The Dissertation Committee for Bryan Nankervis certifies that this is the approved version of the following dissertation:

Predicting Sex Differences in Performance on the SAT I Quantitative Section: How Content and Stereotype Threat Affect Achievement

Committee:

_____
Philip Uri Treisman, Supervisor

_____
O.L. Davis

_____
Susan Empson

_____
Anthony Petrosino

_____
Claire Ellen Weinstein

Predicting Sex Differences in Performance on the SAT I Quantitative Section: How Content and Stereotype Threat Affect Achievement

by

Bryan Nankervis, B.A; B.A.; M.S.

Dissertation

Presented to the Faculty of the Graduate School of

the University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2006

Predicting Sex Differences in Performance on the SAT I Quantitative Section: How Content and Stereotype Threat Affect Achievement

Publication No. _____

Bryan Nankervis, Ph.D.

The University of Texas at Austin, 2006

Supervisor: Philip Uri Treisman

Males consistently outperform females on the SAT-I quantitative section by about one third of a standard deviation. Previous research suggests this is due to a complex mix of biological, sociological, and psychological factors. This study examines 12th-grade male and female performance on NAEP items and uses this data to predict performance gaps on the SAT-I quantitative section. Study results suggest that sex differences in performance are due not only to the construction of the test, but also the environment in which the exam is administered. This research has far-reaching implications for the design and administration of standardized mathematics tests, which have historically exhibited large gaps in performance between the sexes. This research has implications in particular for the SAT, which is used for determining admission to many colleges and the awarding of scholarships.

Table of Contents

List of Figures

Chapter 1: Introduction

This study, which was based on analyses of released National Assessment of Educational Progress (NAEP) items, measured the extent to which males and females[1] exhibited differences in performance on certain types of mathematical problems on standardized tests. This study also sought to clarify what mathematical abilities and cognitive processes may have contributed to these performance differences. This clarified understanding was in turn used to predict performance gaps on the SAT I quantitative section. Performance differences between the sexes across content areas, problem types, and various instruments have previously been documented (Gallagher & Kaufman, 2005; Hyde, Fennema, & Lamon, 1990; Paek, 2002; Willingham & Cole, 1997), and in some situations these differences may be due to a lack of opportunity to learn about particular mathematical topics on the part of females. But it has also been demonstrated that stereotype threat can contribute to performance gaps by causing females to falter in the face of complex numerical reasoning tasks or to forego tedious calculations and opt to guess at an answer while their male counterparts are unaffected (Quinn & Spencer, 2001; Spencer, Steele, & Quinn, 1998). Evidence from these studies points to the broad category of "difficult word problems" as potentially increasing the effects of stereotype threat, but these studies have yet to pin down the specific varieties of problems covered by this nebulous phrase or the degree to which content, ability

[1] This study usually uses the terms *males* and *females* for clarity and to address all age groups covered in this study—because at age 17 or 18, students are no longer *boys* and *girls*, but they are also not quite *men* and *women*.

classification (procedural, conceptual, or problem-solving), or difficulty level contribute to overall gaps in performance.

As part of my dissertation, I reviewed existing research literature on possible biological, sociological, and psychological (e.g., stereotype threat and related issues) reasons for differences between the sexes in performance across mathematical content areas and ability classifications. This literature review included investigations into stereotype threat, motivation theory, the effect of cognitive capacity, the effect of the difficulty level of tasks, and some of the different strategies employed by males and females when working mathematics problems found on standardized tests such as the SAT I. This study then used the NAEP Data Explorer (formerly the NAEP Data Tool) (Department of Education, 2005) to examine the different effects that certain varieties of problems have on male and female performance. This NAEP data was used to predict performance gaps between males and females based upon the variety and frequency of particular test items on the SAT I quantitative section (these performance gaps are consistently about one third of a standard deviation). Using the NAEP data, I further demonstrate that the performance gaps are likely an artifact not only of the SAT I quantitative test's construction, but also of the environment of its administration.

In particular, this study found the largest performance gaps between males and females in the measurement and geometry strands. Males also performed significantly better than females on problem-solving items across all content strands and on more difficult items. Further, by using this NAEP analysis as a basis for predicting

differences in performance between males and females, over half of the gender gap is accounted for on ten released SAT I quantitative instruments.

This research has far-reaching implications for how certain mathematics topics are taught and for the design and administration of standardized mathematics tests, which have historically produced large achievement gaps between the sexes. The research further has implications for the use of such exams in determining admission to colleges and the awarding of scholarships.

Chapter 2: Review of the Relevant Literature

This literature review examines gaps in performance between the sexes, possible reasons for divergence in performance (biological and sociological), differences between the sexes' performance in particular mathematical content areas and problem types, differences in use of mathematical problem-solving strategies, and stereotype threat's effect on performance (including the effects of priming stereotype threat, how Dweck's motivation theory modulates the effects of stereotype threat, differences in strategies used by males and females under conditions of stereotype threat, how cognitive capacity and cognitive interference interact with stereotype threat, and how stereotype threat affects performance on problems of varying difficulty levels). The review then examines how the format and context of the testing instrument can affect performance, what differential item functioning can show about gender bias on standardized tests (including related differences in DIF findings for gender subgroups in content areas, for abilities, and at varying complexity levels), and the history of the SAT and its present validity in predicting college success. Finally, the literature review looks at the use of NAEP item analyses to predict performance differences between males and females.

Gaps in performance between the sexes

Since the rise of the mental testing movement in the early 1900s, differences have been documented between the sexes in the performance of mathematics-related tasks. In a review of studies up through the 1950s, it was found that for late elementary and secondary students, females generally outperform males in terms of computation, while

males outperform significantly in tests of math reasoning, induction, and number series completion. Further, this 1950s review of the literature showed that at the college level males significantly outperform females on problems that require "restructuring" or revision of the problem-solving strategy (Anastasi, 1958). "In general, girls surpass boys in those school subjects depending largely upon verbal abilities, memory, and perceptual speed and accuracy. Boys excel in those subjects that call into play numerical reasoning and spatial aptitudes" (Anastasi, 1958, p. 493). These findings were supported in Maccoby and Jacklin's 1974 study, which documented that from about age 11 on, girls become increasingly superior to boys in verbal ability, while boys from around the age of 12 to 13 have better visual-spatial and mathematical ability.

Possible reasons for divergence in performance

The literature shows three main lines of reasoning to explain the differences in male and female performance on mathematical tasks: biological and sociological (covered here) and psychological (covered later in the sections on stereotype threat). While biological arguments have in recent literature largely been discredited, they still exert an influence in contemporary discourse and thus cannot be ignored.

Biological

In terms of sex differences in performance on cognitive tests, arguments have been made that males for biological reasons have a higher aptitude for mathematics than females, which would account for performance gaps on tests such as the SAT I quantitative section. One claim is that females are more emotional and males are more

rational (Baron-Cohen, 2003). This contention is based on an experiment in which infants (defined as neonates, meaning "yet to be influenced by social and cultural factors") were observed as they were given the option of two things to look at: a mechanical object and a human being. Boys preferred the inanimate object, while girls more often chose to gaze upon the living being (Connellan, Baron-Cohen, Wheelwright, Batki, & Ahluwalia, 2000). This experiment's findings, however, have not been replicated, and they contradict accepted and replicated results from previous similar experiments where no differences were found in attention to people or objects (e.g., Maccoby & Jacklin, 1974).

Hormones may play a role in enhancing some abilities, as levels of testosterone in males have been shown to be causally linked to spatial-skills performance (Janowsky, Oviatt, & Orwoll, 1994), and female-to-male transsexuals have demonstrated marked increases in their spatial abilities after being given large amounts of testosterone (Van Goozen, Cohen-Kettenis, Gooren, Frijda, & Van De Pol, 1995).

Another argument is that males have a greater variability in mathematical performance than females; this variability leads to more males at the far right of the statistical distribution (e.g., at high levels of performance) than females. Studies conducted in the late 1970s and early 1980s with nearly 50,000 junior high students who had taken the SAT I quantitative section early as part of talent searches, revealed that boys exhibited a "substantial" sex difference in mathematical reasoning, and further suggested, because the participants were adolescents, that genetic differences (e.g., deficiencies in female genetics) were the reason for this difference (Benbow &

Stanley, 1980; 1983). Feingold (1992) found that 12th-grade males also exhibited more variable performance than females on the SAT I quantitative section, and suggested that this greater variability, combined with the medium effect size of central tendencies favoring males (moderate differences in average scores between the sexes), could lead to even greater effect sizes in the right tails of the ability distributions for the sexes (larger differences in average scores for high-ability males and females).

However, this domination by males at the far right of the ability distribution (e.g., at the high end of ability in mathematics performance) has diminished over the last two decades in the U.S. and internationally (Feingold, 1994; Monastersky, 2005; Willingham & Cole, 1997), which suggests that sociocultural factors may play a role in gender differences in achievement. In general, evidence for a purely biological basis for differences in mathematical performance between the sexes has been described as "weak, at best" (Wilder, 1997, p. 14).

Sociological

Another argument is that sex-related differences in mathematics performance are due mainly to environmental influences. In this view, differential socialization of boys and girls via prejudicial treatment, social norms, and the expectations of parents, teachers, and fellow students, results in the divergent development of boys and girls and sex-role stereotyping (Baker & Jones, D. P., 1992; Eccles & Jacobs, 1986; Fennema & Peterson, 1985).

For example, in the area of spatial abilities (where males have a documented edge), it has been shown that taking part in related activities (especially those specific

to completing spatial tasks) is essential to the development of the spatial ability, and that females tend to participate in these learning situations less often than males (Baenninger & Newcombe, 1989). Further, it has been demonstrated that when females are given the opportunity to train with visual-spatial tasks, their performance improves (Vasta, Knott, & Gaze, 1996). In addition, international studies have shown that sex differences in performance on mathematics tasks decrease as females are provided more access to advanced training and better jobs (Baker & Jones, D. P., 1992).

Studies have shown that females generally earn better grades than males in required mathematics courses at the secondary level (Gallagher & Kaufman, 2005; Xie & Shauman, 2003), while at the college level, mathematics grades for the sexes are typically the same (Bridgeman & Wendler, 1991), and women enroll as mathematics majors in about the same numbers as men (Chipman, 2005). L. V. Jones's 1987 study of high school seniors' performance on mathematics tests found, however, that achievement was highly dependent on previous coursework completed, and a 2004 Achieve study found that high school geometry courses, which include measurement topics, are necessary for graduation in only thirteen states, while Algebra II is required in only four states. Further, the gender gap in performance on the SAT I quantitative section is significantly smaller in the states that require geometry, and even smaller in those that require Algebra II (College Board, 2006).

So, if subjects stereotypically associated with the male domain (such as mathematics) are not required, then it may be that many females are simply not exposed to the material, which highlights the possible importance of providing an *opportunity to*

learn. Therefore, the evidence suggests that sex differences in mathematics performance are not immutable. Instead, rather than a dichotomy of nature (biological) or nurture (sociological) as possible reasons for divergence in performance, it would seem that differences in performance could be a result of a unique and complex blending of influences for each individual.

Differences between the sexes' performance in particular mathematical content areas and problem types

Sixth- and seventh-grade girls perform better in problems of number sense, problems of estimation, and those that involve patterns, while their male counterparts perform better in geometry and ratio/proportion, and on problems that employ figures (Lane, Wang, & Magone, 1996). These differences in performance continue through high school—for example, studies of students taking college entrance exams (e.g., the SAT and ACT) revealed that females perform better on algebra items that involve familiar algorithms or computation, while males perform better in geometry, mathematical reasoning, word problems, and items including figures, graphs, or tables (Doolittle & Cleary, 1987; Harris & Carlton, 1993).

In a review of studies up to 1985, the observed trend was that among high school students, males were somewhat better than females in solving word problems, and females were better or at least equal in computational skills (Stage, Kreinberg, Eccles, & Becker, 1985). A 1990 meta-analysis of 100 studies revealed "a slight female superiority in computation, no gender difference in understanding of concepts, and a slight male superiority in problem solving" (Hyde, Fennema, & Lamon, p. 147).

In general, and in spite of relatively similar male and female mathematical abilities at the elementary level, males' ability in geometry seems to accelerate past and beyond that of females by the end of high school (Leahy & Guo, 2001). Differences in the ability of spatial visualization may be one key to differences in mathematical trajectories for the sexes. Males have been shown to outperform females on geometric tasks involving spatial visualization, but males and females exhibit roughly equal performance in logical reasoning ability and in the use of geometric problem-solving strategies (Battista, 1990). Males exceed females by some of the largest differences in performance on items that involve mental rotation of three-dimensional objects (Casey, Nuttal, Pezaris, & Benbow, 1995; Willingham & Cole, 1997), and males also perform better than females on items based on measurement of two-dimensional and three-dimensional objects (perimeter, area, surface area, volume), which also involve spatial visualization (Garner & Engelhard, 1999; Li, Cohen, & Ibarra, 2004).

Higher ability in spatial visualization may also play a role in a higher ability to retrieve mathematics facts. College males (undergraduates from the 1996–97 academic year) have been shown to be faster at math-fact retrieval on an achievement test (the Computer-based Academic Assessment System, now known as the Cognitive Aptitude Assessment System) (Royer, Tronsky, Chan, Jackson, & Marchant, 1999). It has been suggested that this higher retrieval speed is due to males having greater flexibility in their choice of problem-solving strategies, because they have the option of employing a spatial approach that might be more appropriate for some items. On the same items,

someone with limited spatial ability might be forced to rely upon a perhaps less effective approach (Casey et al., 1995).

Another possibility to explain the difference in male and female mathematics performance is that math-fact retrieval ability itself plays a pivotal role in solving more complex problems that require more cognitive load. That is, those who can retrieve information more quickly or more effectively would perform better on items requiring recall of a mixture of concepts and/or procedures (Paek, 2002; Royer et al., 1999). Further, it has been demonstrated that on SAT I quantitative items, males take less time to solve problems than females (Paek, 2002). But, regardless of whether math-fact retrieval ability is an effect that favors males (due to higher ability in spatial visualization) or is a cause of their better performance on certain mathematics tasks, it certainly would provide an advantage on a timed test.

Differences in use of mathematical problem-solving strategies

At the elementary level, girls have been found to prefer more algorithmic strategies, while boys prefer more abstract strategies (Fennema, Carpenter, Jacobs, Franke, & Levi, 1998). A study of high school students working problems from the quantitative section of the SAT I (Gallagher, 1992) found that females are more apt to rely on standard algorithms traditionally presented in classrooms, while males are more inclined to use insight. Females demonstrate less prior knowledge than males and use fewer mathematical strategies on SAT mathematics items, even when both groups have similar backgrounds in terms of mathematics courses taken and grades received (Byrnes & Takahira, 1993; Paek, 2002). It has also been shown that on SAT

mathematics items, males use unconventional strategies (including logic, estimation, or insight) significantly more often than females, and females use conventional strategies (such as an algorithm, assigning values to variables, or "plugging in" numbers to a formula) significantly more often than males (Gallagher & De Lisi, 1994; Paek, 2002).

These findings would certainly tend to suggest that males would have the advantage in situations where problems cannot be solved by the more traditional strategies presented in school, and would help to explain why females generally have better grades in school but underperform on tests such as the SAT I quantitative section. One study that used course grades rather than standardized tests to measure differences in mathematics achievement between the sexes noted that "when differences are found, they almost always favor girls, and these differences are quite consistent across samples of varying selectivity for junior high through university mathematics courses" (Kimball, 1989, p. 199). This study's findings, which are contrary to the findings from the SAT mathematics test analyses, suggest that situational variables (such as the testing environment or method of administration) might play a greater role in performance gaps than genetic differences between males and females.

