first number is the first category, and the other is the second, see pages 56–59.

algebra, and for the category other. However, the results should be interpreted with caution because the algebraic manipulations category had only three items.

Topic of the Context

In 2003, the FCAT had just one item on the topic of population. In order to use post hoc analysis in ANOVA, I decided to move this item into topic other. All other topics remained

Table 22:
Mean DIF Values for Content Category in Study 3

Content category	Number of items	*STAND P-DIF* (SD)	*MH D-DIF* (SD)
Number	12	.000 (.033)^{a}	−0.017 (0.422)^{a}
Geometrical measurement			
Formula/scale	13	.001 (.043)^{b}	0.052 (0.619)^{b}
Modified formulas	9	−.017 (.018)^{c}	−0.235 (0.220)^{c}
Informal algebra			
Proportionality	18	−.017 (.037)^{d}	−0.194 (0.562)^{d}
Rate	7	−.007 (.015)^{e}	−0.066 (0.246)^{e}
Patterns			
Function	10	.007 (.040)	0.050 (0.493)^{g}
Pictorial	6	.013 (.024)	0.166 (0.328)
Setting up/translating	10	.016 (.029)	0.173 (0.371)
Functions			
Variable	2	.063 (.016)	0.946 (0.236)
Interpretation	7	.011 (.034)	0.149 (0.486)
Algebraic manipulations	3	.073 (.022)^{abcdef}	1.064 (0.278)^{abcdefg}
Other	11	−.004 (.032)^{f}	−0.035 (0.416)^{f}
Total	108	.002 (.036)	0.031 (0.502)

Note. Two entries with the same superscript in the same column are significantly different according to the Tukey test, $p < .05$.

the same. An ANOVA with respect to topic indicated that the *STAND P-DIF* and *MH D-DIF* were significantly different ($F(6, 101) = 2.955$ and $F(6, 101) = 3.196$, $p < .01$,

Table 23:
Mean DIF Values for Topic in Study 3

Topic	Number of items	Mean *STAND P-DIF* (SD)	Mean *MH D-DIF* (SD)
No context	13	.012 (.028)	0.131 (0.354)
Sports/recreation/transportation	15	$-.023$ (.045)^a	-0.299 (0.540)^a
Physical sciences	12	.001 (.024)	0.015 (0.317)
Retail/currency/business	27	.008 (.037)	0.113 (0.539)
Social studies	4	.049 (.035)^a	0.783 (0.449)^a
Measuring	20	$-.002$ (.033)	-0.039 (0.477)
Other	8	.001 (.031)	0.031 (0.468)
Total	107	.002 (.036)	0.032 (0.502)

Note. Two entries with the same superscript in the same column are significantly different according to the Tukey test, $p < .05$.

respectively). Although this time only two topics had different means, I report all means in Table 23 to show the direction of difference for each topic.

A two-way ANOVA indicated a significant main effect of item category on the value of *MH D-DIF*,[1] $F(11, 65) = 1.950$, $p = .049$. There was no main effect of item topic on *MH D-DIF*, and no significant interaction between subcategory and topic. However, the results were not reliable since Levene's test of equality of error variances for independent variables was significant ($p = .012$).

[1]The results for *STAND P-DIF* were similar.

Other Characteristics

The results for those characteristics that demonstrated significant *STAND P-DIF* mean differences are summarized in Table 24. The results for the other characteristics are not reported. Two of the characteristics that had significantly different means in Study 2 had significant differences in *STAND P-DIF* means in this study: estimated answers and items with or without linear speed. Converting units had significant difference in means in Study 1, and the difference was significant again in this study. The new characteristic that moved to significant status is items with or without visual stimuli. Estimated answers was ranked significant in all three studies. Linear speed was significant in two studies and was not tested separately from rate in Study 1: In that study, rate had significant differences in means.

DIF Items

Only one Category B DIF item favored male students in 2003. This item was not released. The number of Category B DIF items favoring female students went from one in 2001 to three in 2002 and five in 2003. Only two items were released. First, I describe the item favoring male students, then I describe items favoring female students, and finally I examine DIF items from the previous year that appear on the FCAT in 2003.

The Grade 10 DIF item favoring male students was classified by SSS in the data analysis and probability strand. I classified this item as the first category of informal algebra. A table in the item described wins and losses for softball teams and asked about how many games one team should win to be tied with winning team if it would continue to win at the current rate. Although softball is a sport that women play, I think that the topic still favored male students. Mathematically, the item requires the student to find the number of games that the winning team wins by using a simple proportion and then subtracting the number of games that the other team had already won. The item was a GR. It is very likely that more female students just skipped the item than male students did: 39% of female students and

Table 24:
STAND P-DIF Means for Item Characteristics That Contributed to DIF in Study 3

Item characteristic	Number of items	Mean (SD)	t (p value)	Effect size
Visual stimuli				
No	49	.010 (.039)	2.06	
Yes	59	−.005 (.033)	(.042)	0.42
Estimated answers				
No	90	.006 (.038)	2.53	
Yes	18	−.017 (.021)	(.013)	0.64
Converting units				
No	102	.004 (.036)	2.28	
Yes	6	−.030 (.025)	(.024)	0.96
Linear Speed				
No	102	.004 (.036)	2.36	
Yes	6	−.032 (.036)	(.02)	1.02

55% of male students did this item correctly. Unfortunately, actual students' responses for GR items were not available to me.

Two of the female DIF items were DIF in the previous years. One Grade 10 item from 2001 about a marathon runner was discussed on page 51; another one, a Grade 8 item from 2002, was algebraic manipulations item in retail/currency/business context. This item was briefly described on page 71. The second Grade 8 item was a one-step algebraic manipulations item in a social science context. I would expect this item to be at the easy level of difficulty,

but only 58% of the female students and 50% of the male students solved it. The item was a GR type.

Two Grade 9 DIF items were from the function category. One was to find the slope of the line when two points are given; this is the interpretation category of function. The second item was from the variable category of function: finding a dependent or independent variable given the other one. Mathematically, this item is similar to the DIF item favoring male students in 2001 about population of a town shown on pages 70–71. In both items, the dependent variable is given, and the independent variable should be found. However, the topic was different: population for the DIF item favoring male students and retail/currency/business for the item favoring female students.

I discuss the last DIF item favoring female students in more detail. The Grade 10 item was released:

> Max works at a factory that manufactures fiberglass tanks. He needs to make a right circular cylindrical fiberglass tank that has a diameter of 6 meters and a height of 8 meters. What will be the volume, in cubic meters, of this cylinder?
> (Source: FDOE[2] (FCAT, 2006a, p. 52))

This item required the student to slightly alter the basic formula provided in the reference sheet. The formula was given with the radius; in the item, the diameter was provided. The item was solved correctly by roughly the same percentage of male and female students, 51%, which was rounded down for the female students and up for the male students. Nevertheless, the item was a DIF favoring female students. The Figure 4 shows the empirical item response functions for female and male students. This is a good example to show that the difference in percent correct responses between female and male students is not the same as DIF. At the same time it probably shows that not all DIF items are biased. A plausible reason that the female students did better on this item than the male students matched on the total score

[2]The Florida Comprehensive Assessment Test (FCAT) Mathematics items appear by permission of the Florida Department of Education, Office of Assessment, Tallahassee, Florida 32399-0400.

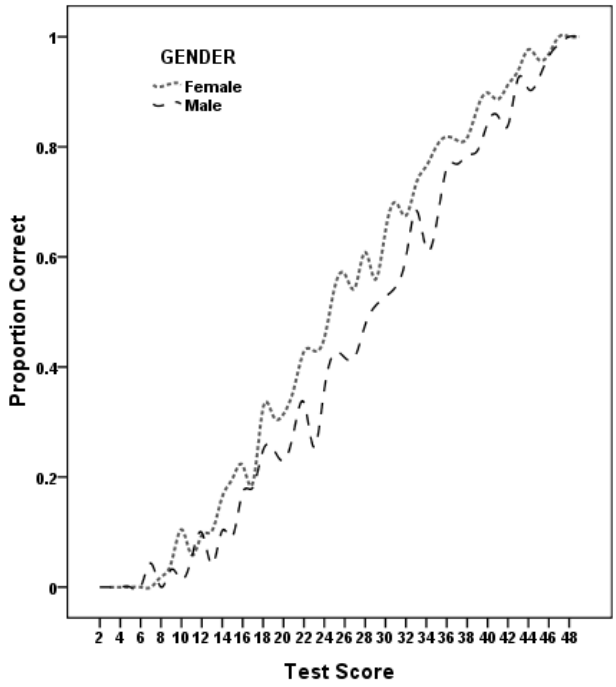


Figure 4: Empirical IRF for a female B-DIF item. (In both groups, 51% of students answered the item correctly.)

could be that they were paying more attention to details and noticing that the diameter was given, not the radius.

Now, I take a look at how this study DIF items behaved on the previous years' tests and vice versa. Of five DIF items favoring female students, one item appeared in 2001 and 2002, and one item in 2002. The first item about a marathon runner, discussed on page 51, was a DIF item in 2001 and 2003. In 2002, it did not make a list, but still highly favored female students. The other item was on the FCAT for 2 years and was a DIF both times. This Grade 8 item on solving an equation in retail/currency/business was briefly described on page 71.

The DIF item favoring male students in 2003 had not appeared on the test before (at least on the forms I analyzed). However, some DIF items from previous years appeared in 2003, and they were not a DIF. One item, a Grade 10 item about a boat race, was briefly

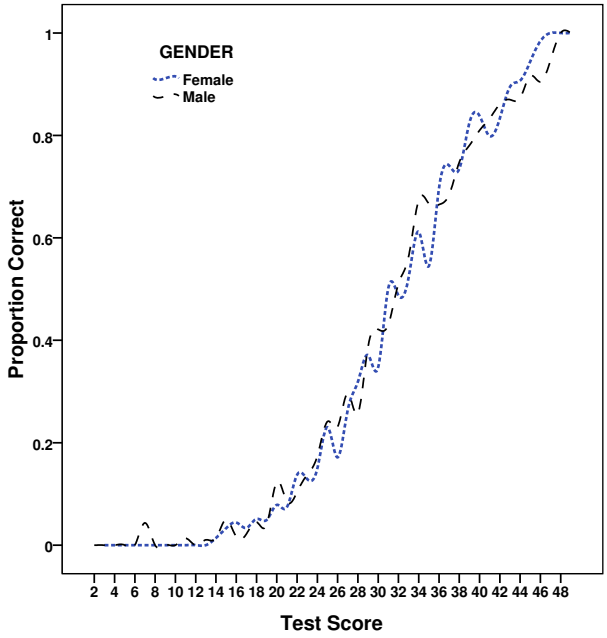


Figure 5: Empirical IRF for a non-DIF item with 32–41% correct responses.

discussed on page 70. In 2003, this item was not flagged by *STAND P-DIF*; male students' performance dropped on this item from 71 to 66%. Female students' percent correct at 56% was far below that of male students, but the item did not show significant DIF. Another Grade 10 item about the area of the deck of cargo ship (see page 69) was also a DIF only in 2002 but not the year before or after. It moved to a very slight positive value of *STAND P-DIF* in 2003, although the performance of female students was below that of male students in all three years: 36 to 45% in 2001, 28 to 40% in 2002, and 32 to 41% in 2003. The response functions of female and male students on this item in 2003 are shown in Figure 5.

Another item, about acceleration of a car, was discussed in detail on pages 67–69. Although the item was not a DIF in 2001 and 2003, it was flagged by *STAND P-DIF* for additional review. Male students performed better than female students overall (55–67%) and matched on the total score (*STAND P-DIF* $= -.05$).

Conclusion

This study again confirmed that the topic of the item is important. Although there was a shift in 2003 to more FCAT items in the retail/currency/business topic, which is clearly a neutral topic, there were still significant differences between the sports/recreation/transportation and the social studies topics. There were also significant difference between some categories. However, the result may be not reliable, because very few items were in the algebraic manipulations category. Of the other characteristics related to content, estimated answers, converting units, and linear speed concept were challenging for female students. A good sign, however, was that the effect size indices went down from the previous years. Visual stimuli in the item affected the female students' performance in this study, but not in the previous ones. However, figures, tables, and graphs tested separately did not show significant differences in means. As I mentioned before, some researchers had found that visual stimuli benefited male students (Harris & Carlton, 1993; Mendes-Barnett & Ercikan, 2006; Ryan & Chiu, 2001), whereas other researchers drew a different conclusion that an item with real-world object was less likely to be biased against female students (A. S. Cohen & Ibarra, 2005). Overall, the existing significant differences became smaller and more subtle. The test designer clearly moved to more gender-neutral items.

After reviewing the DIF items I concluded that although some item properties contribute to DIF, those properties are affected by many other factors, and the results of any study on DIF should be interpreted very cautiously. For example, the matching is done on the total test score, which means that other items on the test influence the DIF indices calculations. Nevertheless, one needs to study DIF items. Some properties appear to be contributing factors in many studies; test designers and educators should know about these properties to design fair tests and to pay more attention to teaching concepts that found to be challenging to some demographic groups.

CHAPTER 6

STUDY 4: DIF CHARACTERISTICS ACROSS YEARS AND GRADES

In this chapter I summarize the results of the previous three studies and describe the changes from year to year and from grade to grade for properties and characteristics that potentially can contribute to DIF.

Item Characteristics Across Years

My overview starts with categories. Figure 6 shows the *STAND P-DIF* mean values for several categories. *STAND P-DIF* values range from -1 to 1. An item is flagged for fairness review when $|STAND\ P\text{-}DIF| > .05$. A negative value of *STAND P-DIF* means that the item favors male students, and a positive value favors female students. The categories shown in the figure are geometrical measurement/modified formula, informal algebra/proportionality and informal algebra/rate, algebraic manipulations, and other. All these categories indicated different performance by male and female students. In 2001, a Tukey post hoc test failed to point out the significant differences in the *STAND P-DIF* means for categories,[1] although an ANOVA showed that the means were different. In 2002, the algebraic manipulations category had a mean different from the other four categories shown in Figure 6. In 2003, the algebraic manipulations mean was different from those of the categories mentioned above and also from those of the number and geometrical measurement/formula/scale categories. However, the algebraic manipulations category had only three items, so little weight should be given to this result.

