4-24-2009

Gender Differences on the American Mathematics Competition AMC 8 Contest

Melissa A. Desjarlais
University of Nebraska at Lincoln, Melissa.Desjarlais@valpo.edu

GENDER DIFFERENCES ON THE AMERICAN MATHEMATICS COMPETITION AMC 8 CONTEST

by

Melissa A. Desjarlais

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Doctor of Philosophy

Major: Educational Studies

Under the Supervision of Professor David Fowler

Lincoln, Nebraska

May, 2009

GENDER DIFFERENCES ON THE AMERICAN MATHEMATICS COMPETITION AMC 8 CONTEST

Melissa A. Desjarlais, Ph. D.

University of Nebraska, 2009

Adviser: David Fowler

This study examines gender differences on the American Mathematics Competition AMC 8 contest between 2003 and 2007 by comparing the performances of male and female United States eighth grade students after controlling for ability. During these years 183,857 males and 178,857 females participated in the contest. Research on gender differences frequently measures impact which is a difference in performance between two groups that can often be explained by different ability distributions. In contrast, differential item functioning (DIF) is a difference in performance after controlling for ability. Three types of analyses were performed to compare the performances. First, statistical analyses identified items with impact, DIF, and uniform or nonuniform DIF. Differences in proportion correct were used to identify impact and type of DIF while the Mantel-Haenszel procedure was used to identify items with gender DIF. Second, substantive analyses placed the items into multiple categories based on NCTM, Gierl, and Harnisch's classifications of mathematics problems. Third, subtest analyses used these categories to look for gender differences in terms of impact and DIF on subsets of the contest. While a majority of the items favored males in terms of impact, after controlling for ability, few items demonstrated gender DIF. None of the hypotheses of differing abilities in males and females suggested by earlier studies were supported by the subtest analyses.

ACKNOWLEDGMENTS

Writing my dissertation and finishing my doctorate while not living in Nebraska creates many challenges, and successfully overcoming them during the past two years, while also fulfilling the role of a faculty member, would have been much more difficult if it were not for the encouragement, support, and aid of many people.

I would like to thank my supervisory committee for their guidance, support, and flexibility. Thanks to my advisor, David Fowler. This journey began with a suggestion by my advisor to analyze data from the American Mathematics Competition. Finishing my degree at a distance can increase the responsibilities of an advisor, and I want to express my appreciation to him for facilitating communication with my committee and for obtaining signatures and submitting required forms.

I am grateful to Steve Dunbar for giving me permission to use the AMC data and for providing electronic copies of the items included in this dissertation. He was also very willing to answer my questions, about the data and the competition, and to include me in an AMC 8 committee meeting. I also want to thank Del Harnisch for his assistance in analyzing the data. Having conversations via his virtual office was very beneficial both in terms of questions answered and encouragement given.

I want to thank the faculty and graduate students in the mathematics and education departments at UNL. I had many wonderful opportunities and experiences, both in teaching and outreach, that helped me develop my philosophies about teaching and learning and enhanced my interest in mathematics education. Teaching the math content courses for pre-services teachers were especially meaningful experiences, and I enjoyed working with Patience, Cheryl, Pari, and Raegan. The research meetings with Rick, Josh, and Pari were also helpful. You all have helped me become a better teacher and researcher.

The Department of Mathematics and Computer Science at Valparaiso University (VU) is a key reason why these past two years of working and writing have been successful ones. The collegiality of the department is exemplified by a great sense of humor, a willingness to help when needed, and a propensity for celebrating special events with food. The gentle reminders to focus on my research and unflagging support while writing and teaching were instrumental in helping me finish as planned. I am blessed by the friends I have found in this department.

I would especially like to thank Rick Gillman for being protective of my time while still encouraging me to pursue outreach opportunities related to my research area and for always being willing to listen and offer advice. Thanks to Jerry Wagenblast for his mentoring and for always being willing to talk about teaching and learning, and to Michael Glass for his willingness to discuss my research and help with computer issues. And I am grateful to Sara, Shane, and Daniel, who began their careers at VU the same time I did; the transition to becoming a faculty member was easier and more enjoyable because of sharing it with you.

A few words of appreciation are also due to those colleagues from math, CS, psychology with whom I have spent many enjoyable Friday afternoons. The combinations of backgrounds and personalities led to many interesting conversations, ranging from insightful to "incite"ful, with both consensus and contention; no issue was too insignificant to be discussed and debated. I have many fond memories of Fridays full of camraderie and affectionate teasing. And special thanks to David, for being the perennial instigator and always finding a way to make me laugh.

Finally, I would like to thank my family for their love, support, and patience during the lengthy journey to finish my degree. I am grateful for growing up in a home with two mathematics teachers, where mathematics and education were highly valued, since it lead me to choose a career as a mathematics educator.

Contents

List of Tables

Chapter 1

Introduction

1.1 Statement of the problem

The International Mathematics Olympiad (IMO), begun in 1959, is an international mathematics competition for high school students which takes place each year in a different country. Although the United States began participating in 1974, the team first had a female member in 1998, and since then no more than one out of the six team members has been female (AMC Director, personal communication, May 30, 2008). The gender disparity is very prominent at this level, yet an IMO team member's journey to this competition typically begins with participation in the American Mathematics Competition AMC 8 contest, and, as will be shown, the gender ratio is very different at that level.

This gender disparity, especially at advanced levels of mathematics is one reason why the mathematics education community has been interested in gender differences in mathematics performance. Many studies, especially in the last 30 years, have been conducted looking for gender differences on mathematics achievement tests such as the Scholastic Aptitude Test. While some have found gender differences, the type and

magnitude of the differences vary. Many of these studies have measured either overall performance on the entire test or performance on a portion of the test based on the content of the items. Other studies have done analyses on individual items looking for differences in performance when comparing students of equal ability. Despite the multitude of studies that have been done, little research has been conducted looking for gender differences in mathematics competitions such as the American Mathematics Competition contests. This study will examine gender differences in performance on the American Mathematics Competition AMC 8 contest.[1]

Through conducting studies on gender differences in mathematics, researchers can gain knowledge about how to help both males and females be successful in mathematics. Identifying differences in how males and females answer questions on mathematics tests can aid in developing teaching methods to improve performance on these tests, encourage interest in studying mathematics, and enhance future success in mathematics. In the United States a significant emphasis is currently placed on preparing students to be successful in science and mathematics since "U.S. industry faces a dire need for employees with the advanced math ability required to conduct the leading edge research and design required to remain globally competitive" (www.mathcounts.org). Both males and females should have the opportunity to develop the skills needed to be successful in mathematics.

One method of preparing students to be successful in mathematics is to get them excited about the subject through participation in mathematics competitions such as the AMC 8 contest. In fact, the American Mathematics Competition "is dedicated to the goal of strengthening the mathematical capabilities of our nation's youth" and they "believe that one way to meet this goal is to identify, recognize and reward ex-

[1]In the remainder of this document, the American Mathematics Competition AMC 8 contest will be referred to as the AMC 8 or the AMC 8 contest.

cellence in mathematics through a series of national contests" (www.unl.edu/amc/). The contests include the AMC 8, AMC 10, AMC 12, American Invitational Mathematics Exam (AIME), and the United States of America Mathematical Olympiad (USAMO). Students who perform well may proceed to the Mathematics Olympiad Summer Program (MOSP) and possibly become a member of the United States team for the International Mathematics Olympiad (IMO).

The AMC 8 is just the beginning of a journey which can take a student to the IMO. Knowledge gained through research at the beginning of a process can lead to influences at the end; that is, knowledge gained about gender differences in performance on the AMC 8 could be used to develop methods to help both males and females to be successful on any of the American Mathematics Competition contests.

This study will address the deficiency in research studies on gender differences in mathematics competitions, such as the AMC 8 contest. The American Mathematics Competition (www.unl.edu/amc/) website provides some general descriptive statistics by gender such as number of students participating, mean score, and number of students who received a particular score, yet in-depth data analyses have not been done looking for gender differences on the AMC 8 contest. In particular, differential item functioning analysis techniques, such as the Mantel-Haenszel procedure, have not been applied to the contest items. Neither have subtest analyses been performed looking at gender differences on groups of items based on their content.

The results from these analyses would be beneficial to both the American Mathematics Competitions AMC 8 Committee, which helps write the items for the contest, and middle school teachers and parents, who help students prepare to participate in the American Mathematics Competitions. The Committee benefits by having data analyses that show trends in performance by the United States eighth grade students and reinforce the committee's ability to select items that primarily do not demonstrate

gender differences. The people who help the students prepare for the AMC 8 contest can see the types of item, based on content, that tend to appear on AMC 8 contest, and the areas where students tend to struggle, and they can use this information to guide their preparations.

1.2 Purpose statement

The purpose of this study is to identify gender differences on the American Mathematics Competition AMC 8 contest from 2003 to 2007 by comparing the performances of male and female United States eighth grade students after controlling for ability.

1.3 Research questions or hypotheses

The purpose of this research can be refined by a series of research questions. Differential item functioning (DIF) will be defined in the next section.

- How do the performances of males and females differ on individual items based on impact, type of DIF, and DIF?
- How do the performances of males and females differ based on the content or cognitive skills associated with the items?
- What patterns are found among items which demonstrate gender DIF?

The differences in performance will compared by considering individual items on the contest and items grouped according to particular classifications, such as content, to be described later in the Methods chapter. The classifications will also be used to describe patterns among items with gender DIF.

Using results from previous studies on gender DIF in mathematics (Gierl et al., 2003; Ryan and Chiu, 2001; Linn and Hyde, 1989; Hyde et al., 1990), hypotheses can be formed based on whether it is expected that males or females will be more likely to answer correctly items which include specific mathematical content or require particular cognitive skills. Table 1.1 indicates which gender is hypothesized to be more likely to answer items correctly.

Table 1.1: Research hypotheses based on content and cognitive skills

Males	Females
Geometry	Algebra
Multiple solution paths, shortcuts	Routine solutions
Spatial	Memorization
Figures, graphs, and tables	Significant verbal content

These hypotheses will be evaluated by using both qualitative and statistical analyses with items which have been identified as containing each type of content or requiring each type of cognitive skill.

1.4 Definition of Terms

Every November the American Mathematics Competition AMC 8 contest is administered to middle school students throughout the United States and approximately twenty other countries. The AMC 8 contest is open to students in eighth grade or younger. It consists of twenty-five multiple choice questions, each with only one possible answer, which "can be solved using material normally associated with the mathematics curriculum for students in eighth grade or below" (AMC 8 Solutions Pamphlet). The students have 40 minutes to answer the questions, and they are allowed scratch paper, graph paper, rulers, erasers, and calculators which are acceptable

to use when taking the SAT, although no problem requires the use of a calculator (AMC 8 Instructions). Some items are included in Appendix E. Note that beginning in 2008, students are no longer allowed the use of a calculator (AMC Director, personal communication, January 5, 2009).

The AMC 8 data will be analyzed to examine differential item function (DIF), yet it is important to make two clarifications with DIF. The first clarification is between impact and DIF. These analyses involve identifying two groups of interest: the focal group is the focus of analysis and the reference group is a basis for comparison for the focal group. Impact is the difference in performance between two groups, and it can often be explained by different ability distributions between the two groups. For example, more members of the focal group may answer an item correctly as compared to the reference group, but this may be due to the fact that the members of the focal group have had more experience solving that type of problem; that is, they have higher ability compared to the reference group. This difference in performance between two groups is impact.

On the other hand, DIF is differences in item functioning after two groups have been matched with respect to some attribute, such as ability; in particular, DIF is an *unexpected* difference in performance between groups which are supposed to be the same according to the given attribute (Dorans and Holland, 1993). For example, if two groups have been matched with respect to ability, yet the focal group has a higher performance on the problem, then the item demonstrates DIF. In this study, the focal group is females and the reference group is males, and the attribute which will be used to match the groups is ability, as measured by the total score on the AMC 8. Dorans and Holland (1993) use Simpson's Paradox (Simpson, 1951) to show why "we should compare the comparable, as is done in DIF analysis;" an example illustrating Simpson's Paradox is included in Appendix A.

The second clarification comes from the fact that DIF can further be divided into two types: uniform DIF and nonuniform DIF. Uniform DIF occurs when there is no relationship between group membership and the matching attribute; that is, the probability of answering an item correctly is greater for one group than the other uniformly over all levels of the matching attribute. In this research, this would mean that one gender would have a larger proportion correct than the other at all ability levels. Nonuniform DIF occurs when there is a relationship between group membership and the matching attribute. For example, the probability of answering an item correctly is greater for females as compared to males for some ability levels while the opposite is true for other ability levels.

The Mantel-Haenszel procedure (Holland and Thayer, 1988) is one method of identifying items which demonstrate uniform DIF. In this study the procedure is used to identify items for which there is an association between gender and answering an item correctly, after controlling for ability. The procedure gives a direction and magnitude of the gender difference and has an associated test of significance. More description, including advantages and disadvantages of using this procedure are described in the Methods chapter, and the details of using the procedure are included in Appendix B.

Chapter 2

Literature Review

2.1 Review of the previous literature

2.1.1 Impact

Many studies have been done to explore gender differences in mathematics related to certain content areas. To be aligned with the research hypotheses, this review will focus on gender differences in geometry, spatial reasoning, computation, algebra, and measurement. Before reviewing studies in these areas, a brief description of the meta-analysis method will be provided.

Meta-analysis is a statistical process which allows results from different studies to be combined or compared. As described by Hyde (1990), the first step is to find studies which report data on a particular research question. The second step is to compute an effect size, d, for each study. For research on gender differences, $d = (\overline{x}_M - \overline{x}_F)/\sigma/s$ where $\overline{x}_M$ is the mean male score, $\overline{x}_F$ is the mean female score, and s is the within-groups standard deviation as computed by $s = \sqrt{(\sigma_M^2 + \sigma_F^2)/2}$ where σ_M, σ_F are the variances for males and females, respectively. The value of d indicates the distance

between the male and female means in terms of standard deviation units. Note that negative values indicate gender differences in favor of females. The third step is to average the d values from all of the studies.

After determining the values of d, they need to be interpreted. One method of interpretation, offered by Jacob Cohen, is that an absolute value of 0.20 is small, a value of 0.50 is medium, and a value of 0.80 is large, although these values are somewhat arbitrary (Cohen, 1969, cited in Hyde, 1990). A second alternative is to compare the effect sizes to ones that have been obtained in other studies in a similar field (e.g. mathematics) or a different field (e.g. psychology). Another interpretation is described in the section on the Gender Similarities Hypothesis

Friedman (1989) performed a meta-analysis on studies published between 1974 and 1987 which examined gender differences in mathematical tasks. When computing effect sizes, she subtracted the male mean from the female mean, so that positive values reflect gender differences which favor females. The 98 studies involved students ranging from first grade to twelfth grade. Friedman found an mean effect size of -0.024, and since the confidence interval contained zero, she stated that it is not possible to say with 95% confidence that there are gender differences in school-aged children. Two conclusions that she drew from the meta-analyses were that gender differences favoring males are decreasing and the average gender difference was very small.

Hyde et al. (1990) did a meta-analysis on 100 studies published between 1967 and 1987. Within the 100 studies there were 259 effect sizes, and 131 (51%) of them favored males, 17 (6%) were zero, and 111 (43%) favored females. The average effect size over all of the studies was 0.20, but when the studies using SAT data (which included a large number of participants and hence had a large influence on the data) were omitted, the effect size was 0.15. While the value indicates gender differences favoring males, the magnitude is small according to an earlier interpretation. Hyde

et al. did notice changes in effect sizes based on age. At the elementary and middle school levels, the effect size was small and favored females. At the high school, college, and adult levels, the effect sizes favored males and became progressively larger: $d = 0.29, d = 0.41, d = 0.59$, respectively. Hyde et al. highlight the fact that when gender differences occur, they are in critical areas such as problem solving, where the effect sizes favor males.

In addition to being used to identify overall gender differences, meta-analysis techniques can be used to explore gender differences based on mathematical content. For example, when gender differences are found in geometry, they frequently favor males. Hyde et al. (1990) examined effect sizes in 100 studies and found small differences in favor of males. Hanna (1986) used data from the Second International Mathematics Study (SIMS) of the International Association for the Evaluation of Educational Achievement to study gender differences among Grade 8 students in Ontario, Canada. Male students answered more geometry items correctly, although there was no differences in the number of items answered incorrectly. Engelhard (1990) also used data from SIMS and found that among 13-year-olds in the United States, there were significant differences favoring males on the geometry items. In contrast, Berberoglu (1995) found that geometry items favored females for high schools students in Turkey.

A content area closely related to geometry is measurement. In fact, when classifying items, Garner and Engelhard (1999) grouped geometry and measurement together. They found that, based on mean scores on these items, males performed better. Using data from SIMS, Hanna (1986) found that Grade 8 males gave more correct responses than females on measurement items.

Another related content area is spatial reasoning. Linn and Petersen (1985) conducted a meta-analysis of spatial reasoning, and they divided abilities into spatial relations, mental rotation, and spatial visualization. They found that mental rota-

tions and spatial visualization were easier for males. In a review of studies of spatial abilities, Linn and Hyde (1989) claim that gender differences in spatial abilities are declining. Nuttall et al. (2005) suggest that gender differences in spatial abilities are crucial to understanding gender differences in mathematics achievement. They found that males do better than females on items involving mental rotation, and, on some mathematics tests, mental rotation ability is a stronger mediator of gender differences than math self-confidence or math anxiety.

While research seems to indicate some advantage for males on geometry, measurement, and spatial items, the opposite is often true for computation and algebra items; that is, females have the advantage on these items. Hyde et al. (1990) found gender differences in computation favoring females, although there were no gender differences in algebra. When considering mean scores on multiple choice items, Garner and Engelhard (1999) found that females scored better on algebra, while males scored better on computation.

2.1.2 Differential Item Functioning

Recently individuals in mathematics education and educational testing have been interested in Differential Item Functioning (DIF). After identifying items which exhibit DIF, the next step is to identify characteristics of items that are associated with DIF. The studies in this subsection describe results related to gender DIF on mathematics achievement tests, and many of the results are similar to the impact results.

In general, DIF studies show that geometry items favor males. Doolittle and Cleary (1987) found that geometry items on a form of the ACT were more difficult for high school females. Another study (Harris and Carlton, 1993) used data from high school students who took the SAT to examine gender differences, and their

results agreed with those of Doolittle and Cleary (1987). In contrast to these studies, Berberoglu (1995) found that geometry items favored females among high school students in Turkey, yet it was suggested these contradictory results could be due to a country effect.