Stereotype threat's effect on performance

A landmark study that introduced the term *stereotype threat* suggests a plausible explanation for some of the differences in male and female mathematics achievement when measured by classroom test grades and by standardized tests. Stereotype threat is defined in this study as a feeling of "being at risk of confirming, as self-characteristic, a

negative stereotype about one's group." Further, stereotype threat "may interfere with the intellectual functioning of these students [affected by stereotype threat], particularly during standardized tests" (Steele & Aronson, 1995, p. 797). This initial study focused on black students who were found to be burdened by the stereotype of having less ability than white students in terms of general intellectual aptitude.

In the study, the researchers, working with a group of black and white students, explicitly "primed" the stereotype for one subgroup of students (the experimental group) before the students took the SAT verbal test. The researchers did not prime the stereotype in a control group. Specifically, before the experimental group took the SAT verbal test, the researchers primed the group by telling them about historical diagnostic differences—that is, that historically blacks had scored worse on this test than whites, and that the test was diagnostic of ability. As a result of this priming, the blacks in the "ability-diagnostic" group underperformed (as measured by their previous SAT scores) in relation to whites in the ability-diagnostic group. But the blacks in the control or "nondiagnostic" group—in which the students saw the test as not diagnostic of their intellectual ability and the students were not told that blacks had historically performed less well—did not underperform as compared to the white students in the group (Steele & Aronson, 1995). Steele and Aronson document in their article that the test scores of students in the experiments were statistically adjusted for differences in the students' previous SAT scores and that there were still differences between black and white subgroup scores in the absence of stereotype threat, but that these differences were similar to the differences found in the students' prior SAT performance.

What is significant about Steele and Aronson's study is that under stereotype threat conditions, these differences in black and white student performance grew. It is not Steele and Aronson's contention, then, that stereotype threat is solely responsible for performance gaps between black and white students, but rather that stereotype threat can contribute to these gaps.

The findings of Steele and Aronson's 1995 study are relevant to this study of performance gaps between the sexes on standardized mathematics tests, because females may also deal with stereotype threat. In particular, females are susceptible to the stereotype threat that they are publicly perceived to be less able in mathematics than males. Steele and Aronson's 1995 study itself notes the broad applicability of stereotype threat: "This threat can befall anyone with a group identity about which some negative stereotype exists, and for the person to be threatened in this way, he [or she] need not even believe this stereotype" (Steele & Aronson, 1995, p. 798).

In later experiments by a variety of researchers, the effects of stereotype threat on female mathematics performance was confirmed among college females who excelled at math and identified strongly with the subject (Spencer et al., 1998). In this study, on an extremely difficult test (composed from the advanced GRE [Graduate Record Examinations] in mathematics), females underperformed compared to males when informed before they took the test of historic sex differences in test performance, but females achieved as well as males when told that the test was gender-insensitive. The results of this experiment contradicted the hypothesis of female genetic deficiency suggested previously by Benbow and Stanley in the early 1980s (Spencer et al., 1998).

Additionally, this experiment's findings confirmed what Steele had concluded in an earlier study: that "stereotype threat may be a possible source of bias in standardized tests, a bias that arises not from item content but from group differences in the threat that societal stereotypes attach to test performance" (Steele, 1997, p. 622).

A test item can be linguistically biased if it uses terms that are unfamiliar to some subgroups, while the overall content of an item might be biased if it refers to situations that are less familiar for certain subgroups (Hambleton & Rodgers, 1995). While bias in item content has been confirmed and addressed by test developers over the years, the effects of stereotype threat are situational (dependent on the testing environment) and cannot be remedied by simply changing the language or premise of specific test items.

One study (Shih, Pittinsky, & Ambady, 1999) examined the effects of racial and gender stereotype threat combined. While Asian-American women are stereotyped positively in mathematics due to their ethnicity, they are stereotyped negatively because of their sex. On a standardized mathematics test (SAT I quantitative section) given to Asian-American women, those whose *ethnicity* was primed performed better, and those whose *gender* was primed performed worse, than a control group for whom neither stereotype was primed (Shih et al., 1999).

This boost in achievement due to a positive stereotype was also observed in previous studies for white students when race was primed (Steele & Aronson, 1995) and for males when gender was primed (Spencer et al., 1998). This potential positive effect of certain stereotyping is supported by social cognitive theory: an earlier (1989) analysis of "human agency in social cognitive theory" noted that "Indeed, people who

believe strongly in their problem-solving capabilities remain highly efficient in their analytic thinking in complex decision making situations, whereas those who are plagued by self-doubts are erratic in their analytic thinking," and "when faced with difficulties, people who are beset by self-doubts abort their attempts prematurely and quickly settle for mediocre solutions, whereas those who have a strong belief in their capabilities exert greater effort to master the challenge" (Bandura, 1989, p. 1176).

This boost in performance for test-takers who are not part of the negatively stereotyped subgroup has been referred to as *stereotype lift.* A meta-analysis of 43 studies found that merely representing tests as diagnostic of ability was enough to induce stereotype lift in non-negatively stereotyped subgroups. Further, the average effect size of stereotype lift (d =.24) was half that of stereotype threat (d =.48) in these studies (Walton & Cohen, 2002). Accordingly, the magnitude of gender performance gaps on standardized tests may be twofold: the sum of the differences resulting from the underachievement by females and overachievement by males.

The effects of priming stereotype threat

Stereotype threat can be primed in various ways. One is the condition of *evaluative scrutiny.* In this situation test-takers know their results will be available to others, such as parents, teachers, administrators, or colleges. On standardized tests such as the SAT or ACT, some degree of evaluative scrutiny is always present. Another way to induce stereotype threat is the condition of *identity salience.* This condition can be thought of as "the likelihood that the identity will be invoked in diverse situations" (Hogg, Terry, & White, 1995, p. 257). Participating in a mixed testing environment (for

example, males and females) or having to identify one's sex prior to a standardized test, as is the norm, can invoke identity salience.

The effects of these two causes of identity salience have been measured in studies conducted by the Educational Testing Service, commonly known as ETS, and the College Board. In one experiment, researchers administered the mathematics section of the GRE general test to males and females on an individual basis (GRE Board, 1999). In this study, while all students identified their sex prior to the test, the gap in scores between males and females was less than half of the gap from the regular administration of the GRE general test (d =.40 versus d =.97) that same year, in which students tested in a mixed environment.

Another study measured the effect on performance of identifying one's sex before a test (College Board, 1998a). The experiment focused on students taking the Advanced Placement Calculus AB exam, and all students took the test in a mixed environment of males and females. Yet for those who indicated their sex on a standard background information sheet *before* the test, the performance gap effect size (d = .41) between males and females was more than triple that for the males and females who identified their sex *after* the test (d = .12). While the difference in average scores between males and females for those who did not indicate their sex until after the test ($\bar{x}_m$ = 38.08 vs. $\bar{x}_f$ = 35.63) was not significant (p = .097), the performance gap for those who did identify their sex prior to the exam ($\bar{x}_m$ = 40.44 vs. $\bar{x}_f$ = 32.20) was significant (p < .0001). This study, along with the ETS study (GRE Board, 1999), clearly demonstrates

that the condition of identity salience, while subtle in nature, can play a major role in inducing stereotype threat.

Stereotype threat can also be primed more overtly by directly telling students before a test that differences in achievement between subgroups have been observed on previous administrations of the same test. Studies by Spencer et al. (1998) looked at differences in performance between male and female students on items from the advanced GRE in mathematics under just such overt priming conditions. While all students took the test in a mixed environment (triggering identity salience as well), those in the experimental group were explicitly primed for stereotype threat. The performance gap between males and females for the experimental group of students was about one third larger than for the control group of students, who experienced only the condition of identity salience (Spencer et al., 1998). It would seem from these results that overt priming of stereotype threat adds stress to the anxiety brought on by identity salience.

In this same group of studies (Spencer et al., 1998), the researchers sought to find out what would happen when they told an experimental mixed group of students in advance that the test they were about to take was insensitive to sex differences. As with the previous study, all the students were subject to identity salience from the mixed testing environment, but those who were *unprimed* (experimental group) exhibited no differences in achievement between males and females, while in the control group (students who were told nothing prior to the test) males outperformed females. In a more recent study, McGlone and Aronson (2006) found that priming females in a

mixed testing situation to consider positive personal attributes (such as their status as students at a select school) resulted in improved scores on the Vandenberg (Vandenberg & Kuse, 1978) mental rotation test. Seemingly, these types of interventions (e.g., telling students that the test is insensitive to sex differences; priming females to consider positive personal attributes before the test) help to ease the condition of identity salience, thereby lessening the effects of stereotype threat.

How Dweck's motivation theory modulates the effects of stereotype threat

According to Dweck (1999), a person's view of his or her intelligence can be characterized in one of two ways: those who think of their intelligence as fixed are known as *entity* theorists, while those who see their intelligence as malleable are referred to as *incremental* theorists. These types of implicit theories (how people think about their intelligence) influence people's achievement goals (Dweck, 1999; Dweck & Elliott, 1983; Elliott & Dweck, 1988). Examples of achievement goals in education include

- mastery goals—where an individual desires to master course material;
- performance-approach goals—where an individual is concerned solely with obtaining a good grade; and
- performance-avoidance goals—where an individual simply wants to avoid doing poorly in a course.

Incremental theorists usually adopt mastery goals and are concerned with improving or developing their competence, while *entity* theorists generally adopt

performance approach goals and are focused more on proving or demonstrating their competence (Ames & Archer, 1988; Dweck, 1986).

Compared to those who are performance-oriented, those who are mastery-oriented generally persist longer on tasks, intensify their efforts under adverse conditions rather than give up (Elliott & Dweck, 1988), and less often develop learned helplessness deficits (Stipek & Kowalski, 1989). While mastery-oriented people typically embrace challenges and use setbacks as feedback toward future success, those who are performance-oriented shy away from challenges and experience increased anxiety and self-doubt in the face of setbacks. As a result, in problem-solving situations, performance-oriented people (also known in Dweck's parlance as helpless-oriented people) may become less able to strategize, and resort to simply guessing at answers (Dweck, 1999).

Therefore, it seems that since entity theorists are more prone to "helplessness deficits" and increased stress or anxiety, which can lead to an inability to strategize, then those who hold an incremental view of their own intelligence would generally be less susceptible than entity theorists to the effects of stereotype threat in problem-solving situations. These implicit theories about one's intelligence may play a role in the gender variety of stereotype threat, as one study of high-ability middle school students found that girls were more likely than boys to be entity theorists (Leggett, 1985). This finding may illuminate why another study found that stereotype threat lowered females' mathematical achievement by hindering their ability to set up problem-solving strategies (Quinn & Spencer, 2001).

Differences in strategies used by males and females under conditions of stereotype threat

A study by Quinn and Spencer (2001) attempted to replicate the Gallagher and De Lisi (1994) experiment, which looked at the kinds of strategies employed by males and females when solving SAT I mathematics word problems. As noted earlier, Gallagher and De Lisi found that females relied more on conventional strategies, while males relied more on unconventional strategies. Quinn and Spencer (2001) further found that a significant number of females were unable to come up with any problem-solving strategy at all.

In the second part of Quinn and Spencer's study, when stereotype threat was primed by evaluative scrutiny and identity salience, females underperformed compared to males on a test of all mathematical word problems, but not on a test of the same problems reduced to their numerical equivalents (that is, no words, just equations). Quinn and Spencer concluded that stereotype threat lowered women's mathematics achievement by hindering their capacity to generate and use problem-solving strategies (2001). They suggested that women had a diminished *cognitive capacity* as a result of having to deal with the effects of stereotype threat. On the surface, the results of this part of Quinn and Spencer's study might seem contradictory, since women, due to their documented edge in verbal skills, would seem to be favored in the domain of word problems. In this part of the study, however, the reason for women's underperformance seemed to lie in difficulty with the processes and strategies of converting word problems into equations (Quinn & Spencer, 2001).

How cognitive capacity and cognitive interference interact with stereotype threat

The human mind has been described as a "limited-capacity information processor" (Weinstein, 2005), and the attempt to control one's mental state can be broken down into the two processes of *operating* and *monitoring* (Wegner, 1994). Wegner notes that the *operating* process "promotes the intended change by searching for mental contents consistent with the intended state" of mind, while the *monitoring* process "tests whether the operating process is needed, by searching for mental contents inconsistent with the intended state" of mind. Both processes working together provide the individual's mental control. Further, the operating process requires greater cognitive capacity than the monitoring process. Since these processes share the same mental space, at times—for example, when the mind is working on difficult or complex tasks—they must compete for limited mental capacity. Under conditions that reduce capacity, "the monitoring process may supersede the operating process and thus enhance the person's sensitivity to mental contents that are the ironic opposite of those that are intended." The operating process is more susceptible to distractions—either in the form of stress or from multitasking. "Anything that distracts the person's attention from the task of mental control will undermine the operating process and so enhance the effect of the monitoring process" (Wegner, 1994, p. 40).

Arousal plays a major role in many learning theories and is affected by levels of anxiety and stress. The Yerkes-Dodson Law, which describes the psychological effects of physiological arousal on performance, says that with appropriate levels of arousal, performance increases, but that when levels of arousal are too low or too high,

performance decreases. The law (it is referred to as a law because the results of Yerkes and Dodson's experiments have been replicated numerous times since the original study) suggests that an inverted U shape describes the relationship between arousal and performance (Yerkes & Dodson, 1908).

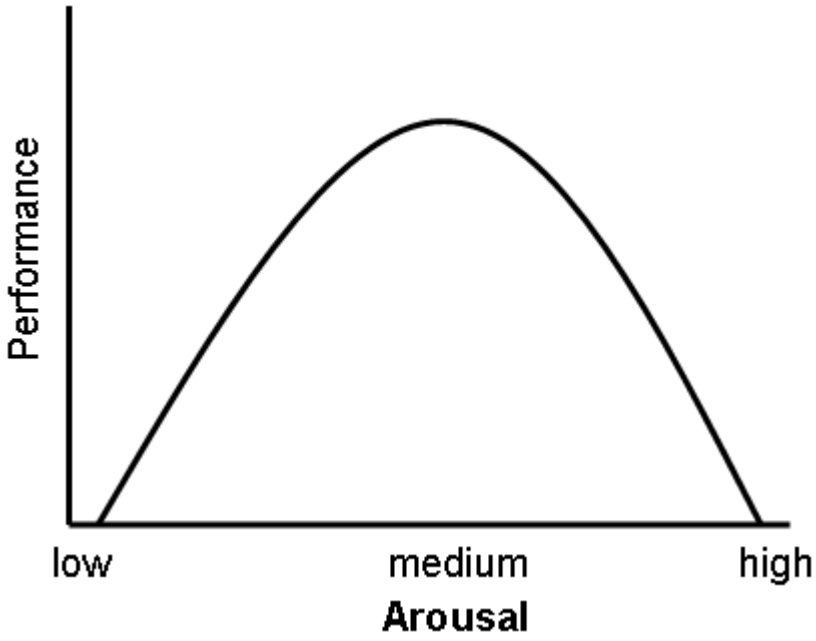

Figure 1. Graph of Yerkes-Dodson Law principle

As can be seen in the image, at low levels of arousal, the subject (in our case, a learner) is simply not motivated enough to perform well, while moderate levels of arousal are conducive to optimal functioning. Overarousal, however, caused, for example, by anxiety or stress, can negatively affect performance physiologically, cognitively, and psychologically.