[1] An ANOVA was run on categories that were slightly different from the category classification introduced in the second study.

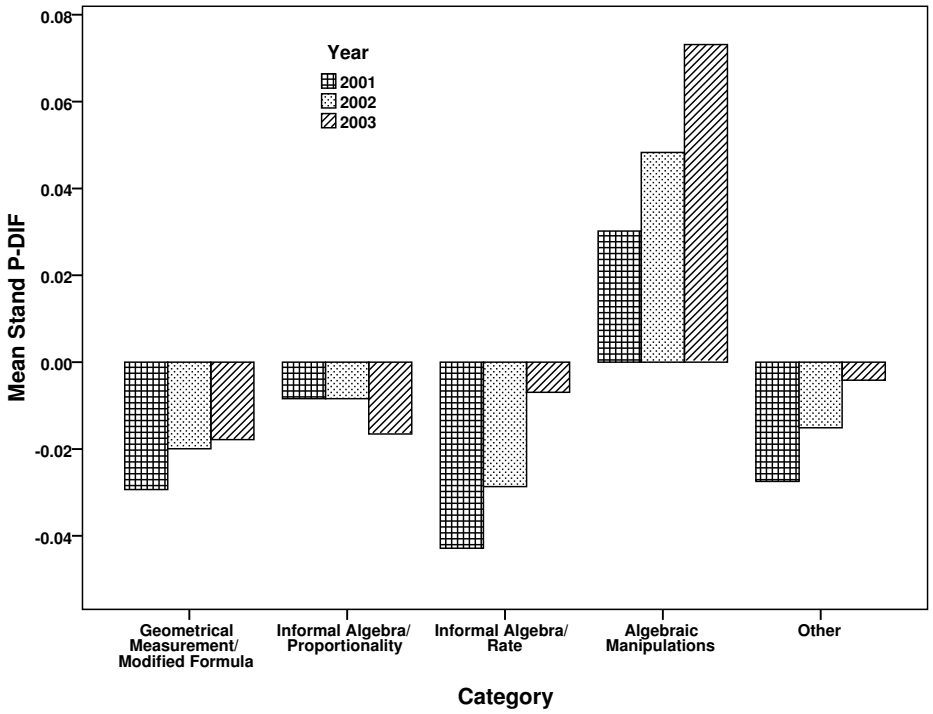


Figure 6: *STAND P-DIF* for selected categories in 2001–2003. An item with a positive index value favors female students. Items with $|STAND\ P\text{-}DIF| > .05$ are flagged for fairness review.

Although categories had different means, topics showed bigger differences. Figure 7 shows the means for all topics across the years. In 2003, only one item had the topic of population. In that analysis, the item topic was coded as other. For this overview, however, I recoded it as population. As can be seen from the figure, the topics of sports/recreation/transportation and population heavily favored male students. In 2003, these topics had less effect on the *STAND P-DIF*. The mean for the social studies topic soared in that year. This topic generally favored female students. In 2003, there were only four items on this topic, and one of them was a highest magnitude DIF item, which is probably the reason for the very high mean.

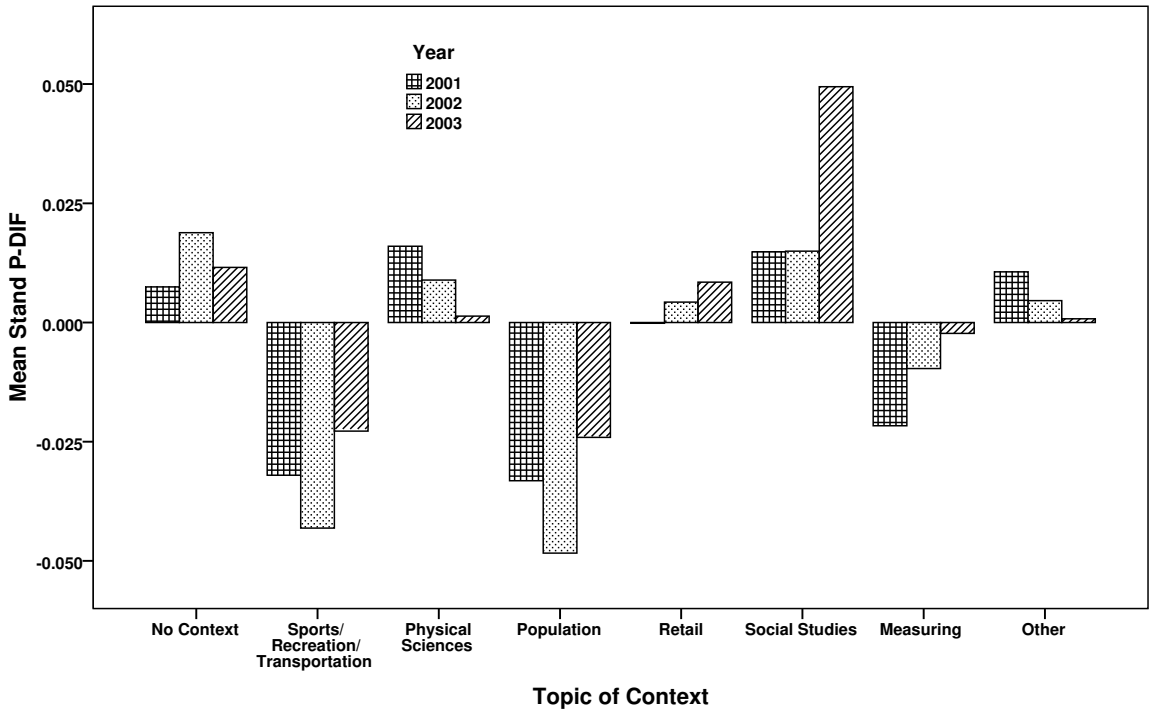


Figure 7: *STAND P-DIF* for topics in 2001–2003.

Two characteristics, estimated answers and linear speed, appeared in all three studies[2] as significant contributors to gender-related DIF. The *STAND P-DIF* means across the years are shown in Figure 8. These characteristics belong directly to the subject tested. They fit the dimension that test makers intended to measure.

Items with estimated answers usually have the words *approximately* or *estimate*. There were 16 such items in 2001, 21 items in 2002, and 18 items in 2003. The total number of items was 107 to 108. Although the significant differences existed in all 3 years, there was a positive trend: the effect size indices went down from 1.14 in 2001 to 0.76 in 2002 and 0.64 in 2003. I cannot say whether this trend was the result of better teaching of estimation or the result of other factors such as very easy items or very difficult items.

[2]Although linear speed was not tested in Study 1, for this overview, I tested this characteristic on 2001 data.

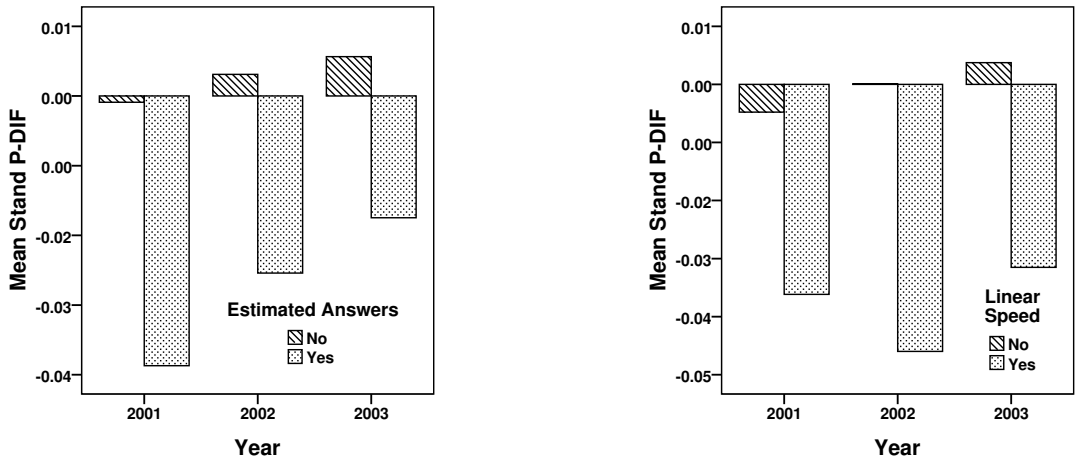


Figure 8: Items with estimated answers and linear speed in 2001–2003.

Items with linear speed were not as frequent as items with estimated answers. Only 6 items with linear speed appeared in each of three years. Although the number of items was small, the results were consistent over the years. The effect size indices were high in all 3 years: 0.75 in 2001, 1.28 in 2002, and 1.02 in 2003. In 2002, the average of *STAND P-DIF* mean was very close to $-.05$, the threshold for an item to be flagged for additional review. At the same time, the other rates characteristic was not a contributing factor to DIF.

Two more characteristics appeared more than once on the list of significant contributors to gender-related DIF: converting units and vehicles in the statement of the item. Figure 9 shows their *STAND P-DIF* means in 2001–2003. Converting units appeared on 5 items in 2001, and the *STAND P-DIF* mean was significantly different from those of other items, with effect size 1.17. There were 8 items in 2002 and no significant difference, and 6 items in 2003 with a significant difference again and effect size 0.96. The number of items with vehicles went down from 17 in 2001, to 11 in 2002, and 8 in 2003. This characteristic was significant in 2001 and 2002, with effect size indices 1.19 and 1.35, respectively.

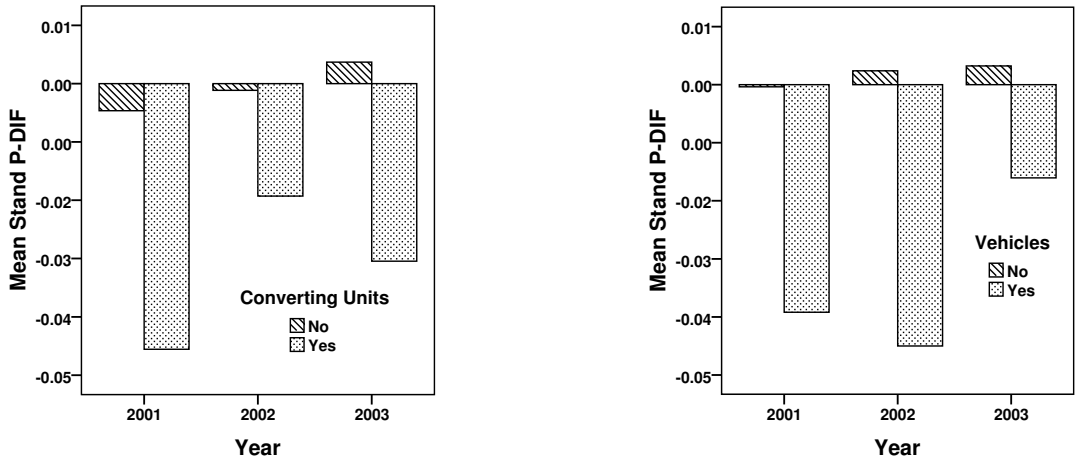


Figure 9: Items with converting units and vehicles in the context in 2001–2003.

Three other characteristics appeared once on the list of contributors to gender-related DIF. They are noncomputed vs. computed solution, visual stimuli, and context vs. noncontext items. I already discussed noncontext items as a part of the topic. The *STAND P-DIF* means for other two characteristics in 2001–2003 are shown in Figure 10.

To conclude the discussion about changes over the years, I look at those 19 items that appeared on the FCAT in all 3 years. The majority of the items demonstrate negligible changes in DIF indices. I briefly discuss those items that showed more than just negligible changes. Several of them were discussed in previous chapters since they were DIF in 2003 or before. One item favoring female students about a marathon runner was shown on page 51. This item was DIF in 2001 and 2003, but not in 2002, although it was flagged by *STAND P-DIF*. Three items favoring male students about boat sailing, area of a deck of a cargo ship, and car acceleration were discussed on pages 67–70. These three items had changes in

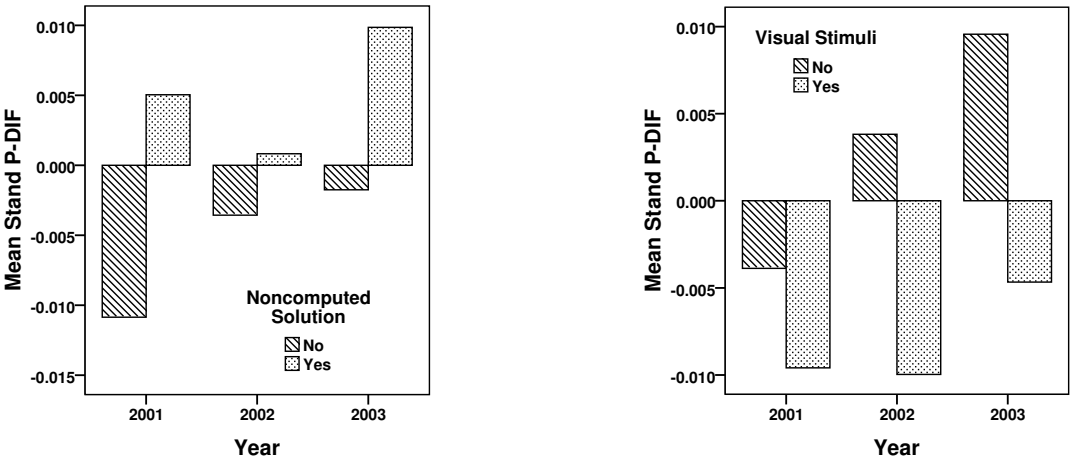
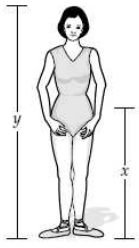


Figure 10: Items with noncomputed solutions and visual stimuli in 2001–2003.