Items with algebra content frequently have been shown to be easier for females. Using items from a form of the ACT, Doolittle and Cleary (1987) found that items with intermediate algebra or algebraic operations were easier for females. Engelhard (1990) used the Mantel-Haenszel procedure with data from SIMS, and he found among 13-year-olds in the United States, females were more likely to answer algebra items correctly. Using data from a high school graduation test, Garner and Engelhard (1999) found that females performed better than males on multiple-choice items containing algebra content.

Computational items have also been associated with gender differences favoring females. Engelhard (1990) found significant differences in favor of females on computational items taken from SIMS data. Doolittle and Cleary (1987) found that females found items on a form of the ACT with computational aspects to be less difficult as compared to males. Once again, Berberoglu (1995) had contradictory results when he found males having an advantage with computational skills.

2.1.3 Differential Bundle Functioning

As described earlier, diferential item functioning is differences in performance on an item by two groups after they have been matched with respect to some attribute, such as ability. An extension of this idea is differential bundle functioning (DBF). DBF occurs when a collection of DIF items which have something in common, such as content, collectively are easier for one group than another. DBF was first presented

by Roussos and Stout (1996). The simultaneous item bias test (SIBTEST) can be used to detect DBF within particular content areas (Shealy and Stout, 1993). The next four studies identify certain content areas which demonstrate DBF.

While many of the earlier studies described involved identifying gender DIF among high school students, Ryan and Fan (1996) examined gender DIF and DBF among eighth grade students. They formed hypotheses based on previous research and tested the hypotheses using SIBTEST on four different mathematics tests. They hypothesized that arithmetic, geometry, and applied (story problems) would be easier for males while algebra and computation problems would be easier for females. Their hypotheses for algebra, geometry, and computation were each supported on three of the four tests, and their hypotheses for applied problems were supported on all four tests. Their hypothesis for arithmetic items were not confirmed; rather than the items being easier for males, they were easier for females on two of the tests.

Mendes-Barnett and Ercikan (2006) also formed hypotheses related to DBF yet they included hypotheses related to not only the content of the items but also the cognitive complexity or context of the problem. They hypothesized that items with problem solving, geometry content, high cognitive complexity, word problems, or visuals included would be easier for males. They also hypothesized that problems involving sequences and series, polynomials, quadratic systems, and exponents and logarithms would exhibit no DBF while items with low cognitive complexity would be easier for females.

The analyses supported four of the hypotheses, and seven bundles demonstrated DBF. The hypotheses about problem solving, problems with high complexity, word problems, and problems with visuals included favoring males were confirmed. The hypotheses about geometry items being easier for males and low complexity problems being easier for females were not confirmed. There were four hypotheses about areas

not exhibiting DBF, and within three of these areas there was DBF: favoring females on polynomials and quadratic relations, and favoring males on logarithms and exponents. It is worth noting, that while there was no DBF for geometry, many of the individual geometry items did have high levels of DIF.

Ryan and Chiu (2001) used a list of attributes developed by Harnisch and his colleagues (included in Appendix D) to form categories and look for DBF. Using results from earlier research on gender differences, Ryan and Chiu chose nine attributes to form categories of items for analysis and formed hypotheses regarding which gender would find those types of items easier. They hypothesized that problems including attributes 5 (word problems), 11 (figures/graphs present), 12 (construction of graphs/figures), 10 (higher order thinking), 14 (test-taking skills), or geometry content would be easier for males while problems including attribute 9 (algebra operations) would be easier for females.

Their results agreed with previous findings in the categories with items having figures or graphs present, higher order thinking in algebra, and geometry content with all of the results indicating the items being more difficult for females. There were also significant results for the word problems being more difficult for females, and the magnitudes of the values were substantially larger than the other results.

A study by Gierl, Bisanz, Bisanz, and Boughton (2003) examines gender DIF on mathematics achievement tests using a DIF analysis framework based on a multidimensional model for DIF proposed by Roussos and Stout (1996). The first part is a substantive analysis using a taxonomy developed by Gallagher, De Lisi, Holst, McGillicuddy–De Lisi, Morely, and Cahalan (2000). The taxonomy includes content and cognitive skills expected to produce gender differences in mathematics. The second part is a statistical analysis to test DIF hypotheses; Gierl et al. performed multiple statistical analyses. During the substantive analysis of the achievement test,

Gallagher et al.'s taxonomy was modified (see Appendix C) by splitting one of their categories into four categories that Gierl et al. claim to be mutually exclusive; these are categories 3 through 6 in the list in the appendix.

Gierl et al. (2003) describe two outcomes related to gender differences in mathematics. They found that males perform better than females on items which include spatial content. They did not find substantial support for the other items in the modified taxonomy. In some analyses they found that females performed better than males on items requiring memorized material, but the differences were small. On the other hand, some analyses showed that males performed better than females on items with significant verbal content, which contradicts the modified taxonomy. These results lead Gierl et al. to suggest that the modifed taxonomy may not be sufficient to understand cognitive reasons for gender differences in mathematics achievement.

2.1.4 Gender Similarities Hypothesis

Some studies have shown that gender differences are small (Hyde et al., 1990) or are decreasing in value (Linn and Hyde, 1989). Aligned with these results, Hyde (2005) offers a very different hypothesis: The Gender Similarities Hypothesis. She claims that males and females are similar on most, but not all, psychological variables. While studies of differences in mathematics, considered a type of cognitive variable, are included, she also applies her hypothesis to psychological variables in areas such as communication, social and personality variables, psychological well-being and motor behaviors. The gender similarities hypothesis (Hyde, 2005) states that most gender differences are in the close-to-zero ($d \leq 0.10$) or small ($0.11 < d < 0.35$) range, a few are in the moderate range ($0.36 < d < 0.65$), and very few are large ($d = 0.66 - 1.00$) or very large ($d > 1.00$), where d is the effect size computed as described earlier.

Hyde conducted a meta-analysis of 46 studies with 124 effect sizes among the categories of psychological variables mentioned above. She found that 78% of the effect sizes are in the small or close-to-zero ranges. Hyde states that the small magnitude of the effect sizes is "even more striking given that most of the meta-analyses addressed the classic gender differences questions–that is, areas in which gender differences were reputed to be reliable, such as mathematics performance" (p. 586). Thus, the meta-analyses support the gender similarities hypothesis.

In her description of the hypothesis, Hyde emphasized the importance of context. Some gender differences appear to be associated with particular situations so that if certain situational aspects are minimized or removed, then the gender differences decrease if not essentially disappear. One example related to context is stereotype threat. One of the stereotypes associated with mathematics is that males are better at math than females. The possibility of being reduced to gender stereotypes can lead to a state which could diminish women's math performance. This situation is referred to as stereotype threat. Conditions involving stereotype threat can be manipulated so that males and females perform equally well on a mathematics test. Another example came from social-role theory which claims that heroic or chivalrous acts are associated with males roles while nurturing acts are associated with female roles. The results from one study suggest that the size of the gender difference in helping behavior can vary based on the social context in which the behavior is measured (Eagly & Crowley, 1986, cited in Hyde, 2005).

While outside the context of this study, it is worth mentioning that Hyde also describes some costs of focusing on gender differences rather than gender similarities. She describes costs within the areas of work, parenting, and relationships, but the ones related to mathematics performance will be highlighted. Hyde mentions the stereotype that boys are better at math than girls are and includes examples of meta-

analyses that support the gender similarities hypothesis, and hence contradict the stereotype. She states that costs associated with this stereotype are that parents and teachers may overlook mathematically talented girls since they do not expect girls to be mathematically talented and parents may have lower expectations of mathematics performance for their daughters. Since research has shown a relationship between parents' expectations for mathematical success and the children's self-confidence and performance in mathematics, Hyde claims that girls may find their confidence to succeed mathematically undermined by parents' and teachers' beliefs.

2.2 Summary of major themes

Four major themes will be highlighted in this section: impact, differential item functioning, differential bundle functioning, and the gender similarities hypothesis. Many studies of gender differences in mathematics involve comparing mean scores for males to mean scores for females (Garner and Engelhard, 1999) and this is a measure of impact. Earlier in this chapter meta-analysis studies (Friedman, 1989; Hyde et al., 1990) were described which compared effect sizes across multiple studies to compare and look for trends in gender differences. The studies often found small effect sizes, some of which increased with age. Gender differences are often examined within a particular content area, such as algebra (favoring females) or spatial ability (favoring males). These differences are often measured on an entire test or a subset of the test, yet no steps are taken to control for ability, a potentially confounding variable.

There are now a large body of studies (Doolittle and Cleary, 1987; Harris and Carlton, 1993; Engelhard, 1990) which focus on individual items on a test to look for instances of differential item functioning. When two groups of participants are matched with respect to an attribute, such as ability, differences in performance

are then due differences in the way the item functions for the two groups, rather than due to differences in ability. The Mantel-Haenszel procedure or Item Response Theory methods are used to identify items with gender DIF. Some studies (Doolittle and Cleary, 1987) indicate gender DIF favoring females with algebra or computation items and favoring males with geometry and cognitively complex items, while other studies (Berberoglu, 1995) indicate gender DIF favoring females in geometry.

A natural extention of differential item functioning is differential bundle functioning where items with similar content are bundled together to look for gender differences on the bundle of items as a whole. Hypotheses are formed, often using results from earlier research, as to the direction of the gender differences. For example, it may be hypothesized that geometry items or items with spatial content are easier for males than females, while computational items or items with low cognitive complexity may be easier for females. A computer program, SIBTEST, is used to evaluate the hypotheses by determining the direction and magnitude of the DBF.

Returning to studies measuring impact, based on trends and consistencies among meta-analysis studies, a gender similarities hypothesis (Hyde, 2005) has been suggested that males and females are more similar than different on most psychological variables. Hyde conducted a meta-analysis where she found that a majority of the studies had effect size less than 0.35, which she consideres to be small effect sizes. In addition to small effect sizes which suggest gender similarities, she suggests that emphasizing gender differences rather than gender similarities can have deterimental effects on individuals associated with stereotypes related to professed gender differences.

2.3 How present study will extend literature

While there have been many studies on gender differences in mathematics based on performances on mathematics achievement tests or using methods to identify differential item functioning, there has been very little research on gender differences on mathematics competitions or applying DIF methods to mathematics competitions.

Articles which have been written about mathematics competitions often focus more on the qualitative benefits of mathematics competitions (Riley and Karnes, 1998) or how they can can be used to help student develop exceptional talent (Campbell and Wu, 1996). There have been some recent studies that have statistically examined mathematics competitions, yet they only include one of the two aspects of examining gender differences among the participants and evaluation using DIF techniques.

One study (Gleason, 2008) evaluated mathematics competitions using items response theory. The study examined two high school mathematics competitions, consisting of multiple choice questions, to attempt to answer three questions: (1) What does the instrument measure? (2) How much information is provided by the instrument? and (3) What types of items provide the most information? Thus, this evaluation was of the questions included in the competition, not the performance of the participants.

Leder et al. (2006) analyzed data from the Australian Mathematics Competition (optional participation) and the Victorian Certification of Education (required partipation) looking for gender differences. The sample consisted of grade 12 students who participated between 2002 and 2004. For both sets of data, there were more males than females receiving the top awards or reaching the top achievement levels, yet neither mean scores nor performance on individual items were compared.

Another study (Andreescu et al., 2008) analyzed the performance of students with exceptional talent in mathematical problems solving by comparing the top performers on three mathematics competitions: the William Lowell Putnam Mathematical Competition, International Mathematical Olympiad, and the USA Mathematical Olympiad. The number of female participants and females among the top performers were compared across multiple countries. While this study did have a gender component, the emphasis was on trying to explain the small numbers of women in the high levels of mathematics, and that the lack of women with an appropriate level of ability was not a sufficient explanation.

Thus, one study used DIF methods to analyze mathematics competitions, but there was no analysis of gender differences in performance, and the other studies had a gender component, but there was no use of DIF methods. This study addresses the deficiency in the literature by using DIF methods to analyze gender differences in performance on a mathematics competition, specifically the American Mathematics Competition AMC 8. The Mantel-Haenszel procedure, described in the next chapter, will be applied to each item to identify differential item functioning related to gender. The differences are then considered in the context of content of the items.

Chapter 3

Methods

3.1 Sample and site

The sample in the study consisted of 183,359 males and 178,857 females who were in the eighth grade in the United States when they participated in the AMC 8 during one of the years from 2003 to 2007. The AMC 8 was administered at each of the students' respective schools, and the data from the administrations during these years was obtained from the American Mathematics Competition office in Lincoln, Nebraska.

3.2 Access and permissions

This research received Institutional Review Board approval from the Board at the University of Nebraska–Lincoln. Permission was obtained from the Director of the American Mathematics Competition to analyze the data from the AMC 8 contests from 2003 through 2007.

3.3 Instruments and their reliability and validity

3.3.1 American Mathematics Competition AMC 8

As described earlier, each year middle school students around the world participate in the AMC 8 contest. Prior to this study, neither the reliability nor the validity of the AMC 8 had been measured in any formal statistical sense, partially due to the method of administration of the contest, but informal measures had been made by the American Mathematics Competition AMC 8 committee. Reliability has been measured by considering the consistency of the mean scores during the past five years, and validity has been established through discussions among the committee members.

Reliability is a measure of the extent to which scores are stable and consistent, and some forms of reliability include test-retest reliability, alternate forms reliability, and internal consistency reliability. The AMC 8 is administered at multiple schools throughout the United States and other countries on the same day in November. The questions are not released prior to the date of the contest and after that date, that year's questions are available to be used to prepare for future contests. There is also only one set of questions. As a result, the same questions cannot be reused for reliability measurement. As a weak measure of reliability, the mean scores of the AMC 8 have not varied by a large amount. Table 3.1 shows the mean scores and standard deviations from 2003 to 2007.

Table 3.1: Mean Scores on the AMC 8 Contest from 2003 to 2007

Year	n	Mean	S. D.
2003	80256	10.64	4.10
2004	76895	10.24	3.68
2005	71154	10.26	3.88
2006	70960	10.36	3.75
2007	69060	10.14	3.86

Another form of reliability, internal consistency reliability, can be measured by a variety of formulas based on correlations between items on two halves of a test. When there is only one administration of a test, the items can be split into two halves, and correlations can be made between the two halves, and this is called split-half reliability. A typical way of splitting a test is into even- and odd-numbered items. Since this method only uses information from half of the test, the Spearman-Brown formula can be used to estimate the reliability for a full length test (Thorndike, 1951). Cronbach's alpha (Cronbach, 1951) is another measure of internal consistency, which uses all of the items on a test, and it is equivalent to the Kuder-Richardson 20 (KR 20) test since the data is dichotomous .

One concern with the KR 20 test is if the test is speeded, then the value computed may not be accurate. According to Rindler (1979), the criteria that ETS uses to determine whether a test is speeded are the following: (1) virtually all students respond to at least 75% of the items, and (2) at least 80% of the students should reach the last item. Table 3.2 contains the percentage of students who answered 19 items (not necessarily the first 19 items) and who reached the 25th item.

Table 3.2: Speededness on the AMC 8 Contest from 2003 to 2007

Year	Items	Percentage	Item	Percentage
2003	19	97.4	25	92.5
2004	19	95.8	25	90.5
2005	19	96.0	25	89.3
2006	19	94.2	25	88.2
2007	19	96.3	25	91.5

As can be seen from the table, each year more than 80% of the students answered the last item, yet fewer than 98% of the students answer at least 75% of the items. Since percentages less than 98% may not be appropriately described as "virtually

all," the contest could be described as slightly speeded. These percentages mean that one should be cautious when interpreting the values of Cronbach's alpha (which is equivalent to KR-20).

Table 3.3 contains the values for Cronbach's alpha, the correlation between the two forms consisting of odd- and even-numbered items, and the Spearman-Brown value for two forms of unequal length (since there are an odd number of items).

Table 3.3: Internal Consistency on the AMC 8 Contest from 2003 to 2007

Year	Cronbach's α	Corr. btw. forms	Spearman-Brown
2003	.724	.587	.740
2004	.650	.474	.643
2005	.687	.549	.709
2006	.696	.537	.699
2007	.712	.569	.725

There is some debate as to what values of reliability coefficients are acceptable. A classic text which is frequently cited (Nunnally, 1970) gives 0.70 as a cut-off for Cronbach's alpha. Another researcher (Garson, 2008) states that a cut-off of 0.60 for Cronbach's alpha is acceptable in exploratory research, a cut-off of 0.70 should be used for adequate reliability, and 0.80 should be used for a good reliability. For the Spearman-Brown split half reliability, Garson (2008) states that a common rule of thumb is 0.80 for adequate reliability and 0.90 for good reliablity, although, 0.60 may again be used as a cut-off in exploratory research. Since the values for Cronbach's alpha and Spearman-Brown split-half coefficient are between .65 and .73 (and the test is slightly speeded), one may want to cautiously accept that the AMC 8 contests have some internal consistency.

While reliability is a measure of how stable and consistent scores are, validity is a measure of the extent to which the score make sense, are meaningful, and allow

a researcher to draw conclusions from the sample to a larger population. According to Garson (2008), a test "may be reliable but not valid, but it cannot be valid without being reliable. That is, reliability is a necessary but not sufficient condition for validity." The AMC 8 has been shown to have moderate levels of reliability and now the validity will be discussed. Three forms of validity are content validity, criterion-referenced validity, and construct validity. Content validity has been informally measured with the AMC 8 contest, but criterion-related and construct validity are difficult to statistically measure because of lack of available information about the students who take the AMC 8.

Criterion-related validity is a measure of how well the scores on the AMC 8 relate to a particular outcome or predict some future outcome, and it is difficult to statistically measure this for the AMC 8. An example of an outcome is performance in a particular mathematics course such as algebra, yet the students' performance in algebra is not available and hence could not be correlated with their performance on the AMC 8. While the primary purpose of the AMC contests is not to predict future performance and despite the difficulty of statistical measurement, Steve Dunbar, the AMC Director, feels that "all of the AMC contests have a high criterion-validity with respect to future mathematical success and more generally with academic success in high school and beyond" (AMC Director, personal communication, May 30, 2008).

The second form of validity, construct validity, is determined by considering if the scores are significant, meaningful, useful, and have a purpose, and this form of validity again is challenging to assess with the AMC 8 data (Creswell, 2005). As with criterion validity, construct validity can be evaluated by correlating the AMC 8 scores with another measure, but information about other measures as applied to the students who have taken the AMC 8 is not currently available. Another measure of construct validity would be how useful are the AMC 8 scores for making decisions,

such as by the schools or teachers about the students who have participated in the AMC 8 contest, yet it is unknown to what extent the scores are used by anyone to make decisions. Construct validity also includes how well the scores from the sample of students who have taken the AMC 8 can be used to generalize to a larger population of students. The data analyzed included the scores of all the United States eighth graders who took the AMC 8 during the five years from 2003 to 2007 instead of a subset, yet since the students were primarily self-selected rather than a random sample of United States eighth graders, the generalizability of the results to all United States eighth graders is limited.