Overarousal can disrupt physiological states when the body reacts to stress by releasing hormones from the adrenal medulla (adrenaline) and the adrenal cortex (cortisol). This hormone release in turn increases heart rate, respiration, and blood pressure, which leads to hyperactivity (Selye, 1956). One study of African Americans under stereotype threat conditions revealed just such increases in blood pressure, which were accompanied by underperformance (compared to a control group with little or no stereotype threat) on verbal test items (Blascovich, Spencer, Quinn, & Steele, 2001).

In addition to these *physiological* symptoms, *cognitive activity* can be disrupted via forgetfulness, confusion, and an inability to concentrate as a result of overarousal (Broadbent, Cooper, FitzGerald, & Parkes, 1982). Overarousal can also disrupt *emotional states* through anxiety. It has been shown that overanxious persons entertain more task-irrelevant thoughts than nonanxious persons, and that these thoughts are frequently focused on negative personal characteristics, most notably during ability assessments such as achievement tests (Wigfield & Eccles, 1989). Further, anxiety has been shown to be partially responsible for gender differences in achievement, causing females to underperform (Osborne, 2001).

In particular, overarousal in the form of anxiety as a result of stereotype threat conditions has been shown to accompany the underperformance of females on standardized mathematics tests (Harder, 1999; Spencer et al., 1998). Further, it has been demonstrated that anxiety and evaluation apprehension are significantly related to females' performance on such tests (Spencer et al., 1998). Bandura states in a 1989 article on human agency that "threat is a relational property concerning the match

between perceived coping capabilities and potentially aversive aspects of the environment," and "those who believe they cannot manage potential threats experience higher levels of stress and anxiety arousal" (p. 1177).

Different levels of arousal in a testing situation—for example, a situation with males and females under conditions of stereotype threat—would suggest differences in the application of the Yerkes-Dodson Law to the male and the female subgroup. If we assume moderate levels of arousal—levels conducive to optimal performance—for all participants coming in to a mathematics test, and add to the females' arousal levels as a result of stereotype threat, we should expect their performance to fall off, while the males' performance remains in the optimal range.

A recent study supports the notion that stereotype threat triggers increased arousal in females, thereby causing them to exceed optimum arousal levels and lowering their performance in mathematics (Ben-Zeev, Fein, & Inzlicht, 2005). This study's results showed a slight increase in mathematics performance for those females who were given the opportunity to attribute their overarousal under stereotype threat conditions to a benign source,[2] rather than to the stereotype. Specifically, Ben-Zeev et al. write "Women's performance deficits in the presence of men were attenuated when they were given an opportunity to misattribute their arousal to an external source" (p. 179).

[2]Ben-Zeev et al. describe the details of the experiment: "Participants in the control group were told that the subliminal noise [that would play in the background as they took the test] would have no discernible physical effects on them. In contrast, those in the misattribution condition were told that the noise was associated with a number of side effects, and that previous participants had noted an increase in arousal, nervousness, and heart rate. They were told not to be alarmed if they felt these side effects and were assured that any such side effects would be temporary. Next, all participants were given the math test" (p. 178).

The study further showed a decrease in mathematics performance for those females not under stereotype threat conditions whose overarousal was also diverted by misattribution. If some females' misattribution of the cause of their overarousal actually changed their arousal levels, this change would support the validity of the Yerkes-Dodson Law, as near optimum levels of arousal were returned to by those in the stereotype threat condition (that is, by those females in the study's mixed-sex testing[3]), while those not subjected to stereotype conditions (females in same-sex testing) underperformed due to either a less-than or greater-than optimum arousal level as a result of the misattribution.

There is also evidence to support the notion of decreased performance for females under stereotype threat conditions (because available capacity for the operating process is reduced due to an increase in the monitoring process) when looking at the area of *working memory capacity*. Working memory capacity can be thought of as the ability to maintain focus on a particular task while disregarding competing irrelevant thoughts (Engle, 2001). It is strongly related to performance on complex cognitive tasks such as problem solving. One study found that a measure of working memory capacity correlates significantly with scores on both the quantitative and the verbal sections of the SAT I (Turner & Engle, 1989). That is, those with more working memory capacity tended to perform better on these SAT I sections. Furthermore, there is strong evidence that "stereotype threat reduces an individual's performance on a complex cognitive test

[3]"Female participants were randomly assigned to either same-sex (three female students) or minority (one female student with two male confederates) conditions" (Ben-Zeev et al., p. 178).

because it reduces the individual's working memory capacity" (Schmader & Johns, 2003, p. 449).

Another subtly different explanation for underperformance on tasks under conditions of high anxiety is that elevated levels of anxiety, rather than *reducing* working memory capacity, *distract* a portion of working memory by shifting focus to task-irrelevant operations. In this theory of *cognitive interference,* intrusive thoughts that accompany anxiety disrupt task performance (Eysenck, 1992). That is, the monitoring process interferes with the operating process instead of displacing it. For example, the condition of evaluative scrutiny can create stress in an individual, and preexisting preoccupations (such as stereotype threat) can manifest themselves in the form of intrusive, self-focused, task-irrelevant thoughts that interfere with task-oriented cognitive processes and thereby create performance deficits (Sarason & Sarason, 1981; Sarason, 1984).

The general notion that stereotype threat interferes with—or reduces—cognitive capacity is supported by the results of earlier mentioned studies on sex differences in use of strategies on standardized mathematics test problems. In particular, these study results indicate that females rely more on standard classroom-presented algorithms, while males more often use nontraditional steps or short cuts (Gallagher, 1992; Paek, 2002). This difference might suggest that while females were limited to standard algorithms as a result of diminished cognitive capacity (caused by anxiety induced by stereotype threat), males had either more cognitive capacity available or a less

distracted focus, which allowed them access to other, more direct methods to solve problems.

Females have also demonstrated less prior knowledge and fewer strategies when working mathematics problems in mixed testing situations (e.g., males and females in the same group) than did males who had similar mathematical backgrounds (Byrnes & Takahira, 1993); this difference again could suggest that females in these situations may have been the victim of reduced working memory capacity or cognitive interference. Another study of strategy use, based on mathematics items from the SAT I and the GRE general test, found that males had more *strategy flexibility,* and, as a result, outperformed females (Gallagher et al., 2000). This greater flexibility demonstrated by males could also suggest that they had more cognitive capacity available than females, because females were losing capacity due to stereotype threat conditions from mixed testing situations.

How stereotype threat affects performance on problems of varying difficulty levels

Along with stereotype threat, the rigor associated with a standardized test plays a major role in affecting the amount of cognitive capacity available. It has been documented that gender gaps in mathematics achievement widen as the difficulty level of problems increases (Feingold, 1988). The work by Spencer et al. (1998), described earlier, provides further support for this idea. Part 1 of this study focused on males and females who had performed well, both in a college calculus course (receiving a final grade of B or better) and on the SAT I or ACT quantitative sections (scoring above the 85th percentile). The control group was given a relatively easy (for their background)

test (composed from the quantitative section of the GRE general exam), while the experimental group was given a difficult test (composed from the advanced GRE exam in mathematics).

Although identity salience (from mixed gender testing) was a condition for females in both the "easy" and the "difficult" test groups, females underperformed compared to males only on the more rigorous test (Spencer et al., 1998). For these high-ability students, the general GRE quantitative section may not have been as challenging, and as a result this exam might have made fewer demands on their cognitive capacity, which allowed the females to deal effectively with any intrusions by stereotype threat.

The Yerkes-Dodson Law (1908) implies that as complexity increases, the inverted U shape (or "inverted teacup," in Yerkes-Dodson's words) slides to the left on the horizontal arousal scale (see figure 1). That is, if we hold arousal constant within optimum levels to perform well on a relatively easy task, then performance (at this constant arousal level) will decrease as the complexity of the task increases. Therefore, the amount of cognitive capacity available to males, who are unaffected by female-male stereotype threat, appears to be dependent upon the complexity of the task, while for females, the amount of cognitive capacity available seems to be dependent on both the complexity of the problem *and* the added level of arousal because of stereotype threat.

Further, all other things being equal, since the arousal level is higher for females affected by stereotype threat than for males who are not, it logically follows that performance gaps should widen on more difficult standardized mathematics tests

because females are further to the right on the arousal scale than males and if the "inverted teacup" moves left, the females' performance will drop off faster than the males due to the concave-down nature of the curve (see Figure 1), as was found by Feingold (1988).

How the format and context of the testing instrument can affect performance

The format and context of an exam can affect student performance. One study in Great Britain revealed a boost in male performance when multiple-choice tests were administered instead of traditional essay or free-response exams for secondary-level academic qualification (Murphy, 1977), while another study found that females outperformed males on multiple-choice tests that focused on algebra (Garner & Engelhard, 1999).

In terms of context, while many computational mathematics problems on standardized tests are non-contextual (meaning they do not rely on context to be solved), problem-solving items are often heavily contextual (for example, word problems provide a context for a mathematical problem). Problems set in mathematics or science contexts have been shown to provide an advantage to males, while those set in contexts of homemaking or human relationships have been shown to provide an advantage to females (Dwyer, 1979; Lawrence, Curley, & McHale, 1988).

Carol A. Dwyer, an Educational Testing Service researcher, describes the double standard sometimes found in the test-construction (SAT I) process:

> In the case of verbal tests, where females have typically performed better than males, test specifications prescribe inclusion of material on which males might be

> expected to perform better, such as material with a science context. Similar adjustments are not usually made in mathematics tests, where males' performance is often higher than that of females. This may be an example of inadvertent sexism on the part of test makers and researchers; the need for balancing is more readily perceived where males are at a disadvantage, and steps are taken to remedy the condition. No parallel steps are taken when it is females who are at a disadvantage—perhaps because the situation seems so familiar to all of us that we do not think to question it (Dwyer, 1979, p. 346).

What differential item functioning can show about gender bias/impact on standardized tests

Differential item functioning (DIF) is a procedure for detecting systematic performance differences on an item between examinees with equal ability who are from different subgroups (Holland & Wainer, 1993). To perform differential item functioning, a focal group's and a reference group's[4] performances on a given test item are compared to see whether an item exhibits DIF. When an item displays differential item functioning, it means that the probability that individuals will succeed on that item is statistically different for various subgroups. DIF may indicate bias. If the subgroups, however, are not of equal ability, then the difference in item performance is referred to as *impact* (Holland & Wainer, 1993). For example, if one subgroup has not had the same *opportunity to learn* tested material as another subgroup, a finding of DIF would indicate impact. While DIF can identify test items that elicit unexpected differences in

[4] Where the *focal group* is usually the group of interest, and the *reference group* is the comparison group.

performance between subgroups, DIF cannot specifically point to the underlying causes of those differences as noted by The American Educational Research Association, American Psychological Association, and National Council on Measurement in Education in 1999.

> Although DIF procedures may hold some promise for improving test quality, there has been little progress in identifying the causes or substantive themes that characterize items exhibiting DIF. That is, once items on a test have been statistically identified as functioning differently from one examinee to another, it has been difficult to specify the reasons for the differential performance or to identify a common deficiency among the identified items (Standards for Educational and Psychological Testing, 1999, p. 78).

Furthermore, it can be difficult to predict the *differential test functioning* (DTF—compares the functioning of *sets of items* by combining the individual DIF analyses) based solely upon differential item functioning analysis of the instrument's items (Gierl, Bisanz, Bisanz, & Boughton, 2002). This difficulty is due in part to the effects of *amplification* and *cancellation*. *Amplification* is the result of several items that exhibit differential item functioning that favors one subgroup combining to create a larger differential test functioning. *Cancellation* occurs when items favoring one subgroup negate the effects of items favoring another subgroup, resulting in a smaller or neglible differential test functioning (Nandakumer, 1993).

Differences in DIF findings for gender subgroups in content areas, for abilities, and at varying complexity levels

In a mathematics performance assessment of middle school children, items favoring boys were found in the areas of geometry, ratio/proportion, and problems with figures, while items favoring girls were found on problems involving number sense, estimation, and patterns (Lane, Wang, & Magone, 1996). In two university mathematics placement tests, items favoring males were found in the areas of geometry, measurement, and problem solving, while items favoring females were found in the areas of number sense and procedural algebra (Li, Cohen, & Ibarra, 2004). Items in the SAT I quantitative section have been shown to exhibit DIF favoring males on more complex items and favoring females on less complex items (Harris & Carlton, 1993).

History of the SAT and its present validity in predicting college success

In 1900, the presidents of several northeastern universities formed the College Entrance Examination Board for the purpose of standardizing the admissions process. Not long after, IQ ("intelligence quotient") tests were devised by the French psychologist Alfred Binet and eventually administered by Harvard professor Robert Yerkes (of Yerkes-Dodson Law fame) to almost two million army recruits. Carl C. Brigham, an associate of Yerkes, developed the first SAT (originally called the Scholastic Aptitude Test and later, the Scholastic Assessment Test) and administered it to high school students in 1926. In 1934, Harvard began using the SAT for awarding scholarships, and a year later required it of all incoming students. By the end of the

1930s all Ivy League schools were using the SAT. In 1948, Educational Testing Service (ETS) was formed to administer the exam on a larger scale (the College Board contracts with Educational Testing Service to help develop and administer the test), and by 1957 more than half a million students were taking the SAT annually. In 1968, the University of California System began requiring all students take the exam as part of the application process (Lemann, 1999). Today, about 1.5 million students take the SAT annually (College Board, 2006).

In Hyde, Fennema, and Lamon's 1990 meta-analysis of gender differences in mathematics performance, the overall effect size favoring males in all studies of mathematics tests they examined (100 studies total)—except the SAT I quantitative section—was $d = 0.15$. However, the SAT I had an effect size favoring males of $d = 0.40$. In making decisions about student admissions and scholarships, colleges use SAT scores, but the validity of the SAT I quantitative section in predicting the postsecondary performance of female students has been questioned. Studies have shown that on the SAT I quantitative section, males consistently scored a third of a standard deviation (or more) higher than females who later took the same college mathematics courses as the males and earned the same grades (Bridgeman & Wendler, 1991; Wainer & Steinberg, 1992).

The College Board has offered various reasons for differences in the performance of males and females on the SAT I quantitative section. These reasons include

- A larger proportion of males than females take advanced mathematics and science courses in high school and are therefore better prepared.

- Larger proportions of males than females major in mathematics-related fields and therefore take more rigorous mathematics courses.
- Since more females take the SAT, their average score is lowered because a greater proportion of females that aspire to college are from a lower socioeconomic status and are therefore more disadvantaged in terms of parental income and education. (College Board, 1998b)

Since SAT test developers employ a variety of mathematics problems, each with its own subset of strategies that can be called on in the solving process, it should be expected that some item types will create more of a performance difference between males and females than others. Although at the item level, SAT performance differences due to DIF on particular problems may be small and may have effect sizes that are considered small according to Cohen (1992), if these types of problems occur on an instrument with high frequency, the overall difference and effect size could be larger. Nandakumar (1993) has demonstrated that items exhibiting moderate levels of DIF can combine and amplify for a more substantial total DIF and suggests that "the decision to remove/add items should not be based on item level analysis alone but should consider the effect of such items at the test level" (Nandakumar, 1993, p. 308).

In terms of controlling for DIF on the SAT I, the College Board makes the following claim:

> Calculations determine the likelihood that differences in performance on any question result from overall ability differences or something inherent in the question. Questions that clearly perform differently for any group are carefully

> reviewed and nearly always eliminated from the pool of potential test questions. A number of additional analyses and quality control procedures are implemented at the question and test level to ensure tests are fair to all groups (College Board, 1998b, p. 3).