DIF indices. One other item was never a DIF item. It was flagged by *STAND P-DIF* twice, in 2001 and 2002, but not in 2003. The item was released by FDOE:[3]

> Artists have traditionally studied human proportions to draw human figures realistically. When drawing a female figure like the one in this picture, the realistic ratio of the distance from the hip to the toe (x) to the height of the woman (y) is 0.613. An artist is creating a 9-inch-high drawing of a woman. What should be the approximate distance in inches from the hip to the toe?
>
> (Source: FDOE (FCAT, 2005, p. 18))



It was a geometrical scale item, a type of item on which female students usually perform well. However, this item had the word *approximate*, which may be the reason that the female

[3]The Florida Comprehensive Assessment Test (FCAT) Mathematics items appear by permission of the Florida Department of Education, Office of Assessment, Tallahassee, Florida 32399-0400.

students did not do well on it. Although the item was not flagged by *STAND P-DIF* in 2003, the female students' performance on it was well below the male students' performance: 53 to 63% in 2001, 57 to 66% in 2002, and 57 to 65% in 2003.

Item Characteristics Across Grades

In the previous studies I looked at the test characteristics within a year, combining Grade 8, 9, and 10 items. In this section I briefly report on item characteristics by grade, combining items from 2001 to 2003.

Only a few items were repeated across the grades the same year, whereas there were many repeated items within a grade across years. Table 25 shows the number of items that appeared on the FCAT test once, twice, or three times at the same grade. I used all items in the analysis. Since many items were repeated, these results should be interpreted with caution.

Table 25:
Repeated FCAT Items by Grade From 2001–2003

Number of	Number of items		
appearances	Grade 8	Grade 9	Grade 10
1	57	51	51
2	15	19	17
3	5	6	8
Total	107	107	109

The results by grade in the content categories are summarized in Figure 11. I had to remove several categories to fit in one figure. The categories that are not shown are formula/scale of geometrical measurement, setting up/translating, and other. None of the three changed much from year to year, and none benefited either gender. The formula/scale mean

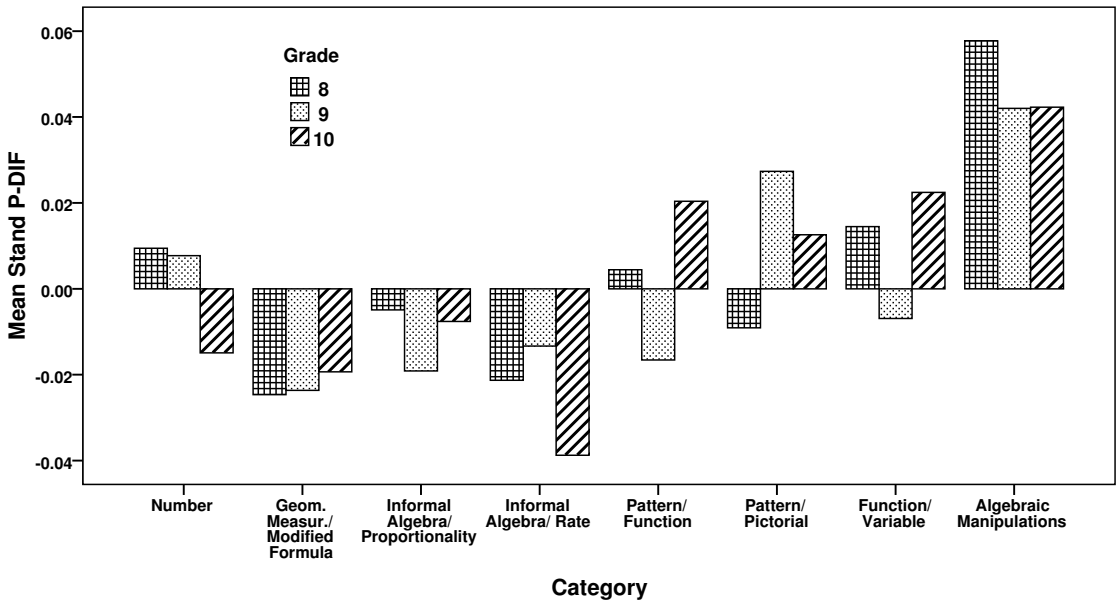


Figure 11: *STAND P-DIF* for categories by grade.

was slightly negative, the mean for setting up/translating was slightly positive, and the mean for other was slightly tilted toward male students. Some categories consistently benefited the same gender across the grades. These categories were modified formula of geometric measurement, both categories of informal algebra, and algebraic manipulations. Only the last of these categories benefited female students.

Topics were more consistent across the grades than categories. The results for the context topic are presented in Figure 12. There were clearly topics that benefited female students, such as social studies and no-context, and topics that benefited male students, such as sport/recreation/transportation, population, and maybe measuring. The topics of retail and physical sciences were close to neutral.

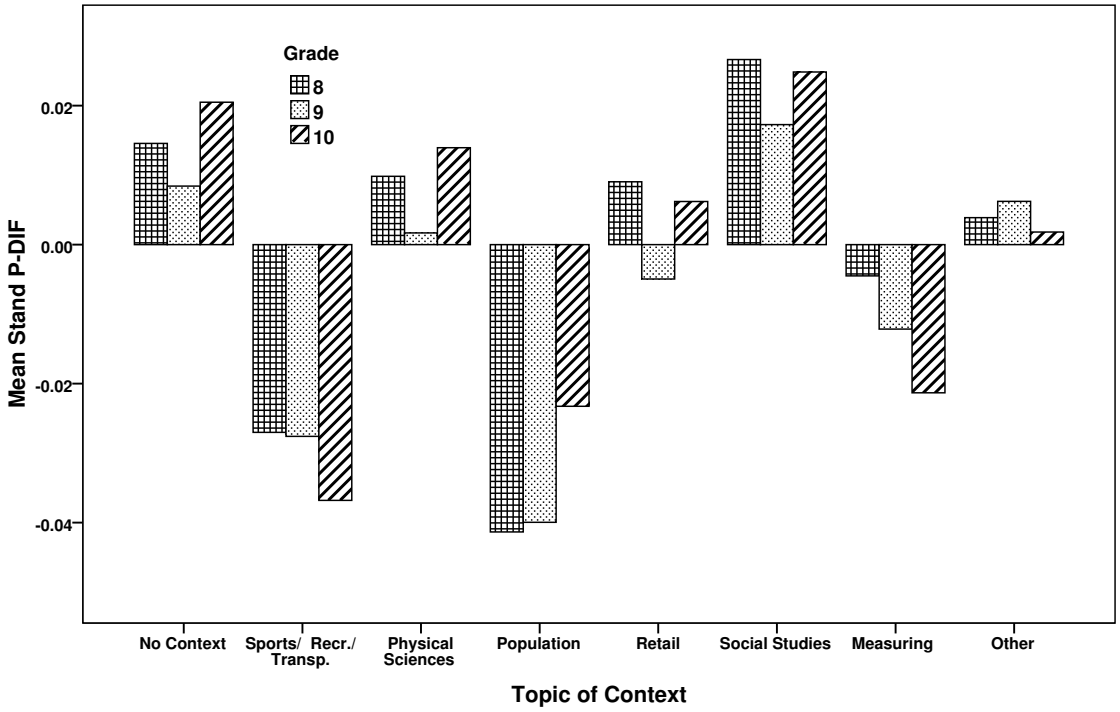


Figure 12: *STAND P-DIF* for topics by grade.

At Grade 8, the *STAND P-DIF* means for categories were significantly different ($F(10, 96) = 3.558$, $p < .001$). A Tukey post hoc test showed that the algebraic manipulations category mean was significantly different from that for geometrical measurement (formula/scale and modified formula) and informal algebra (proportionality and rate). The *STAND P-DIF* means for topics were significantly different, too ($F(7, 99) = 3.169$, $p < .01$). But only the mean of the retail/currency/business topic was different from the mean of the sports/recreation/transportation. Estimated answers, converting units, items with vehicles were factors contributing to DIF. One new characteristic appeared to be a contributor to DIF: one-step items versus more-than-one-step items. Results on characteristics are reported in Table 26.

Table 26:
Characteristics of Items by Grade

Characteristic	Mean difference[a] (SE)		
	Grade 8	Grade 9	Grade 10
Estimated answers	$-.037^{**}$(.011)	$-.023^{**}$(.007)	$-.030^{**}$(.010)
Converting units	$-.033^{*}$ (.016)	$-.038^{**}$(.011)	$-.012$ (.018)
Context	$-.015$ (.011)	$-.017^{*}$ (.008)	$-.028^{*}$ (.012)
Vehicles	$-.032^{**}$(.010)	$-.006$ (.014)	$-.059^{**}$(.010)
Noncomputed solution	.009 (.008)	$.019^{**}$(.006)	.004 (.009)
One step[b]	$.020^{*}$ (.009)	$-.010$ (.013)	$-.006$ (.019)
Linear speed	$-.036$ (.020)	$-.009$ (.012)	$-.062^{**}$(.015)
Other rates[c]	$-.003$ (.009)	$-.010$ (.008)	$-.030^{**}$(.010)
Visual stimuli	$-.007$ (.007)	.002 (.006)	$-.024^{**}$(.008)
Multiple choice	$-.009$ (.007)	.007 (.006)	$-.021^{*}$ (.008)

[a] Difference between the *STAND P-DIF* means of items with characteristic and without.
[b] Test performed on items with computed solutions.
[c] Items with linear speed were excluded from the analysis.
* $p < .05$.
** $p < .01$.

In Grade 9, the *STAND P-DIF* means for categories were significantly different ($F(10, 94) = 2.426$, $p < .05$), however, Levene's test for homogeneity of variances was significant ($p < .01$); therefore, a Tamhane post hoc test was used. The function/interpretation category was significantly different in means from the geometrical measurement/modified formulas and informal algebra/proportionality categories. The *STAND P-DIF* means for topics were significantly different ($F(7, 98) = 5.06$, $p < .001$). The mean for the topic of

population was significantly different from 4 topics: no context, physical sciences, social studies and other. The mean for the sports/recreation/transportation topic was different from those of the no context and social studies topics. Of the characteristics, estimated answers, converting units, noncomputed solutions, and context items appeared to be factors contributing to DIF. Items involving linear speed and vehicles were not. An interesting fact is that there was only one algebraic manipulations item across all 3 years.

In Grade 10, the *STAND P-DIF* mean for the algebraic manipulations category was different from the mean for the rate category of informal algebra, the function/interpretation category, and the category of other ($F(10, 97) = 2.6, p < .01$). Only one item was coded as function/finding variables in all 3 years. Among the topics, the mean for sports/recreation/transportation was different from the means for no-context, physical sciences, retail/currency business, and social studies ($F(7, 101) = 4.293\, p < .001$). The characteristics contributing to DIF grew in number in Grade 10: linear speed, other rates, multiple choice, estimated answers, context, vehicles, and visual stimuli.

There was more variation across grades than across years. Only estimated answers appeared to be a contributor to DIF at all three grades. Converting units, vehicles, and context were contributing factors in two grades. Noncomputed solutions, linear speed, one-step items, other rates, multiple choice, and visual stimuli appeared only once. Although the results across the years and across the grades differed in some instances, there were categories, topics, and characteristics that behaved similarly. Algebraic manipulations favored female students across all grades and years. Informal algebra/rate was challenging to them across all years and grades. Topics were the most consistent in both analyses, as were items with estimated answers.

CHAPTER 7

STUDY 5: DIF ITEMS IN THE TRENDS IN INTERNATIONAL MATHEMATICS AND SCIENCE STUDY

Method

The Trends in International Mathematics and Science Study (TIMSS) is an international assessment to measure trends in mathematics and science learning. TIMSS is conducted by the International Association for the Evaluation of International Achievement (IEA), an independent international cooperative of national research institutions and government agencies. The aim of TIMSS is to improve the teaching and learning of mathematics and science by providing data about students' achievement relative to different curricula and instructional practices. TIMSS collects data on curriculum, students, teachers, and school principals through extensive questionnaires. These data, along with student assessment, give policy makers, curriculum specialists, and researchers a dynamic picture of educational practices around world and help them to devise policies for educational reforms (Mullis, Martin, Gonzalez, & Chrostowski, 2004). TIMSS 2003 is the third in a continuing cycle of international mathematics and science assessments conducted every 4 years. Nearly 50 countries participated in the 2003 study.

The goals of the present study were to compare the characteristics of DIF items in TIMSS 2003 for U.S. students with the characteristics on the FCAT and to see how the classification devised for the FCAT works with a different type of assessment test. According to the *TIMSS 2003 Technical Report* (Martin, Mullis, & Chrostowski, 2004), the gender differences in TIMSS 2003 were negligible in many countries at Grade 8. However, there were variations

across the countries. In some countries, female students significantly outperformed male students. In the United States, male students had higher achievement than female students.

Sample

The U.S. sample for TIMSS had 8912 students: 4629 female students and 4283 male students. There were 12 different forms of the test. I ran the analyses separately on each form. The sample size was around 740 students per form. The forms had from 29 to 60 mathematics items, and on average around 50 items. Two forms, 7 and 8, did not have any released items; I did not analyze them. Each released item was on the test on two or three different forms. However, some items were marked as "end of session" and the others as "regular session" in the description of scaling. I used raw data with the actual students' responses that I recoded to correct-incorrect responses. Constructed-response items with a correct-incorrect answer were treated the same as multiple-choice items. Constructed-response items with partial credit were counted as correct if at least partial credit was given (Michaelides, 2008). There are more complicated techniques to deal with polytomous items; however, they were not necessary for the present study. I ran an analysis on two forms with the partial credit items coded as incorrect and another analysis with them coded as correct. The difference in impact on indices of other items was negligible, so I decided to use the latter approach. However, any interpretation of the DIF status of constructed response items with partial credit should be done carefully.