The third type of validity, content validity, is the extent to which the questions are representative of all possible questions that could be asked, in this case questions about middle school mathematics, and it has been assessed with regards to the AMC 8 contest. There is a committee and a panel of readers who are involved in choosing the questions to be part of the AMC 8 contest each year, and they include middle-school teachers and people who teach pre-service and in-service middle school teachers. To informally address content validity, there are discussions among these individuals about whether a particular question is appropriate based on middle school curriculum (AMC Director, personal communication, May 30, 2008). The questions which are chosen to be on the AMC 8 are those for which there is a sufficient amount of agreement about the appropriateness of the questions which implies that the content validity of the AMC 8 is at a satisfactory level.

While reliability and validity are important measures of any test used to make decisions about students and there are many statistical tests that can be used to compute reliability and validity coefficients, determining whether those coefficients are "good" are typically based on rules of thumb or common practice rather than hard and fast rules. Occasionally factors not directly related to the test, such as

time, money, or purpose of the test, influence whether "good" coefficients are accepted rather than investing more resources to obtain better coefficients. Informal measures of reliability and validity have been made through conversations among committee members and opinions of the AMC director who has extensive experience with the American Mathematics Competitions. The reliability and validity of the AMC 8 contests are also a product of the combined experiences of those who have been involved in the contests over the years; these individuals have the experience to help them know what works well and they use this to design new contests each year. While some of the aspects of the AMC contests make it difficult to compute formal statistical analyses, internal consistency has been measured using a variety of formulas, and the results lend support to the informal conclusions of the American Mathematics Competition AMC 8 committee members.

3.3.2 Mantel-Haenszel Procedure

The Mantel-Haenszel procedure is one method of measuring differential item functioning (DIF), which is present when an item functions differently for examinees in two matched groups. In this study, it is determined whether there are AMC 8 contest items that function differently among males and females who are otherwise similar in their overall score. Mantel and Haenszel (1959) first described the procedure for studying matched groups, then Holland (1985) and Holland and Thayer (1985) adapted the procedure to be used to assess DIF (Dorans and Holland, 1993). It can be used to measure the association between two dichotomous variables, such as gender and a correct or incorrect response to an item, after controlling for a confounding variable associated with the variables, such as mathematical ability. At each ability level m, a 2×2 contingency table is constructed that includes the number of people

who answered the question correctly and incorrectly for the focal and the reference group. When considering gender, males typically are the reference group while females are the focal group. Table B.1 in Appendix B shows a typical 2×2 contingency table and the table below will be used in the following description.

Table 3.4: The 2×2 Contingency Table for ability level m.

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

The Mantel-Haenszel procedure has an associated null hypothesis expressed in terms of the odds ratio. Let p_1, p_2 be the probability of success for two groups and $q_i = 1 - p_i$. Then the odds ratio (OR) is defined as:

$$\text{OR} = \frac{p_1/q_1}{p_2/q_2}.$$

The null DIF hypothesis for the Mantel-Haenszel procedure is

$$H_0 : \frac{R_{rm}}{W_{rm}} = \frac{R_{fm}}{W_{fm}}, m = 1, \ldots, M.$$

That is, the odds of answering an item correctly for some ability level is the same for both the focal group and the reference group across all M levels of ability.

Mantel and Haenszel (1959) developed a chi-square test of the null DIF hypothesis against an alternative hypothesis referred to as the constant odds ratio hypothesis:

$$H_\alpha : \frac{R_{rm}}{W_{rm}} = \alpha \frac{R_{fm}}{W_{fm}}, m = 1, \ldots, M, \alpha \neq 1.$$

The parameter α is the *common odds ratio* since under H_α the value of α is the same

across all ability levels. The corresponding chi-square test is distributed approximately with one degree of freedom.

Test developers at ETS use a delta metric (with a mean of 13 and a standard deviation of 4) in their analyses, and hence Holland and Thayer (1985) converted α into the delta metric with the following formula:

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

Note that positive values indicate that the focal group, in this case females, are more likely to answer an item correctly.

After computing the MH D-DIF values, items need to be classified, based on the magnitude of the values, for further analysis. According to Dorans and Hollands (1993), ETS classifies levels of DIF demonstrated by items by placing the items in one of three categories: negligible DIF (A), intermediate DIF (B), and large DIF (C). Items are classified in category A if either MH D-DIF is not statistically different from zero or the magnitude of the MH D-DIF value has an absolute value less than 1. Items are classified in category C if the MH D-DIF value both has absolute value greater than 1.5 and is statistically significantly larger than 1.0 in absolute value. The remaining items are placed in category B.

Compared to other methods of identifying DIF, the Mantel-Haenszel procedure has multiple advantages, yet there is one significant disadvantage. Some of the advantages of the Mantel-Haenszel procedure are that it is computationally simple, easy to implement, and has an associated test of significance (Rogers and Swaminathan, 1993). A disadvantage of the procedure is that it is designed to detect uniform DIF and therefore it may not detect nonuniform DIF (Swaminathan and Rogers, 1990; Rogers and Swaminathan, 1993; Narayanam and Swaminathan, 1996).

There are multiple ways to address this disadvantage. While none of these methods are used with this study, they offer directions for further research. There is a modified Mantel-Haenszel procedure that has been shown to have some ability at detecting non-uniform DIF (Mazor et al., 1994). The second way is to apply the Breslow-Day test which uses the odds ratio to identify non-uniform DIF (Penfield, 2003). A third method is logistic regression (Swaminathan and Rogers, 1990) which detects both uniform and nonuniform DIF, although this method is neither computationally simple nor easy to implement.

3.4 Procedures of data collection

Electronic text files were obtained from the American Mathematics Competition office in Lincoln, Nebraska. For each year 2003 through 2007, the files included the students' response on each question and their gender, age, grade, and location where they took the test. The data for students who were in the eighth grade in the United States was selected for analysis. SPSS was used for data organization and analysis.

3.5 Analysis of the data

In this study, the focal group was females while the reference group was males. The level of mathematical ability was measured by the total score on the AMC 8. Positive values of MH D-DIF favor females while negative values favor males. The significance level was $p = .01$ for a chi-square test with one degree of freedom. Standard practice is to divide participants into five groups, and the students were divided in this way according to the range of scores in Table 3.5.

Table 3.5: Ability levels based on total score on the AMC 8

Ability level	Range of scores
Low	0-5
Med-low	6-10
Med	11-15
Med-high	16-20
High	21-25

At each ability level, the counts of males and females and the means by gender were computed. For some of the students, the gender was not included. These students comprised less than two percent of the total number of students participating each year and were omitted from the data analysis.

3.5.1 Statistical Analysis

First basic descriptive analyses were done for each of the five years of data. After finding the percent of participation by gender each year; the mean, median, and standard deviation were computed for males and females. The effect size d (Cohen, 1969) was computed as $d = (\overline{x}_M - \overline{x}_F)/\sigma$ where $\sigma = \sqrt{(\sigma_M^2 + \sigma_F^2)/2}$, the difference in means divided by the square root of the average of the variances.

Three types of statistical analyses were done to identify DIF and impact. Each of them focused on individual items, and there were a total of 125 items from the five years. The first analysis involved the impact as measured by the proportion of males and females who answered an item correctly. The proportion correct for each gender was computed and impact was measured as the male proportion correct minus the female proportion correct. The second analysis was used to determine the existence of non-uniform DIF by computing the proportion correct for males and females at each of the five ability levels. To determine the statistical signficance of the impact values,

a two-way χ^2 test was applied to each item both over all ability levels and within each ability level to test the null hypothesis that there is no relationship between gender and answering the item correctly. For the third analysis, the Mantel-Haenszel procedure was applied to each item to determine if there was an association between gender and answering an item correctly after controlling for mathematical ability.

3.5.2 Substantive analysis

The next group of analyses involve two steps: first, the AMC 8 items were placed into categories according to certain classification schemes; next, the number of items in each category were counted. The items were classified according to impact, type of DIF, MH D-DIF, NCTM content standards, Gierl et al.'s modified taxonomy, Harnisch et al.'s attributes, length of the stem of the item, and the gender of any names included in the stem.

The first three classification methods were based on impact, type of DIF, and MH D-DIF. The three categories of impact were: no impact if the impact was not significant, impact favoring males if the value was positive, or impact favoring females if the value was negative. For type of DIF, items for which at least three of the ability levels had significant values were placed into one of three DIF categories: nonuniform DIF, uniform DIF favoring males, or uniform DIF favoring females. An item was classified as demonstrating nonuniform DIF if at least 1 ability level favored one gender while the remaining ability levels favored the other gender. An item was classified as uniform DIF favoring one gender if all of the ability levels with significant impact values favored that gender.

The items also were placed into one of three categories based on MH D-DIF values: negligible DIF, MH D-DIF favoring males, or MH D-DIF favoring females. Recall

that ETS classifies items with MH D-DIF values between -1 and 1 as having negligible DIF. Since only two items out of the 125 fell in this category, a finer classification was used in this research, and items with MH D-DIF values between -0.5 and 0.5 were considered to have negligible DIF. Positive values of MH D-DIF at least 0.5 indicate the item favors females, while negative values at most -0.5 indicate that the item favors males. This classification method was used to identify items with gender DIF.

The next classification methods involved classifying the items based on the content, cognitive charactistics, and length of the questions. First, the 125 mathematics items were classified in terms of mathematical content by using the National Council of Teachers of Mathematics (NCTM) five content standards: Number and Operations, Algebra, Geometry, Measurement, and Data Analysis and Probability (www.nctm.org).

Gallagher et al. (2000) developed a taxonomy of content and cognitive characteristics of items which may account for gender differences in mathematics. For example, an item which has multiple solution paths, one of which may be a shortcut, is expected to favor males while an item which requires using memorized material, such as a definition or a formula, is expected to favor females. Gierl et al. (2003) modified this taxonomy to include more specific and mutually exclusive categories. The AMC 8 items also were classified according to this modified taxonomy, included in Appendix C, the second classification method.

The third classification method was based on a list of twenty attributed developed by Harnisch et al. which are included in Appendix D. One attribute is that the item may contain figures, graphs or tables, and another attribute is that solving the item may involve generating a figure or a table. Each item was further classified by length based on the number of words and lines used to state the question; an item was classified as long if it either included at least 50 words or covered at least 4 lines.

Based on conversations with members of the AMC 8 committee, a final classification method was included. When names are included in the stem of an item, the committee attempts to balance the names by gender. Thus, the items were classified into three categories based on the gender(s) represented by the names: male, female, or both. Note that the classifications were not mutually exclusive; that is, a particular item may be classified as a geometry item, a measurement item, having multiple solution paths, and including a diagram.

After all of the items were classified according to the eight methods, the items in each classification scheme were counted. These numbers of items give a basic blueprint for the types of questions that typically or rarely appear on an AMC 8 test. The items classified according to the NCTM content standards, Gierl et al.'s modified taxonomy, Harnisch et al.'s attributes, length, and gender of names were used for subtest analyses.

3.5.3 Subtest analysis

The three methods of classifications (NCTM, Gierl et al., and Harnisch et al.) of the items were also used for subtest statistical analyses of means and proportions correct by gender and ability level. First, for each student, the number correct for a particular category was computed. Then the mean male score and the mean female score was found. Then, the impact was computed as described earlier. Since the number of items varied between categories, the impact was divided by the number of items in each category. This allowed values from multiple years and multiple categories to be compared.

These computations are measures of impact, and earlier the importance to differentiate between impact and DIF was stressed. To attempt to compare items within a

category and include DIF results, for each category the number of items which demonstrated negligible DIF, DIF favoring males, and DIF favoring females were counted. For example, of the 14 geometry items on the 2003 AMC 8 contest, 12 of them had negligible DIF while the remaining two were split between DIF favoring males and DIF favoring females. The subtest analyses were applied to five classification methods: NCTM, Gierl et al., Harnisch et. al, length, and gender of names.

These subtest analyses were done for certain categories within the classification methods. Only a subset of the categories from each classification were chosen and this was primarily based on the number of items in the categories. For some items, there were too few items to be able to draw meaningful conclusions, and for one category, the number of items was large enough that any results for that category would not likely be different from analyses for the entire test. For the NCTM standards, the three categories were Numbers and Operations, Geometry, and Measurement. For Gierl et al.'s modified taxonomy, the categories were Spatial, Routine-Familiar, and Memorization. For Harnisch et al.'s attributes, the attributes 1, 10, and 11 were chosen. The first attribute involved many basic number operations, attribute 10 was related to higher mental processes, and attribute 11 identified items with figures, tables, or graphs.

Finally, the twenty-four items which were identified as having non-negligible DIF were placed in one table with their correponding classifications. These items were qualitatively explored for patterns and similarities within each classification category. The specific items with gender DIF are included in Appendix E.

Chapter 4

Results

This chapter contains the results of the data analysis on the American Mathematics Competitions AMC 8 Contests for 2003 through 2007. The first section contains the descriptive analysis of all the data and most of the tables displaying the results are included in this section. The second section contains inferential analysis to address the research questions and hypotheses. The final section contains the remaining tables.

4.1 Descriptive analysis of all data

The descriptive analysis of the data is divided into three subsections. The first subsection includes the statistical analyses used to determine impact and MH D-DIF for each of the items, and the results are displayed in a table for each of the five years. Next are the results from the substantive analyses where the items were placed into categories based on multiple classification methods; again, the results are displayed in tables for each year. In the third subsection, the subtest analysis results are organized in tables by classification method, with tables for impact values and distributions of DIF items by gender.

4.1.1 Statistical analysis

Basic descriptive data about the participants and the AMC 8 contest scores for the five years from 2003 to 2007 are included in Table 4.24. Each year there were more males than females, although the difference in percentages was never more than 3%, and the percentages for each gender had a range of at most 1% over the five years. The means, medians, and standard deviations were consistent during the five years. Each year the male mean score was greater than the female mean score, and the means were statistically significant ($p < 0.001$). The medians were typically less than the means, although during 2005 the medians were greater than the means for both genders. The male scores had larger standard deviations as compared to the female scores during the five years, but the ranges of standard deviations were less than 0.5 for both males and females. The effect sizes were consistently small.

Tables 4.1 through 4.5 include summaries by year of statisical analyses for impact and MH D-DIF. [1] To aid the reading of the tables, cells with significant impact values favoring females are colored pale pink while cells with values favoring males are colored light blue; in the MH D-DIF column, values with magnitudes at least 0.5 are colored. After identifying the item number, the next column is the overall impact which is based on proportions of males and proportions of females who answered the item correctly. Significance was determined using a two-way χ^2 test. Note that negative values indicate the impact favors females, and the rows of the table are sorted by increasing value of impact.

The next five columns include the impact values for each ability level; the ability levels medium-low, medium, and medium-high are represented by M-L, med, and M-H, respectively. These columns will be referred to as the ability impact columns. A

[1] A careful explanation of reading information from the table is included for the first table while only highlights are included for the remaining tables.

two-way χ^2 test identified significant impact values, and these were used to determine the type of DIF for each item. Items having at least three ability levels with significant impact were classified as uniform DIF favoring males, uniform DIF favoring females, or nonuniform DIF. An item has uniform DIF if all of the significant values favor the same gender; otherwise the item has nonuniform DIF.

The last column includes the MH D-DIF values. Using ETS criteria, MH D-DIF values between -1 and 1 indicate negligible DIF. Using this criteria, only two items have non-negligible DIF.[2] In this study, MH D-DIF values between -0.5 and 0.5 are considered negligible DIF. In contrast to the impact analyses, for MH D-DIF, positive values favor females. Significance was determined using the Mantel-Haenszel procedure.

Table 4.1 contains the impact and MH D-DIF results for the 2003 AMC 8 contest. All of the impact values are statistically significant ($p < 0.01$) except for item 14. Two items (1, 17) have impact favoring females while the remaining twenty-two items favor males. Five items (1, 17, 14, 8, 16) demonstrate uniform DIF favoring females, and five items (22, 19, 24, 15, 3) demonstrate uniform DIF favoring males. This can be seen in the table in the rows where at least three of the cells in the ability impact columns have the same color. Four items (12, 11, 4, 7) show nonuniform DIF. In the table, each of these items have 1-2 cells of one color and 1-2 cells of the other color (among the ability impact columns). Of the items with nonuniform DIF, only item 7 was identified as having gender DIF using the classification criteria of this study. Among the remaining three items, only one has a significant MH D-DIF value, and it is between -0.3 and -0.2. Recall that the Mantel-Haenszel procedure is not adept at identifying items with nonuniform DIF.

[2]Item 3 on the 2003 AMC 8 has gender DIF favoring males, and item 9 on the 2007 AMC 8 has gender DIF favoring females.