Still, on the SAT I quantitative section some items that exhibit DIF favoring males (such as measurement or geometry problems) are retained, because they are considered legitimate or necessary areas in which all students should be tested. Yet, even though statistically different performance by subgroups on particular items is generally controlled for on these exams, there still exists a large overall gap in performance between males and females on the SAT I quantitative section. Even though, then, the differences at the item level are not statistically significant, the performance gap may result from the accumulated effects (amplification) of DIF items that favor males throughout the SAT I quantitative section, coupled with higher levels of stress associated with the high-stakes and speeded nature of the exam.

Using National Assessment of Educational Progress item analyses to predict performance differences between males and females

The National Assessment of Educational Progress (NAEP), known as "the Nation's Report Card," has been assessing American students in multiple subjects, including mathematics, since 1969. Test development specialists and subject matter experts at ETS (the same people who create the SAT) construct tests and test items for NAEP (Department of Education, 1997). The NAEP reports data for students' achievement at the 4th, 8th, and 12th grades, and its website provides item analyses, via

the NAEP Data Explorer, for released mathematics problems used in the 1990, 1992, and 1996 (as well as other years) administrations of the national test. These item analyses are broken down by grade (4, 8, and 12), content (number sense, statistics, measurement, geometry/spatial, and algebra), ability (procedural, conceptual, and problem solving), and difficulty (easy, medium, and hard). The assessment results are disaggregated by sex, ethnicity, economic background, and public or private school. In mathematics at the 12th-grade level, ten to twenty thousand students are sampled with each administration (Department of Education, 2005).

A recent study of student performance using the NAEP Data Explorer found that performance gaps for 12th-grade males and females were largest in the content areas of geometry, measurement, and number operations (McGraw, Lubienski, & Strutchens, 2006). Further, these performance gaps were predominant in the upper end of score distributions, which would suggest that more difficult items led to the performance gap between high-performing males and females.

For students taking the NAEP, it lacks the level of evaluative scrutiny and high stakes associated with the SAT. But the NAEP does involve the condition of identity salience for triggering stereotype threat, in that students must indicate their sex prior to the test and must take the test in a mixed environment. Therefore, while NAEP analyses of males and females at the 12th-grade level may be useful in predicting performance differences due to content, ability, and/or difficulty levels on the SAT I quantitative section, the analyses fail to capture the full extent of differences that may result from the effects of stereotype threat.

Chapter 3: Methodology

This chapter details an analysis of performance differences between males and females on NAEP mathematics items in content areas, ability classifications, and at different difficulty levels. It then describes how gender gaps in performance on released versions of the SAT I quantitative section are predicted based upon the NAEP analysis.

Analysis of performance differences between males and females on NAEP items

This study used the National Assessment of Educational Progress Data Explorer (formerly the NAEP Data Tool) to analyze 12th-grade male and female performance on NAEP mathematics items across content areas, ability classifications, and difficulty levels. NAEP brakes down released items by various demographics, and, in particular, by the percentage of males and females selecting each possible answer from the multiple-choice format. From these analyses, I generated a percentage of males and females who got each item correct and incorrect. I determined significant male–female differences in performance on items by using a two proportion (based on percentage correct) z-test. I calculated effect sizes using Cohen's (1988) *d*, which is computed by taking the difference between two times the arcsine of the square root of the percentage correct for each sex.

This method is suggested by Cohen for working with proportions (percentage of each subgroup answering correctly), because the more familiar method of dividing the mean difference by the standard deviation would result in the same effect size if, for example, the percentage correct for two groups was 80% versus 60%, or 40% versus 20%, even though only one third more were correct in the former case while twice as

many were correct in the latter case (Cohen, 1988). I employed similar procedures for calculating significant differences and effect sizes for item categories using the *average percentage correct* for males and females across all items within the category (an example of a category would be hard algebra/conceptual).

Additionally, I calculated expected values for males and females for each NAEP item category. For this calculation, I computed the expected value for a subgroup by multiplying the number of points the item is worth by the probability of success (percentage correct for that subgroup) and then subtracting the number of points a student could lose (if any, for answering incorrectly) multiplied by the probability of failure (percentage incorrect for that subgroup). I then calculated a difference in expected value (DEV) between males and females by subtracting the female expected value from the male expected value for each category of items. A positive DEV for a category represents the number of points on average that males would be expected to outscore females on an item from that particular category, while a negative DEV would reflect the number of points on average that females would be expected to outscore males. To help identify any emergent trends in mathematical performance for the sexes I constructed a table for each ability classification (content versus difficulty), difficulty level (content versus ability), and content area (ability versus difficulty) (See Appendix 1).

Predicting differences in performance between males and females on the SAT I quantitative section

Using NAEP criteria, I then coded items from ten released versions of the SAT I quantitative section (College Board, 1997; the released exams spanned 1994 to 1997) by problem content area, ability classification, and difficulty level. The coding of SAT items required strict adherence to NAEP guidelines for content and ability classification so that I could use NAEP category DEVs to predict performance differences between males and females on the SAT I quantitative section. Many SAT I items, however, overlap the provided NAEP descriptions and could be classified in different categories. Even NAEP representatives conceded that "expert reviewers of the 1990 assessment often were unable to agree on the best placement of some items in the framework matrix" (National Assessment Governing Board, 2002, p. 15). Thus, I used NAEP content descriptions, along with any additional criteria realized from the actual NAEP classification, to code items (I provide these content descriptions and additional criteria in Appendix 2). Inter-coder reliability for content classifications was 96%. I also used NAEP ability descriptions in coding; these ability descriptions are provided in Appendix 3. Inter-coder reliability for ability classifications was 90%.

NAEP classifies each item's level of difficulty according to the percentage of all students who answered the item correctly. Items for which 0 up to 40% of all participants get the correct answer are classified as hard, items for which 40 up to 60% get the correct answer are classified medium difficulty, and items for which 60% or above get the correct answer are deemed easy. I used this NAEP classification system

for coding SAT I quantitative items for difficulty levels. This classification was possible because released SAT I quantitative items provided by the College Board are broken down in difficulty level by quintiles (e.g., each item is grouped into a difficulty level according to whether 0 to 20%, 20 to 40%, 40 to 60%, 60 to 80% or 80 to 100% of students get the correct answer) (College Board, 1997). Therefore, the lower two, middle, and upper two quintiles classifying level of difficulty for items from the SAT I quantitative section match exactly with NAEP's hard, medium, and easy classifications.Thus, coding the difficulty level of the SAT I quantitative items was straightforward.

The SAT I quantitative section has sixty items. In accordance with scoring procedures on the SAT, I calculated expected values for the first fifty items on the ten released SAT I quantitative instruments (these first fifty are multiple choice/comparison problems) by multiplying the percentage of each subgroup who got the item correct (I generated these percentages in my analysis of the NAEP items based upon the item category) by ten points (the approximate value of each SAT I quantitative item) and subtracting the percentage who got the item incorrect multiplied by 2.5 points (the College Board scoring procedure includes a penalty intended to discourage guessing on the SAT). For the last ten items on the ten released SAT I quantitative instruments (which require grid-in answers[5]) the expected value was simply the percentage of each subgroup who got the item correct multiplied by ten points (the value of a SAT I

[5] A "grid-in" is a type of item that—in contrast to a multiple-choice item, which provides several answers from which students choose—requires students to calculate the answer to the mathematical problem and fill that answer into a grid. This type of item is also known as a Student-Produced Response, or SPR, question.

quantitative item), since students are not penalized for wrong answers in this section of the SAT I quantitative section. Once I had thus coded the expected values for all the items (600 total) on each of the ten released SAT I quantitative section instruments, I assigned a difference in expected value (DEV) to each item.

I then totaled the DEVs to quantify the accumulation of performance differences on all items across each of the ten released SAT I quantitative instruments. Some item categories (for example, easy-difficulty algebra/problem solving and hard-difficulty measurement/conceptual) found in these SAT I quantitative sections were not represented in the NAEP data and were therefore assigned a neutral DEV (0.00). (In Appendix 4, I provide the coding of each released SAT-I quantitative section along with the item DEVs.) I compared the average of the total DEVs for each of the ten released SAT I quantitative instruments (see Appendix 5) to the actual gender gaps in average scores on the SAT I quantitative section over the same span of years as the released exams (1994 to 1997). In addition, I calculated the total predicted difference of expected value (TDEV) of particular item categories on the SAT I quantitative section by multiplying that item category's DEV by the average frequency of items (that is, by how many times a type of item appeared in each instrument) from that category per instrument across all ten released tests (see Appendix 6). This TDEV quantifies the effect that each category of item on the SAT I quantitative section has on male and female performance.

Chapter 4: Results

This section begins with a report of NAEP analysis results from each of the content areas, ability classifications, and difficulty levels in order to identify any emergent trends within these sub-categories. Next, I will present total DEVs from particular item categories to determine their relative effect on the performance of males and females on SAT I quantitative items. The section concludes with an analysis of the released SAT I quantitative section exams by content, ability, and difficulty and a predicted difference in performance for males and females based on the total DEVs of the ten instruments.

Content Areas

Algebra

Males performed significantly better on algebra problem-solving items (DEV = .825, d = .20, $p < .001$) and generally would be expected to outscore females by approximately one fifth of a point per algebra item (DEV = .201, d = .05).

Data Analysis, Statistics, and Probability

Males performed significantly better than females on procedural (DEV = .763, d = .19, $p < .001$) and conceptual (DEV = .6, d = .08, $p < .05$) items. In general, males performed significantly higher than females on data analysis, statistics, and probability items (DEV = .492, d = .08, $p < .05$) and would be expected to outscore females by approximately one half point per item.

Geometry and Spatial Visualization

Males performed significantly better on procedural (DEV = .667, d = .08, $p < .01$) and problem solving (DEV = .778, d = .12, $p < .001$) items. Males were also significantly favored on medium/hard items (DEV = .647, d = .10, $p < .01$). In general, males performed significantly better on geometry items (DEV = .518, d = .10, $p < .05$) and would be expected to outscore females by approximately one half point per item.

Measurement

Males performed significantly better on conceptual (DEV = 1.075, d = .31, $p < .001$) and problem solving (DEV = .554, d = .08, $p < .05$) items as well as medium/hard items (DEV = .95, d = .12, $p < .05$). Medium-difficulty measurement/conceptual items had the largest difference in expected value and effect size (DEV = 1.875, d = .33, $p < .001$) of all item categories. In general, males performed significantly better on measurement items (DEV = .578, d = .11, $p < .01$) and would be expected to outscore females by more than one half point per item. Measurement items exhibited the largest content area differences in performance.

Number Sense and Operations

Males performed significantly better on procedural (DEV = .513, d = .14, $p < .001$) and medium/hard (DEV = .914, d = .13, $p < .01$) items. In general, males performed significantly better on number sense and operations items (DEV = .205, d = .07, $p < .05$) and would be expected to outscore females by approximately one fifth of a point per item.

Ability Classifications

Procedural

Males performed significantly better on number sense and operations (DEV = .513, d = .14, $p < .001$), geometry (DEV = .667, d = .14, $p < .01$), and data analysis/statistics/ probability (DEV = .763, d = .14, $p < .001$) items. In terms of difficulty, males performed significantly better on medium (DEV = .396, d = .07, $p < .05$) and hard (DEV = .375, D = .07, $P < .05$) items. In general, males would be expected to outscore females by approximately one quarter point per item (DEV = .26, d = .06).

Conceptual

Males performed significantly better on measurement (DEV = 1.075, d = .18, $p < .001$) and data analysis/statistics/probability (DEV =.6, d = .09, $p < .05$) items. Males performed significantly better on medium items (DEV = .929, d = .16, $p < .001$). In general, males would be expected to outscore females by approximately one third of a point per item (DEV =.324, d = .06).

Problem solving

Males performed significantly better on measurement (DEV = .554, d = .10, $p < .05$), geometry (DEV = .778, d = .13, $p < .001$), and algebra (DEV = .825, d = .13, $p < .001$) items. Males performed significantly better on medium (DEV = .97, d = .18, $p < .001$) and hard (DEV = .491, d = .08, $p < .05$) items. In general, males performed significantly better on problem solving items (DEV = .519, d = .12, $p < .001$) and would be expected to outscore females by more than one half point per item. Problem solving items exhibited the largest ability classification differences in performance.

Difficulty levels

Easy

Males performed significantly better on data analysis/statistics/probability items (DEV = .609, d = .16, $p < .05$). In general, there was virtually no difference between males and females on easy items (DEV = .039, d = .01).

Medium

Males performed significantly better on number sense and operations (DEV = 1.313, d = .24, $p < .001$), measurement (DEV = 1.5, d = .25, $p < .001$), geometry (DEV = .669, d =.13, $p < .01$), and data analysis/statistics/probability (DEV = .688, d = .12, $p < .01$) items. Males performed significantly better on procedural (DEV = .396, d = .07, $p < .05$), conceptual (DEV = .929, d = .16, $p < .001$), and problem solving (DEV = .97, d = .17, $p < .001$) items. In general, males performed significantly better on medium items (DEV = .792, d = .14, $p < .001$) and would be expected to outscore females by approximately four fifths of a point per item. Medium difficulty items exhibited the largest difficulty level performance differences.

Hard

Males performed significantly better on number sense and operations (DEV = .516, d = .08, $p < .01$), measurement (DEV = .4, d = .05, $p < .05$), and geometry (DEV = .625, d = .08, $p < .01$) items. Males performed significantly better on procedural (DEV = .375, d = .07, $p < .05$) and problem solving (DEV = .491, d = .07, $p < .05$) items. In general, males performed significantly better on hard items (DEV = .422, d = .06, $p < .05$) and would be expected to outscore females by approximately two fifths of a point per item.

Total DEV of Item Categories on the SAT I quantitative section

The relative effect on male and female performance for item categories is listed in Appendix 6. Categories with a DEV = 0 are neutral and their relative frequency on an instrument would not affect the overall gender gap (e.g. easy-difficulty measurement/procedural and medium-difficulty algebra/conceptual categories). Conversely, regardless of the magnitude of a category's DEV, if it does not appear on the SAT, then it will have no impact (e.g. hard-difficulty geometry/procedural category).

By far, the greatest predicted difference in performance favoring males is in the general category of problem solving at the medium or hard difficulty level, which accounts for nine of the twenty-two categories that reflect an advantage for males. These are spread across all content areas and average over eleven items (out of sixty) per test and combine for a projected total DEV of 7.64 points. The two more difficult number sense and operations procedural categories (medium and hard), together, average six items per test and account for a total DEV of 6.96 points. Finally, the single category of medium-difficulty measurement/conceptual has the largest DEV (1.875) of all item categories and averages 2.6 problems per instrument, accounting for a total DEV of 4.875 points per instrument.

At the bottom of the list in Appendix 6 are the categories which have the biggest predicted difference in performance favoring females. While none of these categories exhibit DEVs that significantly favor females, most notable is the category of number sense and operations/conceptual problems which average three items per test and have

a total DEV of -1.5 points as well as the general category of algebra procedural items, which have a combined average of 7.8 items per instrument and a total DEV of -1.3 points.

Prediction of gender gaps on the SAT I quantitative section

Analysis of the ten released SAT I quantitative section instruments revealed that on average 35% of items were number sense and operations, 20% were measurement, 10% were geometry and spatial sense, 12% were data analysis/statistics/probability, and 23% were algebra/functions items. In terms of ability, 37% of items were classified as procedural, 37% were conceptual, and 26% were problem solving. The breakdown by difficulty was 34% easy items, 39% medium, and 27% hard. The average total DEV for the released SAT I quantitative section tests was 20.25, which would suggest an expected difference in scores for males and females of about 20 points (compared to the actual gap of 35 points on the SAT I quantitative section from 1994 to 1997). 86% of the items found on the SAT I quantitative section released tests were matched with like categories from NAEP (the rest were item categories not covered by NAEP), and 57% of the actual gap found on the SAT I quantitative section was predicted. This suggests that sex differences in performance on the SAT I quantitative section are not solely due to item content.