Classification

There were 100 unique released items counting separately scored parts as separate items. I classified 75 as algebra or algebra-related. Table 27 shows the distribution of items by TIMSS content domain and category classification. I used the same classification as in Studies 2 and 3. The content classification worked well. I found that the TIMSS content domain classification was better aligned with my category classification than the FCAT classification.

Table 27:
Distribution of Selected TIMSS Items by TIMSS Content Domain and Content Category

	TIMSS content domain					
Content category	Number	Measurement	Geometry	Algebra	Data analysis & probability	Total
Number	12					12
Geometrical measurement		4/8				4/8
Informal algebra	16/2	3/2		1/0	1/0	21/4
Pattern				2/5		2/5
Setting-up translation				3		3
Functions			1/0	0/2	0/1	1/3
Algebraic manipulations				11		11
Total	31	17	1	24	2	75

Note. For subdivided categories, the first number is for the first subcategory, and the other for the second, see pages 56–59.

Many items that I classified as number were quite different from so-called number items on the FCAT; nevertheless, they satisfied the criteria for this category. Because TIMSS was conducted at Grade 8 only, it is not surprising that many items were in the first several categories. The algebra domain was represented mostly by patterns and algebraic manipulations items. Although Table 27 shows five pictorial patterns, three were parts of the same item. Nevertheless, that is more than there were on the FCAT. There were very few TIMSS

Table 28:
Distribution of Items by Topic in TIMSS 2003 and FCAT

		FCAT		
Topic	TIMSS	2001	2002	2003
No context	43	12	11	13
Sports/recreation/transportation	6	19	17	15
Physical sciences	1	16	16	12
Population		8	8	
Retail/currency/business	5	19	19	27
Social studies	2	7	9	4
Measuring	4	19	20	20
Other	16	8	7	8
Total	75	108	107	107

items in the informal algebra/rate category, whereas the informal algebra/proportionality category had the largest number of items. In the FCAT, almost all measurement strand items were geometrical measurement, either formula/scale or formulas/altered formula. In contrast, TIMSS had many non-geometrical-measurement items, which I classified into the informal algebra/proportionality category. All released items on TIMSS clearly fit one of the categories, and not a single item was classified as other.

The classification of topics did not work as well as the content classification. The majority of the TIMSS items either did not have a context or were classified as other topic. Table 28 shows the distribution of topic on the TIMSS 2003 and on the FCAT in 2001–2003. Almost all items were short.

Those characteristics that gave trouble to female students on the FCAT were almost absent from TIMSS. From 75 items, 3 items asked for estimated answers; 4 required converting units, and those units were metric. Only 4 items referred to vehicles, and 2 items involved linear speed. Fewer TIMSS items had visual stimuli than on the FCAT. Almost half of the FCAT items had either figure, table, or graph. Only a quarter of the TIMSS items had visual stimuli, mostly figures. This characteristic was a contributing factor to DIF only in the 2003 FCAT.

The analysis in this study is different from that of the previous studies. I had access to released items only, so testing the means for *STAND P-DIF* and *MH D-DIF* was not a sound method. I ran an analysis on all data to find items that were DIF by the ETS classification for *MH D-DIF*. The released DIF items are discussed and compared with the DIF items from the FCAT.

Although there were only 75 unique items in TIMSS, in my data these items were repeated one or more times from different forms of the test. As I mentioned above, the number of mathematics items on each form varied. The *MH D-DIF* index depends on the other items as well as on the number of items on the test. In other words, the *MH D-DIF* is not stable from test to test because it is not an intrinsic characteristic of the item.

Results

Compared with the FCAT, TIMSS had many more DIF items, and their magnitude was greater. By the ETS classification, there are three categories of DIF: negligible, Category A DIF (or simply A DIF); intermediate, Category B DIF; and large, Category C DIF. Counting repeated items, there were 198 items. Of these, 7 items were Category C DIF: 6 favoring male students, and 1 favoring female students. The number of Category B DIF items was also large: 8 favoring male students, and 8 favoring female students. Some of the items were DIF on more than one form. Overall, with repetitions, 23 DIF items favored male students and 12 items favored female students (some statistics on these items can be

Table 29:
Distribution of DIF Items on TIMSS and FCAT by Category

	Male students			Female students		
	TIMSS		FCAT	TIMSS		FCAT
Category	C DIF	B DIF	B DIF	C DIF	B DIF	B DIF
Number	2	3			1	
Geometrical measurement		2	3			1
Informal algebra	3	3	4	1	1	
Pattern					2	2
Setting-up/translation					1	
Function	1		1		1	1
Algebraic manipulations					2	3
Other			1			

Note. Each item appears in the table just once. Items in both DIF categories for a given gender are counted as C DIF only.

found in Appendix C). DIF items favoring male students were more concentrated in the algebra-related group, whereas for female students DIF items were mostly in the algebra group. The ratio of DIF items in the algebra-related group to DIF items in the algebra group was 13 to 1 for male students and 3 to 6 for female students. The distribution of DIF items from the TIMSS and the FCAT by categories is given in Table 29. Similar to the results from the FCAT study, female students in TIMSS did relatively worse than male students on informal algebra and geometrical measurement and better on algebraic manipulations. In addition to the findings of the FCAT studies, male students performed relatively better on the number and female students on the pattern categories. On the FCAT,

these categories did not have any differences in means. As I mentioned above, TIMSS items in the number category were different in scope from those on the FCAT. On TIMSS, the number category had more items with fractions and decimals. On the FCAT, the number category had mostly scientific notation and simple calculation items. My study of the FCAT did not find a difference in performance on items with fractions, but my study of the TIMSS did. There were more items with fractions favoring male students than female students. In contrast, female students performed relatively better than male students on patterns.

Table 30 shows the distribution of DIF items on the TIMSS and on the FCAT by characteristic. There were no big differences between the TIMSS and the FCAT if one considers that some characteristics such as converting units or estimated answers were barely present on the TIMSS, and items on the FCAT were predominantly with a context. Items with fractions were discussed above. I included all items with fractions in this characteristic whether they were from number or algebraic manipulations. There were no DIF items on the FCAT with a noncomputed solution, although that characteristic benefited female students and was significant in 2 out of 3 years. On the TIMSS, female students and male students had the same number of DIF items with a noncomputed solution.

As I mentioned above, my context topic classification did not work well with the TIMSS items. Most of the topics I had to classify as other. Male students had five DIF items and female students had one DIF item with that topic. One of the two remaining DIF context items favoring female students had retail/currency business, and the other was in the topic of sports/recreation/transportation. The latter topic favored male students in the FCAT study. Among DIF items favoring male students, two had this topic, and there was one item in each of the following topics: measuring and physical sciences. Statistics about no-context items are reported in Table 30.

One interesting detail came out when I compared DIF items to the items in the Kilpatrick et al. (2007) study on U.S. students' performance on TIMSS algebra items. Kilpatrick et al. looked at those items on which U.S. students did well and did poorly in absolute terms

Table 30:
Distribution of DIF Items on TIMSS and FCAT by Characteristic

	Male students			Female students		
	TIMSS		FCAT	TIMSS		FCAT
Characteristic	C DIF	B DIF	B DIF	C DIF	B DIF	B DIF
Visual stimuli	2	2	5		1	
Noncomputed solution	1	1			2	
Fractions	3	3			1	
Ratio, percent	4		2		1	
No context	1	4			6	
Estimated answers			5			
Converting units			3			
Vehicles		1	6			1
Linear speed		1	2			
Other rates			2			

Note. Each item appears in the table just once. Items in both DIF categories for a given gender are counted as C DIF only.

and relative to the performance of students from other countries.[1] An item was classified as absolute high-performance if more than 75% of the students answered it correctly. An item was classified as relative high-performance if American students' performance was the first or second among seven countries that Kilpatrick et al. chose as a set of systems representing comparably developed countries. An item was classified as absolute low-performance if less than 25% of the students answered it correctly. The U. S. students had relatively low-

[1]Although the study covered three years of TIMSS in Grade 4 and 8, I refer only to 2003 Grade 8 items.

Table 31:
Frequency of DIF Items in Algebra for Low and High Performance Levels on TIMSS

Classification	Number of items	Favoring female students
High performance		
Absolute	2	
Relative	3	2
Low performance		
Absolute	3	2
Relative	4	1

Note. Classification of items based on Kilpatrick et al. (2007).

performance on an item when it ranked last among the seven countries. It is interesting that many of these items in both the high- and low-performance categories were DIF items. This result is reported in Table 31. It is not surprising that the DIF items favored female students because only algebra items were studied by Kilpatrick et al., and DIF items favoring male students were concentrated more in algebra-related items (see Table 29). The surprise is having DIF items among the absolute low-performance items because when percent correct is high or low, it is difficult to detect differences in performance. The remaining items in the absolute low-performance group and relative high-performance group were close to being B DIF favoring female students ($MH\ D\text{-}DIF = 0.993$ and $MH\ D\text{-}DIF = 0.916$, respectively).

As I mentioned above, there were 23 unique DIF items. I discuss only the ones that I found to be the most representative or that seem controversial. I start with the algebra-related items. Students should be comfortable with content topics that come before algebra. These items may be a key to understanding sources of gender DIF. According to Table 29,

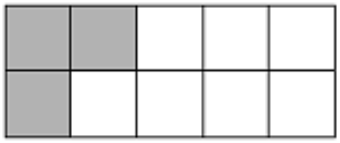
these items mostly favor male students. All released TIMSS 2003 are available at `http://timss.bc.edu/timss2003i/released.html`

Item M01_04 appeared on 3 different forms, and it was always a DIF favoring male students, twice in Category C and once in Category B.

> Alice can run 4 laps around a track in the same time that Carol can run 3 laps. When Carol has run 12 laps, how many laps has Alice run?
>
> a. 9 b. 11 c. 13 d. 16
>
> Source: TIMSS 2003 Mathematics Items: Released Set, Eighth Grade. Copyright by IEA. Available at `http://timss.bc.edu/timss2003i/released.html`

I do not have a plausible explanation of why female students performed worse on this item than male students with the same total score on the test. According to the *TIMSS 8th Grade Mathematics Concepts and Mathematics Items Book* (2003), only 48% of the U.S. students solved this item correctly, just one percentage point above the international average. There were no similar proportion items on the FCAT. On the geometrical measurement scale items, female students did relatively well, but other types of proportion were mostly absent from the FCAT.

Another surprising item, M01_01, was twice C DIF favoring male students:

> In the figure, how many MORE small squares need to be shaded so that $\frac{4}{5}$ of the small squares are shaded?
>
> 
>
> Source: TIMSS 2003 Mathematics Items: Released Set, Eighth Grade. Copyright by IEA. Available at `http://timss.bc.edu/timss2003i/released.html`

This item was easier than the previous one; 61% of the students solved it correctly. I would not claim that female students do not understand fractions, but the evidence points toward fractions as a source of difficulty for them. Of 198 items (with repetitions), 36 items had fractions. One third of these items were either DIF (9 items) or very close to being DIF (3 items) favoring male students. There were only 2 DIF items and 1 close-to-DIF item with fractions that favored female students. Among the items without fractions, the distribution of DIF and close-to-DIF items was more even: 17 $(14 + 3)$ favoring male students, and 16 $(10 + 6)$ favoring female students. A t test on released items showed that the *STAND P-DIF* and *MH D-DIF* means were significantly different for items with and without fractions ($p < .001$).

The item M04_06 was the fraction item that had the most disturbing results:

> Write a fraction that is less than $\frac{4}{9}$.
>
> Source: TIMSS 2003 Mathematics Items: Released Set, Eighth Grade. Copyright by IEA. Available at `http://timss.bc.edu/timss2003i/released.html`

This item appears on the DIF list twice as Category B DIF favoring male students. It did not favor either gender on the third form. It raises uneasy questions about conceptual understanding of fractions by U.S. students and female students, in particular. Overall, 69% of U.S. students answered this item correctly, and it was well above the international average. Nevertheless, the concept behind this item is very basic. Without understanding it, students are unlikely to understand anything in dealing with fractions. I would expect more students to answer this question correctly, and I did not expect this item to be high level DIF favoring either group.

Differences in performance on items with fractions is connected to differences in performance on items with percent. Of 10 items with percent, 3 items were DIF favoring male students, and 7 did not favor either gender. Item M01_13 was a Category C DIF item favoring male students in the U.S. sample:

> At a play, $\frac{3}{25}$ of the people in the audience were children.
>
> What percent of the audience was this?
>
> a. 12% b. 3% c. 0.3% d. 0.12%
>
> Source: TIMSS 2003 Mathematics Items: Released Set, Eighth Grade. Copyright by IEA. Available at `http://timss.bc.edu/timss2003i/released.html`

Across three forms, 69% of U.S. female students and 74% of U.S. male students answered this item correctly. On the regular forms, the difference was larger, and on the third form, on which the item was scaled as *end of session* (Martin et al., 2004, p. 248), the female students did slightly better. Although this item was classified as C DIF on one form only, it is clear that the concept of percent was not mastered as well by female students as it was by male students. The most popular distracter for all students was *b*.

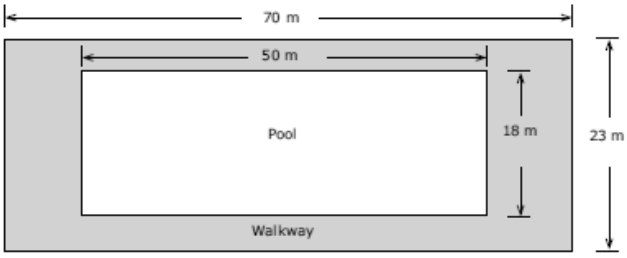
There were more DIF items with fractions and percent that favored male students. I presented only the most basic here. To be fair to the female students, I should report that on one item with fractions (M01_11), they performed better than the male students with the same total score, and this item was on the DIF list twice:

> In a group of children, 16 have birthdays during the first half of the year, and 14 have birthdays during the second half of the year. What fraction of the group have birthdays during the first half of the year?
>
> a. $\frac{14}{30}$ b. $\frac{14}{16}$ c. $\frac{16}{14}$ d. $\frac{16}{30}$ e. $\frac{30}{16}$
>
> Source: TIMSS 2003 Mathematics Items: Released Set, Eighth Grade. Copyright by IEA. Available at `http://timss.bc.edu/timss2003i/released.html`

This item was solved correctly by 78% of female students and 70% of male students on one of the forms where this item was a DIF. The most popular distracter was *c*, 15 and 17%, respectively.