Table 4.1: Impact and MH D-DIF for the 2003 AMC 8 Contest

	Impact						
Item	Overall	Low	M-L	Med	M-H	High	MH D-DIF
1	-0.20^{*}	-0.063^{*}	-0.059^{*}	-0.037^{*}	-0.017^{*}	0.002	0.755^{*}
17	-0.013^{*}	-0.035^{*}	-0.036^{*}	-0.058^{*}	-0.065^{*}	-0.020	0.474^{*}
14	-0.003	-0.016	-0.026^{*}	-0.057^{*}	-0.067^{*}	-0.012	0.451^{*}
8	0.011^{*}	-0.035^{*}	-0.045^{*}	-0.027^{*}	-0.008	-0.005	0.389^{*}
25	0.021^{*}	-0.007	0.002	-0.002	-0.015	0.020	0.018
16	0.021^{*}	-0.039^{*}	-0.021^{*}	-0.021^{*}	-0.002	0.008	0.228^{*}
21	0.022^{*}	0.006	0.003	-0.009	$-0.028^{\dagger}$	-0.029	0.050
23	0.023^{*}	-0.002	0.001	0.002	0.001	-0.010	-0.020
20	0.026^{*}	0.018^{*}	0.006	0.004	0.028^{*}	0.015	-0.197^{*}
22	0.030^{*}	0.021^{*}	0.018^{*}	0.015^{*}	-0.013	-0.025	-0.284^{*}
12	0.040^{*}	$-0.014^{\dagger}$	-0.017^{*}	0.000	0.043^{*}	0.008	0.064
2	0.042^{*}	0.014	-0.002	-0.004	0.019	-0.010	-0.008
18	0.042^{*}	0.008	0.001	0.006	0.030^{*}	0.023	-0.115
13	0.045^{*}	-0.012	-0.006	0.001	0.011	0.023	0.024
11	0.046^{*}	-0.021^{*}	-0.016^{*}	0.019^{*}	0.061^{*}	0.016	-0.073
4	0.047^{*}	-0.007^{*}	0.023^{*}	0.013^{*}	0.005	-0.008	-0.225^{*}
19	0.050^{*}	-0.003	0.004	0.022^{*}	0.071^{*}	0.078^{*}	-0.204^{*}
9	0.050^{*}	-0.013	0.001	$0.015^{\dagger}$	0.013	0.002	-0.066
6	0.057^{*}	0.025^{*}	0.010	-0.005	-0.009	-0.005	-0.045
10	0.058^{*}	0.000	-0.013^{*}	0.003	0.015	0.004	0.050
5	0.061^{*}	0.011	0.050^{*}	0.010^{*}	-0.005	0.005	-0.406^{*}
24	0.076^{*}	$0.018^{\dagger}$	0.020^{*}	0.048^{*}	0.081^{*}	0.028	-0.489^{*}
15	0.096^{*}	0.005	0.037^{*}	0.078^{*}	0.045^{*}	-0.010	-0.601^{*}
7	0.099^{*}	-0.008	-0.044^{*}	0.079^{*}	0.038^{*}	0.001	-0.569^{*}
3	0.124^{*}	0.077^{*}	0.127^{*}	0.051^{*}	$0.007^{\dagger}$	0.004	-1.234^{*}

Significance levels: $^{*}p < 0.01$, $^{\dagger}p < 0.05$

Five of the items (1, 17, 11, 24, 3) have significant impact at four of the ability levels, while the impact value at the high ability level is not significant. Only item 19 has a significant impact value at the high ability level. Using the ETS criteria, item 3 has intermediate DIF favoring males, and it also has the largest impact value favoring males. Of the four items (1, 15, 7, 3) with non-negligible DIF, one item favors females and three items favor males.

Table 4.2: Impact and MH D-DIF for the 2004 AMC 8 Contest

	Impact						
Item	Overall	Low	M-L	Med	M-H	High	MH D-DIF
14	0.002	0.011	−0.002	−0.018*	−0.061*	0.019	0.108*
13	0.004	0.001	−0.012*	−0.047*	−0.073*	−0.033	0.389*
21	0.006†	0.010	−0.018*	−0.030*	−0.001	0.007	0.247*
10	0.007	−0.059*	−0.036*	−0.026*	−0.011	−0.002	0.372*
11	0.008†	−0.036*	−0.042*	−0.031*	−0.017*	0.030†	0.428*
4	0.012*	−0.026†	−0.043*	−0.011	−0.021†	0.012	0.296*
24	0.012*	0.002	0.005	−0.006	−0.003	−0.016	0.003
8	0.015*	−0.033*	−0.027*	−0.008	0.001	0.018	0.190*
15	0.015*	0.002	−0.018*	−0.040*	−0.040*	−0.036	0.287*
2	0.023*	−0.026*	−0.008	−0.004	−0.010	0.016	0.081
25	0.024*	0.011†	0.016*	0.005	0.001	0.052	−0.238*
19	0.026*	0.003	−0.005	−0.005	0.000	0.014	0.043
5	0.027*	0.007	0.008	−0.031*	−0.036*	0.016	0.104*
1	0.031*	0.031*	0.025*	0.010*	0.000	−0.002	−0.656*
18	0.032*	0.015	−0.005	0.001	0.013	0.014	−0.004
12	0.035*	0.003	0.005	0.029*	0.024	−0.066	−0.210*
17	0.038*	0.010	0.012*	−0.005	−0.021	−0.049	−0.047*
3	0.043*	0.007	0.013†	0.012*	0.003	0.002	−0.160
22	0.049*	0.011	0.014*	0.026*	0.058*	0.080†	−0.408*
7	0.052*	0.001	0.001	−0.005	−0.004	0.000	0.021
20	0.068*	−0.002	0.017*	0.045*	0.036*	−0.025	−0.392*
23	0.073*	0.003	0.024*	0.061*	0.082*	0.025	−0.591*
16	0.082*	0.020*	0.029*	0.068*	0.117*	0.039	−0.830*
9	0.104*	0.011	0.043*	0.088*	0.057*	0.005	−0.678*
6	0.125*	0.033*	0.093*	0.076*	0.036*	0.005	−0.912*

Significance levels: $*p < 0.01$, $^{\dagger}p < 0.05$

The results for the 2004 AMC 8 are in Table 4.2. Except for three items, all impact values are statistically significant and favor males. Four items have uniform DIF favoring females and seven items have uniform DIF favoring males. Only one item has nonuniform DIF, and it has negligible DIF. Three items have significant impact at four ability levels. Five items have non-negligible DIF, and all of them favor males.

Table 4.3: Impact and MH D-DIF for the 2005 AMC 8 Contest

	Impact						
Item	Overall	Low	M-L	Med	M-H	High	MH D-DIF
14	-0.024^{*}	-0.027^{*}	-0.032^{*}	-0.085^{*}	-0.065^{*}	-0.019	0.585^{*}
11	-0.017^{*}	-0.083^{*}	-0.073^{*}	-0.030^{*}	-0.009	0.000	0.632^{*}
19	$-0.012^{\dagger}$	-0.028^{*}	-0.053^{*}	-0.057^{*}	-0.014	0.017	0.522^{*}
15	0.000	-0.003	-0.001	-0.030^{*}	-0.021	-0.040	0.170^{*}
25	0.006	$-0.016^{\dagger}$	-0.011^{*}	-0.011	-0.012	0.025	0.147^{*}
23	0.012^{*}	0.012	-0.008	-0.016^{*}	-0.010	0.021	0.126^{*}
9	0.015^{*}	-0.008	-0.031^{*}	-0.021^{*}	0.004	-0.018	0.240^{*}
20	0.014^{*}	0.002	-0.018^{*}	-0.002	0.004	0.057	0.099
3	0.017^{*}	-0.010	-0.018^{*}	-0.033^{*}	-0.012	0.012	0.232^{*}
2	0.019^{*}	$-0.024^{\dagger}$	-0.019^{*}	-0.019^{*}	-0.025^{*}	-0.008	0.272^{*}
8	0.023^{*}	-0.028^{*}	-0.028^{*}	$-0.012^{\dagger}$	-0.008	-0.032	0.234^{*}
22	0.025^{*}	0.004	0.004	0.006	0.034^{*}	0.106^{*}	-0.127^{*}
21	0.027^{*}	0.013	0.001	0.004	-0.005	-0.025	-0.035
4	0.046^{*}	-0.006	-0.007	-0.005	-0.011	-0.006	0.083
24	0.046^{*}	$-0.006^{\dagger}$	-0.007^{*}	-0.005	-0.011	-0.006	-0.128^{*}
13	0.047^{*}	0.007	$0.008^{\dagger}$	0.004	-0.017	-0.027	-0.069
10	0.050^{*}	0.034^{*}	0.040^{*}	0.016^{*}	0.005	0.004	-0.630^{*}
1	0.060^{*}	0.027^{*}	0.034^{*}	0.005	-0.004	-0.004	-0.253^{*}
12	0.062^{*}	0.009	$0.012^{\dagger}$	0.021^{*}	-0.012	0.023	-0.161^{*}
18	0.079^{*}	0.032^{*}	0.037^{*}	0.041^{*}	$0.031^{\dagger}$	0.022	-0.465^{*}
6	0.085^{*}	-0.003	0.028^{*}	0.071^{*}	0.052^{*}	0.009	-0.538^{*}
16	0.088^{*}	$0.012^{\dagger}$	0.031^{*}	0.070^{*}	0.134^{*}	0.072^{*}	-0.897^{*}
7	0.096^{*}	0.022^{*}	0.034^{*}	0.078^{*}	0.070^{*}	0.021	-0.816^{*}
5	0.104^{*}	$0.025^{\dagger}$	0.088^{*}	0.051^{*}	0.023^{*}	0.005	-0.787^{*}
17	0.107^{*}	0.004	0.061^{*}	0.078^{*}	0.053^{*}	$0.033^{\dagger}$	-0.684^{*}

Significant levels: $^{*}p < 0.01, ^{\dagger}p < 0.05$

Table 4.3 displays the results for the 2005 AMC 8 data. All of the impact values are statistically significant except for two. Three items have negative impact values, and twenty items have positive impact values. Of the twelve items with uniform DIF, five favor females and seven favor males; there are no items with nonuniform DIF. Six items have significant impact values for four ability levels. Nine items have non-negligible DIF; three of them favor females while six of them favor males.

Table 4.4: Impact and MH D-DIF for the 2006 AMC 8 Contest

	Impact						
Item	Overall	Low	M-L	Med	M-H	High	MH D-DIF
1	-0.009^{*}	-0.209^{*}	-0.014^{*}	-0.007^{*}	-0.004	-0.003	0.761^{*}
14	-0.002	-0.357^{*}	-0.032^{*}	-0.012^{*}	0.003	-0.004	0.487^{*}
19	0.001	-0.035^{*}	-0.030^{*}	-0.037^{*}	-0.013	-0.012	0.334^{*}
20	0.005	-0.003	-0.026^{*}	-0.050^{*}	-0.038^{*}	-0.013	0.410^{*}
15	0.014^{*}	-0.001	-0.026^{*}	-0.044^{*}	-0.012	0.003	0.326^{*}
24	0.020^{*}	0.010	0.003	-0.004	-0.014	0.097^{*}	-0.019
16	0.024^{*}	-0.002	$-0.013^{\dagger}$	$-0.015^{\dagger}$	-0.004	0.038	0.173^{*}
9	0.025^{*}	-0.015	-0.006	$-0.015^{\dagger}$	-0.019	0.011	0.121
2	0.025^{*}	-0.004	0.017^{*}	0.002	0.006	$0.017^{\dagger}$	-0.219
7	0.025^{*}	0.003	-0.004	-0.028^{*}	-0.050^{*}	-0.027	0.250^{*}
11	0.028^{*}	0.011	0.014^{*}	0.009	-0.018	$-0.078^{\dagger}$	-0.151
25	0.029^{*}	-0.003	0.001	-0.001	0.020	0.032	-0.023
5	0.030^{*}	-0.007^{*}	$-0.006^{\dagger}$	$0.012^{\dagger}$	0.014^{*}	0.004^{*}	-0.017
10	0.035^{*}	0.005	0.003	-0.019^{*}	-0.022	0.001	0.094
8	0.037^{*}	-0.041^{*}	-0.004	0.023^{*}	-0.012	-0.014	-0.027
6	0.040^{*}	0.002	0.002	-0.009	-0.014	-0.004	0.035
22	0.044^{*}	$0.017^{\dagger}$	0.013^{*}	0.023^{*}	0.049^{*}	$0.070^{\dagger}$	-0.367^{*}
18	0.048^{*}	0.010	$0.009^{\dagger}$	0.022^{*}	0.028	0.022	-0.229^{*}
21	0.048^{*}	0.012	0.004	0.024^{*}	0.055^{*}	0.011	-0.269^{*}
23	0.054^{*}	0.018^{*}	0.015^{*}	0.027^{*}	-0.006	0.001	-0.312^{*}
4	0.060^{*}	0.019	0.047^{*}	0.009	0.007	0.008	-0.354^{*}
17	0.072^{*}	0.012	0.007	0.061^{*}	0.092^{*}	-0.006	-0.443^{*}
13	0.082^{*}	0.002	$0.009^{\dagger}$	0.050^{*}	0.070^{*}	0.042	-0.366^{*}
12	0.083^{*}	$0.020^{\dagger}$	0.036^{*}	0.044^{*}	0.029^{*}	0.000	-0.420^{*}
3	0.108^{*}	0.030^{*}	0.052^{*}	0.089^{*}	0.075^{*}	0.011	-0.728^{*}

Significant levels: $^{*}p < 0.01, ^{\dagger}p < 0.05$

The 2006 AMC 8 results are in Table 4.4. All items have significant impact, except for three. Only one item favors females while the remaining twenty-one items favor males. Four items have uniform DIF favoring females, and five items have uniform DIF favoring males. Item 5 is the only one with nonuniform DIF, and it has negligible DIF. Two items have significant impact at all ability levels except high ability. Only two items have gender DIF, one favoring each gender.

Table 4.5: Impact and MH D-DIF for the 2007 AMC 8 Contest

	Impact						
Item	Overall	Low	M-L	Med	M-H	High	MH D-DIF
9	-0.039^{*}	-0.127^{*}	-0.084^{*}	-0.034^{*}	$-0.014^{\dagger}$	-0.007	1.045^{*}
18	0.001	-0.032^{*}	-0.032^{*}	-0.036^{*}	-0.012	0.030	0.311^{*}
3	0.011^{*}	-0.037^{*}	-0.044^{*}	-0.017^{*}	$-0.015^{\dagger}$	0.001	0.365^{*}
5	0.018^{*}	0.003	0.000	0.000	-0.002	-0.011	0.009
1	0.019^{*}	0.004	-0.007	0.001	-0.001	0.003	0.054
25	0.021^{*}	$0.013^{\dagger}$	0.007	0.006	0.006	$0.105^{\dagger}$	-0.161^{*}
20	0.022^{*}	-0.015	-0.015^{*}	-0.005	$0.035^{\dagger}$	$0.051^{\dagger}$	0.096
24	0.024^{*}	0.006	0.000	-0.008	-0.004	-0.044	0.050
21	0.026^{*}	0.009	0.004	-0.006	0.014	-0.014	-0.037
13	0.027^{*}	0.006	0.003	0.009	-0.015	$-0.086^{\dagger}$	-0.069
2	0.027^{*}	-0.015	-0.009	-0.013^{*}	-0.008	-0.015	0.153^{*}
4	0.030^{*}	-0.023^{*}	-0.017^{*}	$-0.013^{\dagger}$	0.008	-0.008	0.160^{*}
10	0.034^{*}	0.004	$-0.007^{\dagger}$	$-0.012^{\dagger}$	-0.016	0.006	0.145^{*}
11	0.035^{*}	-0.001	$-0.011^{\dagger}$	-0.024^{*}	0.008	0.006	0.155^{*}
22	0.042^{*}	0.009	0.012^{*}	0.020^{*}	0.015	$0.094^{\dagger}$	-0.215^{*}
23	0.042^{*}	0.024^{*}	0.019^{*}	$0.012^{\dagger}$	-0.018	0.062	-0.205^{*}
14	0.048^{*}	0.019^{*}	0.010^{*}	0.015^{*}	0.008	-0.003	-0.289^{*}
19	0.050^{*}	0.019^{*}	0.005	0.008	0.049^{*}	0.029	-0.161^{*}
16	0.057^{*}	0.003	0.003	$0.026*$	0.044^{*}	-0.002	-0.194^{*}
8	0.058^{*}	0.025^{*}	0.007	$-0.010^{\dagger}$	-0.005	0.006	-0.026
12	0.068^{*}	-0.003	$0.016^{\dagger}$	$0.019^{\dagger}$	0.004	0.012	-0.166^{*}
15	0.088^{*}	-0.015	0.039^{*}	0.055^{*}	0.020	-0.013	-0.421^{*}
17	0.092^{*}	0.020^{*}	0.038^{*}	0.051^{*}	$0.025^{\dagger}$	0.011	-0.500^{*}
7	0.094^{*}	0.027^{*}	0.054^{*}	0.038^{*}	0.010	0.008	-0.521^{*}
6	0.118^{*}	0.037^{*}	0.076^{*}	0.084^{*}	0.066^{*}	-0.012	-0.915^{*}

Significant levels: $^{*}p < 0.01, ^{\dagger}p < 0.05$

Table 4.5 has the 2007 AMC 8 results. All of the impact values were significant except for item 14. One item had impact favoring females; the remaining 23 items favored males. The nine items with uniform DIF had three favoring females. Item 20 was the only one with nonuniform DIF, and it had non-negligible DIF. Four items were significant at all ability levels except for high ability. Four items had gender DIF, and the one favoring females also had intermediate DIF using the ETS criteria.

4.1.2 Substantive analysis

The first step of the substantive analysis was to classify each item according to multiple classification methods described in the previous chapter. Tables 4.6 through 4.10 include the results of all of the classifications that were applied to each item: impact, type of DIF, MH D-DIF, NCTM content standards, Gierl et al.'s modified taxonomy, Harnisch et al.'s attributes, length of stem, and gender of names included in the stem. As with the previous five tables, the items were sorted by impact.

After identifying the item number, within the next three columns of the table, gender differences were represented with F for female and M for male; if a cell is left blank, then the gender differences were either not significant or negligible. These columns repeat the results from the earlier set of tables.

For the other five columns, a letter or number in a column indicates the item was placed in that category of the corresponding classification scheme. In the NCTM column, N represents number and operations, G represents geometry, and M represents measurement. In the Gierl column, S represents spatial, R represents routine-familiar, and M represents memorization. In the Harnisch column, 1 represents the attribute involving number computations, 10 represents the attribute for higher mental processes, and 11 represents the attribute which includes figures, tables, and graphs. For length, S represents short and L represents long. Note that every item was placed in one of the two categories. For the names, F indicates the item contains female name(s), M indicates the item contains male name(s), and B indicates that the item contains both male and female names. One item on the 2003 AMC 8 included an ambiguous name, and this item was labeled with A in the names column.

Table 4.6 contains the results for the 2003 AMC 8 contest. Classifying the items according to the NCTM content standards identified 15 number and operations items,

Table 4.6: Summary of Classifications for the 2003 AMC 8 Contest

Item	Impact	Type of DIF	MH D-DIF	NCTM	Gierl	Harnisch	Length	Names
1	F	F	F	G	S, R, M	11	S	B
17	F	F				10, 11	L	B
14		F		N	M	1, 10, 11	S	
8	M	F		N, G, M	S, R, M	11	S	B
25	M			G, M	S, R, M	10, 11	L	
16	M	F		N	S, R		L	B
21	M			G, M	S, R, M	10, 11	S	
23	M			N, G	S, R	10, 11	S	
20	M			G, M	S, R, M		S	
22	M	M		N, G, M	S, R, M	1, 10, 11	S	
12	M	NU			S, M	10	S	
2	M			N	M	1	S	
18	M			G	S	11	L	F
13	M			G	S	10, 11	L	
11	M	NU		N	R	1	L	A
4	M			N	R	10	S	M
19	M	M		N	M	1, 10	S	
9	M			N, G, M	S, R, M	11	S	M
6	M			G, M	S, R, M	11	S	
10	M			N, G, M	S, R, M	11	S	F
5	M			N	R	1	S	
24	M	M		G	S	11	L	
15	M	M	M	G	S	11	S	
7	M	NU	M	N	R	1	L	B
3	M	M	M	N	R	1	S	M

14 geometry items, and 8 measurement items; and only two items were not placed in any category. Using Gierl et al.'s modified taxonomy, there are 16 spatial items, 16 routine items, 13 items which required memorized information, and one item not placed in any of these categories. Eight items had attribute 1, 10 items had attribute 10, 15 items had attribute 11, and one item had none of these attributes. Eight items were classified as long items. Ten items included a name; 2 items had female names, 3 items had male names, and 5 items had both male and female names.