Chapter 5: Discussion and Conclusions

This chapter discusses the results of this study from the NAEP analysis of sex differences in performance and the prediction of sex differences on the SAT I quantitative section. The chapter concludes with recommendations for closing performance gaps between males and females on the SAT I quantitative section, implications and limitations of this study, and possible future work.

Analysis of sex differences in performance on NAEP items

The results of my analysis of 12th-grade male and female performance on NAEP mathematics items across content areas, ability classifications, and difficulty levels agrees with findings from past studies across these areas. In my study, males performed significantly better than females on NAEP items that tested ability in data-analysis/statistics/probability, geometry, number sense/operations, and measurement; these performance differences correspond with the findings of Garner and Engelhard (1999); Li, Cohen, and Ibarra (2004); McGraw, Lubienski, and Strutchens (2006); and Willingham and Cole (1997). McGraw, Lubienski, and Strutchens, whose study of NAEP performance focused on total scores (rather than item analyses) from each content area, likewise found the largest male–female performance gaps were in the measurement and geometry strands.

My analysis of NAEP items also found that males performed significantly better on problem-solving items. Again, similar results have been reported by Hyde, Fennema, and Lamon (1990); Quinn and Spencer (2001); and Willingham and Cole (1997). My analysis showed that more difficult (e.g., medium and hard) items

significantly favored males as well, corresponding to results from Feingold (1988) and Spencer, Steele, and Quinn (1998). However, my analysis found the largest performance gaps on medium-difficulty NAEP items (DEV = .792), with hard problems (DEV = .422) exhibiting only about half the difference in performance found on medium-difficulty problems. This finding may be due to a "floor" effect resulting from lower percentages of males and females getting correct answers on the hard problems because a portion of the percentage of correct answers can be attributed to test-takers guessing at answers (that is, if there are five choices on a multiple-choice item, then on average, one out of five people guessing should get the correct answer).

The two released NAEP items with the largest effect sizes—that is, the largest difference between male and female performance—were from the medium-difficulty measurement/conceptual category. A more detailed analysis of these items revealed that females were seemingly less familiar than males with the units of measurement involved. In one of these NAEP problems, students are tasked with estimating the height of their classroom door, using meters as the unit of measurement, while in the other problem, they are asked to match a measurement given in cubic inches to a length, surface area, or volume. Two thirds of the males chose the correct responses (problem one: two meters; problem two: volume) compared to only half of the females.

Solving *measurement* problems like these depends, perhaps, less on innate skills than on experience with the measurement units involved. Thus, in situations such as these, providing females an opportunity to learn (in this case, to become more familiar with those measurement units) could help to close the performance gaps. However, in

my analysis on the various *problem-solving* NAEP items, spanning all content areas, there seemed to be no pattern explaining why females chose wrong answers more often than males. The only common thread throughout these items was their similar levels of complexity and difficulty, which, in previous studies, have been key elements in the manifestation of stereotype threat for females (Feingold, 1988; Quinn & Spencer, 2001; Spencer, Steele, & Quinn, 1998).

The largest effect size of any content area was d = .11 (in the measurement content area); the largest effect size for any single item was d = .36 (in the medium-difficulty measurement/conceptual content area). The majority of single-item effect sizes were less than .20. However, this study demonstrates that these item differences have the potential to combine for larger overall differences in male and female test performance. Differences in some items favoring males, such as measurement problems, may be due to the lack of opportunity to learn the material on the part of females and reflect impact (e.g., a difference in item performance due to the subgroups not being of equal ability). Other items, such as word problems across all content areas, may reflect bias due to the testing environment in conjunction with stereotype threat.

Predicting sex differences on the SAT I quantitative section

My study's NAEP-based prediction of a 20-point gender gap on the SAT I quantitative section falls short of the actual performance gaps (approximately a 35 point gap in each of the years of the ten released tests studied—1994 to 1997). My analysis matched 86% of the SAT I quantitative section items in the ten released tests with

corresponding NAEP categories and 57% of the actual performance gap is accounted for through this matching. Although my analysis of NAEP item performance provides a benchmark for male–female performance gaps on the SAT I quantitative section, my analysis is influenced by the fact that the NAEP is by nature a low-stakes test. This is in contrast to the high-stakes SAT I, the outcome of which affects decisions about college admissions and scholarships. It is quite possible that the extra stress associated with the high-stakes environment of SAT I quantitative section administration contributes to the effects of stereotype threat, thereby causing larger performance gaps. The larger gaps between male and female performance on the actual SAT I quantitative section than was predicted by my NAEP analysis might also result from a more select sample of students (those who are college bound) taking the SAT, than the broader sample of students taking the NAEP. This more select sample of students may therefore be more susceptible to stereotype threat (because they generally are higher ability students), as found by Gallagher and DeLisi (1994).

Although the College Board states that it monitors particular SAT I items for fairness, clearly some impact/bias accumulation over all the items is occurring on the sixty-problem instruments it uses. Further, on the ten released SAT I quantitative sections that I analyzed, not only did males have generally greater effect sizes and predicted DEVs with the items that favored them than did the females, but the ratio of item DEVs that favored males to those that favored females was more than two to one. Thus, the evidence indicates that performance gaps between males and females found on the SAT I quantitative section are in part a function of the effect size of particular

items as well as the frequency with which those items occur. However, my NAEP analysis's predicted performance gap (about 20 points) was short of the actual SAT I quantitative performance gap (35 points). So, the findings of this study, when taken in conjunction with current knowledge on stereotype threat, strongly suggest that—in addition to gaps resulting from the construction of the exam—the remainder of the performance gap between males and females on the SAT I quantitative section may be attributed to the environment of the test's administration, as has already been generally inferred by Hyde, Fennema, and Lamon (1990).

Conclusion

The Educational Testing Service creates items for both the NAEP and the SAT I and administers the SAT I on behalf of the College Board to about 1.5 million college-bound students each year. SAT scores, of course, are used to help universities make decisions about admitting students and as part of the process for awarding scholarships. As current College Board President Gaston Caperton sees it,

> The SAT can be more than an admissions test. It sets a standard that you have to have. And let's think about what an admissions system would be without the SAT. If you don't have some sort of *totally objective* measurement by which admissions people look at students, some criteria, how in the world can you do it on a fair basis? (Hoover, 2006, p.9)(Emphasis added by dissertation's author).

So, if the quantitative section of the SAT *is* truly objective, the call should go out to schools and educators to provide females with the opportunity to learn those skills that are necessary to compete with males. Specialized training on particular skills, such

as spatial reasoning, where females have successfully improved their skills, is rarely available (Vasta, Knott, & Gaze, 1996; Baenninger & Newcombe, 1989).

However, if it is the case that the SAT I quantitative section accurately measures skills needed in college, then why is it that females with similar backgrounds generally make the same grades as males in freshman mathematics courses, even though these females consistently underperform on the SAT? Perhaps the content of the SAT I quantitative section *doesn't* accurately reflect the skills needed for first-year college students.

The typical core curriculum mathematics course required for students at universities is college algebra or its equivalent, which is a content area where females perform on par with males according to NAEP data (in my analysis, college algebra had the smallest content area DEV, .201). My analysis, however, found that about 42% (on each of the ten released tests) of the SAT I quantitative section items were in the measurement, geometry, or data analysis/statistics/probability content areas, all of which are content areas that my analysis, based on NAEP data, has shown to significantly favor males. Typically, measurement and geometry topics are addressed in high school geometry courses. Yet currently, only a handful of states require geometry for high school graduation, and statistics courses are generally optional. Therefore, since these topics are often not encountered by college freshmen, there is little incentive for females to study them in high school other than in preparation for college entrance exams such as the SAT I.

Thus, evidence to date suggests that performance gaps on standardized math tests such as the SAT I quantitative section are influenced by a complex mix of sociological, psychological, and (to a lesser extent) biological factors unique to each individual. To provide a level playing field for both sexes, opportunities to learn (for both sexes) must not only be made available by educators, but also, in some situations, must be required components of the curriculum, so that all students are equitably prepared for taking high-stakes exams that many universities require as part of the admissions process. Further, consideration should be given to modifying the testing environment and providing accommodations in order to diminish the effects of stereotype threat.

Recommendations

Since it seems that performance gaps on the SAT I quantitative section are likely due to how the test is administered (mixed testing and identification of sex prior to test) and how the test is constructed (because there are differential performances by sex on individual items, which combine for larger sex differences), I make the following recommendations:

- The College Board should reevaluate the skills required for today's incoming freshmen to succeed in college, and should ensure that those skills are accurately reflected and weighted in the content of the SAT I quantitative section, thereby making the exam a more accurate predictor of first-year college success for males and females.
- Educators and administrators should provide the opportunity to learn the requisite skills for college to both sexes. In particular, if geometry, measurement, and

statistics topics are deemed necessary for college success, then those courses should become part of the required curriculum for all college-bound students.

- The Educational Testing Service, to lessen the effects of stereotype threat, should consider shifting its administration of the SAT to individualized (or same sex) testing and moving identification of one's sex to after or well in advance of taking the SAT.

Implications

This study identifies particular mathematics topics and abilities where either males or females are seemingly disadvantaged on standardized tests. These identifications can be used by those who construct and administer these exams to modify them to eliminate impact or bias. This information can also help students and their educators when preparing for such exams. If high-stakes decisions about admission to colleges and awarding of scholarships are based (even partially) on the results of a single testing instrument, then the instrument, as well as secondary school curricula for college-bound students, must be aligned with, and accurately reflect, the topics and skills necessary for college.

Limitations of the study

Using NAEP data to indirectly predict performance on the SAT I quantitative section is a crude method, indeed, because basing my analysis on a few aligned NAEP and SAT I items cannot fully represent the broad range of items possible within a certain content/ability/difficulty category. Also, the sample of students taking the NAEP is more diverse (college bound and not college bound) than the sample of those

taking the SAT (college bound), and what constitutes the difficulty level of an item is possibly different for each of the samples. Further, the SAT is a high-stakes exam that affects admission to colleges and selection for scholarships. One would expect higher levels of stress associated with the SAT testing environment than with that of NAEP, and therefore a heightened level of stereotype threat.

Future work

More in-depth studies need to be conducted on the categories of mathematics items on the NAEP that are identified with large differences in performance between males and females (e.g. measurement, geometry, problem-solving, and more difficult problems). While analysis of male and female performance on multiple-choice items can reveal limited insight into student thinking, analyzing the same items in an open-response format (say, under experimental conditions), could shed more light on the cognitive processes of students, especially on the more difficult problem-solving items that span all content areas. Also, it might be appropriate for researchers to take a closer look at differences in performance for males and females on the Preliminary SAT / National Merit Scholarship Qualifying Test (PSAT/NMSQT, taken by high school juniors and constructed by ETS as well), since National Merit Scholarships (which historically are awarded to males more often than females) are based on PSAT/NMSQT scores and SAT I scores.

Appendix 1
Difference in Expected Value
(Male EV – Female EV on multiple choice items)
* p < .05, ** p < .01, *** p < .001

Problem Solving

Content vs. Difficulty	number operation	measurement	geometry	data analysis	algebra	**total**
easy	-.125	-----	-----	-.125	-----	-.125
medium	-----	.938***	.838**	1.125***	1.125***	.97***
hard	.563*	.4	.688**	.25	.625*	.491*
total	.15	.554*	.778***	.275	.825***	.519***

Conceptual

Content vs. Difficulty	number operation	measurement	geometry	data analysis	algebra	**total**
easy	-.531*	-.125	.25	1.188***	-.5**	.017
medium	-----	1.875***	.3125	.25	0	.929***
hard	.25	-----	-----	.1875	.5	.363
total	-.196	1.075***	.268	.6*	.1875	.324

Procedural

Content vs. Difficulty	number operation	measurement	geometry	data analysis	algebra	**total**
easy	.05	0	-----	.7625***	-.0625	.127
medium	1.313***	-----	.875***	-----	-.375	.396*
hard	.75***	-----	.563*	-----	-.125	.375*
total	.513***	0	.667**	.763***	-.203	.26

Easy

Content vs. Ability	Number operation	measurement	geometry	data analysis	algebra	**total**
procedural	.05	0	-----	.763***	-.063	.127
conceptual	-.531*	-.125	.25	1.188***	-.5**	.017
problem solving	-.125	-----	-----	-.125	-----	-.125
total	-.187	-.042	.25	.609*	-.281	.039

Medium

Content vs. Ability	number operation	measurement	geometry	data analysis	algebra	**total**
procedural	1.313***	-----	.875***	-----	-.375	.396*
conceptual	-----	1.875***	.313	.25	0	.929***
problem solving	-----	.938***	.838**	1.125***	1.125***	.97***
total	1.313***	1.5***	.669**	.688**	.188	.792***

Hard

Content vs. Ability	number operation	measurement	geometry	data analysis	algebra	**total**
procedural	.75***	-----	.563*	-----	-.125	.375*
conceptual	.25	-----	-----	.188	.5	.363
problem solving	.563*	.4*	.688**	.25	.625*	.491*
total	.516**	.4*	.625**	.219	.364	.422*

Algebra

Ability vs. Difficulty	procedural	conceptual	problem solving	**total**
easy	-.0625	-.5 **	-----	-.281
medium	-.375	0	1.125***	.225
hard	-.125	.5 *	.625 *	.364
total	-.203	.214	.825***	.201

Data Analysis

Ability vs. Difficulty	procedural	conceptual	problem solving	**total**
easy	.763***	1.188***	-.125	.609*
medium	-----	.25	1.125***	.688**
hard	-----	.188	.25	.219
total	.763***	.6*	.275	.492*

Geometry

Ability vs. Difficulty	procedural	conceptual	problem solving	**total**
easy	-----	.25	-----	.25
medium	.875***	.313	.838**	.669**
hard	.563*	-----	.688**	.625**
total	.667**	.268	.778***	.518*

Measurement

Ability vs. Difficulty	procedural	conceptual	problem solving	**total**
easy	0	-.125	-----	-.042
medium	-----	1.875***	.938***	1.5***
hard	-----	-----	.4*	.4*
total	0	1.075***	.554*	.578**

Number Operations

Ability vs. Difficulty	procedural	conceptual	problem solving	**total**
easy	.05	-.05*	-.125	-.187
medium	1.313***	-----	-----	1.313***
hard	.75***	.25	.563*	.516**
total	.513***	-.196	.15	.205*

Appendix 2

NAEP Content Guidelines

Number Properties and Operations

1) Number sense

- Write, rename, represent, or compare real numbers (e.g., pi, square root of 2, numerical relationships using number lines, models, or diagrams).
- Represent very large or very small numbers using scientific notation in meaningful contexts.
- Find or model absolute value or apply to problem situations.
- Interpret calculator or computer displays of numbers given in scientific notation.
- Order or compare real numbers, including very large or small real numbers.

2) Estimation

- Establish or apply benchmarks for real numbers in contexts.
- Make estimates of very large or very small numbers appropriate to a given situation by: identifying when estimation is appropriate or not, determining the level of accuracy needed, selecting the appropriate method of estimation, or analyzing the effect of an estimation method on the accuracy of results.
- Verify solutions or determine the reasonableness of results in a variety of situations including scientific notation, calculator, and computer results.
- Estimate square or cube roots of numbers less than 1,000 between two whole numbers.