The next two items that I discuss were from the geometrical measurement category. Item M04_07 appeared twice on the DIF list as a Category B DIF favoring male students. On the

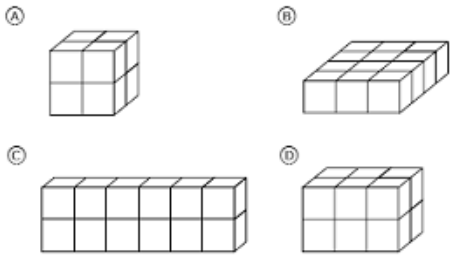
third form, the item was scaled as an *end of session* item. It was only slightly favoring male students on that form:

> A rectangular shaped swimming pool has a paved walkway around it as shown.
>
> 
>
>
> What is the area of the paved walkway?
>
> a. $100\,m^2$ b. $161\,m^2$ c. $710\,m^2$ d. $1,610\,m^2$
>
> Source: TIMSS 2003 Mathematics Items: Released Set, Eighth Grade. Copyright by IEA. Available at `http://timss.bc.edu/timss2003i/released.html`

This item was from TIMSS 1999 and was one of the trend items that helped to link the previous assessments to the current one. Calculators were not allowed before 2003. Therefore, calculators were not permitted on trend items. With a calculator, the simplest strategy would be to calculate the difference in areas of the two rectangles. Without a calculator and taking into account the given dimensions, it might be easier to add two products: 23×20 and 50×5. In both cases, calculating mistakes are possible, but the item was an MC, which might help students spot a mistake and correct it if they had time. I looked at percent correct on the form on which the item was not a DIF, and I found that the male students outperformed the female students on the item, 43 to 35%. About 6% of the students omitted the item. Because it was at the end of the test, the students may not have had time to calculate. They may have eliminated the first two choices as too small, and then chose one of the remaining two. For female students, the choice looked almost random: 35% for the correct choice and 34% for *d*. More male students probably calculated the answer or eliminated answer *d* as too big: 43% chose the correct choice, and 26% chose the main distracter *d*. Widespread

guessing probably skewed the outcome a little. The percentages above are from students who answered the item. I counted omitted answers as incorrect.

Item M02_01 was a B DIF favoring male students:

> All the small blocks are the same size. Which stack of blocks has a different volume from the others?
>
> 
>
>
> Source: TIMSS 2003 Mathematics Items: Released Set, Eighth Grade. Copyright by IEA. Available at `http://timss.bc.edu/timss2003i/released.html`

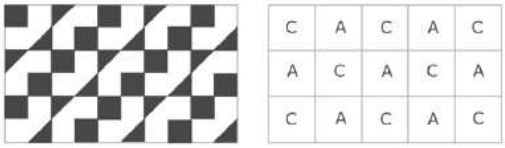
On the form on which this item was a DIF, 54% of male students and 37% of female students answered correctly. The most popular distracter was *b*: 36% of female students and 24% of male students. The performance on this item might suggest that female students are not as good at spatial tasks as male students are. However, the next example might suggest otherwise.

The next item (Figure 13) was a pictorial pattern item. This item was discussed in the Kilpatrick et al. (2007) study as one with relatively low performance. The item is very unusual and challenging. On one form the item was a B DIF favoring female students, and on another form, it did not make the list but was tilted toward female students. Although the difference in percent correct was not great, 49% to 45 for female and male students, respectively, it provides evidence that female students can do relatively well on some spatial tasks.

One more pattern item, M04_04, was a Category B DIF favoring female students. It is not as complex as the previous item, but it involved reasoning and not just solving a routine problem. Although only 45% of the responses were correct , U.S. students performed

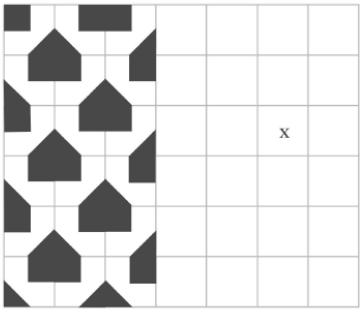
The tiles can be placed on a grid in four different ways. The four ways are shown below with a letter, A, B, C, or D, to identify each one.

These letters can be used to describe tiling patterns. For example, the pattern below can be described by the grid of letters shown next to it.



C	A	C	A	C
A	C	A	C	A
C	A	C	A	C

If the pattern on the grid below was continued, what letter would identify the orientation of the tile in the cell labeled X?



Source: TIMSS 2003 Mathematics Items: Released Set, Eight Grade. Copyright by IEA. Available at `http://timss.bc.edu/timss2003.html`

Figure 13: TIMSS 2003 released item M13_05

relatively well on this item, with only 5 of 49 participating countries performing better (TIMSS 2003 8th grade, 2003, p. 105):

> The numbers in the sequence 7, 11, 15, 19, 23, . . . increase by four. The numbers in the sequence 1, 10, 19, 28, 37, . . . increase by nine. The number 19 is in both sequences. If

> the two sequences are continued, what is the next number that is in BOTH the first and the second sequences?
>
> Source: TIMSS 2003 Mathematics Items: Released Set, Eighth Grade. Copyright by IEA. Available at `http://timss.bc.edu/timss2003i/released.html`

The last TIMSS item I discuss is the only C DIF item favoring female students. The overall U.S. performance on this item was 74%. The item was a DIF on two forms.

> Alice ran a race in 49.86 seconds. Betty ran the same race in 52.30 seconds. How much longer did it take Betty to run the race than Alice?
>
> a. 2.44 seconds b. 2.54 seconds c. 3.56 seconds d. 3.76 seconds
>
> Source: TIMSS 2003 Mathematics Items: Released Set, Eighth Grade. Copyright by IEA. Available at `http://timss.bc.edu/timss2003i/released.html`

This items tested students' skill in subtracting decimals with borrowing. Whether female students were better at understanding the concept of doing the subtraction or whether the male students were not paying attention is not clear.

One last result that I report is the distribution of DIF items by TIMSS classification (Mullis et al., 2004) of cognitive domain. I did not use a complexity classification in my studies. I could not devise my own, and I did not find a good one to use. Nevertheless, I could not pass up the opportunity to see whether there was a relation between cognitive domain and DIF. Table 32 does not show any clear relation, although there may be one. Male students have more DIF items in all cognitive domains but reasoning. The pictorial pattern item that I discussed earlier was coded as using concepts, but I would classify it as reasoning. Then it would be safe to say that the female students performed at least as well on reasoning items as the male students matched on the total score. However, the female students did not perform as well on items from the using-concepts domain. This is reflected in the table: the male students have six times as many DIF items in using concepts as the female students. In other cognitive domains, the ratio is lower. Almost all of the concepts

Table 32:
Distribution of DIF Items by Cognitive Domain on TIMSS

Cognitive	Male students		Female students	
domain	C DIF	B DIF	C DIF	B DIF
Knowing facts and procedures	2	1		2
Using concepts	2	4		1
Solving routine problems	2	2	1	4
Reasoning		1		1

Note. Repeated DIF items appear in the table just once. The higher degree of DIF is reported if an item was in both DIF categories.

items involved fractions, percents, or ratios. Looking closer at the content domain probably makes more sense in understanding gender-related DIF.

Conclusion

TIMSS items are very lean on context, which may help students concentrate on mathematical concepts involved in the items. Most DIF items are usually the result of multidimensionality, measuring not the intended trait but something else. In this sense, TIMSS is a good tool for studying mathematical characteristics of DIF items. A DIF item is not necessary a biased item. If a DIF item favoring one demographic group is measuring the intended trait, then one can conclude that how this concept is taught may benefit one demographic group over another.

The main conclusion from this study is that U.S. female students did not perform as well as U.S. male students with the same total score on very basic items that used concepts of fractions, ratio, and percent. At the same time, the female students performed relatively

well on items involving relations, patterns, and functions. Kilpatrick et al. (2007) wrote in the conclusion of their study on TIMSS algebra items that "algebra is of limited use if it is understood as generalized arithmetic only" (p. 122). I agree with this statement. However, I learned from analyzing the TIMSS DIF items that female students lack conceptual understanding in arithmetic, which may affect their number sense. Maybe the generalized arithmetic facet of algebra is very important for female students. Both TIMSS and FCAT showed that female students are relatively good with algebraic manipulations and functional relationships. Inadequate number sense may hinder their performance in algebra later, when it becomes more complex and involves not just integers as coefficients. Female students did not perform well in the informal algebra category. This category involves mostly modeling arithmetic expressions and doing calculations. Female students are good with calculations. The modeling part is the one that is difficult for them. An inability to set up an arithmetic expression would lead to inability to set up a function, which is probably the main skill required in calculus and many other courses.

My other conclusion from this study is that the classification I developed in the previous studies worked well for content categories and did not work at all for topics. The characteristics that contributed to DIF in the FCAT studies were mostly absent from the TIMSS items. To determine which characteristics contribute to gender-related DIF, one should study different assessments.

CHAPTER 8

DISCUSSION

The findings of the five studies reported in this dissertation confirmed the results of other studies on gender-related differential item functioning (DIF) as well as provided new evidence on characteristics of items that contribute to DIF. I applied slightly different approaches to analyzing the data from the Florida Comprehensive Assessment Test (FCAT) and Trends in International Mathematics and Science Study (TIMSS). In studies on the FCAT, I went from characteristics to items, whereas for TIMSS I started from the DIF items in an attempt to understand what characteristics might contribute to DIF. The two approaches complemented each other. Comparing these two tests with respect to gender-related DIF produced interesting results that can help assessment, research, and potentially instruction.

The results on content categories largely confirmed previous findings that female students perform better on algebra items and male students on items on geometry, measurement, and number and operations (Gallagher & De Lisi, 1994; Harris & Carlton, 1993; McGraw et al., 2006; Mendes-Barnett & Ercikan, 2006; Willingham & Cole, 1997). However, the present analyses broke strands into smaller categories, and the differences have been pinpointed more precisely to particular content. In the algebra strand, only the algebraic manipulations category favored female students consistently. The means of DIF indices for algebraic manipulations were significantly different from those in other categories, and there were DIF items in this category. Other categories, such as functions or translating word problems into equations or functions did not show significant differences in performance on the FCAT. The patterns category did not have significantly different means from other categories on the FCAT. Mendes-Barnett and Ercikan had similar results. However, items with patterns were

disproportionally present on the list of DIF items benefiting female students on the FCAT and TIMSS.

I analyzed very few geometry items, only items that could be classified as geometrical measurement. Only one category of geometrical measurement showed some differences in performance between male and female students. No items on basic area, or volume were DIF items. The other category of geometrical measurement, which requires one to use more than one formula or alter an existing basic formula, was somewhat troubling to female students. Three DIF items benefiting male students were on the FCAT (counting 2001 to 2003 altogether), and three were on the TIMSS.

Items from the number and operations strand went mainly into two categories according to the classification in the present study: number and informal algebra. The results from the FCAT and TIMSS differ with respect to this strand. On the FCAT, female students had difficulty with informal algebra word problems. Although the means of DIF indices for informal algebra were not always different from those of other categories, almost half of all DIF items favoring male students were from the informal algebra category. The analysis did not detect any differences in performance in the number category. The results from the TIMSS confirmed that the informal algebra category is challenging for female students. This category has the largest number of DIF items favoring male students among all categories. However, unlike the FCAT, number category items in TIMSS also appeared to be challenging for female students. Almost a third of the male DIF items were from this category. Dealing with percents and fractions appeared to be challenging for female students.

Analyses of the FCAT showed that the choice of a context topic for an item is important. Not many studies have been devoted to the issue of the topic of test items. Several researchers have concluded that students do better on items with context (Koedinger et al., 2008; Koedinger & Nathan, 2004; Nathan & Koedinger, 2000a, 2000b). Some researchers have found that "male" and "female" topics can contribute to DIF (A. S. Cohen & Ibarra, 2005; Gallagher & De Lisi, 1994; Gallagher et al., 2000). Kaminski et al. (2008) concluded that learning a concept in a context hinders transfer to a different context. However, this

study was about learning not testing. There should not be any confusion between teaching in a context and a topic of a context on a test item. Blum and Niss (1991) wrote that

> there are abbreviated and restricted links between mathematics and reality which are much more frequently found: On the one hand a direct application of already developed "standard" mathematical models to real situations with a mathematical content, on the other hand a "dressing up" of purely mathematical problems in the words of another discipline or of everyday life. Such problems often give a distorted picture of reality. (p. 40)

A topic for a test item is more about "dressing up." The topic of an item should not hinder performance of any demographic group. On FCAT 2001, three out of four DIF items favoring male students dealt with bicycle races. Although it could be a coincidence, it still raises questions. At the same time, the relatively large number of items in retail/currency/business is problematic too. Although the topic is neutral with respect to gender and relates well to everyday life, it might impede transfer to other topics. Test designers should use a variety of topics; however, science should be a priority because exposure to science topics might have double benefit: help students become accustomed to science and transfer mathematical concepts to novel situations. There also should be a balance between content and context. If item content is difficult for some group, then the item context should be familiar. I understand that test designers need to monitor many groups at the same time; however, balance is important for fair assessment.

Several researchers (Hyde et al., 1990; Mendes-Barnett & Ercikan, 2006; Ryan & Chiu, 2001) found that male students did relatively better on word problems. Whether male students learn better to ignore the topic of an item on the test or transfer knowledge better between topics is not clear. The lean context of the TIMSS items did not hinder the performance of either gender.