Table 4.7: Summary of Classifications for the 2004 AMC 8 Contest

Item	Impact	Type of DIF	MH D-DIF	NCTM	Gierl	Harnisch	Length	Names
14				G, M	S, R, M	10, 11	S	
13		F				1, 10	L	B
21	M			N	S, R, M	1, 10, 11	S	
10		F			R		S	M
11	M	NU		N	S, M	1, 10	L	
4	M	F			R		S	B
24	M			G, M	S, R, M	10, 11	S	
8	M			N	M		S	
15	M	F		G	S	11	L	
2	M				S, R		S	
25	M			G, M	S, M	10, 11	S	
19	M			N		10	S	
5	M						S	
1	M	M	M	N, M	R	1	S	
18	M			N		10	L	B
12	M			M		10	L	F
17	M						S	
3	M			N, M	R	1	S	M
22	M	M		N		1, 10	S	
7	M			N	R	1	L	
20	M	M		N		1, 10	S	
23	M	M	M	G	S	11	S	F
16	M	M	M	N	R	1	L	
9	M	M	M		M	1	S	
6	M	M	M	N	R	1	S	F

The 2004 AMC 8 results are in Table 4.7. There were 12 items with number and operations content, 5 geometry items, and 6 measurement items. Seven items were not in any of these categories while none were placed in all three. There were 8 spatial items, 11 routine items, 7 items requiring memorization; and 8 items were in none of these categories. Eleven items had attribute 1, 11 items had attribute 10, 6 items had attribute 11, and 6 items had none of the attributes. Seven were long items, and eight items contained a name: 3 female, 2 male, and 3 with both.

Table 4.8: Summary of Classifications for the 2005 AMC 8 Contest

Item	Impact	Type of DIF	MH D-DIF	NCTM	Gierl	Harnisch	Length	Names
14	F	F	F			10	L	
11	F	F	F	N	R	1	L	B
19	F	F	F	G, M	S, R, M	10, 11	S	
15				G, M	S, M	10	S	
25				G, M	S, M	1, 10, 11	L	
23	M			G, M	S, R, M	10, 11	S	
9	M			G, M	S, R, M	11	S	
20	M			N, G	S		L	B
3	M			G	S, M	11	S	
2	M	F		N	R	1	S	M
8	M	F		N	M	1, 10	S	
22	M			N		1, 10	L	
21	M			G	S	11	S	
4	M			G, M	S, R, M		S	
24	M			N			L	
13	M			G, M	S, R, M	10, 11	S	
10	M	M	M	M		1	L	M
1	M			N	R		S	F
12	M			N			S	M
18	M	M		N	M	1	S	
6	M	M	M	N		1	S	
16	M	M	M	N		10	L	
7	M	M	M	G, M	S, R, M	1, 11	S	M
5	M	M	M	N			S	
17	M	M	M	G, M	S, R M	11	S	F

Table 4.8 shows the results for the 2005 AMC 8.There were 12 number and operations items, 12 geometry items, and 10 measurement items. While one item was not placed in any of these categories, no item was placed in all three. Eight items were not placed in any of the Gierl categories, and of the remaining items, 12 were spatial, 9 were routine, and 12 required memorization. There were 19 items with Harnisch attributes, and nine were in each of the three categories. There were eight long items. Eight items contained names: 2 female, 4 male, and 2 with both types.

Table 4.9: Summary of Classifications for the 2006 AMC 8 Contest

Item	Impact	Type of DIF	MH D-DIF	NCTM	Gierl	Harnisch	Length	Names
1	F	F	F	N	R		S	F
14		F			R		S	B
19		F		G, M	S, R, M	10, 11	S	
20		F					S	F
15	M			N, M		10	L	B
24	M			N	S	1, 10, 11	S	
16	M			N		10	S	B
9	M			N		1	S	
2	M						S	M
7	M			N, G, M	R, M	1	S	
11	M			N	M		S	
25	M			N	S, M	1, 10, 11	L	M
5	M	NU		G, M	S, R, M	11	S	
10	M			G, M	S, M	1, 11	S	M
8	M			N	R	1, 11	S	
6	M			G, M	S, R, M	11	S	
22	M	M		N	S, M	10, 11	L	
18	M			N, G, M	S, M	1	L	
21	M			G, M	S, R, M		L	
23	M	M		N			L	
4	M			G	S, R	11	S	F
17	M			N	S, M	1, 11	S	M
13	M	M		M	R, M	10	L	B
12	M	M		N	R	1	S	F
3	M	M	M	M	R		S	F

The 2006 AMC 8 results are summarized in Table 4.9. There are 14 number and operations items, 8 geometry items, 10 measurement items, and 3 do not have any of these types of content. There are 11 spatial items, 12 routine items, 12 memorization items, and 6 items that were not placed in any of these categories. Eight items had none of the three attributes, but there were 9 items with attribute 1, 7 items with attribute 10, and 10 items with attribute 10. Seven items were classified as long. Thirteen items contained names: 5 female only, 4 male only, and 4 with both.

Table 4.10: Summary of Classifications for the 2007 AMC 8 Contest

Item	Impact	Type of DIF	MH D-DIF	NCTM	Gierl	Harnisch	Length	Names
9	F	F	F	G	S	10, 11	S	
18		F		N	R	1	S	
3	M	F		N	R, M	1	S	
5	M				R		L	M
1	M				R, M	1	L	F
25	N			N, G, M	S, M	1, 10, 11	L	
20	M	NU		N		1	L	
24	M			N	R, M	1	L	
21	M				R		S	
13	M			N	S, M	11	S	
2	M			N	S,R, M	1, 11	L	
4	M	F			R		S	M
10	M			N		1	S	
11	M			G	S	10, 11	L	
22	M	M		G, M	S, M	1	L	
23	M	NU		G, M	S, R, M	11	S	
14	M	M		G, M	S, R, M		S	
19	M			N	M	10	S	
16	M			G, M	S, M	11	S	F
8	M			G, M	S, R, M	11	S	
12	M			N, G, M	S, M	1, 11	S	
15	M			N		1, 10	S	
17	M	M	M	N	R	1	S	
7	M	M	M	N	M	1	S	
6	M	M	M	N	R	1	L	

Table 4.10 contains the results for the 2007 AMC 8 contest. There were 14 number and operations items, 9 geometry items, 7 measurement items, and 4 items not placed in any of these categories. All except for three of the items were classified using the Gierl taxonomy: 11 spatial items, 13 routine items, and 14 items requiring memorized material. Aside for 4 items without any of the attributes, there were 14 items with attribute 1, 5 items with attribute 10, and 9 items with attribute 11. Only four items contained names: 2 with male names and 2 with female names.

The second part of the substantive analysis was to count the number of items in each of the categories for all of the classification methods. Table 4.11 shows the number of items in each of seven different ranges of MH D-DIF values. As can be seen, the majority of the items either have small MH D-DIF values or are not significant. For this study, items with MH D-DIF value between -0.5 and 0.5 are considered to have negligible DIF. Twenty-four items have non-negligible DIF and will be used to address the research questions and research hypotheses.

Table 4.11: Distribution of MH D-DIF Values

	2007	2006	2005	2004	2003	Total
$1 \leq x \leq 1.5$	1	0	0	0	0	1
$0.5 < x < 1$	0	1	3	0	1	5
$0 < x < 0.5$	5	6	6	8	4	29
not stat. sig.	8	8	7	7	10	40
$-0.5 < x < 0$	8	8	3	5	7	31
$-1 < x \leq -0.5$	3	1	6	5	2	17
$-1.5 \leq x \leq -1$	0	0	0	0	1	1

Table 4.12 contains the number of items in each category when classified by impact, type of DIF, and MH D-DIF. For the impact classification, a majority of the items have impact favoring males and only seven items have impact favoring females. When classified based on the type of DIF, slightly less than 50% of the items, with some form of DIF, demonstrate uniform or nonuniform DIF with approximately half of the items having uniform DIF favoring males. There are 22 items (37.3%) with uniform DIF favoring females and only 7 items (5.6%) with nonuniform DIF. Over eighty percent of the items have negligible DIF, yet for the remaining twenty-four items, three times as many favor males as compared to females.

During the process of classifying the items according to type of DIF, twenty items were identified as having significant gender differences at four of the five ability levels. For example, the gender differences could be significant at the low, medium-low,

Table 4.12: Statistical Analysis Classification

		2003	2004	2005	2006	2007	Total
Impact	Not significant	1	3	2	3	1	10
	Males	22	22	20	21	23	108
	Females	2	0	3	1	1	7
Type of DIF	Non-uniform	4	1	0	1	1	7
	Males	5	7	7	5	6	30
	Females	5	4	5	4	4	22
MH D-DIF	Negligible DIF	21	20	16	23	21	101
	Males	3	5	6	1	3	18
	Females	1	0	3	1	1	6

medium, and medium-high ability levels but not at the high ability level. Out of these items, only two of them were not significant at the low ability level while the other 18 items were not significant at the high ability level.

The next three classification methods are based on the content of the items. Recall that any particular item could be classified in more than one category. The first method used the NCTM content standards, and the number of items in each content category is contained in Table 4.13. Since the number of items with algebra or data analysis and probability content was fewer than five most years, these two content standards were not included in the subtest analyses. The number of number and operations items included in each contest were quite consistent during the five years while there was more variation in the number of geometry and measurement items.

Table 4.13: NCTM Classifcation

	2003	2004	2005	2006	2007	Total
Number & Operations	14	12	12	14	14	66
Algebra	3	1	3	4	3	14
Geometry	14	5	12	8	9	48
Measurement	8	6	10	10	7	41
Data Analysis & Prob.	2	4	0	2	7	15

Gierl et al. presented a taxonomy of content and cognitive skills which may be associated with gender differences in mathematics, the second classification method, and the results from the classifications of the AMC 8 contest items using this taxonomy are shown in Table 4.14. Items with multiple solutions paths or spatial content may be easier for males while items with a verbal component, involving routine solutions, or requiring recall of memorized information may be easier for females.

Table 4.14: Gierl et al. Classification

	2003	2004	2005	2006	2007	Total
Multiple solution paths	4	2	3	3	1	11
Spatial	16	8	12	11	11	57
Verbal	5	8	5	6	1	25
Routine–Unfamiliar	0	0	1	1	1	3
Routine–Familiar	16	11	9	12	13	61
Memorization	13	7	12	12	14	58

Two additional categories were related to the solutions of items requiring information about traditionally gender-based activities which would be more likely to be familiar to males (e.g. sports) or to females (e.g. intepersonal relationships), hence making it easier for one gender to answer the item. None of the 125 items were classified in either of these two categories. There are very few items each year classified as having multiple solution paths, a verbal component, or a routine solution in an unfamiliar situation, so these categories were not included in the subtest analyses.

The third classification method used a list of 20 attributes developed by Harnisch et al., and the number of items having each attribute are shown in Table 4.15. For example, attribute 7 is "Recalls and interprets knowledge based on definitions, properties, or relations from arithmetic, algebra, and geometry. Performs computions in arithmetic, geometry, signed numbers, absolute values, medians, and modes," and the largest number of items have this attribute.

Most of the attributes have fewer than 30 total items or fewer than 5 items in any given year, and a few attributes have zero items over the five years. For example, two attributes, 19 and 20, are related to trigonometry which is not typically included in the middle school mathematics curriculum, and hence not on the AMC 8 contest. Because of the small numbers, it is difficult to do meaningful subtest analyses. On the other hand, if the number of items is too large, such as for attribute 7, then the results may not be very different from those for the entire test. For these reasons, only attributes 1 (computations), 10 (higher mental processes), and 11 (includes figures, tables, or graphs) were chosen for subtest analyses.

Table 4.15: Harnisch et al. Attributes

	2003	2004	2005	2006	2007	Total
1	8	11	9	9	14	51
2	2	3	1	1	1	8
3	2	6	1	1	7	17
4	8	3	7	6	6	30
5	5	5	6	5	4	25
6	5	3	7	6	3	23
7	14	18	18	22	17	89
8	0	2	1	1	2	6
9	0	0	0	0	0	0
10	10	11	9	7	5	42
11	15	6	9	10	9	49
12	2	1	3	3	4	13
13	1	2	2	0	2	7
14	2	1	2	2	0	7
15	4	6	5	3	3	21
16	1	1	2	0	1	5
17	6	6	3	6	0	21
18	4	9	1	0	0	14
19	0	0	0	0	0	0
20	0	0	0	0	0	0

The items were also classifed by length, the fourth classification, based on the number of words to state the questions, and the results are included in Table 4.16.

Recall that an item was classifed as long if it either included at least 50 words or covered at least 4 lines. Out of the 125 items, less than one third of them were classified as long items, yet the number of long items each year was quite consistent.

Table 4.16: Classification by Length

	2003	2004	2005	2006	2007	Total
Long	8	7	8	7	9	39
Short	17	18	17	18	16	86

The final classification involved the gender of any names included in the stem of the item. After a dicussion with members of the AMC 8 Committee where it was mentioned that they attempt to balance the number of items with male names and with female names, the items were placed into one of three categories: male names, female names, and both types of names. There was one item which contained the name Lou and did not include any pronouns; this name was considered ambiguous and was not counted among the items with names. Considering the items with only one type of name, the numbers are rather balanced for each year except for 2005.

Table 4.17: Gender of Names included in Stem

	2003	2004	2005	2006	2007	Total
Male	3	2	4	4	2	15
Female	2	3	2	5	2	14
Both	5	3	2	4	0	14

Looking at the classifications over five years can show patterns or trends in the types of problems which typically are included on the AMC 8 contests, and they can be used to prepare for future contests. Based on the number of items in particular categories, the classifications can also be used to choose categories for subtest analyses which will then be used to evaluate the research questions and research hypotheses.

4.1.3 Subtest analysis

Two types of subtest analyses were applied to the results of the classifications; one computed impact values for certain content areas and the other considered the distributions of items with negligible and gender DIF. Table 4.18 has the results of the subtest analyses with the NCTM, Gierl, and Harnisch classifications. The impact values are all positive, indicating the impact favors males, and range from 0.018 to 0.056. Only one value is less than 0.020 and two values are more than 0.050 with almost half of the 45 values between 0.030 and 0.039. In comparison, the impact values for the individual items range from −0.039 to 0.125 with almost half of the 125 values between 0.018 and 0.056; a majority of the values are between 0 and 0.050.

Table 4.18: Impact by Classification Method

NCTM	2003	2004	2005	2006	2007
Number & Operations	.047	.045	.048	.036	.049
Geometry	.039	.024	.032	.036	.038
Measurement	.034	.025	.037	.040	.049
Gierl	2003	2004	2005	2006	2007
Spatial	.038	.020	.032	.039	.035
Routine–Familiar	.045	.035	.030	.039	.039
Memorization	.029	.024	.037	.038	.042
Harnisch	2003	2004	2005	2006	2007
1: Computations	.056	.052	.040	.041	.047
10: Mental Processes	.026	.025	.018	.027	.030
11: Figures & Tables	.033	.022	.036	.037	.032

First, consider the values for the NCTM categories. The number and operations values are primarily between 0.040 and 0.049 and the geometry values are primarily between 0.030 and 0.039, with each category having one smaller value, while the measurement values have a slightly larger range. For the Gierl categories, the values for spatial items are primarily between 0.030 and 0.039 with one smaller value and

the values for routine items are primarily between 0.030 and 0.039 with one larger value, while the memorization items have a greater variety of values. The Harnisch attributes values are less consistent. The values for attribute 1 range from 0.040 to 0.056 and for attribute 10 range from 0.018 to 0.030, while the values for attribute 11 are primarily between 0.030 and 0.039 with one smaller value. While there is some variety for each category, the range of all of the values is relatively small since over 90% of the items are between 0.020 and 0.050.

None of the calculations for Table 4.18 account for the confounding variable of ability. To try to control for ability and consider gender differences within a classification scheme, the classification schemes are viewed within the context of MH D-DIF. Tables 4.19–4.23 show the number of items within each classification method that have negligible DIF or gender DIF favoring males or females.

Table 4.19: MH D-DIF by NCTM Standard

Numbers & Operations	2003	2004	2005	2006	2007	Total
Negligible	12	9	8	13	11	53
Male	2	3	3	0	3	11
Female	0	0	1	1	0	2
Geometry	2003	2004	2005	2006	2007	Total
Negligible	12	4	9	8	8	41
Male	1	1	2	0	0	4
Female	1	0	1	0	1	3
Measurement	2003	2004	2005	2006	2007	Total
Negligible	8	5	6	9	7	35
Male	0	1	3	1	0	5
Female	0	0	1	0	0	1

Table 4.19 contains the number of items that demonstrate non-negligible DIF for the three of the NCTM content standards. Of the 66 number and operations items, 80% have negligible DIF while 17% have gender DIF favoring males and 3% have gender DIF favoring females. Of the 48 geometry items, 85% have negligible DIF

with the number of items demonstrating DIF favoring males or females each less than 10%. Approximately 85% of the measurement items have negligible DIF with 12% having gender DIF favoring males and 2% favoring females. For each content area, over 80% of the items have negligible DIF. For the items which do have non-negligible DIF, they are basically balanced between the genders for geometry, but for numbers and operations and measurement, approximately five times as many items have gender DIF favoring males as compared to females.

Table 4.20: MH D-DIF by Gierl et al.'s Modified Taxonomy

Spatial	2003	2004	2005	2006	2007	Total
Negligible	14	7	9	11	10	51
Male	1	1	2	0	0	4
Female	1	0	1	0	1	3
Routine	2003	2004	2005	2006	2007	Total
Negligible	13	8	6	10	11	48
Male	2	3	1	1	2	9
Female	1	0	2	1	0	4
Memorization	2003	2004	2005	2006	2007	Total
Negligible	12	6	9	12	13	52
Male	0	1	2	0	1	4
Female	1	0	1	0	0	2

The DIF distributions for the Gierl categories are in Table 4.20. Of the 58 spatial items, 88% have negligible DIF while items with gender DIF favoring each gender are each less than 7%. For the 61 routine items, almost 80% have negligible DIF while 15% have gender DIF favoring males and 7% have gender DIF favoring females. Among the 58 memorization items, 90% have negligible DIF and items with gender DIF favoring each gender are each less than 7%. For each category, a majority of the items have negligible DIF. For the remaining items, the spatial items are basically balanced between the genders, but the routine and memorization items have about twice as many items with gender DIF favoring males as compared to females.