3) Number operations

- Perform computations with real numbers including common irrational numbers or the absolute value of numbers.
- Describe the effect of multiplying and dividing by numbers including the effect of multiplying or dividing a real number by: zero, or a number less than zero, or a number between zero and one, or one, or a number greater than one.
- Solve application problems involving numbers, including rational and common irrationals, using exact answers or estimates as appropriate.

4) Ratios and proportional reasoning

- Use proportions to model problems.
- Use proportional reasoning to solve problems (including rates).
- Solve problems involving percentages (including percent increase and decrease, interest rates, tax, discount, tips, or part/whole relationships).

5) Properties of number and operations

- Solve problems involving factors, multiples, or prime factorization.
- Use prime or composite numbers to solve problems.
- Use divisibility or remainders in problem settings.
- Apply basic properties of operations.
- Provide a mathematical argument about a numerical property or relationship.

6) Additional

- LCM
- GCF
- Rounding
- Compound interest
- Sequences with numbers
- Estimation of exponential growth

Measurement

1) Measuring physical attributes

- Estimate or compare perimeters or areas of two-dimensional geometric figures.
- Estimate or compare volume or surface area of three-dimensional figures.
- Solve problems of angle measure, including those involving triangles or other polygons or parallel lines cut by a transversal.
- Solve mathematical or real-world problems involving perimeter or area of plane figures such as polygons, circles, or composite figures.
- Solve problems involving volume or surface area of rectangular solids, cylinders, cones, pyramids, prisms, spheres, or composite shapes.
- Solve problems involving indirect measurement such as finding the height of a building by finding the distance to the base of the building and the angle of elevation to the top.
- Solve problems involving rates such as speed, density, population density, or flow rates.

2) Systems of measurement

- Select or use appropriate type of unit for the attribute being measured such as volume or surface area.
- Solve problems involving conversions within or between measurement systems, given the relationship between the units.
- Determine appropriate accuracy of measurement in problem situations (e.g., the accuracy of measurement of the dimensions to obtain a specified accuracy of area) and find the measure to that degree of accuracy.
- Construct or solve problems (e.g., number of rolls needed for insulating a house) involving scale drawings.
- Compare lengths, areas, or volumes of similar figures using proportions.

Geometry

1) Dimension and shape

- Use two-dimensional representations of three-dimensional objects to visualize and solve problems involving surface area and volume.
- Give precise mathematical descriptions or definitions of geometric shapes in the plane and in three-dimensional space.
- Draw or sketch from a written description plane figures (e.g., isosceles triangles, regular polygons, curved figures) and planar images of three-dimensional figures (e.g., polyhedra, spheres, and hemispheres).
- Describe or analyze properties of spheres and hemispheres.

2) Transformation of shapes and preservation of properties

- Recognize or identify types of symmetries (e.g., point, line, rotational, self-congruences) of two- and three-dimensional figures.
- Give or recognize the precise mathematical relationship (e.g., congruence, similarity, orientation) between a figure and its image under a transformation.
- Perform or describe the effect of a single transformation on two- and three-dimensional geometric shapes (reflections across lines of symmetry, rotations, translations, and dilations).
- Describe the final outcome of successive transformations.
- Justify relationships of congruence and similarity, and apply these relationships using scaling and proportional reasoning.

3) Relationships between geometric figures

- Apply geometric properties and relationships in solving multistep problems in two and three dimensions (including rigid and nonrigid figures).
- Represent problem situations with geometric models to solve mathematical or real-world problems.
- Use the Pythagorean theorem to solve problems in two- or three-dimensional situations.
- Describe and analyze properties of circles (e.g., perpendicularity of tangent and radius, angle inscribed in a semicircle).
- Analyze properties or relationships of triangles, quadrilaterals, and other polygonal plane figures.
- Describe or analyze properties and relationships of parallel, perpendicular, or intersecting lines, including the angle relationships that arise in these cases.

4) Position and direction

- Describe the intersections of lines in the plane and in space, intersections of a line and a plane, or of two planes in space.
- Describe or identify conic sections and other cross sections of solids.
- Represent two-dimensional figures algebraically using coordinates and/or equations.
- Use vectors to represent velocity and direction.

5) Mathematical reasoning

- Make, test, and validate geometric conjectures using a variety of methods including deductive reasoning and counterexamples.

6) Additional

- Slope, distance, and midpoints

Data Analysis and Probability

1) Data representation

- Histograms, line graphs, scatterplots, box plots, circle graphs, stem and leaf plots, frequency distributions, and tables.
- Read or interpret data, including interpolating or extrapolating from data.
- For a given set of data, complete a graph and then solve a problem using the data in the graph (histograms, scatterplots, line graphs).
- Solve problems by estimating and computing with univariate or bivariate data (including scatterplots and two-way tables).
- Given a graph or a set of data, determine whether information is represented effectively and appropriately (bar graphs, box plots, histograms, scatterplots, line graphs).
- Compare and contrast the effectiveness of different representations of the same data.

2) Characteristics of data sets

- Calculate, interpret, or use mean, median, mode, range, interquartile range, or standard deviation.
- Recognize how linear transformations of one-variable data affect mean, median, mode, and range (e.g., effect on the mean by adding a constant to each data point).
- Determine the effect of outliers on mean, median, mode, range, interquartile range, or standard deviation.
- Compare two or more data sets using mean, median, mode, range, interquartile range, or standard deviation describing the same characteristic for two different populations or subsets of the same population.
- Given a set of data or a scatterplot, visually choose the line of best fit and explain the meaning of the line. Use the line to make predictions.
- Use or interpret a normal distribution as a mathematical model appropriate for summarizing certain sets of data.
- Given a scatterplot, make decisions or predictions involving a line or curve of best fit.
- Given a scatterplot, estimate the correlation coefficient (e.g., Given a scatterplot, is the correlation closer to 0, .5, or 1.0? Is it a positive or negative correlation?).

3) Experiments and samples

- Identify possible sources of bias in data collection methods and describe how such bias can be controlled and reduced.
- Recognize and describe a method to select a simple random sample.
- Make inferences from sample results.
- Identify or evaluate the characteristics of a good survey or of a well-designed experiment.

4) Probability

- Analyze a situation that involves probability of independent or dependent events.
- Determine the theoretical probability of simple and compound events in familiar or unfamiliar contexts.
- Given the results of an experiment or simulation, estimate the probability of simple or compound events in familiar or unfamiliar contexts.
- Use theoretical probability to evaluate or predict experimental outcomes.
- Determine the number of ways an event can occur using tree diagrams, formulas for combinations and permutations, or other counting techniques.
- Determine the probability of the possible outcomes of an event.
- Determine the probability of independent and dependent events.
- Determine conditional probability using two-way tables.
- Interpret probabilities within a given context.

Algebra

1)Patterns, relations, and functions

- Recognize, describe, or extend arithmetic, geometric progressions, or patterns using words or symbols, including square roots.
- Express the function in general terms (either recursively or explicitly), given a table, verbal description, or some terms of a sequence.
- Identify or analyze distinguishing properties of linear, quadratic, inverse ($y = k/x$) or exponential functions from tables, graphs, or equations.
- Determine the domain and range of functions given various contexts.
- Recognize and analyze the general forms of linear, quadratic, inverse, or exponential functions (e.g., in $y = ax + b$, recognize the roles of a and b).
- Express linear and exponential functions in recursive and explicit form given a table or verbal description.

2) Algebraic representations

- Translate between different representations of algebraic expressions using symbols, graphs, tables, diagrams, or written descriptions.
- Analyze or interpret relationships expressed in symbols, graphs, tables, diagrams, or written descriptions.
- Graph or interpret points that are represented by one or more ordered pairs of numbers on a rectangular coordinate system.
- Perform or interpret transformations on the graphs of linear and quadratic functions.
- Use algebraic properties to develop a valid mathematical argument.
- Use an algebraic model of a situation to make inferences or predictions.
- Given a real-world situation, determine if a linear, quadratic, inverse, or exponential function fits the situation
- Solve problems involving exponential growth and decay.

3) Variables, expressions, and operations

- Write algebraic expressions, equations, or inequalities to represent a situation.
- Perform basic operations, using appropriate tools, on algebraic expressions (including grouping and order of multiple operations involving basic operations, exponents, roots, simplifying, and expanding).
- Write equivalent forms of algebraic expressions, equations, or inequalities to represent and explain mathematical relationships.

4) Equations and inequalities

- Solve linear, rational, or quadratic equations or inequalities.
- Analyze situations or solve problems using linear or quadratic equations or inequalities symbolically or graphically.
- Recognize the relationship between the solution of a system of linear equations and its graph.
- Solve problems involving more advanced formulas [e.g., the volumes and surface areas of three dimensional solids; or such formulas as: $A = P(1 + r)^t$, $A = Pe^{rt}$].
- Given a familiar formula, solve for one of the variables.
- Solve or interpret systems of equations or inequalities.

5) Additional

- Use trigonometric functions and identities.
- Find angle degree measure using inverse trig functions.
- Right triangle relationships for 30/60 and 45/45 triangles.
- Solve/simplify equations/expressions involving complex fractions.

Appendix 3

NAEP Mathematical Abilities

Conceptual understanding

Students demonstrate conceptual understanding in mathematics when they provide evidence that they can recognize, label, and generate examples of concepts; use and interrelate models, diagrams, manipulatives, and varied representations of concepts; identify and apply principles; know and apply facts and definitions; compare, contrast, and integrate related concepts and principles; recognize, interpret, and apply the signs, symbols, and terms used to represent concepts. Conceptual understanding reflects a student's ability to reason in settings involving the careful application of concept definitions, relations, or representations of either.

Procedural knowledge

Students demonstrate procedural knowledge in mathematics when they select and apply appropriate procedures correctly; verify or justify the correctness of a procedure using concrete models or symbolic methods; or extend or modify procedures to deal with factors inherent in problem settings. Procedural knowledge encompasses the abilities to read and produce graphs and tables, execute geometric constructions, and perform noncomputational skills such as rounding and ordering. Procedural knowledge is often reflected in a student's ability to connect an algorithmic process with a given problem situation, to employ that algorithm correctly, and to communicate the results of the algorithm in the context of the problem setting.

Problem solving

Students demonstrate problem solving in mathematics when they recognize and formulate problems; determine the consistency of data; use strategies, data, models; generate, extend, and modify procedures; use reasoning in new settings; and judge the reasonableness and correctness of solutions. Problem-solving situations require students to connect all of their mathematical knowledge of concepts, procedures, reasoning, and communication skills to solve problems.

Appendix 4

SAT Coding (* No NAEP items in this category)

SAT #1	content	ability	difficulty	DEV	
1	M	P	E	0	
2	N	C	E	-.531	
3	A	P	E	-.0625	
4	N	C	E	-.531	
5	A	C	E	-.5	
6	M	P	E	0	
7	A	PS	E	0	*
8	N	C	M	0	*
9	M	P	M	0	*
10	M	C	M	1.875	
11	A	P	E	-.0625	
12	N	C	M	0	*
13	G	PS	M	.838	
14	N	P	M	1.313	
15	A	C	H	.5	
16	N	C	M	0	*
17	D	PS	M	1.125	
18	N	PS	M	0	*
19	M	PS	H	.4	
20	A	PS	M	1.125	
21	A	C	H	.5	
22	N	P	H	.75	
23	M	PS	H	.4	
24	N	C	H	.25	
25	M	C	H	0	*
1	N	P	E	.05	
2	N	C	E	-.531	
3	N	C	E	-.531	
4	M	P	E	0	
5	M	C	M	1.875	
6	D	C	E	1.188	
7	A	C	E	-.5	
8	N	P	M	1.313	
9	M	PS	H	.4	
10	A	C	M	0	
11	A	C	M	0	
12	N	P	M	1.313	
13	M	PS	H	.4	
14	A	PS	H	.625	
15	M	PS	H	.4	
16	G	PS	E	0	*
17	A	P	M	-.375	
18	D	PS	E	-.125	
19	A	P	E	-.063	
20	A	C	M	0	
21	N	P	M	1.313	
22	N	PS	M	0	*
23	N	PS	M	0	*
24	G	PS	H	.688	
25	D	PS	H	.25	
1	M	C	E	-.1	
2	N	P	E	.04	
3	G	C	E	.2	
4	D	P	E	.65	
5	G	C	M	.25	
6	N	C	M	0	*
7	D	P	H	0	*
8	D	C	H	.15	
9	N	C	H	.2	
10	G	PS	H	.55	
Total				19.27	

SAT #2	content	ability	difficulty	DEV	
1	A	P	E	-.063	
2	N	C	E	-.531	
3	A	P	E	-.063	
4	A	PS	E	0	*
5	D	C	E	1.188	
6	G	C	E	.25	
7	G	C	E	.25	
8	A	P	E	-.063	
9	A	C	M	0	
10	A	P	M	-.375	
11	G	PS	M	.838	
12	N	P	M	1.313	
13	M	PS	M	.938	
14	M	PS	M	.938	
15	N	P	M	1.313	
16	N	PS	M	0	*
17	G	P	M	.875	
18	A	PS	M	1.125	
19	N	C	H	.25	
20	A	P	M	-.375	
21	D	PS	H	.25	
22	D	PS	H	.25	
23	D	C	H	.188	
24	A	PS	H	.625	
25	M	PS	H	.4	
1	D	P	E	.763	
2	A	C	E	-.5	
3	M	C	E	-.125	
4	N	C	M	0	*
5	M	C	E	-.125	
6	M	PS	M	.938	
7	N	P	M	1.313	
8	D	P	E	.763	
9	A	C	M	0	
10	M	PS	H	.4	
11	N	C	H	.25	
12	A	C	M	0	
13	A	C	M	0	
14	A	PS	M	1.125	
15	N	P	H	.75	
16	N	PS	E	-.125	
17	M	P	E	0	
18	A	P	E	-.063	
19	N	C	E	-.531	
20	G	P	E	0	*
21	A	P	M	-.375	
22	G	PS	M	.838	
23	N	PS	H	.563	
24	G	PS	H	.688	
25	D	PS	H	.25	
1	A	P	E	-.05	
2	N	C	E	-.4	
3	M	C	E	-.1	
4	N	PS	E	-.1	
5	D	P	E	.65	
6	A	P	M	-.3	
7	M	PS	M	.75	
8	N	C	H	.2	
9	A	PS	H	.5	
10	N	P	H	.6	
Total				18.06	

SAT #3	content	ability	difficulty	DEV	
1	A	P	E	-.063	
2	N	P	E	.05	
3	A	P	E	-.063	
4	N	C	E	-.531	
5	M	C	E	-.125	
6	N	P	E	.05	
7	N	C	E	-.531	
8	M	C	M	1.875	
9	N	P	E	.05	
10	N	C	H	.25	
11	M	C	M	1.875	
12	N	PS	M	0	*
13	A	P	M	-.375	
14	N	PS	E	-.125	
15	D	C	H	.188	
16	D	PS	M	1.125	
17	G	C	M	.313	
18	N	P	M	1.313	
19	N	C	M	0	*
20	M	PS	M	.938	
21	D	C	H	.188	
22	M	C	H	0	*
23	A	P	H	-.125	
24	A	PS	H	.625	
25	N	PS	H	.563	
1	N	P	E	.05	
2	N	P	E	.05	
3	N	C	E	-.531	
4	G	C	E	.25	
5	A	C	E	-.5	
6	N	C	M	0	*
7	N	P	M	1.313	
8	N	C	M	0	*
9	M	C	M	1.875	
10	N	P	M	1.313	
11	N	P	M	1.313	
12	N	P	M	1.313	
13	M	C	H	0	*
14	N	C	H	.25	
15	M	PS	H	.4	
16	M	C	E	-.125	
17	A	P	E	-.063	
18	N	C	E	-.531	
19	N	PS	M	0	*
20	N	P	M	1.313	
21	A	P	M	-.375	
22	N	P	M	1.313	
23	G	PS	H	.688	
24	A	PS	H	.625	
25	M	PS	H	.4	
1	N	P	E	.04	
2	M	P	E	0	
3	G	P	M	.7	
4	D	PS	M	.9	
5	M	PS	M	.75	
6	N	PS	M	0	*
7	M	PS	M	.75	
8	A	P	H	-.1	
9	N	P	H	.6	
10	M	PS	H	.32	
Total				21.76	