Related to a topic, but tested as a separate characteristic on the FCAT, was the mention of vehicles in the statement of the item. Although items with speed require one to use vehicles

sometimes, nevertheless, the object with a speed could be a runner, a bird, or a sound. If an item is a geometrical ratio item, why should it be an airplane model or a boat model? For finding area, why use the deck of a cargo ship? On the FCAT, vehicles in the statement of the item appeared to negatively affect the performance of female students.

Several characteristics tested in the analyses on the FCAT that affected the performance of female students were mathematical. One of them concerned items with estimated answers. Estimation is a very important skill. It is related to and depends on number sense and helps students' understanding of underlying concepts. "To the person without number sense, arithmetic is a bewildering territory in which any deviation from the known path may rapidly lead to being totally lost" (Dowker, 1992, p. 52). Good estimation skills make it possible to catch mistakes and correct them once made. Although the FCAT did not actually test estimation, the persistent underperformance of female students on items with words such as *approximately* and *reasonable estimate* raises a concern that female students were not comfortable with these items and perhaps with estimation in general. It may also explain why female students tended to perform worse on informal algebra items than male students with the same total score: Those items were testing number sense. The better performance on MC items by male students, although not significantly better in the present studies, can also be a result of better estimation skills. Quickly estimating answers and finding the correct one on an MC item gives the student more time to work on more complicated items.

The main implication from this result is that estimation strategies should be taught at school. Studies on estimation strategies show that most such strategies used by mathematicians (Dowker, 1992) and about a quarter used by college students (Levine, 1982) were unlikely to be taught at school; teaching more estimation strategies at school would help all students. Test designers should continue using items with estimation on tests and maybe in greater number to encourage educators to pay more attention to this very important skill. The *Principles and Standards for School Mathematics* (National Council of Teachers of Mathematics [NCTM], 2000) emphasize skills in computational fluency and estimation for

every grade level. Unfortunately, the book does not discourage the use of calculators even in elementary school. Although studies on short-term use of calculators did not show impact on calculation skills, problem solving, or conceptual development, there have been no studies on long-term use of calculators (NMAP, 2008). My personal experience in teaching introductory level mathematics courses in a four-year research institution convinced me that the use of calculators should be discouraged at schools. The survey of algebra teachers by the National Mathematics Advisory Panel (NMAP) "indicated that the use of calculators in prior grades was one of their concerns" (p. 50). If one can calculate an answer on a calculator, there is no need to estimate.

Linear speed was another characteristic on the FCAT that gave trouble to female students. Students should start learning the concept of linear speed early in elementary school. The NCTM (2000) recommends discussing quantitative change in Grades K–2. In Grades 3–5, "students should have opportunities to study situations that display different patterns of change—change that occurs at a constant rate, such as someone walking at a constant speed, and rates of change that increase or decrease" (p. 163). In Grades 6–8, the major focus in algebra is to learn to use functions in modeling patterns of quantitative change. Unfortunately, the extensive example NCTM uses is about a phone plan, not about linear speed. In Grades 10–12, the examples go beyond basic linear speed. There is a possibility that this simple topic is overlooked by teachers. Finding rates other than linear speed was not a contributing factor to DIF, suggesting that linear speed is not sufficiently discussed in class. Male students may have an advantage over female students on this concept because of their interest in cars. Items with linear speed and rate in general were not common on TIMSS.

Converting nonmetric units was troubling for female students. For Grades 6–8, the NCTM (2000) measurement strand proposes "understanding both metric and customary system of measurement" (p. 399) as the first expectation, and converting units within the same system as the second. Unfortunately, there is no mention of converting between the systems. I

would agree that both systems should be taught in school; however, having items involving converting nonmetric units on the test is questionable.

As I mentioned before, several researchers (Mendes-Barnett & Ercikan, 2006; Ryan & Chiu, 2001) found that male students benefit more from visual stimuli in the item than female students do. A. S. Cohen and Ibarra (2005) suggested that visual stimuli help female students. In these studies, visual stimuli appeared on the list of factors contributing to DIF in the 2003 FCAT and in Grade 10. More DIF items favoring male students had visual stimuli than those favoring female students on the FCAT and TIMSS.

There is also disagreement among researchers concerning computational tasks. Hyde et al. (1990) found that female students are better at computational tasks, whereas Harris and Carlton (1993) concluded that female students are better when a noncomputed solution is required. On standardized state tests, a noncomputed solution usually means matching the function or equation to the statement of the item. In the present studies, noncomputed solution items favored female students on the FCAT, although not every year. On the TIMSS study, the same number of DIF items with a noncomputed solution favored female and male students. Whether female students are thoroughly checking all possibilities before choosing an answer, unlike male students, who tend to estimate or guess, or whether there are other reasons, is not clear to me.

The current studies did not confirm or refute the result of the Bielinski and Davison (1998) study that female students outperform male students on easy items, whereas male students outperform female students on difficult items. There is some indirect evidence that contradicts the finding that female students outperform male students on the easiest items. In the absolute low-performing algebra items on the TIMSS, two items were DIF favoring female students and no DIF items were among the absolute high-performance items, although overall there were more DIF items favoring male students than female students.

In reflecting on the findings of the present studies, I found it necessary to be cautious in interpreting the results. It is clear that there are many different variables that may affect

performance of students on a test, and some groups of students are affected differently from other groups. I analyzed several characteristics in the studies, and I found that some of them had an influence on DIF indices year after year. These characteristics should be studied more. Several characteristics never appeared on the list of contributors to DIF. Although it is probably premature to discard those characteristics as unimportant in their contribution to DIF, test designers should probably not worry much about them. In addition to characteristics that were always on the list or never on the list, there were characteristics that contributed to DIF in one year but not in other years. Test designers and teachers should be aware of these characteristics, and more research should be done to determine whether the characteristics themselves or their interactions with other characteristics are contributing to DIF. The DIF indices for an item are not stable. They depend on the other items on the test as well as on the number of items. When an item is repeatedly on the DIF list, then there is something about it that one should carefully study. It is not to say that if the item appears on the list of DIF items just once, one should not pay attention. All items that were flagged by either *MH D-DIF* or *STAND P-DIF* should be reviewed.

Implications for Item Development and Research

Underrepresentation of women in mathematics and mathematics-related careers is an important issue for our country. At a time when more and more jobs are becoming technologically advanced, the country needs a well-educated workforce to stay competitive in the world. Despite the large gains women have made in education, there has been no substantial increase in the percent of women in high-level technological jobs. Why do more women not pursue advanced careers in mathematics, science, and engineering? It is a complex question. Many researchers are studying this issue from different angles. I hope that my findings contribute to understanding this complex issue. Studying gender-related DIF items can contribute to finding early signs of difference in performance and concepts that are elusive for one or the other gender as well as to address the issue of fairness in assessment.

Item Development

Test development plays a large and important part in the field of educational and psychological measurement. Many studies have been done on gender-related DIF. However, not many studies have investigated the topic of a mathematical item. My study addressed that issue. The main implication from the study for item developers is the importance of the item context. Because the topic of an item on a mathematics test is mainly "dressing up" a mathematical problem, test developers should strive to make topics as neutral as possible for all demographic groups. I suggest using either school-related topics or science and social studies topics. The first is familiar to every student, and the other two are important for education. Balancing item content and item context is another suggestion. Female students performed relatively well on algebraic manipulation items. These items can have a context that is not very familiar to female students. However, with geometrical measurement items, on which students need to apply several basic formulas, the use of an airplane model is perhaps not appropriate. A familiar topic might help female students perform better on such items.

Another suggestion for test developers is to inform educators about DIF items for all demographic groups. The information would be a better statistic than proportion correct because DIF studies compare groups by matching them on many levels according to their ability, even though a simplistic criterion such as total score is used. Teachers know their students better than educational researchers; they may help researchers identify the reasons of why an item is a DIF item.

Research

The findings of the study may give new directions of future research on gender-related DIF as well as on gender differences. One of the directions I see arises from finding challenging content categories for female students. It is clear that female students' foundations in number sense are not sound. That many of the DIF items favoring male students were in the informal algebra category on both tests, and that many of the DIF items were in the number category

on the TIMSS, as well as female students' uneasiness with estimated answers on the FCAT, support this conclusion. Studies on strategy use in the late grades of elementary school and early grades of middle school could help one understand what concepts are elusive for female students. The DIF studies on test assessment in middle school can help identify early differences in informal algebra. Categories can be subdivided to reflect different concepts. On a broader scale, if mathematics educators and educational measurement researchers study mathematical concepts together, that might open a new direction in research on DIF.

Differential Concept Functioning

The main reason for DIF as seen by educational measurement researchers is the multidimensionality of the item, which distorts the measurement of the intended dimension. What if the item is unidimensional and still a DIF item? I suggest that the focal and reference groups that are matched on their total score, and presumably have the same ability, have differential concept functioning.

These studies showed that female students struggled with fractions more than male students did. Although there are many concepts in the topic of fractions, it is possible to isolate the concrete concepts that are challenging to female students. Some items can be dropped from a test because of context or language. However, an item cannot be thrown out because a particular mathematics concept is more challenging for one demographic group than another. After identifying a particular concept that demonstrates differential concept functioning, researchers should study how this concept is taught at school. Although there are many different curricula, it would be possible to classify different ways of teaching this particular concept. Subsequent studies could link the classification with results on statewide tests. Case studies of the concept could help to identify ways to improve instruction. From studying large assessment tests to case studies, the collaboration between mathematics educators and educational measurement researchers is very important. Studies of differential

concept functioning can help all demographic groups and can be extended to latent groups such as groups with different learning styles.

Randomized Studies

Previous studies on particular item characteristics have reached different conclusions. One of these characteristics is visual stimuli. Some researchers found that visual stimuli benefited female students; other studies had an opposite conclusion. I suggest doing studies in which specific item characteristics such as visual stimuli are systematically varied. On a large assessment test, the same item can be given to students with and without a figure on different forms. A Category B-DIF item about a car's acceleration (see pages 67–69) could be given on a test in three different forms: only text, information only on the graph, and in the form it was given on the FCAT, both text and a graph. Several items in different content categories can help identify whether a graph helps or hinders the performance of a particular demographic group. Many item characteristics can be studied by administering different forms of an item to randomly chosen groups. In addition to different types of visual stimuli (figure, table, and graph), item characteristics such as the size and the type of the numbers or different units, distracting information can be studied this way. Well-designed studies can identify which of these characteristics contribute to DIF and potentially help educators improve instruction.

The present study does not have immediate implications for instruction. However, with additional studies such work can make a big contribution to improving instruction. I have tremendous respect for all public school teachers. Being a college teacher for many years, I hesitate to teach them how to teach before more research is done on the subject.

Limitations

This study has several limitations. One is not using the cognitive complexity of items. My attempts to devise the classification were not successful. Many cognitive complexity classifications are based on Bloom's (1956) taxonomy , its recent modifications (Anderson, Krath-

wohl, & Bloom, 2001; Marzano & Kendall, 2007), or similar classifications (Webb, 2007). Bloom's taxonomy is difficult to use because it requires an inference about the skill, knowledge, and background of the students responding to the item. I considered classifications from TIMSS, NAEP, and the FCAT. They did not work the way I wanted the cognitive complexity classification to work. I tested the TIMSS classification, and I did not see any patterns.

Another limitation is not testing linguistic complexity. I attempted to research this topic, but found it too complex to study in a short time. I feel that linguistic complexity could be useful for the FCAT, but not for the TIMSS, on which items were concise and clear.

One more limitation is not doing analysis on the National Assessment of Educational Progress (NAEP) data. I gave up after several attempts to get the data. However, it would be a project for more than one person. Just preparing the students' data for 3 years of the FCAT and 12 forms of the TIMSS, not counting building two data files for items, was an enormous task. But a NAEP analysis would be a nice addition to the study, as it considered our "National Report Card."

Conclusion

Education will always have room for improvement. New times will create new challenges. One of the challenges that faces our country now is a deficit in the well-educated work force. Understanding the reasons more women are not pursuing careers in mathematics and related fields can help address the challenge. I hope that my study helps to take a step in the direction of understanding this phenomenon.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67–91.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Mathematics Competition. (2008). *The MAA American Mathematics Competition.* Retrieved October 7, 2008, from `http://www.unl.edu/amc`

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* New York: Longman.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 3–24). Hillsdale, NJ: Erlbaum.

Ansell, E., & Doerr, H. (2000). NAEP findings regarding gender: Achievement, affect, and instructional experience. In E. A. Silver & P. A. Kenney (Eds.), *Results from the seventh mathematics assessment of the National Assessment of Educational Progress* (pp. 73–106). Reston, VA: National Council of Teachers of Mathematics.

Barley, S., & Orr, J. (1997). *Between craft and science: Technical work in U.S. settings.* Ithaca, NY: Cornell University.

Battista, M. T. (1990). Spatial visualization and gender differences in high school geometry. *Journal for Research in Mathematics Education*, *21*, 47–60.

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.

Benbow, C. P., & Stanley, J. C. (1983). *Academic precocity: Aspects of its development.* Baltimore: Johns Hopkins University Press.

Bielinski, J., & Davison, M. L. (1998). Gender differences by item difficulty interactions in multiple-choice mathematics items. *American Educational Research Journal*, *35*, 455–476.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals* (1st ed.). New York,: Longmans, Green.

Blum, W., & Niss, M. (1991). Applied mathematical problem solving, modelling, applications, and links to other subjects: State, trends and issues in mathematics instruction. *Educational Studies in Mathematics*, *22*, 37–68.

Boaler, J. (2003). When learning no longer matters: Standardized testing and the creation of inequality. *Phi Delta Kappan*, *84*, 502–506.