Table 4.21: MH D-DIF by Harnisch et al.'s Attributes

1: Computations	2003	2004	2005	2006	2007	Total
Negligible	6	7	6	9	11	39
Male	2	4	2	0	3	11
Female	0	0	1	0	0	1
10: Mental Processes	2003	2004	2005	2006	2007	Total
Negligible	10	11	6	7	4	38
Male	0	0	1	0	0	1
Female	0	0	2	0	1	3
11: Diagrams & Tables	2003	2004	2005	2006	2007	Total
Negligible	13	5	6	10	8	42
Male	1	1	2	0	0	4
Female	1	0	1	0	1	3

Table 4.21 contains the MH D-DIF results for three of the Harnisch attributes. Out of the 51 items with attribute 1, 76% have negligible DIF with 22% having gender DIF favoring males and 2% having gender DIF favoring females. For the 42 items with attribute 10, 90% have negligible DIF while items with gender DIF favoring males and females are each less than 7%. There are 49 items with attribute 11, and 86% of them have negligible DIF, 8% have gender DIF favoring males, and 6% have gender DIF favoring females. A majority of the items identified as having each attribute have negligible DIF. Among the items which have non-negligible DIF, they are basically balanced between the genders for attribute 11, three times as many items with attribute 10 favor females as compared to males (the only example of more items favoring females than males), and eleven items favor males as compared to one item favoring females for the items with attribute 1.

Table 4.22 contains the distribution of gender DIF item based on the length of the item. Among the 39 long items, 82% have negligible DIF and no more than 2 items in any year have gender DIF, although across the five years, there are about twice as many items favoring males rather than females. For the remaining short items, a

Table 4.22: MH D-DIF by Length of Stem

Long	2003	2004	2005	2006	2007	Total
Negligible	7	6	4	7	8	32
Male	1	1	2	0	1	5
Female	0	0	2	0	0	2
Short	2003	2004	2005	2006	2007	Total
Negligible	14	14	12	16	13	69
Male	2	4	4	1	2	13
Female	1	0	0	1	1	4

majority of them (80%) have negligible DIF. There was at most 1 item in any year with gender DIF favoring females, and over the five years, there are about three times as many items favoring males as compared to females.

Table 4.23: MH D-DIF by Names

Male names	2003	2004	2005	2006	2007	Total
Negligible	2	2	2	4	2	12
Male	1	0	2	0	0	3
Female	0	0	0	0	0	0
Female names	2003	2004	2005	2006	2007	Total
Negligible	2	1	1	3	2	9
Male	0	2	1	1	0	4
Female	0	0	0	1	0	1
Both types of names	2003	2004	2005	2006	2007	Total
Negligible	3	3	1	4	0	11
Male	1	0	0	0	0	1
Female	1	0	1	0	0	2

Table 4.23 shows the DIF distribution among the items which contain names. For the 15 items with a male name, 80% have negligible DIF, 20% have gender DIF favoring males, and none of the items have gender DIF favoring females. Among the 14 items with female names, 64% have negligible DIF, 29% have gender DIF favoring males, and 7% have gender DIF favoring females. For the 14 items with both types of

names, 79% have negligible DIF, 7% have gender DIF favoring males, and 14% have gender DIF favoring females. None of the categories contain large numbers of items with non-negligible DIF, but in two cases, two categories have more items with gender DIF favoring males rather than females and only for the items with both types of names are there more items with gender DIF favoring females as compared to males, albeit only one more item.

4.2 Analysis to address questions and hypotheses

This section uses earlier descriptive analyses–statistical, substantive, and subtest–summarized in the tables of the previous section to address the research questions and research hypotheses. The first subsection answers the three research questions, and the second subsection evaluates the eight research hypotheses.

4.2.1 Research Questions

The first research question addresses differences in performance based on impact, type of DIF, and DIF, and Table 4.12 contains information about the number of items in each of these three categories. First, consider gender differences using impact, but recall that impact does not account for the confounding factor of ability. Among the 125 items, 108 (86.4%) of them have impact favoring males while only 7 (5.6%) of the items have impact favoring females, with the remaining items having no significant difference in performance. While impact is a measure of gender differences that compares the performances of all of the males to all of the females, the type of DIF is a measure of gender differences that compares their performances at different ability levels. Males could perform better than females at all ability levels, or only at some ability levels, and these differences are obscured within the impact value.

To avoid this possibility, gender differences should also be measured using the type of DIF. Among the 59 items with significant differences in performances for at least 3 ability levels, 30 (50.8%) items have uniform DIF favoring males. This means that whenever there is a significant difference, it is always in favor of males. In comparison, 22 (37.3%) items have uniform DIF favoring females. While this is not close to half, it is greater than the 5.6% of the items with impact favoring females. There are also 7 (11.9%) items which have nonuniform DIF: at some ability levels males perform better while at other ability levels, females perform better. The type of DIF items are more evenly distributed, in terms of gender, than the impact items.

The third measure of gender differences, MH D-DIF, statistically controls for the confounding variable of ability. Using the Mantel-Haenszel procedure, 101 (80.6%) items have negligible DIF and this number is almost as large as the number of items with impact favoring males. As measured in this study, less than 20% of the items demonstrate gender differences, after controlling for ability, which is very different from the impact results. Although there are fewer items with gender differences as measured by MH D-DIF values, there are still 18 items favoring males as compared to 6 items favoring females. Note that using ETS criteria identified only 2 out of the 125 items as having non-negligible DIF.

In terms of impact, a majority of the items favor males and only a small number favor females. The numbers of items are more evenly distributed for the type of DIF, although, again, a majority of the items favor males. Using the Mantel-Haenszel procedure and controlling for ability results in fewer than 20% of the items demonstrating gender DIF. Among these items, three times as many favor males rather than females. Thus, for each measure of gender differences, more items favor males as compared to females, yet the absolute number of items is much smaller for MH D-DIF as compared to impact or type of DIF.

The second research question considers differences in performance based on the content of the items. The items were placed in categories by using five classification methods: NCTM content standards, Gierl et al.'s modified taxonomy, Harnsich et al.'s list of attributes, length of the stem, and the gender of any names included in the stem. Categories within each classification were chosen for two types of subtest analyses to measure gender differences.

First, the gender differences can be measured by using impact values from different categories, as shown in Table 4.18. All of the values are positive which indicates that the impact favors males within each content area. For a particular content area, there is usually little variability since most of the items have values within an interval of length 0.010. There are some differences when comparing content areas. For example, number and operations has larger impact values than geometry or measurement, and measurement has more variability than the other two categories. Among the Gierl items, spatial and routine-familiar items have most of the values between 0.030 and 0.039 while memorization has more variability. For the Harnisch attributes, attribute 1 has the largest impact value, and attribute 10 has the smallest. There is minimal variability among attributes 10 and 11 and more variability with attribute 1.

While there is variability in the content-based impact values, comparing this to the variability among all 125 items, the content area variability seems less distinctive. Table 4.25 compares the ranges and distributions of the impact values over all 125 items and over the items classified by content. The impact for the 125 items ranges from -0.039 to 0.129, as seen in the first three rows of the table, while the impact based on the classification of the items by content ranges from 0.018 to 0.056, as seen in the middle three rows of the table. The last three rows of the table show the distribution of the 125 items between 0.010 and 0.059. The interval between 0 and 0.050 contains 56% of the impact values for the 125 items and 95.6% of the content-

based impact values; that is, there is more variability among the impact values for the 125 items as compared to the variability of the content-based impact values.

Comparing content-based impact values to overall impact values shows that the two sets of results are very consistent. For example, since 86.4% of the 125 items have impact favoring males, it is not surprising that all of the content area impact values favor males. Since these values do not control for ability, to measure gender differences within certain content areas while also controlling for ability, the items from content areas should be viewed within the context of the MH D-DIF values.

The second way to measure gender differences based on the content of the items is by counting the number of items with gender DIF within the various content areas, as shown in Tables 4.19, 4.20, and 4.21. First, note that for each of the content areas, the majority of the items have negligible DIF, ranging from 80.3% to 90.5% of the items within any given content area. For items within the geometry, spatial, memorization, mental processes, and diagrams and tables categories; each gender has at most 4 items with DIF favoring that gender, and in each instance, the two genders differ by at most 2 items. For these content areas, the gender differences are minimal. For number and operations, the gender DIF favors males on 11 items and females on 2 items, and for measurement the ratio is 5 to 1. For routine items, the ratio is 9 to 4, and for spatial items, the ratio is 11 to 1. These four categories indicate some gender differences when considering content area and controlling for ability.

When considering gender differences within particular areas, as compared to individual items, some differences have similar distributions and magnitudes while other gender differences are minimized. The content-based impact values all favor males, yet the magnitudes of the values are smaller than many of the impact values for individual items. When looking at MH D-DIF within a category, a majority of the items have negligible DIF. For five categories, the number of items favoring each gender are

similar; for four categories, a larger number of items favor males rather than females.

The third research question addresses patterns among items with gender DIF. Table 4.26 contains information associated with the 24 items which demonstrate non-negligible DIF, and the items are sorted with increasing values of MH D-DIF. The first 6 items have gender DIF favoring females while the other 18 items have gender DIF favoring males. Except for item 7 on the 2003 AMC 8, the items are consistent with regards to gender in terms of impact, type of DIF, and MH D-DIF; that is, all of the items with female gender DIF also have impact favoring females and uniform DIF favoring females and similarly for the items with male gender DIF.

Since there are only 6 items, it is difficult to find patterns or consistencies among the items with gender DIF favoring females. No category contains more than three items other than the length category, although four categories do contain three items. Two-thirds of the items are short, yet this agrees with the overall distribution of short items. Only three of the items contain a name in the stem, which makes it difficult to draw conclusions, but note that none of the three items had only female names.

Among the 18 items which demonstrate gender DIF favoring males, there is more potential for patterns or consistencies. For the NCTM classification, the number of number and operations item is greater than the sum of the other two categories. For the Gierl et al. classification, a majority of the items are routine and there are only two spatial items. Using the Harnisch et al. attributes, over 75% of the items have attribute 1, and only a few have attribute 11, and one item has attribute 10. Over two-thirds of the items are short which is consistent with the overall distribution of short items. Eight of the items included names in the stem, and half of them only had a female name which is a contrast to the items with female gender DIF.

4.2.2 Research Hypotheses

Of the eight research hypotheses, three are not evaluated using data from this study, due to small numbers of items in those categories. These three hypotheses are related to algebra content, multiple solution paths or shortcuts, and significant verbal content. The five remaining hypotheses address items with geometry content; spatial content; figures, tables, or graphs; involving memorized material; and routine solutions.

The first of these hypotheses states that males will be more likely to answer geometry items correctly. The impact values for the geometry items range from 0.024 to 0.039, and being all positive, favor males. In contrast, out of the 7 geometry items with gender DIF, 4 items favor males while 3 items favor females. Among the 24 items with gender DIF, 50% of the 6 items with gender DIF favoring females are geometry items as compared to only 22% of the items with gender DIF favoring males.

The second hypothesis is that males will be more likely to answer correctly spatial items. The impact values for these items range from 0.020 to 0.039 and all favor males. Out of the 7 spatial items with gender DIF, 4 items favor males, and 3 items favor females. For the items with gender DIF, spatial items reprsent 50% of the items with female gender DIF and 11% of the items with male gender DIF.

The third hypothesis is that males will be more likely to correctly answer items which contain figures, tables, or graphs. The impact values all favor males and range from 0.022 to 0.037. Among the 7 items in this category with non-negligible DIF, 4 items favor males and the remaining items favor females. Considering the 24 items with non-negligible DIF, 50% of the items with gender DIF favoring males and 22% of the items with gender DIF favoring males are items with figures, tables, or graphs.

The fourth hypothesis is that females will be more likely to answer items correctly which require the recall of memorized material. The impact values range from 0.024

to 0.042 and all favor males. Among these items with non-negligible DIF, 4 favor males and 2 favor females. For the 24 items with gender DIF, one-third (33%) of the items favoring females, and two-ninths (22%) of the items favoring males are memorization items.

The fifth hypothesis is that females will be more likely to correctly answer items with routine solutions. The impact values all favor males and range from 0.030 to 0.045. Among these items, 9 favor males and 4 favor females. Out of the items with non-negligible DIF, two-thirds of the items with gender DIF favoring females, and half of the items with gender DIF favoring males require routine solutions.

In each of these five cases, the results are similar. The impact values occur within a small range and all favor males, yet these values do not control for ability. When considering the items within the respective category with gender DIF, there are more items favoring males than females, but when considering the items within all of the items which have gender DIF favoring males or females, the percentage of items is greater for females than for males. In terms of impact and number of items with non-negligible DIF within a category, the numbers favor males, yet none of the numbers are large. In terms of all of the items with non-negligible DIF, the numbers favor females. These results suggest there is minimal support to the hypotheses and that further analyses should be done using the data.

4.3 Tables and figures to display the data

Although described in earlier sections, these tables are included here because they are better displayed in a landscape format. Table 4.24 includes basic descriptive data for the AMC 8 contests. Table 4.25 has the comparisons of the ranges of impact values. Table 4.26 contains the classifications for the 24 items with gender DIF.

Table 4.24: AMC 8 Contest Scores by Gender

	2003		2004		2005		2006		2007	
	Males	Females	Males	Females	Males	Females	Males	Females	Males	Females
N	39607	39338	38255	37390	35278	34709	35733	34053	34486	33367
Percent	50.2%	49.8%	50.6%	49.4%	50.4%	49.6%	51.2%	48.8%	50.8%	49.2%
Mean	11.16	10.11	10.69	9.78	10.73	9.78	10.8	9.9	10.63	9.62
Median	11	10	10	10	11	10	10	10	10	9
Std. Dev.	4.296	3.813	3.831	3.465	4.032	3.643	3.986	3.429	4.073	3.551
d	0.259		0.249		0.247		0.242		0.264	

Table 4.25: Number of Items by Impact

Overall: entire range	$-0.039 - 0$	n.s.	$0 - 0.049$	$0.050 - 0.099$	$0.100 - 0.129$
Number	7	10	70	31	7
Percent	5.6	8	56	24.8	5.6
Classification	$0.010 - 0.019$	$0.020 - 0.029$	$0.030 - 0.039$	$0.040 - 0.049$	$0.050 - 0.059$
Number	1	9	22	11	2
Percent	2	20	49	24	4
Overall: subset	$0.010 - 0.019$	$0.020 - 0.029$	$0.030 - 0.039$	$0.040 - 0.049$	$0.050 - 0.059$
Number	10	18	11	25	14
Percent	12.8	23.1	14.1	32.1	17.9

Table 4.26: Items with Non-negligible MH D-DIF

Year	Item	Impact	Type of DIF	MH D-DIF	NCTM	Gierl	Harnisch	Length	Names
2007	9	F	F	1.046	G	S	10, 11	S	
2006	1	F	F	0.761				S	M
2003	1	F	F	0.755	G	S, R, M	11	S	B
2005	11	F	F	0.632	N	R	1	L	B
2005	14	F	F	0.585			10	L	
2005	19	F	F	0.522	G, M	S, R, M	10, 11	S	
2007	17	M	M	-0.500	N	R	1	S	
2007	7	M	M	-0.521	N	M	1	S	
2005	6	M	M	-0.538	N		1	S	
2004	1	M	M	-0.565	M	R	1	S	
2003	7	M	NU	-0.569	N	R	1	L	B
2004	23	M	M	-0.591	G		11	S	F
2003	15	M	M	-0.601	G		11	S	
2005	10	M	M	-0.630	M		1	L	M
2004	9	M	M	-0.678		M	1	S	
2005	17	M	M	-0.684	G, M	S, M	11	S	F
2006	3	M	M	-0.728	M	R		S	F
2005	5	M	M	-0.787	N			S	
2005	7	M	M	-0.816	G, M	S, R, M	1, 11	S	M
2004	16	M	M	-0.830	N	R	1	L	
2005	16	M	M	-0.897	N		10	L	
2007	6	M	M	-0.915	N	R	1	L	
2004	6	M	M	-0.912	N	R	1	S	F
2003	3	M	M	-1.234	N	R	1	S	M

Chapter 5

Discussion

5.1 Summary of major results

This study explored gender differences on the American Mathematics Competitions AMC 8 contest using statistical, substantive, and subtest analyses. The statistical analyses were based on impact computed from differences in proportion correct, types of DIF as measured by impact within ability levels, and the Mantel-Haenszel procedure for detecting DIF. These analyses were applied to each item on every contest. The substantive analyses involved classifying the items using various methods and counting the number of items in each category. Certain categories were chosen, based on their size, for the subtest analyses. These analyses measured impact and identified the distribution of items with gender DIF within each of the chosen categories.

After doing the statistical analyses, the substantive analyses placed items in one of three categories related to each of the three statistical analyses. The first statistical analysis was to measure the impact of each item by subtracting the female proportion correct from the male proportion correct. Items were placed into one of three categories: no impact, impact favoring males, and impact favoring females. Out of

the 125 items, 10 had no significant impact and 7 had impact favoring females, and the remaining items had impact favoring males.

For the second statistical analysis, the number of ability levels with significant impact values was used to place each item into one of three categories: uniform DIF favoring males, uniform DIF favoring females, or nonuniform DIF. Fifty-nine items had some type of DIF. Seven of the items had nonuniform DIF while the items with uniform DIF consisted of 30 items favoring males and 22 items favoring females.

The third statistical analysis used the Mantel-Haenszel procedure to identify items with DIF by computing an MH D-DIF value, and the items were placed into one of three categories if there was negligible DIF, gender DIF favoring males, or gender DIF favoring females. There were 24 items with gender DIF: 18 favoring males and 6 favoring females.

Looking at the pattern of items classified by impact, type of DIF, and MH D-DIF over the five years of data, certain patterns and consistencies emerge along with a few inconsistencies. First, it should be noted that there were only 2 out of the 125 items which would be identified as having non-negligible DIF according to ETS criteria. In this study these criteria were modified slightly so that there would be more items among which to look for patterns and consistencies in gender differences.