SAT #4	content	ability	difficulty	DEV	
1	A	P	E	-.0625	
2	N	C	E	-.531	
3	M	C	E	-.125	
4	A	P	E	-.063	
5	G	C	E	.25	
6	N	P	E	.05	
7	A	PS	E	0	*
8	M	P	E	0	
9	G	PS	M	.838	
10	N	PS	E	-.125	
11	N	PS	M	0	*
12	D	P	M	0	*
13	N	P	M	1.313	
14	M	PS	M	.938	
15	A	PS	M	1.125	
16	N	PS	M	0	*
17	A	P	M	-.375	
18	M	PS	M	.938	
19	N	PS	M	0	*
20	G	PS	M	.838	
21	A	P	M	-.375	
22	A	PS	H	.625	
23	N	P	H	.75	
24	D	PS	H	.25	
25	N	P	H	.75	
1	N	C	M	0	*
2	G	C	E	.25	
3	N	PS	E	-.125	
4	N	C	M	0	*
5	G	PS	E	0	*
6	N	PS	E	-.125	
7	A	PS	M	1.125	
8	A	PS	M	1.125	
9	D	C	M	.25	
10	M	C	M	1.875	
11	A	P	H	-.125	
12	M	C	H	0	*
13	A	PS	H	.625	
14	M	C	M	1.875	
15	N	C	H	.25	
16	A	P	E	-.063	
17	M	P	E	0	
18	A	P	E	-.063	
19	M	PS	M	.938	
20	N	PS	M	0	*
21	A	PS	M	1.125	
22	A	C	M	0	
23	M	PS	M	.938	
24	D	PS	H	.25	
25	D	PS	H	.25	
1	A	P	E	-.05	
2	N	P	E	.04	
3	G	C	E	.2	
4	A	C	E	-.4	
5	M	C	M	1.5	
6	D	P	M	0	*
7	D	C	M	.2	
8	N	PS	H	.45	
9	M	PS	H	.32	
10	D	PS	H	.2	
Total				19.84	

SAT #5	content	ability	difficulty	DEV	
1	N	P	E	.05	
2	A	P	E	-.063	
3	A	C	E	-.5	
4	G	PS	E	0	*
5	N	P	M	1.313	
6	A	P	E	-.063	
7	M	PS	E	0	*
8	N	C	E	-.531	
9	N	PS	M	0	*
10	A	P	M	-.375	
11	D	PS	M	1.125	
12	N	C	M	0	*
13	A	PS	M	1.125	
14	A	P	M	-.375	
15	N	P	M	1.313	
16	G	PS	M	.838	
17	A	PS	H	.625	
18	M	C	M	1.875	
19	A	P	M	-.375	
20	M	C	M	1.875	
21	A	C	H	.5	
22	M	P	H	0	*
23	A	C	H	.5	
24	G	C	H	0	*
25	D	PS	H	.25	
1	N	C	E	-.531	
2	M	C	E	-.125	
3	A	PS	M	1.125	
4	A	P	E	-.063	
5	A	C	M	0	
6	A	P	E	-.063	
7	N	C	M	0	*
8	A	C	M	0	
9	M	C	M	1.875	
10	N	P	M	1.313	
11	M	C	H	0	*
12	N	C	H	.25	
13	A	C	M	0	
14	M	P	H	0	*
15	G	C	H	0	*
16	A	P	E	-.063	
17	A	P	E	-.063	
18	G	PS	E	0	*
19	N	P	M	1.313	
20	A	C	M	0	
21	M	P	M	0	*
22	N	P	M	1.313	
23	D	C	M	.25	
24	D	C	H	.188	
25	N	PS	H	.563	
1	A	P	E	-.05	
2	D	C	E	.95	
3	M	C	E	-.1	
4	D	C	M	.2	
5	D	PS	H	.2	
6	N	PS	M	0	*
7	G	PS	M	.67	
8	M	PS	M	.75	
9	D	C	H	.15	
10	M	PS	H	.32	
Total				19.48	

SAT #6	content	ability	difficulty	DEV	
1	N	P	E	.05	
2	D	C	E	1.188	
3	N	C	E	-.531	
4	G	C	E	.25	
5	N	C	E	-.531	
6	M	P	E	0	
7	N	C	E	-.531	
8	D	P	E	.763	
9	G	P	M	.875	
10	N	PS	E	-.125	
11	G	C	M	.313	
12	N	P	M	1.313	
13	N	C	M	0	*
14	G	P	M	.875	
15	N	P	M	1.313	
16	M	PS	H	.4	
17	A	C	H	.5	
18	G	PS	H	.688	
19	A	P	H	-.125	
20	N	P	H	.75	
21	M	PS	H	.4	
22	N	P	M	1.313	
23	N	C	H	.25	
24	N	C	H	.25	
25	M	PS	H	.4	
1	N	P	E	.05	
2	N	P	E	.05	
3	N	C	E	-.531	
4	N	P	M	1.313	
5	A	C	M	0	
6	M	C	M	1.875	
7	N	C	E	-.531	
8	G	C	M	.313	
9	G	C	H	0	*
10	N	P	M	1.313	
11	N	C	H	.25	
12	N	C	H	.25	
13	A	P	E	-.063	
14	N	P	M	1.313	
15	D	P	H	0	*
16	A	P	E	-.063	
17	N	P	M	1.313	
18	A	P	E	-.063	
19	M	C	M	1.875	
20	N	C	M	0	*
21	M	PS	M	.938	
22	N	PS	H	.563	
23	A	PS	H	.625	
24	M	PS	H	.4	
25	D	PS	H	.25	
1	A	P	E	-.05	
2	D	P	E	.65	
3	N	P	M	1.05	
4	A	C	E	-.4	
5	A	C	M	0	
6	G	PS	M	.67	
7	D	C	M	.2	
8	N	P	M	1.05	
9	M	PS	M	.75	
10	G	C	H	0	*
Total				25.40	

SAT #7	content	ability	difficulty	DEV	
1	A	P	E	-.063	
2	M	C	E	-.125	
3	D	P	E	.763	
4	N	P	E	.05	
5	M	P	E	0	
6	A	P	E	-.063	
7	A	P	E	-.063	
8	N	PS	E	-.125	
9	A	P	E	-.063	
10	G	C	M	.313	
11	N	PS	M	0	*
12	G	C	E	.25	
13	N	PS	M	0	*
14	N	P	M	1.313	
15	N	PS	M	0	*
16	A	PS	M	1.125	
17	M	C	M	1.875	
18	N	P	M	1.313	
19	M	C	M	1.875	
20	D	C	H	.188	
21	G	C	H	0	*
22	N	PS	H	.563	
23	N	PS	H	.563	
24	D	PS	H	.25	
25	M	PS	H	.4	
1	A	P	E	-.063	
2	M	P	E	0	
3	D	C	E	1.188	
4	N	C	E	-.531	
5	M	C	M	1.875	
6	N	C	E	-.531	
7	A	P	M	-.375	
8	N	P	M	1.313	
9	N	P	M	1.313	
10	M	P	H	0	*
11	A	P	M	-.375	
12	N	P	H	.75	
13	G	PS	H	.688	
14	A	P	H	-.125	
15	M	C	H	0	*
16	M	C	E	-.125	
17	N	P	E	.05	
18	N	P	E	.05	
19	A	P	M	-.375	
20	M	C	M	1.875	
21	D	C	H	.188	
22	M	PS	M	.938	
23	N	PS	M	0	*
24	D	C	H	.188	
25	D	C	H	.188	
1	N	P	E	.04	
2	D	C	E	.95	
3	N	P	E	.04	
4	G	C	E	.2	
5	N	PS	M	0	*
6	M	PS	M	.75	
7	D	C	M	.2	
8	N	PS	H	.45	
9	M	PS	H	.32	
10	A	C	H	.4	
Total				21.79	

SAT #8	content	ability	difficulty	DEV	
1	A	P	E	-.063	
2	N	C	E	-.531	
3	A	P	E	-.063	
4	G	C	E	.25	
5	N	P	E	.05	
6	A	PS	E	0	*
7	N	P	E	.05	
8	M	PS	M	.938	
9	D	P	E	.763	
10	N	PS	M	0	*
11	N	P	E	.05	
12	G	C	M	.313	
13	N	P	M	1.313	
14	N	P	M	1.313	
15	N	C	H	.25	
16	A	P	M	-.375	
17	G	C	M	.313	
18	N	PS	M	0	*
19	N	P	M	1.313	
20	M	PS	H	.4	
21	N	P	M	1.313	
22	M	P	M	0	*
23	A	P	H	-.125	
24	D	C	H	.188	
25	G	C	H	0	*
1	N	P	E	.05	
2	G	C	E	.25	
3	N	C	E	-.531	
4	A	P	E	-.063	
5	G	C	E	.25	
6	N	P	E	.05	
7	M	C	E	-.125	
8	A	P	M	-.375	
9	G	C	H	0	*
10	N	C	H	.25	
11	D	C	M	.25	
12	A	P	H	-.125	
13	G	C	M	.313	
14	N	C	H	.25	
15	N	P	H	.75	
16	N	C	E	-.531	
17	M	P	E	0	
18	N	PS	M	0	*
19	N	P	E	.05	
20	A	PS	M	1.125	
21	M	C	M	1.875	
22	M	PS	M	.938	
23	A	P	H	-.125	
24	N	P	H	.75	
25	M	PS	H	.4	
1	A	P	E	-.05	
2	N	PS	E	-.1	
3	A	P	E	-.05	
4	M	P	M	0	*
5	D	P	E	.65	
6	D	P	M	0	*
7	N	PS	H	.45	
8	N	PS	H	.45	
9	M	PS	H	.32	
10	D	C	H	.15	
Total				15.15	

SAT #9	content	ability	difficulty	DEV	
1	N	P	E	.05	
2	N	P	E	.05	
3	N	C	E	-.531	
4	A	P	E	-.063	
5	M	C	E	-.125	
6	N	C	E	-.531	
7	M	P	E	0	
8	D	PS	M	1.125	
9	M	P	M	0	*
10	A	P	M	-.375	
11	D	P	E	.763	
12	A	P	M	-.375	
13	M	C	H	0	*
14	N	C	M	0	*
15	N	P	M	1.313	
16	G	PS	H	.688	
17	N	C	H	.25	
18	M	PS	H	.4	
19	N	C	M	0	*
20	N	P	M	1.313	
21	N	P	H	.75	
22	D	P	M	0	*
23	M	PS	H	.4	
24	D	C	H	.188	
25	N	PS	H	.563	
1	N	P	E	.05	
2	A	P	E	-.063	
3	N	C	E	-.531	
4	A	C	E	-.5	
5	M	C	E	-.125	
6	A	P	M	-.375	
7	A	C	M	0	
8	A	C	M	0	
9	A	C	M	0	
10	N	P	M	1.313	
11	M	C	M	1.875	
12	N	C	H	.25	
13	N	C	H	.25	
14	N	C	H	.25	
15	G	PS	H	.688	
16	A	P	E	-.063	
17	N	P	E	.05	
18	M	C	E	-.125	
19	A	C	E	-.5	
20	N	P	M	1.313	
21	M	C	M	1.875	
22	G	P	M	.875	
23	A	C	H	.5	
24	N	C	H	.25	
25	D	PS	H	.25	
1	N	P	E	.04	
2	A	P	E	-.05	
3	M	C	E	-.1	
4	A	PS	E	0	*
5	A	PS	M	1.125	
6	G	C	M	.313	
7	N	P	H	.75	
8	G	PS	H	.688	
9	N	C	H	.25	
10	N	PS	H	.563	
Total				16.93	

SAT #10	content	ability	difficulty	DEV	
1	N	P	E	.05	
2	D	P	E	.763	
3	A	P	E	-.063	
4	N	P	E	.05	
5	N	C	E	-.531	
6	M	P	E	0	
7	N	P	M	1.313	
8	A	P	M	-.375	
9	N	P	M	1.313	
10	M	C	M	1.875	
11	N	C	M	0	*
12	A	PS	M	1.125	
13	A	P	M	-.375	
14	A	P	M	-.375	
15	G	P	M	.875	
16	D	P	E	.763	
17	D	C	M	.25	
18	N	C	H	.25	
19	G	C	H	0	*
20	A	P	H	-.125	
21	N	P	H	.75	
22	M	C	H	0	*
23	N	P	M	1.313	
24	M	C	H	0	*
25	M	PS	H	.4	
1	A	C	E	-.5	
2	N	C	E	-.531	
3	N	C	E	-.531	
4	G	C	E	.25	
5	N	P	E	.05	
6	A	P	H	-.125	
7	D	C	E	1.188	
8	G	C	M	.313	
9	A	P	M	-.375	
10	M	C	H	0	*
11	M	C	M	1.875	
12	N	P	M	1.313	
13	M	C	M	1.875	
14	D	C	H	.188	
15	A	C	M	0	
16	A	P	E	-.063	
17	D	P	E	.763	
18	N	P	E	.05	
19	N	P	E	.05	
20	M	C	M	1.875	
21	D	C	M	.25	
22	M	C	M	1.875	
23	N	P	H	.75	
24	D	P	M	0	*
25	M	PS	H	.4	
1	A	P	E	-.05	
2	N	P	E	.04	
3	A	C	E	-.4	
4	N	PS	E	-.1	
5	M	C	M	1.5	
6	N	C	M	0	*
7	M	PS	H	.32	
8	G	C	M	.25	
9	N	PS	H	.45	
10	M	PS	H	.32	
Total				22.51	

Appendix 5

Table A. National Mean SAT/SAT I Scores for College-Bound Seniors, 1972-2001* (Recentered Scale)

Year	Verbal			Math		
	Male	Female	Total	Male	Female	Total
1972	531	529	530	527	489	509
1973	523	521	523	525	489	506
1974	524	520	521	524	488	505
1975	515	509	512	518	479	498
1976	511	508	509	520	475	497
1977	509	505	507	520	474	496
1978	511	503	507	517	474	494
1979	509	501	505	516	473	493
1980	506	498	502	515	473	492
1981	508	496	502	516	473	492
1982	509	499	504	516	473	493
1983	508	498	503	516	474	494
1984	511	498	504	518	478	497
1985	514	503	509	522	480	500
1986	515	504	509	523	479	500
1987	512	502	507	523	481	501
1988	512	499	505	521	483	501
1989	510	498	504	523	482	502
1990	505	496	500	521	483	501
1991	503	495	499	520	482	500
1992	504	496	500	521	484	501
1993	504	497	500	524	484	503
1994	501	497	499	523	487	504
1995	505	502	504	525	490	506
1996	507	503	505	527	492	508
1997	507	503	505	530	494	511
1998	509	502	505	531	496	512
1999	509	502	505	531	495	511
2000	507	504	505	533	498	514
2001	509	502	506	533	498	514

*For 1972-1986 a formula was applied to the original mean and standard deviation to convert the mean to the recentered scale. For 1987-1995 individual student scores were converted to the recentered scale and then the mean was recomputed. From 1996-1999, nearly all students received scores on the recentered scale. Any score on the original scale was converted to the recentered scale prior to computing the mean. For 2000 and 2001, all scores are reported on the recentered scale.