Bridgeman, B., & Lewis, C. (1996). Gender differences in college mathematics grades and SAT-M scores: A reanalysis of Wainer and Steinberg. *Journal of Educational Measurement*, *33*, 257–271.

Burton, N. (1996). Have changes in the SAT affected women's mathematics performance? *Educational Measurement: Issues and Practice*, *15*(4), 5–9.

Byrnes, J. P. (2005). Gender differences in math. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 73–98). New York: Cambridge University Press.

Byrnes, J. P., Hong, L., & Xing, S. (1997). Gender differences on the math subtest of the Scholastic Aptitude Test may be culture-specific. *Educational Studies in Mathematics*, *34*, 49–66.

Caplan, J. B., & Caplan, P. J. (2005). The perseverative search for sex differences in mathematics ability. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 25–47). New York: Cambridge University Press.

Carr, M., & Alexeev, N. (2008). *Developmental trajectories of mathematic strategies: Influence of fluency, accuracy and gender.* Manuscript submitted for publication.

Carr, M., & Jessup, D. L. (1997). Gender differences in first-grade mathematics strategy use: Social and metacognitive influences. *Journal of Educational Psychology*, *89*, 318–328.

Carr, M., Jessup, D. L., & Fuller, D. (1999). Gender differences in first-grade mathematics strategy use: Parent and teacher contributions. *Journal for Research in Mathematics Education*, *30*, 20–46.

Casey, M. B., Nuttall, R. L., & Pezaris, E. (1997). Mediators of gender differences in mathematics college entrance test scores: A comparison of spatial skills with internalized beliefs and anxieties. *Developmental Psychology*, *33*, 669–680.

Charbonneau, L. (1996). From Euclid to Descartes: Algebra and its relation to geometry. In N. Bednarz, C. Kieran, & L. Lee (Eds.), *Approaches to algebra: Perspectives for research and teaching* (pp. 15–37). Dordrecht: Kluwer.

Chipman, S. F. (2005). Research on the women and mathematics issue: A personal case history. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 1–24). New York: Cambridge University Press.

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

Cohen, A. S., & Ibarra, R. A. (2005). Examining gender-related differential item functioning using insights from psychometric and multicontext theory. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological*

approach (pp. 143–171). New York: Cambridge University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

College Board. (2008). *SAT percentile ranks for males, females, and total group.* Retrieved October 7, 2008, from `http://www.collegeboard.com/`

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

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

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

Dowker, A. (1992). Computational estimation strategies of professional mathematicians. *Journal for Research in Mathematics Education*, *23*, 45–55.

Dwyer, C. A., & Johnson, L. M. (1997). Grades, accomplishments, and correlates. In W. W. Willingham & N. S. Cole (Eds.), *Gender and fair assessment* (pp. 127–156). Mahwah, NJ: Erlbaum.

Fennema, E., Carpenter, T. P., Jacobs, V. R., Franke, M. L., & Levi, L. W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher*, *27*(5), 6–11.

Fisher, M. (2008). *Study: No gender differences in math performance.* Retrieved October 5, 2008, from `http://www.news.wisc.edu/15412`

Florida Comprehensive Assessment Test (FCAT). (2005). *FCAT released tests, Grade 10.* Retrieved October 7, 2008, from `http://fcat.fldoe.org/fcatrelease.asp`

Florida Comprehensive Assessment Test (FCAT). (2006a). *FCAT released tests, Grade 10.* Retrieved October 7, 2008, from http://fcat.fldoe.org/fcatrelease.asp

Florida Comprehensive Assessment Test (FCAT). (2006b). *FCAT released tests, Grade 9.* Retrieved October 7, 2008, from http://fcat.fldoe.org/fcatrelease.asp

Florida Department of Education. (1996). *The Sunshine State Standards.* Retrieved August 23, 2008, from http://www.fldoe.org/bii/curriculum/sss/sss1996.asp

Florida Department of Education. (2001). *Mathematics test item specifications: Grade 9–10.* Retrieved October 7, 2008, from http://fcat.fldoe.org/histpub.asp

Florida Department of Education. (2005). *FCAT handbook: A resource for educators.* State of Florida, Department of State. Retrieved October 7, 2008, from http://fcat.fldoe.org/handbk/fcathandbook.asp

Florida Department of Education. (2008). *Florida Comprehensive Assessment test (FCAT).* Retrieved October 7, 2008, from http://fcat.fldoe.org/

Gallagher, A. M. (1992). *Cognitive patterns of gender differences on mathematics admission tests* (Tech. Rep. No. ETS RR 92-2). Princeton, NJ: Educational Testing Service.

Gallagher, A. M., & De Lisi, R. (1994). Gender differences in Scholastic Aptitude Test: Mathematics problem solving among high-ability students. *Journal of Educational Psychology*, *86*, 204–211.

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-DeLisi, A. V., Morely, M., & Cahala, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology*, *75*, 165–191.

Gallagher, A. M., & Kaufman, J. C. (2005). Gender differences in mathematics: What we know and what we need to know. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 316–331). New York: Cambridge University Press.

Garner, M., & Engelhard, G. J. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education*, *12*,

29–51.

Geary, D. C., Saults, S., Liu, F., & Hoard, M. (2000). Sex differences in spatial cognition, computational fluency, and arithmetical reasoning. *Journal of Experimental Child Psychology*, *77*, 337–353.

Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice*, *24*(1), 3–14.

Glod, M. (2008, March 14). Panel urges schools to emphasize core math skills. *The Washington Post*, A06.

Halpern, D. F., Wai, J., & Saw, A. (2005). Why female are sometimes greater than and sometimes less than males in math achievement. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 48–72). New York: Cambridge University Press.

Hanson, B. A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics*, *23*, 244–253.

Harris, A. M., & Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education*, *6*, 137–151.

Henderson, D. L. (2001, April). *Prevalence of gender DIF in mixed format high school exit examinations.* Paper presented at the annual meeting of the American Educational Research Association, Seattle.

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

Huang, G., Taddese, N., & Walter, E. (2000). *Entry and persistence of women and minorities in college science and engineering education.* Retrieved October 7, 2008, from `http://nces.ed.gov/pubSearch/pubsinfo.asp?pubid=2000601`

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

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science*, *321*, 494–495.

Ibarra, R. A. (1996). *Beyond affirmative action: Reframing the context of higher education.* Madison: University of Wisconsin Press.

Ibarra, R. A., & Cohen, A. S. (1997). Multicontextuality: A hidden dimension in testing and assessment. In *The GRE, FAME Report Series* (Vol. 3; pp. 13–16). Princeton, NJ: Educational Testing Service.

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). Learning theory: The advantage of abstract examples in learning math. *Science*, *320*, 454–455.

Kilpatrick, J., & Gieger, J. L. (2000). The performance of students taking advanced mathematics courses. In E. A. Silver & P. A. Kenney (Eds.), *Results from the seventh mathematics assessment of the National Assessment of Educational Progress* (pp. 377–409). Reston, VA: National Council of Teachers of Mathematics.

Kilpatrick, J., Mesa, V., & Sloane, F. (2007). U.S. algebra performance in an international context. In T. Loveless (Ed.), *Lessons learned : What international assessments tell us about math achievement* (pp. 85–126). Washington, DC: Brookings Institution Press.

Kimball, M. M. (1989). A new perspective on women's math achievement. *Psychological Bulletin*, *105*, 198–214.

Klein, S. S., Kramarae, C., & Richardson, B. (2007). Examining the achievement of gender equity in and through education. In S. S. Klein et al. (Eds.), *Handbook for achieving gender equity through education (2nd ed.).* (pp. 1–13). Mahwah, NJ: Erlbaum.

Klein, S. S., Richardson, B., Grayson, D. A., Fox, L. H., Kramarae, C., Pollard, D. S., et al. (Eds.). (2007). *Handbook for achieving gender equity through education* (2nd ed.). Mahwah, NJ: Erlbaum.

Koedinger, K. R., Alibali, M. W., & Nathan, M. J. (2008). Trade-offs between grounded and abstract representations: Evidence from algebra problem solving. *Cognitive Science*, *32*, 366–397.

Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *Journal of the Learning Sciences*, *13*, 129–164.

Lacampagne, C., Campbell, P. B., Damarin, S., Herzig, A. H., & Vogt, C. M. (2007). Gender equity in mathematics. In S. S. Klein et al. (Eds.), *Handbook for achieving gender equity through education* (2nd ed., pp. 235–253). Mahwah, NJ: Erlbaum.

Langenfeld, T. E. (1997). Test fairness: Internal and external investigations of gender bias in mathematics testing. *Educational Measurement: Issues and Practice*, *16*(1), 20–26.

Levine, D. R. (1982). Strategy use and estimation of college students. *Journal for Research in Mathematics Education*, *13*, 350–359.

Lewin, T. (2008, March 14). Report urges changes in the teaching of math in U.S. schools. *The New York Times*, p. 20.

Li, Y. (2002). *Detecting differences in item response as a function of item characteristics.* Unpublished master's thesis, University of Wisconsin, Madison.

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing*, *4*, 115–136.

Livingston, S. A. (2006). Item analysis. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 421–441). Mahwah, NJ: Erlbaum.

Lubienski, S. T. (2000). Problem solving as a means toward mathematics for all: An exploratory look through a class lens. *Journal for Research in Mathematics Education*, *31*, 454–82.

Lubienski, S. T., McGraw, R., & Strutchens, M. (2004). NAEP findings regarding gender: Mathematics achievement, student affect, and learning practices. In P. Kloosterman & F. K. Lester (Eds.), *Results and interpretations of the 1990 through 2000 mathematics*

assessments of the National Assessment of Educational Progress (pp. 305–336). Reston, VA: National Council of Teachers of Mathematics.

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

Martin, M., Mullis, I., & Chrostowski, S. (Eds.). (2004). *TIMSS 2003 technical report.* Retrieved October 7, 2008, from `http://timss.bc.edu/timss2003.html`

Marzano, R. J., & Kendall, J. S. (2007). *The new taxonomy of educational objectives* (2nd ed.). Thousand Oaks, CA: Corwin Press.

McGraw, R., Lubienski, S. T., & Strutchens, M. E. (2006). A closer look at gender in NAEP mathematics achievement and affect data: Intersections with achievement, race/ethnicity, and socioeconomic status. *Journal for Research in Mathematics Education*, *37*, 129–150.

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education*, *19*, 289–304.

Mervis, J. (2007). U.S. expert panel sees algebra as key to improvements in math. *Science*, *318*, 1534–1535.

Mervis, J. (2008). Expert panel lays out the path to algebra — and why it matters. *Science*, *319*, 1605–1605.

Michaelides, M. P. (2008). An illustration of a Mantel-Haenszel procedure to flag misbehaving common items in test equating. *Practical Assessment, Research and Evaluation*, *13*(7). Retrieved October 17, 2008, from `http://pareonline.net/pdf/v13n7.pdf`

Mullis, I., Martin, M., Gonzalez, E., & Chrostowski, S. (2004). *Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades.* Retrieved October 7, 2008, from `http://timss.bc.edu/timss2003.html`

Nathan, M. J., & Koedinger, K. R. (2000a). An investigation of teachers' beliefs of students' algebra development. *Cognition and Instruction*, *18*, 209–237.

Nathan, M. J., & Koedinger, K. R. (2000b). Teachers' and researchers' beliefs about the development of algebraic reasoning. *Journal for Research in Mathematics Education*, *31*, 168–190.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through Grade 8 mathematics: A quest for coherence.* Reston, VA: Author.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment* (J. Pellegrino, N. Chudowsky, & R. Glaser, Eds.). Washington, DC: National Academy Press.

National Science Foundation. (2006). *Women, minorities, and persons with disabilities in science and engineering.* Retrieved October 25, 2008, from `http://www.nsf.gov`

National Science Foundation. (2008). *ADVANCE: Increasing the participation and advancement of women in academic science and engineering careers.* Retrieved October 25, 2008, from `http://www.nsf.gov`

No Child Left Behind Act of 2001. (2001). Retrieved October 25, 2008, from `http://www.ed.gov/nclb`

Nuttall, R. M., Casey, M. B., & Pezaris, E. (2005). Spatial ability as a mediator of gender differences on mathematics tests. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 121–142). New York: Cambridge University Press.

Park, A. (2008, July 24). The myth of the math gender gap. *Time*. Retrieved October 25, 2008, from `http://www.time.com/time/`

Pinzur, M. I. (2003, May 23). Dump the test, thousands demand. *The Miami Herald*, 1B.

Pinzur, M. I., & Ovalle, D. (2003, May 16). FCAT scores hit 5-year high. *The Miami Herald*, 1A.

Post, T. R., Behr, M. J., & Lesh, R. (1988). Proportionality and the development of prealgebra. In A. F. Coxford (Ed.), *The ideas of algebra, K–12* (1988 Yearbook of the National Council of Teachers of Mathematics, pp. 78–90). Reston, VA: NCTM.

Quaid, L. (2008, July 24). Math study finds girls are just as good as boys. *USA Today*. Retrieved October 7, 2008, from `http://www.usatoday.com/news/nation/2008-07-24-1490518258_x.htm`

Radford, L. (1996). The roles of geometry and arithmetic in the development of algebra: Historical remarks from a didactic perspective. In N. Bednarz, C. Kieran, & L. Lee (Eds.), *Approaches to algebra: Perspectives for research and teaching* (pp. 39–53). Dordrecht: Kluwer.

Radford, L. (2001). The historical origins of algebraic thinking. In R. Sutherland, T. Rojano, A. Bell, & R. C. Lins (Eds.), *Perspectives on school algebra* (pp. 13–36). Dordrecht: Kluwer.

Reed, D., Fox, L. H., Andrews, M. L., Betz, N., Evenstad, J. P., Harris, A., et al. (2007). Gender equity in testing and assessment. In S. S. Klein et al. (Eds.), *Handbook for achieving gender equity through education* (2nd ed., pp. 155–169). Mahwah, NJ: Erlbaum.