Considering only the magnitudes of the values and not the direction indicated by the signs, items with small impact values often demonstrate uniform DIF or MH D-DIF values favoring females, while items with larger impact values correspond to items favoring males. That is, in terms of impact, the gender differences favoring males are often larger than the gender differences favoring females.

Frequently, items with gender DIF also have impact or uniform DIF favoring the same gender. That is, if an item had gender DIF favoring females, it usually also had impact and uniform DIF favoring females. Sorting the items in increasing values of

impact in Tables 4.1 through 4.5 helped to highlight this pattern.

All of the items with nonuniform DIF also had impact values favoring males, but only one of these seven items had gender DIF and it was gender DIF favoring males. A drawback to using the Mantel-Haenszel procedure to identify items with DIF is that it is not adept at identifying items with nonuniform DIF. This was supported by the results of this study.

One inconsistency, related to Simpson's paradox, is evident among the items with uniform DIF favoring females. Twenty-one items had uniform DIF favoring females, yet eight of these items also had impact favoring males. On these items, as a group, more males than females answered the item correctly, yet within some of the ability levels, more females than males answered the item correctly. These results support the claim that it is important to control for confounding variables such as ability since some gender differences could be obscured using only impact values.

After classifying the items in terms of impact, type of DIF, and MH D-DIF, the items were classified using the NCTM content standards, Gierl et al.'s modified taxonomy, Harnisch et al.'s attributes, length of stem, and gender of names included in the stem of the item. Recall that the categories were not mutually exclusive. The specific distributions of items within the first three classification categories are included in Tables 4.13, 4.14, and 4.15. Some categories had a consistent number of items each year while others had more variability with some years having only five items and other years having fourteen items.

The number of items over the five years was used to choose three categories from each classification scheme for the subtest analyses. For the impact subtest analyses, three categories from each of the NCTM, Gierl, and Harnisch classifications were chosen, while for the MH D-DIF distributions, all five classification methods were used. The impact analyses all favored males while the DIF distributions had mixed

results. The length and name classifications were not used in the impact subtest analyses but were considered in the DIF distributions and when looking for patterns among the items with gender DIF.

5.2 Relationship of results to existing studies

The results of this study support some conclusions found by earlier studies, yet there are some instances when these results contradict existing studies. During the statistical analysis involving type of DIF, some items were identified as having significant gender differences at four of the five ability levels. Among the twenty such items, two were not significant at the low ability level and eighteen were not significant at the high ability level. Thus for 14% of the 125 items, there were significant differences at four ability levels, yet there were no significant differences at the high ability level. This is in contrast to some earlier studies (Benbow and Stanley, 1980, 1983; Hyde et al., 1990) which found significant and meaningful gender differences among the high-performing individuals.

As mentioned earlier when describing the Mantel-Haenszel procedure, there are multiple research studies (Swaminathan and Rogers, 1990; Rogers and Swaminathan, 1993; Narayanam and Swaminathan, 1996) that suggest the Mantel-Haenszel procedure does not function well at identifying nonuniform items with gender DIF. The substantive analysis in this study identified seven items with nonuniform DIF, yet only one of these items was identified by the Mantel-Haenszel procedure and classification criteria used in this study as having gender DIF. These results suggest that another method of identifying items with nonuniform DIF should be used in combination with the Mantel-Haenszel procedure; alternatives are described in the next section.

Many studies find gender differences associated with particular mathematical content areas. For example Hyde et al. (1990) and Linn and Hyde (1989) found that males do better than females on spatial items, while females do better on computational items (Hyde et al., 1990; Linn and Hyde, 1989). Other studies Garner and Engelhard (1999); Engelhard (1990) have found that females do better on algebra items.

In this study, among the items classified by the NCTM content standards: there were too few algebra items for meaningful statistical analyses, the number and operations items often involved computations, and the measurement items often had related characteristics of geometry. The impact values for the number and operations, geometry, and measurement items all favored males with the values ranging from 0.024 to 0.049. In terms of DIF, a majority of each type of item had non-negligible DIF, and the ratio of items with gender DIF favoring males to items favoring females was 11 to 2 for number and operations, 4 to 3 for geometry, and 5 to 1 for measurement.

When comparing these results to earlier research, if the numbers and operations items are considered computational items, the results are very different from the Hyde studies since both impact and DIF measures show results favoring males. While the results for geometry favor males, as agrees with earlier research, the magnitudes of the impact values are small.

Ryan and Chiu (2001) examined differential item functioning within certain mathematics content areas using Harnisch et al.'s classification. They chose nine attributes and formed hypotheses related to which attribute categories would contain items easier for men or easier for women. In particular, they hypothesized that items with figures or graphs present (attribute 11) and items requiring higher order thinking (attributed 10) would be easier for men. Their analyses supported their hypothesis about items with figures or graphs, but the results for the higher order thinking items

were not significant.

As mentioned earlier, the impact by classification scheme values in this study for attributes 10 and 11 favor males, yet the values were small. For the DIF results, out of the forty-two items identified as requiring higher mental processes, one has MH D-DIF values favoring males and three have MH D-DIF values favoring females. There were forty-nine items which contain figures, tables or graphs and four had gender DIF favoring males and three had gender DIF favoring females. These results are similar to those of Ryan and Chiu.

Gierl et al. (2003) used a modified taxonomy from Gallagher et al. (2000) to examine gender differences. The modified taxonomy contained content and cognitive characteristics that research has shown to be associated with gender differences. For example, spatial items are expected to produce gender differences favoring males while familiar, routine items and items requiring memorized material are expected to produce gender differences favoring females. Gierl et al. (2003) found that males perform better than females on spatial items and females performed slightly better than males on items requiring memorized material. They stated that the other characteristics of the taxonomy were either rarely observed among the items or they were not related to gender differences. Gierl et al. claimed that the the taxonomy may not be entirely adequate to understand gender differences in mathematics since they found only one cognitive skill related to gender differences, when controlling for ability.

The content-based impact values for this study were all positive, indicating impact favoring males, with the values ranging from 0.20 to 0.45. For the fifty-sevn spatial items, four had gender DIF favoring males while 3 favored females, and the fifty-eight memorization items had four items favoring males and two items favoring females. For both types of items, a majority of them had no gender DIF and the remaining items had similar numbers favoring males and favoring females. In contrast, out of

the sixty-one routine-familiar items, nine favored males and four favored females. The results from this study differ both from those used to develop the modified taxonomy and from some of Gierl et al.'s results.

Meta-analysis studies compute an effect size for many studies and then use this effect size to compare results between the studies. This effect size d is based on the differences in means found for males and females. These studies can be used to find effect sizes associated with content areas or changes in gender differences based on age, and they often focus on differences. Hyde (2005) suggests that there is support for a gender similarities hypothesis, that males and females are similar on most, but not all, psychological variables. Using the d value for effect size computed in meta-analysis studies, the hypothesis states that most psychological gender differences are in the close-to-zero ($d \leq 0.10$) or small ($0.11 < d < 0.35$) range, few are in the moderate range ($0.36 < d < 0.65$), and very few are large ($d = 0.66 - 1.00$) or very large ($d > 1.00$). Computing d values for each year of AMC 8 data, they all are in the small range which supports the gender similarities hypothesis.

5.3 Limitations of the study

This study has some limitations that influence its results and generalizability. A major limitation is that the Mantel-Haenszel procedure is not the best method to identify items which have nonuniform DIF. Studies have shown (Swaminathan and Rogers, 1990; Rogers and Swaminathan, 1993; Narayanam and Swaminathan, 1996) that while the Mantel-Haenszel procedure is very good at identifying items with uniform DIF, it is less adept at identifying items with nonuniform DIF. As mentioned in the previous section, there is evidence that this occurred within this study.

Another limitation is that the results from this study only apply to eighth graders

in the United States who participate in the AMC 8 contest. Students participating in the contest tend to be above-average students, and they are self-selected by their interest in participating. At some schools, all students participate while at other schools only interested students participate. Sometimes students spend months preparing while other students do little or no preparation at all. Due to these differences in the students' participation and their levels of preparation, the results from this study should not be generalized to a typical eighth grader.

The study also only considers students responses to the items. There is no information about in what ways and to what extent students prepared for the contest or what strategies they used while solving the problems. This study examined gender differences based on the content of the items, but strategy use can be related to gender differences in performance (Gallagher et al., 2000; Gallagher and Lisi, 1994) and there currently is no information available on what strategies students use to solve items on the AMC 8 contest.

Other limitations are related to the categorization of the items and the interpretation of the results. Some items are easy to classify based on their content, and most people familiar with the material would agree with the classification. Other items are more difficult to classify, such as whether or not an item is routine to solve, and the classifications were based on one individual's opinion. It may have been more appropriate to have two or three raters who would discuss differences in classifications. Ryan and Chiu (2001) suggest that assigning an item to more than one category can be a tradeoff. While this may be appropriate because items have many characteristics, they suggest interpreting results becomes more difficult when two categories with significant results contain the same item.

5.4 Implications for future research

There are many directions for future research both related to this set of data and data from other American Mathematics Competition contests. First, other statistical techniques could be applied to this data to identify gender differences in performance. Second, the American Mathematics Competitions include the AMC 10 and AMC 12, in addition to the AMC 8. Results from these contests could also be examined for gender differences. Finally, the focus has been on answering items correctly, yet gender differences among incorrect answers could be explored.

Many statistical methods are available for identifying differential item functioning which can identify both uniform and nonuniform DIF. One such possibility is a modified Mantel-Haenszel procedure (Mazor et al., 1994) which has been shown to identify items with nonuniform DIF better than the procedure used in this study. Penfield (2003) has developed a method that combines the Mantel-Haenszel procedure with a Breslow-Day statistic. Other methods such as logistic regression (Swaminathan and Rogers, 1990) or item response theory techniques (Thissen et al., 1993) could also be used.

For the subtest analyses, aside from the computations of impact within a classification category, the other analyses were more subjective and less statistical. Analysis of covariance could be used to compare gender groups with ability as a covariate. Differential bundle functioning, as described in the literature review, could be used to identify gender differences within particular content areas.

Overall, few gender differences were found on the AMC 8, and those that were found were rather small by ETS standards. The percentages of females participating each year are close to 50%. On the other hand, there are very few females who become members of the United States team which participates in the IMO. This suggests that

something occurs between 8th and 12th grade. Also, meta-analysis studies (Hyde et al., 1990) suggest that gender differences are minor in elementary grades but they are more prevelant in high school. For these two reasons, further research should be done with older students who participate in mathematics competitions. The analyses applied to the AMC 8 contest could also be applied to the AMC 10 and AMC 12 contests.

On the AMC 8 contests, no points are given for incorrect answers, which suggests that the particular wrong answer chosen is insignificant next to the selection of a wrong answer. Marshall (1983) claims that differences in proportion correct can obscure differences in proportions of males or females choosing particular distractors. Marshall (1983), using a scheme developed by Radatz (1979), and Green et al. (1989), analyzing differential distractor functioning, describe methods of examining gender differences in choice of distractors.

5.5 Overall significance of the study

The results of this study have both theoretical and practical significance. As described earlier, there are few studies which examine gender differences in mathematics competitions, so this study addresses a deficiency in the literature. By ETS standards, there were only two items with a significant relationship between gender and choosing a correct answer. These are favorable theoretical results.

The results also have practical significance. The substantive analysis provides a blueprint for the types of problems which typically appear on AMC 8 contents, and the subtest analyses indicate where males or females typically struggle. These results can be used to develop preparation methods to help males and female be successful on future AMC contests.

These practical applications can motivate further study since the answers found regarding gender differences in answering an item correct on the AMC 8 can lead to using other statistical methods for identifying gender differences and to questions about gender differences in differential bundle functioning and differential distractor functioning. These methods and questions can be applied to both the AMC 8 and other AMC contests, for, as Thorstein Veblen states: "The outcome of any serious research can only be to make two questions grow where only one grew before."

Appendix A

Simpson's Paradox

Table A.1 (after (Dorans and Holland, 1993)) illustrates Simpson's Paradox.

Table A.1: Summary of the Performance of Two Hypothetical Groups on an Imaginary Item

Ability	Group A			Group B		
	N	N_c	N_c/N	N	N_c	N_c/N
Low	400	40	.10	1000	200	.20
Medium	1000	500	.50	1000	600	.60
High	1000	900	.90	400	400	1.00
	2400	1400	.60	2400	1200	.50

The first three columns relate to Group A while the last three columns relate to Group B. The first three rows correspond to three different ability levels, and the fourth row is the sum. The symbols, N, N_c, and N_c/N represent the number of people at each ability level, the number who answered the item correctly, and the proportion who answered the item correctly, respectively.

In Group A, 60% of the people answered the item correctly while in Group B, 50% of the people answered the item correctly. Then the impact on this item is $.6 - .5 = .1$ in favor of Group A. Yet, at each ability level, Group B outperformed

Group A on this item by 0.1. The different results from impact and DIF are due to the unequal distributions of ability for the two groups. According to Dorans and Holland, the "imaginary item actually disadvantages Group A, but because Group A is more able than Group B, the overall impact suggests that the item favors Group B" and illustrates the importance of controlling for ability (Dorans and Holland, 1993).

Appendix B

Mantel-Haenszel Procedure

These steps are modified from Rosner (2000).

To assess the association between two dichotomous variables (such as gender and right-wrong response to a test item) after controlling for one or more confounding variables (such as ability), use the following procedure:

Table B.1: Relationship of Gender to Item Response in the ith Stratum.

	Item Score		
	Right	Wrong	Total
Male	a_i	b_i	$a_i + b_i$
Female	c_i	d_i	$c_i + d_i$
	$a_i + c_i$	$b_i + d_i$	n_i

1. Form k strata, based on the level of the confounding variable(s), and construct a 2×2 table relating the two dichotomous variables within each stratum, as shown in Table B.1.

2. Compute the total observed number of units (O) in the $(1,1)$ cell over all strata, where

$$O = \sum_{i=1}^{k} O_i = \sum_{i=1}^{k} a_i$$

3. Compute the total expected number of units (E) in the $(1,1)$ cell over all strata, where

$$E = \sum_{i=1}^{k} E_i = \sum_{i=1}^{k} \frac{(a_i + b - i)(a_i + c_i)}{n_i}$$

4. Compute the variance (V) of O, where

$$V = \sum_{i=1}^{k} V_i = \sum_{i=1}^{k} \frac{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)}{n_i^2(n_i - 1)}$$

5. The test statistic is then given by

$$X^2_{MH} = \frac{(|O - E| - 0.5)^2}{V}$$

6. For a two-sided test with significance level α,

   if $X^2_{MH} > \chi^2_{1,1-\alpha}$ then reject H_0.

   If $X^2_{MH} \leq \chi^2_{1,1-\alpha}$ then accept H_0.

7. The exact p-value for this test is given by $p = Pr(\chi^2_2 > X^2_{MH})$.

8. Use this test only if the variance $V \geq 5$.

9. Which row or column is designated as first is arbitrary. The test statistic X^2_{MH} and the assessment of significance are the same regardless of the order of the rows and columns.

Appendix C

Gierl et al. modified taxonomy

The Modified Gallaher et al. (2000) Taxonomy Outlining the Content and Cognitive Skills Expected to Produce Gender Differences in Mathematics

A. Knowledge and Skills Favoring Males

1. Item Context Favoring Males: Solving the problem requires material more likely to be familiar to males (e.g., items requiring knowledge about traditionally male activitites such as reacing cars or playing football).

2. Shortcults/Multiple solution paths

    a) Multiple solution paths, meaning more than one solution path leads to a correct answer. The uqick solution may be imaginative or insightful (but does not require drawing a picture).

    b) Test-taking skills can contribute to the faster or more accurate solution.

    c) The context looks like a familiar one, but the solution is not one that is generally associated with the context.

3. Spatial

    a) Requires the conversion of a word problem to a spatial representation (i.e., generation of spatial format). Spatial representation is an important part of the problem.

    b) Requries using a given spatial representation (e.g., convert it to a mathematical expression or extract information to be used in solving a problem). Spatial representation is an important part of the problem.

    c) Requires the transformation of information presented in a spatial format to a different spatial format (e.g., a given parabola has to be modified according to some rules). The change has to be produced.

    d) Spatial information must be maintained in "working memory" while other spatial information is being transformed (e.g., maintain a particular shape in working memory so that it can be compared with a transformed shape). Working memory refers to the information we activate and use when solving problems. Working memory can become overloaded, resulting in errors, when there simply are too many pieces of information to keep track of simultaneously. Also, information can be lost from working memory over time.

    e) Multiple solution paths, meaning more than one solution path leads to a correct answer. One or more of the likely solutions involves drawing or using a picture.

B. Knowledge and Skills Favoring Females

1. Item Context Favoring Females: Solving the problem requires material more likely to be familiar to females (e.g., items requiring knowledge about traditinoally female activities such as the cost of family care or interpersonal relationships).

2. Verbal:

    a) Requries the conversion of a owrd problem to an algebraic expression. These items require the conversino only. This category does not include items where a mathematical expression is generated as a step in arriving at a solution to the problem.

    b) Verbal information must be maintained in working memory while additional information is being processes; primarily used for items with heavy verbal load.

    c) Reading math (e.g., using a newly defined function or understanding the properties of an algebraic expression).

3. Application of Routine Mathematical Solutions to New, Unfamiliar Situations

    a) Requires labeling the problem as a specific type of problem solving and/or retrieving a formula or routine that should be known from memory, but is not immediately apparent.

    b) The problem is multi-step and requires accuracy adn a systematic approach. For example, two successive calculations must be done and the second calculation uses information from the first calculation in a new, unfamiliar situation.

4. Application of Routine Mathematical Solutions to Familiar Situations

    a) The context is a familiar one, frequently seen in mathematics course work; the solution path is one that is generally associated with the context.

    b) The problem is multi-step and requires accuracy and a systematic approach. For example, two successive calculations must be done and the second calculation uses information from the first calculation but in a familiar situation.

5. Memorization: Recall of definitions, terms, formulas, and mathematical facts necessary to solve the problem. For example, the item requires the examinee to know the properties of an arithmetic sequence, the eccentricity of a parabola, the radius of a circle, or the properties of conics

6. Symbolic Processes:

    a) Solution requires pure algebraic manipulation or calculation

    b) Questions where two mathematical expressions or quantities must be compared and the values of the two are equal (this type of problem has no verbal element).

Appendix D

Harnisch et al. attributes

Revised Description of Attributes

Adapted from D. Harnisch, K.K. Tatsuoka, and J.L. Wilkins. *Reporting Math Proficiencies Based on New SAT-M Items.* Paper presented at the November 1995 American Evaluation Association Meet, Vancouver, Canada. Included in (citation)

1. Deals with odd and even integers, prime numbers, factors, rational numbers, ordering, ratios, percentages, place value, powers, roots, and averages.