16632-F426.SRSpg2.01/29/01

Appendix 6
Total Predicted DEV for Categories (per instrument)

TDEV = DEV x (Average # items per instrument)
(Other categories were either not found on the SAT or not covered by NAEP)

Category (content/ability/difficulty)	TDEV
NPM	5.909
MCM	4.875
MPSM	1.689
APSM	1.463
MPSH	1.08
NPH	1.05
DPE	.991
DCE	.832
NPSH	.732
GPSM	.67
NCH	.575
DPSM	.563
APSH	.563
GPSH	.55
GPM	.525
ACH	.35
GCE	.35
GCM	.344
DPSH	.325
DCH	.263
DCM	.225
NPE	.18
MPE	0
ACM	0
DPSE	-.013
NPSE	-.125
APH	-.125
MCE	-.2
APE	-.277
ACE	-.55
APM	-.9
NCE	-1.5

References

Achieve (2004). *The expectations gap: A 50-state review of high school graduation requirements.*Achieve, Inc.

Anastasi, A.(1958). *Differential Psychology: Individual and Group Differences in Behavior.* New York: Macmillan.

Ames, C. & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology,* 80, 260-270.

Baenninger, M. & Newcombe, N (1989). The role of experience in spatial test performance: A meta-analysis. *Sex Roles,* 20, 327-344.

Baker, D.P., & Jones, D.P. (1992). Opportunity and performance: A sociological explanation for gender differences in academic mathematics. In J. Wrigley (Ed.), *Education and gender equality* (pp. 193-203) London: Falmer.

Bandura, A.(1989). Human agency in social cognitive theory. *American Psychologist,* 44(9), 1175-1184.

Baron-Cohen, S. (2003). *The essential difference: The truth about the male and female brain.* New York: Basic Books.

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

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

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

Ben-Zeev, T., Fein, S., & Inzlicht, M. (2005). Arousal and Stereotype Threat. *Journal of Experimental Social Psychology,* 41, 174-181.

Blascovich, J., Spencer, S.J., Quinn, D., & Steele, C. (2001). African Americans and High Blood Pressure: The Role Of Stereotype Threat. *Psychological Science,* 12(3), 225-229.

Bridgeman, B. & Wendler, C. (1991). Gender differences in predictors of college mathematics performance and in college mathematics course grades. *Journal of Educational Psychology,*83,(2), 275-284.

Byrnes, J.P. & Takahira, S. (1993). Explaining Gender Differences on SAT-Math Items. *Developmental Psychology,* 29(5), 805-810.

Broadbent, D.E., Cooper, P.E., FitzGerald, P., & Parkes, K.R. (1982). The Cognitive Failures Questionnaire (CFQ) and its correlates. *British Journal of Clinical Psychology,* 21, 1-16.

Casey, M. B., Nuttal, R., Pezaris, E., & Benbow, C. (1995). The influence of spatial ability on gender differences in mathematics college entrance test scores across diverse samples. *Developmental Psychology, 31,*679–705.

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

Cohen, J. (1992). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
The College Board (1997). *10 Real SATs.* Forrester Center, WV: College Board Publications.
The College Board (1998a). *Inquiring About Examinee's Ethnicity and Sex: Effects on AP Calculus AB Examination Performance.* Report No. 98-1.
The College Board (1998b). Research summary, *Office of Research and Development,* RS-04.
The College Board (2006). *2006 college bound seniors: A profile of SAT program test takers.*http://www.collegeboard.com/prod_downloads/about/news_info/cbsenior/yr2006/national-report.pdf
Connellan, J., Baron-Cohen, S., Wheelwright, S., Batki, A., & Ahluwalia, J. (2000). Sex differences in human neonatal social perception. *Infant Behavior & Development, 23,* 113–118.
Department of Education, National Center for Educational Statistics (1997). The NAEP Guide: A description of the content and methods of the 1994 and 1996 assessments. Available at http://nces.ed.gov/pubs/97586.pdf.
Department of Education, National Center for Educational Statistics (2005). National Assessment of Educational Progress data tool, 2005. Available from The Nation's Report Card Web site, http://nces.ed.gov/nations reportcard/naepdata.
Doolittle, A.E. & Cleary, T.A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement,* 24, 157-166.
Dweck, C. (1986). Motivational processes affecting learning. *American Psychologist,* 41, 1040-1048.
Dweck, C. (1999). *Self-theories: Their role in motivation, personality, and development.* Philadelphia: Psychology Press.
Dweck, C. & Elliott, E.S. (1983). Achievement motivation. In P.H. Mussen (Ser. Ed.) & E.M. Heatherington (Vol. Ed.), *Handbook of child psychology: Vol. 4, Socialization, personality, and social development* (4th ed., pp. 643-691). New York: Wiley.
Dwyer, C.A. (1979). The role of tests and their construction in producing apparent sex-related differences. In M.A. Wittig & A.C. Petersen (Eds.) *Sex-Related Differences in Cognitive Functioning: Developmental Issues* (pp. 335-353). New York: Academic Press.
Eccles, J.S. & Jacobs, J.E. (1986). Social forces shape math attitudes and performance. *Journal of Women in Culture and Society,* 11, 367- 380.
Elliott, E.S. & Dweck, C. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology,* 54, 5-12.
Engle, R.W. (2001). What is working memory capacity? In H.L. Roediger III & J.S. Nairne (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 297-314). Washington, DC: American Psychological Association.
Eysenck, M.W. (1992). *Anxiety: The Cognitive Perspective.* Hove, UK: Lawrence Erlbaum Associates.

Feingold, A. (1988). Cognitive Gender Differences Are Disappearing. *American Psychologist,* 43(2), 95-103.
Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research,* 62 (1), 61-48.
Feingold, A. (1994). Gender differences in variability in intellectual abilities: A cross-cultural perspective. *Sex Roles: A Journal of Research*, 30, 81-92.
Fennema, E., Carpenter, T.P., Jacobs, V.R., Franke, M.L., & Levi, L.W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher,* 27, 6-11.
Fennema, E. & Peterson, P. (1985). Autonomous learning behavior: A possible explanation of gender-related differences in mathematics. In L.C. Wilkinson & C.B. Marrett (Eds.) *Gender influences in classroom interaction* (pp. 17-35). Orlando, FL: Academic Press.
Gallagher, A.M. (1992). *Sex Differences in Problem-Solving Strategies Used by High-Scoring Examinees on the SAT-M.* College Board.
Gallagher, A.M. & De Lisi, R. (1994). Gender Differences in Scholastic Aptitude Test-Mathematics Problem Solving Among High-Ability Students. *Journal of Educational Psychology*, 86(2), 204-211.
Gallagher, A.M., De Lisi, R, Holst, P.C., McGillicuddy-DeLisi, A.V., Morely, M. & Cahalan, C. (2000). Gender Differences in Advanced Mathematical Problem Solving. *Journal of Experimental Child Psychology,* 75, 165-190.
Gallagher, A. M., & Kaufman, J. C. (2005). *Gender differences in mathematics.* New York: Cambridge University Press.
Garner, M. & Engelhard, G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education,* 12, 29-51.
Gierl, M.J., Bisanz, J., Bisanz, G.L., & Boughton, K.A. (2002). *Identifying content and cognitive Skills that produce gender differences in mathematics: A demonstration of the DIF Analysis framework.* Paper presented at the Annual meeting of the National Council on Measurement in Education, New Orleans, La.
GRE Board (1999). *Test Difficulty and Stereotype threat on the GRE General Test.* Report No. 96-06R.
Harder, J.A. (1999). *The effect of private versus public evaluation on stereotype threat for women in mathematics.* Dissertation, University of Texas at Austin.
Harris, A.M. & Carlton, S.T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education, 6, 137-151.*
Hambleton, Ronald & Rodgers, Jane (1995). Item bias review. *Practical Assessment, Research & Evaluation*, 4(6)
Holland, P.W. & Wainer, H. (1993). *Differential Item Functioning.* Hove, UK: Lawrence Erlbaum Associates.
Hogg, M., Terry, D. & White, K. (1995). A tale of two theories: A critical comparison of identity theory with social identity theory. *Social Psychology Quarterly, 58*, 255-269.

Hoover, E. (2006). Captain Caperton. *The Chronicle of Higher Education,* 52, week 43 (June 30).
Hyde, J.S., Fennema, E., & Lamon, S.J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin,* 107(2), 139-155.
Janowsky, J.S., Oviatt, S.K., & Orwoll, E.S. (1994). Testosterone influences spatial cognition in older men. *Behavioral Neuroscience,* 108, 325-332.
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, 180-186.
Kimball, M.M. (1989). A new perspective on women's math achievement. *Psychlogical Bulletin,* 105(2), 198-214.
Lane, S., Wang, N., & Magone, M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational Measurement: Issues and Practice,* Winter, 21-27.
Lawrence, I.M., Curley, W.E, & McHale, F.J. (1988). *Differential Item Functioning for Males and Females on SAT-Verbal Reading Subscore Items.* Colleg Board Report No. 88-4.
Leahy, E. & Guo, G. (2001). Gender differences in mathematical trajectories. *Social Forces,* 80, 713-732.
Lemann, N. (1999). *The Big Test.* New York: Farrar, Straus, and Giroux.
Leggett, E. (1985). *Children's entity and incremental theories of intelligence: Relationships to achievement behavior.* Paper presented at the meeting of the Eastern Psychological Association, Boston.
Li, Y, Cohen, A.S., & Ibarra, R.A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing,* 4, 115-136.
Maccoby, E.E., & Jacklin, C.N. (1974). *The Psychology of Sex Differences.* Stanford: Stanford University Press.
McGraw, R., Lubienski, S.T., & Strutchens, M.E. (2006). A closer look at gender in NAEP mathematics achievement and affect data: Intersections with achievement, race/ethnicity, and socioeconomic status. *Journal for Research in Mathematics Education,* 37, 129-150.
McGlone, M.S. & Aronson, J. (2006). Stereotype threat, identity salience, and spatial reasoning. *Applied Developmental Psychology,* in press.
Monastersky, R. (2005). Primed for numbers. *Chronicle of Higher Education.*
Murphy, R.J.L. (1977). *Sex differences in examination performance: Do these reflect differences in ability or sex-role stereotypes?* Paper presented at the International Conference on Sex-Role Stereotyping, Cardiff, Wales, July.
Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF, *Journal of Educational Measurement,* 30, 293-311.
National Assessment Governing Board (2002). *Mathematics Framework for the 2003 National Assessment of Educational Progress.* Washington, D.C.: U.S. Government Printing Office.
Osborne, J.W. (2001). Testing Stereotype Threat: Does Anxiety Explain Race and Sex Differences in Achievement? *Contemporary Educational Psychology,*26,291-310.

Paek, P.L. (2002). Problem solving strategies and metacognitive skills on SAT mathematics items. (Doctoral dissertation, University of California, Berkeley, 2002). Dissertation Abstracts International 63 (09), 3139.

Quinn, D., & Spencer, S. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues,* 57(1), 55-71.

Royer, J.M., Tronsky, L.N., Chan, Y., Jackson, S.J., & Marchant, H. (1999). Math-fact retrieval as the cognitive mechanism underlying gender differences in math test performance. *Contemporary Educational Psychology,* 24, 181-266.

Sarason, I.G. (1984). Stress, anxiety and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology,* 46, 929-938.

Sarason, I.G. & Sarason, B.R. (1981). The importance of cognition and moderator Variables in stress. In D. Magnusson (Ed.), *Towards a psychology of situations: An interactive perspective* (pp. 195-210). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schmader, T. & Johns, M. (2003). Converging Evidence that Stereotype Threat Reduces Working Memory Capacity. *Journal of Personality and Social Psychology,* 85(3), 440-452.

Selye, Hans, M.D. *The Stress of Life*. New York: McGraw-Hill, 1956.

Shih, M., Pittinsky, T.L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science,* 10(1), 80-83.

Spencer, S.J., Steele, C., & Quinn, D.M. (1998). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology,* 35, 4-28.

Stage, E.K., Kreinberg, N., Eccles, J., & Becker, J.R. (1985). Increasing the participation and achievement of girls and women in mathematics, science, and engineering. In S.S. Klein (Ed), *Handbook for Achieving Sex Equity through Education.* Baltimore: Johns Hopkins University Press.

Standards for Educational and Psychological Testing. (1999). Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Steele, C., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology,* 69(5), 797-811.

Steele, C. (1997). How stereotypes shape intellectual identity and performance. *American Psychologist,* 52(6), 613-629.

Stipek, D.J. & Kowalski, P.S. (1989). Learned helplessness in task-orienting versus performance-orienting testing conditions. *Journal of Educational Psychology,* 81, 384-391.

Turner, M.L. & Engle, R.W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language,* 28, 127-154.

Van Goozen, S.H.M., Cohen-Kettenis, P.T., Gooren, L.J.G., Frijda, N.H., & Van De Pol, N.E. (1995). Gender differences in behavior: Activating effects of cross-sex hormones. *Psychoneuroendocrinology,* 20, 343-363.

Vandenberg, S. G. & Kuse, A. R. (1978). Mental rotation, a group test of three-dimensional spatial visualization. *Perceptual & Motor Skills*, 47, 599-604.

Vasta, R., Knott, J.A., &Gaze, C.E. (1996). Can spatial training erase the gender differences on the water-level task? *Psychology of Women Quarterly,* 20, 549-568.

Wainer, H. & Steinberg, L.S. (1992). Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: A bidirectional validity study. *Harvard Educational Review,* 62, 323-336.

Walton, G.M. & Cohen, G.L. (2002). Stereotype lift. *Journal of Experimental Social Psychology,* 39, 456-467.

Wegner, D.M. (1994). Ironic Processes of Mental Control. *Psychological Review,* 101(1), 34-52.

Weinstein, C.E. (2005). In class discussion from **Current Topics in Cognition**, Fall 2005.

Wigfield, A. & Eccles, J.S. (1989). Test Anxiety in Elementary and Secondary School Students. *Educational Psychologist,* 24(2), 159-183.

Wilder, G.Z. (1997). Antecedents of gender differences. In *Supplement to Gender and fair assessment.* Princeton, NJ: Educational Testing Service.

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

Xie, Y., & Shauman, K. (2003). *Women in science: Career processes and outcomes.* Cambridge, MA: Harvard University.

Yerkes, R.M. & Dodson, J.D. (1908) The Relation of Strength of Stimulus to Rapidity of Habit-Formation. *Journal of Comparative Neurology and Psychology,* 18, 459-482.

VITA

Bryan Nankervis was born in Cambridge, England on April 9, 1956, the son of Rosemary Nankervis and Jack Nankervis. After completing his work at Marshall High School, San Antonio, Texas, in 1974, he entered Southwest Texas State University in San Marcos, Texas. He received the degree of Bachelor of Arts in Mathematics from Southwest Texas State University in August 1978. He entered graduate school at Southwest Texas State University and received the degree of Master of Science in Mathematics and Bachelor of Arts in Anthropology in December 1985. He also served in the United States Navy and retired in 1991 as a Lieutenant Commander. He worked as an instructor of mathematics at Southwest Texas State University, Texas Lutheran University, Lamar University-Port Arthur, and Angelina College. In September 2001 he entered the Graduate School of The University of Texas.

Permanent Address: 1203 W. San Antonio, San Marcos, Texas 78666

This dissertation was typed by the author.