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

Royer, J. M., & Garofoli, L. M. (2005). Cognitive contributions to sex differences in math performance. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 99–120). New York: Cambridge University Press.

Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education*, *14*, 73–90.

Schmeiser, C. B. (2006). Epilogue. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 759–760). Mahwah, NJ: Erlbaum.

Sells, L. (1973, May). High school mathematics as the critical filter in the job market. In *Developing Opportunities for Minorities in Graduate Education: Proceedings of the Conference on Minority Graduate Education at the University of California, Berkeley* (pp. 39–47). Berkeley: University of California.

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

Snyder, T. D., Tan, A. G., & Hoffman, C. M. (2006). *Digest of education statistics, 2005* (Tech. Rep. No. NCES 2006030). Washington, DC: National Center for Education Statistics.

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

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

TIMSS 2003 8th grade mathematics concepts and mathematics items book. (2003). Retrieved October 7, 2008, from `http://nces.ed.gov/timss/educators.asp`

Tobias, S. (1976, September). Math anxiety. *MS*(80), 56–59.

Tobias, S. (1978). *Overcoming math anxiety.* New York: Norton.

U.S. Department of Education. (2006, April 18). *President establishes National Mathematics Advisory Panel.* Retrieved October 17, 2008, from `http://www.ed.gov/news/pressreleases/2006/04/04182006a.html`

Usiskin, Z. (1988). Conceptions of school algebra and uses of variables. In A. F. Coxford (Ed.), *The ideas of algebra, K–12* (1988 Yearbook of the National Council of Teachers of Mathematics, pp. 8–19). Reston, VA: NCTM.

Webb, N. L. (2006). Identifying content for student achievement tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 155–180). Mahwah, NJ: Erlbaum.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, *20*, 7–25.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Erlbaum.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003a). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement*, *63*, 51–64.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003b). DIF detection and lnterpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment*, *9*, 61–78.

Zieky, M. (2006). Fairness review in assessment. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 359–376). Mahwah, NJ: Erlbaum.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223–233.

APPENDICES A

MANTEL-HAENSZEL AND STANDARDIZATION PROCEDURES

As described in Dorans and Holland (1993), Mantel and Haenszel (1959) introduced a new procedure for the study of matched groups. Holland and Thayer (1988) subsequently adapted that procedure for use in identifying differential item functioning (DIF). For each item, the data used in the Mantel-Haenszel (MH) method are in the form of a table, where m is a score level. For every mth slice, the contingency table shown in Table 5 on page 37 has the number of correct and incorrect item scores for each group and the total. The group of interest is called the focal group (f), and the group used for comparison is the reference group (r).

The null hypothesis for MH states that getting the item correct is the same for both groups at any given level of the matching variable:

$$H_0 = [R_{rm}/W_{rm}]/[R_{fm}/W_{fm}] \qquad m = 1, \ldots, M$$

Mantel and Haenszel (1959) developed a chi-square test of the null hypothesis, which is tested against an alternative known as the constant odds ratio hypothesis:

$$H_a = \alpha[R_{rm}/W_{rm}]/[R_{fm}/W_{fm}] \qquad m = 1, \ldots, M \text{ and } \alpha \neq 1$$

The parameter α is the common odds ratio, which is the same for all levels of m. An estimate of the constant odds ratio was given by Mantel and Haenszel (1959):

$$\alpha_{MH} = \left[\sum_i R_{rm}W_{fm}/N_{tm}\right] / \left[\sum_i R_{fm}W_{rm}/N_{tm}\right]$$

Holland and Thayer (1988) converted α_{MH} into a difference expressed in the so-called delta metric. This metric is used by ETS for expressing DIF.

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

Positive values of *MH D-DIF* favor the focal group, and negative values favor the reference group. For more information about the MH method, including chi-square test statistics and the standard error formula for *MH D-DIF*, see Dorans and Holland (1993).

The standardization's item discrepancy indices are used to flag items for further visual inspection with help of graphs of empirical item response functions or differences between empirical item response functions for different groups. These indices include *STAND P-DIF* and its delta metric version *STAND D-DIF*. The former index is used more often, although the latter one has a smaller variance and correlates higher with *MH D-DIF*.

$$STAND\ P\text{-}DIF = P_f - P_f^* = \frac{\sum_m N_{fm} P_{fm}}{\sum_m N_{fm}} - \frac{\sum_m N_{fm} P_{rm}}{\sum_m N_{fm}} = \frac{\sum_m N_{fm}(P_{fm} - P_{rm})}{\sum_m N_{fm}},$$

where

$$P_f = \frac{\sum_m N_{fm} P_{fm}}{\sum_m N_{fm}} \qquad \text{and} \qquad P_f^* = \frac{\sum_m N_{fm} P_{rm}}{\sum_m N_{fm}},$$

$P_{fm} = \frac{R_{fm}}{N_{fm}}$ is the proportion correct at score level m in the focal group,

$P_{rm} = \frac{R_{rm}}{N_{fm}}$ is the proportion correct at score level m in the reference group,

N_f is the number of examinees in the focal group, and

N_r is the number of examinees in the reference group. More information on the *STAND D-DIF* index and its standard error can be found in Dorans and Holland (1993).

The MH and standardization procedures can be used with an external or internal matching criterion. In practice, a matching variable for these methods is usually not an external criterion but an internal one, such as the total score on the test. This criterion is readily available and represents in some sense a measure of ability. If a test has many DIF items, then the total score on the test may be a biased criterion (Clauser, Mazor, & Hambleton, 1993). To reduce that bias, a two-stage procedure was suggested by Holland and Thayer (1988). In the first stage, the total score of the test is used as the matching variable, and DIF items are identified by any accepted procedure. In the second stage, the criterion is a purified score, where the scores of all DIF items are removed from the total test score. For each item studied, however, its score should be included (Holland & Thayer,

1988). This inclusion complicates the procedure because the second stage should be run several times for each excluded item separately in order to include its score just in one run. When the studied item is not included in the matching score, however, the MH procedure will not behave correctly if the item is not a DIF item. At the same time, a DIF item can be misidentified if other DIF items are included in the total score. The two-stage procedure is called either *purification of the matching criterion* (Clauser et al., 1993) or *criterion refinement* (Dorans & Holland, 1993). Clauser et al. studied the two-stage procedure for MH on simulated data. Zenisky et al. (2003a) performed studies on a large state assessment test sample in mathematics, language, and science using the standardization procedure.

APPENDICES B

PROGRAM DESCRIPTION

The analyses reported for *STAND P-DIF* and *MH D-DIF* were performed by using a Perl computer program written by Boris Alexeev, a graduate student in mathematics, for this study. Alexeev used the formulas given in Dorans and Holland (1993). For MH, the program computes the chi-square test statistic, *MH D-DIF*, and the standard error for *MH D-DIF*. For standardization, the program computes *STAND P-DIF*, *STAND D-DIF*, and the standard error for *STAND P-DIF*. The program does not compute the standard error for *STAND D-DIF* because of a mistake in Dorans and Holland. Attempts to find the formula in other articles were unsuccessful. *STAND P-DIF* is the index usually used for large-scale testing programs along with ETS's classification guidelines. *STAND D-DIF* was calculated to check its correlation with *MH D-DIF*, which according to Dorans and Holland should be high. the *MH D-DIF* index after the first stage was compared for one data set to results from another computer program that calculates the MH index; the results were exactly the same. The correlation between *STAND D-DIF* and *MH D-DIF* was .99, which provides assurance that the program performed correctly.

The program runs the two-stage procedure for each index at the same time. There is an option to run one stage and manually exclude items that demonstrate high DIF and then run the second stage. In an automatic version, either the boundaries for items retained for the matching variable can be specified, or default values can be used.

In the first stage all indices are computed. Items with $|STAND\ P\text{-}DIF| < a$ and $|MH\ D\text{-}DIF| < b$ are considered "clear" to be included into a new "purified" score that will be used as a matching variable for the second stage. The values a and b can be specified

in the configuration file, or the default values will be used. The default value for *STAND P-DIF* was .04, and for *MH D-DIF* was .8. These values are lower than the usual threshold for flagging a DIF item by 20%. According to study by Zenisky et al. (2003a), items that demonstrate DIF in the first stage are not necessarily DIF items in the second stage, and non-DIF items in the first stage can become DIF items in the second stage. Zenisky et al. stated that this effect is highly correlated with the number of DIF items in Stage 1. Although this claim is probably true, the main reason for items changing status is the distribution of DIF items in the first stage. If an equal number of items with approximately equal magnitude of DIF are removed from both sides of the scale (favoring male students and favoring female students), then *STAND P-DIF* and *MH D-DIF* is not likely to change. However, if deletion of items is one-sided (by number or magnitude), then the indices will move toward that side. As a result, some items can change status. Clearly, this status change may happen for borderline items. Choosing boundaries lower than the threshold for flagging DIF items ensures that no DIF items contribute to the matching score. The remaining items can be considered sufficiently clean to be used in the matching score. If only one index is needed in the purifying procedure, then the boundaries for the other index should be put high, and the results from the output for the undesired index should be ignored. If one does not want to compute *STAND P-DIF*, then one should set the boundary for *STAND P-DIF* at 1.0. All items have an absolute value of this index less or equal 1 and that means the deletion will use only boundaries set for *MH D-DIF*. Then only the results for *MH D-DIF* are reliable.

In the second stage, the program runs k times, where k is the number of items removed from the score. In each cycle, the program runs on $(n-k+1)$ items; $(n-k)$ clean items, and one suspected item. The matching variable is the total score on clean items plus the score on the suspected item. For each clean item, the indices are calculated k times, and they are averaged for the final report, whereas for each removed item, the indices are calculated once.

The computation of indices was done with all multiple-choice and gridded-response items on the test for a particular grade, although not all items were used in further analyses.

APPENDICES C

TIMSS DIF ITEMS

Table 33:
TIMSS DIF Items Favoring Female Students

		Scaling	DIF	*MH*	*STAND*	Proportion correct		
Item	Form	status[a]	category	*D-DIF*	*P-DIF*	Female	Male	All
M01_11	1	R	B	1.261	.081	.78	.70	.74
M01_11	6	R	B	1.062	.067	.72	.65	.69
M01_12	1	R	B	1.186	.070	.68	.61	.65
M04_01	3	R	B	1.189	.060	.23	.20	.22
M04_04	3	R	B	1.110	.089	.49	.42	.46
M04_05	3	R	C	1.573	.105	.76	.68	.72
M04_05	4	R	C	1.519	.099	.81	.70	.75
M04_05	9	E	B	1.474	.106	.74	.64	.69
M09_06	9	R	B	1.212	.061	.21	.19	.20
M10_04	10	R	B	1.314	.104	.66	.59	.62
M13_01	3	R	B	1.158	.065	.85	.79	.82
M13_05	5	E	B	1.057	.081	.42	.37	.40

[a] R – regular scaling, E – end of session.

Table 34:
TIMSS DIF Items Favoring Male Students

		Scaling	DIF	*MH*	*STAND*	Proportion correct		
Item	Form	status[a]	category	*D-DIF*	*P-DIF*	Female	Male	All
M01_01	1	R	C	−1.746	−.119	.53	.66	.59
M01_01	6	R	C	−1.694	−.087	.57	.67	.62
M01_04	1	R	C	−2.297	−.173	.38	.58	.47
M01_04	12	E	C	−1.844	−.136	.42	.53	.47
M01_04	6	R	B	−1.030	−.069	.43	.52	.48
M01_13	6	R	C	−1.696	−.109	.65	.77	.70
M02_01	2	R	B	−1.389	−.116	.37	.54	.45
M02_13	11	E	B	−1.108	−.086	.32	.41	.36
M03_07	3	R	C	−1.891	−.152	.44	.62	.55
M03_07	2	R	B	−1.445	−.084	.44	.61	.52
M03_08	2	R	C	−1.709	−.130	.28	.45	.37
M03_08	10	E	C	−1.627	−.124	.28	.46	.36
M03_08	3	R	B	−1.333	−.094	.28	.42	.35
M03_12	3	R	B	−1.123	−.086	.53	.64	.58
M03_15	3	R	C	−1.748	−.116	.48	.63	.55
M03_15	10	E	B	−1.224	−.064	.41	.55	.47
M04_02	3	R	B	−1.032	−.094	.41	.54	.47
M04_06	4	R	B	−1.460	−.103	.65	.76	.70
M04_06	9	E	B	−1.102	−.074	.62	.72	.67
M04_07	3	R	B	−1.319	−.105	.33	.47	.40
M04_07	4	R	B	−1.103	−.075	.33	.44	.39
M09_07B	9	R	B	−1.145	−.055	.20	.31	.25
M13_08	5	E	B	−1.031	−.074	.73	.82	.77

[a] R – regular scaling, E – end of session.

APPENDICES D

ABBREVIATIONS

ACT	American College Testing
AERA	American Educational Research Association
AMC	American Mathematics Competition
ANOVA	Analysis of variance
APA	American Psychological Association
CRT	Criterion-referenced test
DIF	Differential item functioning
ETS	Educational Testing Service
FCAT	Florida Comprehensive Assessment Test
FDOE	Florida Department of Education
FR	Free response
GR	Gridded response
GRE	Graduate Record Examination
ICC	Item characteristic curve
IEA	International Association
	for the Evaluation of International Achievement
IRF	Item response function
IRT	Item response theory
MAA	Mathematical Association of America
MC	Multiple choice

MH	Mantel-Haenszel
MH D-DIF	Common odds ratio converted into delta metric
NAEP	National Assessment of Educational Progress
NCLB	No Child Left Behind Act
NCME	National Council on Measurement in Education
NCTM	National Council of Teachers of Mathematics
NMAP	National Mathematics Advisory Panel
NRC	National Research Council
NRT	Norm-referenced test
NSF	National Science Foundation
SSS	Sunshine State Standards
STAND P-DIF	Standardized p-difference
TIMSS	Trends in International Mathematics and Science Study