2. Deals with variables (addition and subtraction only), linear equations, linear algebraic expressions, signed numbers, absolute values, and irrational numbers.

3. Deals with higher degree algebraic expressions, functions, sets, simple probability, combinatorics, modes and medians, and exponents with variables.

4. Deals with perimeter, area, and volume for triangles, circles, rectangles, and other geometric objects. In analytic geometry, deals with points and lines in relation to a coordinate system.

5. Translates word problems into arithmetic and algebraic expression(s). Identifies implicit variables and relations. Deals with real-world problems and real-world experiences.

6. Restructures problems into solvable forms. Chooses better, simpler, or quicker strategies to solve problems. Chooses from rules, properties, and theorems the better, simpler, or quickest one to use.

7. Recalls and interprets knowledge based on definitions, properties, or relations from arithmetic, algebra, and geometry. Performs computations in arithmetic, geometry, signed numbers, absolute values, medians, and modes.

8. Applies mathematical rules and properties to solve equations (simultaneous); derives, factors, and computes algebraic expressions.

9. Skill with calculator. Conducts complicated algebraic operations.

10. Applies higher mental processes to solve problems. Sorts problems into implicit component parts and restructures them to make the problem solvable.

11. Works with figures, tables, and graphs.

12. Generates figures or tables for problem solving.

13. Understands the properties of the right triangle.

14. Takes advantage of the form of the test items and other test-taking methods without solving the problem in the manner intended by the item writer. Solves a task by working backward from the multiple-choice options.

15. Works with problems having several steps (explicit or implicit). Establishes subgoals of the problem; orders, prioritizes, and executes the subgoals in a step-by-step fashion.

16. Comprehends sentences with the negation, "at least" comparison, "must be," "could be," and with relations of increasing and decreasing.

17. Keeps track of what a question is asking, paying attention to detail. Identifies constraints. Follows verbally written instructions; reads complex, long sentences.

18. Translates verbal expressions into mathematical expressions where variable term(s), constant(s), and needed operation(s) are readily apparent.

19. Applies the relations between the functions of trigonometric and angles and the functions of trigonometry.

20. Utilizes the graphs to express the function of the trigonometry for problem solving.

Appendix E

Items with Gender DIF

2003 AMC 8

1. Jamie counted the number of edges of a cube, Jimmy counted the number of corners, and Judy counted the number of faces. They then added the three numbers. What was the resulting sum?

(A) 12 **(B)** 16 **(C)** 20 **(D)** 22 **(E)** 26

3. A burger at Ricky C's weighs 120 grams, of which 30 grams are filler. What percent of the burger is not filler?

(A) 60% **(B)** 65% **(C)** 70% **(D)** 75% **(E)** 90%

7. Blake and Jenny each took four 100-point tests. Blake averaged 78 on the four tests. Jenny scored 10 points higher than Blake on the first test, 10 points lower than him on the second test, and 20 points higher on both the third and fourth tests. What is the difference between Jenny's average and Blake's average on these four tests?

(A) 10 **(B)** 15 **(C)** 20 **(D)** 25 **(E)** 40

15. A figure is constructed from unit cubes. Each cube shares at least one face with another cube. What is the minimum number of cubes needed to build a figure with the front and side views shown?

(A) 3 **(B)** 4 **(C)** 5 **(D)** 6 **(E)** 7

2004 AMC 8

1. On a map, a 12-centimeter length represents 72 kilometers. How many kilometers does a 17-centimeter length represent?

(A) 6 **(B)** 102 **(C)** 204 **(D)** 864 **(E)** 1224

6. After Sally takes 20 shots, she has made 55% of her shots. After she takes 5 more shots, she raises her percentage to 56%. How many of the last 5 shots did she make?

(A) 1 **(B)** 2 **(C)** 3 **(D)** 4 **(E)** 5

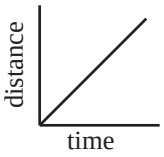
9. The average of the five numbers in a list is 54. The average of the first two numbers is 48. What is the average of the last three numbers?

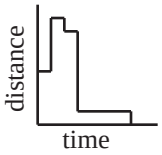
(A) 55 **(B)** 56 **(C)** 57 **(D)** 58 **(E)** 59

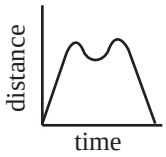
16. Two 600 ml pitchers contain orange juice. One pitcher is $\frac{1}{3}$ full and the other pitcher is $\frac{2}{5}$ full. Water is added to fill each pitcher completely, then both pitchers are poured into one large container. What fraction of the mixture in the large container is orange juice?

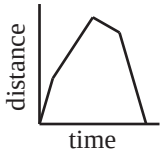
(A) $\frac{1}{8}$ **(B)** $\frac{3}{16}$ **(C)** $\frac{11}{30}$ **(D)** $\frac{11}{19}$ **(E)** $\frac{11}{15}$

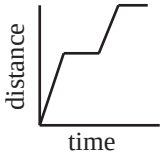
23. Tess runs counterclockwise around rectangular block $JKLM$. She lives at corner J. Which graph could represent her straight-line distance from home?

(A) 

(B) 

(C) 

(D) 

(E) 

2005 AMC 8

5. Soda is sold in packs of 6, 12 and 24 cans. What is the minimum number of packs needed to buy exactly 90 cans of soda?

(A) 4 **(B)** 5 **(C)** 6 **(D)** 8 **(E)** 15

6. Suppose d is a digit. For how many values of d is $2.00d5 > 2.005$?

(A) 0 **(B)** 4 **(C)** 5 **(D)** 6 **(E)** 10

7. Bill walks $\frac{1}{2}$ mile south, then $\frac{3}{4}$ mile east, and finally $\frac{1}{2}$ mile south. How many miles is he, in a direct line, from his starting point?

(A) 1 **(B)** $1\frac{1}{4}$ **(C)** $1\frac{1}{2}$ **(D)** $1\frac{3}{4}$ **(E)** 2

10. Joe had walked half way from home to school when he realized he was late. He ran the rest of the way to school. He ran 3 times as fast as he walked. Joe took 6 minutes to walk half way to school. How many minutes did it take Joe to get from home to school?

(A) 7 **(B)** 7.3 **(C)** 7.7 **(D)** 8 **(E)** 8.3

11. The sales tax rate in Bergville is 6%. During a sale at the Bergville Coat Closet, the price of a coat is discounted 20% from its \$90.00 price. Two clerks, Jack and Jill, calculate the bill independently. Jack rings up \$90.00 and adds 6% sales tax, then subtracts 20% from this total. Jill rings up \$90.00, subtracts 20% of the price, then adds 6% of the discounted price for sales tax. What is Jack's total minus Jill's total?

(A) $-\$1.06$ **(B)** $-\$0.53$ **(C)** \$0 **(D)** \$0.53 **(E)** \$1.06

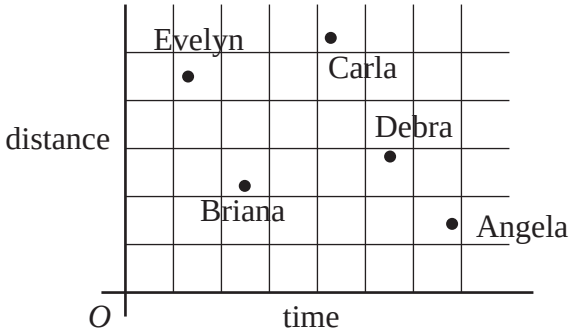
14. The Little Twelve Basketball Conference has two divisions, with six teams in each division. Each team plays each of the other teams in its own division twice and every team in the other division once. How many conference games are scheduled?

(A) 80 **(B)** 96 **(C)** 100 **(D)** 108 **(E)** 192

16. A five-legged Martian has a drawer full of socks, each of which is red, white or blue, and there are at least five socks of each color. The Martian pulls out one sock at a time without looking. How many socks must the Martian remove from the drawer to be certain there will be 5 socks of the same color?

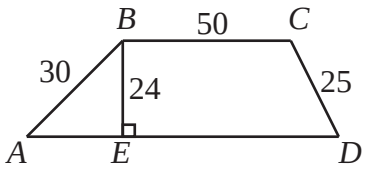
(A) 6 **(B)** 9 **(C)** 12 **(D)** 13 **(E)** 15

17. The results of a cross-country team's training run are graphed below. Which student has the greatest average speed?



(A) Angela **(B)** Briana **(C)** Carla **(D)** Debra **(E)** Evelyn

19. What is the perimeter of trapezoid $ABCD$?



(A) 180 **(B)** 188 **(C)** 196 **(D)** 200 **(E)** 204

2006 AMC 8

1. Mindy made three purchases for \$1.98, \$5.04 and \$9.89. What was her total, to the nearest dollar?

(A) \$10 **(B)** \$15 **(C)** \$16 **(D)** \$17 **(E)** \$18

3. Elisa swims laps in the pool. When she first started, she completed 10 laps in 25 minutes. Now she can finish 12 laps in 24 minutes. By how many minutes has she improved her lap time?

(A) $\frac{1}{2}$ **(B)** $\frac{3}{4}$ **(C)** 1 **(D)** 2 **(E)** 3

2007 AMC 8

6. The average cost of a long-distance call in the USA in 1985 was 41 cents per minute, and the average cost of a long-distance call in the USA in 2005 was 7 cents per minute. Find the approximate percent decrease in the cost per minute of a long-distance call.

(A) 7 **(B)** 17 **(C)** 34 **(D)** 41 **(E)** 80

7. The average age of 5 people in a room is 30 years. An 18-year-old person leaves the room. What is the average age of the four remaining people?

(A) 25 **(B)** 26 **(C)** 29 **(D)** 33 **(E)** 36

9. To complete the grid below, each of the digits 1 through 4 must occur once in each row and once in each column. What number will occupy the lower right-hand square?

1		2	
2	3		
			4

(A) 1 **(B)** 2 **(C)** 3 **(D)** 4 **(E)** cannot be determined

17. A mixture of 30 liters of paint is 25% red tint, 30% yellow tint and 45% water. Five liters of yellow tint are added to the original mixture. What is the percent of yellow tint in the new mixture?

(A) 25 **(B)** 35 **(C)** 40 **(D)** 45 **(E)** 50

Bibliography

AMC 8 instructions. Prepared by the Committee on the American Mathematics Competition.

AMC 8 solutions pamphlet. Prepared by the Committee on the American Mathematics Competition.

American Mathematics Competitions. Retrieved August 27, 2008 from http://www.unl.edu/amc/.

Andreescu, T., Gallian, J. A., Kane, J. M., and Mertz, J. E. (2008). Cross-cultural analysis of students with exceptional talent in mathematical problem solving. *Notices of the AMS*, 55(10):1248–1260.

Becker, B. J. (1990). Item characteristics and gender differences on the SAT-M for mathematically able youths. *American Educational Research Journal*, 27(1):65–87.

Benbow, C. P. and Stanley, J. C. (1980). Sex differences in mathematical ability: Fact or artifact? *Science*, 210(12):1262–1264.

Benbow, C. P. and Stanley, J. C. (1983). Sex differences in mathematical reasoning ability: More facts. *Science*, 222:1029–1031.

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

Campbell, J. R. and Wu, W.-T. (1996). Development of exceptional academic talent: International research studies. *International Journal of Educational Research*, 25(6):479–483.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York.

Creswell, J. W. (2005). *Educational Research (2nd ed.)*. Pearson, Merrill, Prentice, Hall, Upper Saddle River, NJ.

Cronbach, L. J. (1949). *Essentials of Psychological Testing*. Harper & Row, Publishers.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Dar-Nimrod, I. and Heine, S. J. (2006). Exposure to scientific theories affects women's math. *Science*, 314. Supporting online material available,.

Doolittle, A. E. and Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement*, 24(2):157–166.

Dorans, N. J. and Holland, P. W. (1993). Dif detection and description: Mantel-Haenszel and standardization. In Holland, P. W. and Wainer, H., editors, *Differential Item Functioning*. Lawrence Erlbaum Associates, Inc.

Engelhard, Jr., G. (1990). Gender differences in performance on mathematics items: Evidence from the United States and Thailand. *Contemporary Educational Psychology*, 15:13–26.

Ethington, C. A. (1990). Gender differences in mathematics: an international perspective. *Journal for Research in Mathematics Education*, 21(1):74–80.

Fennema, E. and Sherman, J. (1977). Sex-related differences in mathematics achievement, spatial visualization and affective factors. *American Educational Research Journal*, 14(1):51–71.

Fennema, E. and Sherman, J. (1978). Sex-related differences in mathematics achievement and related factors: a further study. *Journal for Research in Mathematics Education*, pages 189–203.

Friedman, L. (1989). Mathematics and the gender gap: A meta-analysis of recent studies on sex differences in mathematical tasks. *Review of Educational Research*, 59(2):185–213.

Gallagher, A. M. and Lisi, R. D. (1994). Gender differneces in Scholastic Aptitude Test–Mathematics problem solving among high-ability students. *Journal of Educational Psychology*, 86(2):204–211.

Gallagher, A. M., Lisi, R. D., Holst, P. C., Lisi, A. V. M.-D., Morely, M., and Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology*, 75:297–334.

Garner, M. and Engelhard, Jr., G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education*, 12(1):29–51.

Garson, G. D. (2008). Reliability analysis. from Statnotes: Topics in Multivariate Analysis. Retrieved August ,27, 2008 from http://www2.chass.ncsu.edu/garson/pa765/statnote.htm.

Gierl, M. J., Bisanz, J., Bisanz, G. L., and Boughton, K. A. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurement*, 40(4):281–306.

Gleason, J. (2008). An evaluation of mathematics competitions using item response theory. *Notices of the AMS*, 55(1):8–15.

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

Hanna, G. (1986). Sex differences in the mathematics achievement of eighth graders in Ontario. *Journal for Research in Mathematics Education*, 17(3):231–237.

Harris, A. M. and Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education*, 6(2):137–151.

Holland, P. W. and Thayer, D. T. (1985). An alternative defintion of the ets delta scale of item difficulty. Research report, Princeton, NJ.

Holland, P. W. and Thayer, D. T. (1988). Differential item performance and the mantel-haenszel procedure. In Wainer, H. and Braun, H. I., editors, *Test Validity*. Erlbaum, Hillsdale, NJ.

Hyde, J. S. (1990). Meta-analysis and the psychology of gender differences. *Signs: Journal of Women in Culture and Society*, 16(1):55–73.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, 60(6):581–592.

Hyde, J. S., Fennema, E., and Lamon, S. J. (1990). Gender differences in mahtematics performance: A meta-analysis. *Psychological Bulletin*, 107(2):139–155.

Hyde, J. S. and Linn, M. C. (2006). Gender similarities in math and science. *Science*, 314:599–600.

Jones, L. V. (1987). The influence on mathematics test scores, by ethnicity and sex, of prior achievement and high school mathematics courses. *Journal for Research in Mathematics Education*, 18(3):180–186.

Leder, G. C., Forgasz, H. J., and Taylor, P. J. (2006). Mathematics, gender, and large scale data: New directions or more of the same? In Novotná, J., Moraová, H., Krává, M., and Stehlíková, N., editors, *Proceedings of the 30th Conference of the International Group for the Psychology of Mathematics Education*, volume 4, pages 33–40. PME.

Lindquist, E. F. (1951). *Educational Measurement*. George Banta Publishing Company.

Linn, M. C. and Hyde, J. S. (1989). Gender, mathematics, and science. *Educational Researcher*, 18(8):17–19, 22–27.

Linn, M. C. and Petersen, A. C. (1985). Emergence and characteriation of sex differences in spatial ability: A meta-analysis. *Child Development*, 56:1479–1498.

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

Marshall, S. P. (1983). Sex differences in mathematical errors: An analysis of distractor choices. *Journal for Research in Mathematics Education*, 14(4):325–336.

Mazor, K. M., Clauser, B. E., and Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement*, 54(2):284–291.

Mehrens, W. A. and Ebel, R. L., editors (1967). *Principles of Educational and Psychological Measurement: A Book of Selected Readings*. Rand McNally & Company.

Mendes-Barnett, S. and Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimesional model approach. *Applied Measurement in Education*, 19(4):289–304.

Narayanam, P. and Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, 20(3):257–274.

Nunnally, J. C. (1970). *Introduction to psychological measurement*. McGraw-Hill Book Company.

Nuttall, R. L., Casey, M. B., and Pezaris, E. (2005). Spatial ability as a mediator of gender differences on mathematics tests. In Gallagher, A. M. and Kaufman, J. C., editors, *Gender Differences in Mathematics*. Cambridge University Press.

Penfield, R. D. (2003). Applying the Breslow-Day test of trend in odds ratio heterogeneity to the analysis of nonuniform DIF. *The Alberta Journal of Educational Research*, 49(3):231–243.

Radatz, H. (1979). Error analysis in mathematics education. *Journal for Research in Mathematics Education*, 10:163–172.

Reyes, L. H. and Stanic, G. M. A. (1988). Race, sex, socioeconomic status, and mathematics. *Journal for Research in Mathematics Education*, 19(1):26–43.

Riley, T. L. and Karnes, F. A. (1998). Mathematics + competitions = a winning formula! *Gifted Child Today Magazine*, 21(4):42–44.

Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement*, 16(4):261–270.

Rogers, H. J. and Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting DIF. *Applied Psychological Measurement*, 17(2):105–116.

Rosner, B. (2000). *Fundamentals of Biostatistics*. Duxbury/Thomson, Pacific Grove, California, fifth edition.

Roussos, L. and Stout, W. (1996). A multidimensionality-based DIF analysis paradigm.

Ryan, K. E. and Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education*, 14(1):73–90.

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

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, 16(3):143–152.

Shealy, R. and Stout, W. (1993). An item response theory model for test bias and differential test functioning. In Holland, P. W. and Wainer, H., editors, *Differential Item Functioning*. Lawrence Erlbaum Associates, Inc.

Simpson, E. H. (1951). Interpretation of interaction contingency tables. *Journal of the Royal Statistical Society, (Series B)*, 13:238–241.

Swaminathan, H. and Rogers, H. J. (1990). Detecting differential item functioning with logistic regression procedures. *Journal of Educational Measurement*, 27(4):361–370.

Thissen, D., Steinberg, L., and Wainer, H. (1993). Detection of differential item function using the parameters of item response theory. In Holland, P. W. and Wainer, H., editors, *Differential Item Functioning*. Lawrence Erlbaum Associates, Inc.

Thorndike, R. L. (1951). Reliability. In Lindquist, E. F., editor, *Educational Measurement*. American Council on Education, Washington.