Sex Differences in the Performance of High-Scoring Examinees on the SAT®-M

Ann M. Gallagher

College Board Report No. 90-3
ETS RR No. 90-27

College Entrance Examination Board, New York, 1990

Ann M. Gallagher is a Research Associate at Educational Testing Service.

The College Board is a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,700 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the programs of the College Board and participate in the determination of its policies and activities.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101-0886. The price is $7.

Printed in the United States of America.

CONTENTS

Figures

Tables

ABSTRACT

Performance of high-scoring males and females on the mathematics section of three forms of the College Board's Scholastic Aptitude Test (SAT-M) was examined to determine how item content, solution strategy, and speededness differentially affect performance. The mathematical and verbal sections of the SAT were also compared for similarities in the performance patterns of high scorers.

Items on three forms of the SAT-M were classified on the basis of content and potential solution strategies. Male and female performance in the top 5 percent was examined using the Mantel-Haenszel Differential Item Functioning (MH-DIF) procedure. A main effect was found for item-solution strategy, but not for item content. Males tended to outperform females on items requiring logical estimation or insight even when subjects were matched on years of mathematics course taking and average mathematics grades.

Conventional measures indicated that the test was not differentially speeded. However, omit rates indicated that females omitted a greater proportion of items requiring estimation. Different patterns by gender were found on the mathematical and verbal sections of the test.

RESEARCH

Previous research has shown consistent differences favoring males in performance on the mathematics sections of many standardized achievement tests administered to adolescents. These studies are in contrast with research examining grades in college-level mathematics classes (Clark and Grandy 1984) or studies investigating the ability to apply learned knowledge (Fennema 1974; Senk and Usiskin 1983) where no differences in performance of males and females were found among high-school-age students. It is unclear whether males as a group are generally more able in mathematics than females, or whether other examinee characteristics and/or test characteristics are the cause of the disparity in test scores.

Early research in the area of gender differences in mathematical performance documented differences that occurred at various ages or in specific areas of mathematics. Maccoby and Jacklin (1974), for example, reviewed studies that were performed between 1961 and 1972. Results of their tabulations indicated that differences favoring males are generally not found for elementary school students but emerge during adolescence and occur fairly consistently for high school and college students. Studies examining subjects in specific ability groups (Benbow and Stanley 1980) found that differences favoring males are greater at the higher end of the continuum than for students of average ability.

Recent efforts to examine gender differences in mathematics have attempted to validate hypotheses concerning the origins of observed differences. The hypotheses that attribute performance to cognitive factors can be grouped into two categories: (1) hypotheses that ascribe differences to males' and females' different background experiences, and (2) hypotheses that attribute differences in mathematics to gender differences in problem-solving style or the approach to learning.

A large number of the studies examining the effects of background on mathematics performance have focused on gender differences in mathematics course taking. It is well documented that males take more mathematics courses than females (Armstrong 1981; Benbow and Stanley 1982; Fennema and Sherman 1977). However, in a review of these and other studies, Kimball (1989) points out that even when course taking is controlled, sex-related differences in mathematics performance are not eliminated.

Several studies have found gender differences in the attention received by males and females in the classroom. Males at all levels receive more attention than females, and the difference appears to be more pronounced for high achievers (Brophy 1985; Eccles and Blumenfeld 1985; Leinhardt, Seawald, and Engle 1979). In a study examining types of teacher attention, Leder (1987) found that females received more attention on product-oriented issues, whereas males received more attention on process-oriented issues.

Another factor that has been found to be related to performance in mathematics is confidence in one's own mathematical ability. Several studies have found that males are more confident of their mathematical ability than females, even for high-performing students (Fox, Brody, and Tobin 1985; Hudson 1986).

Studies testing hypotheses that gender differences in mathematics are related to differences in problem solving or approaches to learning have found that males tend to outperform females on items requiring problem-solving skills or items that require the examinee to set up the problem (Doolittle and Cleary 1987; Marshall 1984; McPeek and Wild 1987; O'Neill, Wild, and McPeek 1989). The same studies also found that females outperformed males on items requiring application of a clearly defined algorithm.

It is hypothesized that these gender differences in mathematical "reasoning," or the ability to apply mathematical principles in order to set up and solve a problem, are the result of males' and females' different approaches to learning mathematics (Grieb and Easley 1984). According to this hypothesis, the learning process for males is fairly independent of classroom assignments, whereas females rely substantially more on algorithms taught in class and procedures outlined by the teacher (Kimball 1989). According to Grieb and Easley (1984), autonomous learning exhibited by males allows them to develop their own solutions to problems without relying on algorithms provided by the teacher. Although this hypothesis presents a logical explanation for gender differences in problem solving and why females, who receive grades that are equal to or higher than those of males in mathematics classes, also tend to receive lower scores on standardized tests of mathematics, there is little empirical evidence to support this hypothesis.

Another hypothesis relating performance on tests of mathematics to problem-solving strategy holds that females rely more on verbal strategies whereas males tend to use deductive, gestalt-type strategies that enable them to see the solution without actually working the problem out. Dorans and Livingston (1987) examined gender differences in mathematical and verbal scores on the SAT. According to this hypothesis, females should attain higher SAT-V scores at each SAT-M score level, but the variance of SAT-V scores would be higher for males than for females. Dorans and Livingston examined data for examinees scoring 600 or above over two administrations of the SAT. Females had higher mean SAT-V scores than males had at each score level on the SAT-M. However, the standard deviation of the males' SAT-V scores was no larger than that of the females.

Finally, the timed nature of standardized tests of mathematical ability may affect males and females differently. Rindler (1979) suggests that time limits may affect response accuracy. Accuracy under time constraints may be affected differently for males and females. Evans (1980) examined the performance of males and females on SAT items. The findings showed no overall sex differences on the mathematical sections and no advantage for females in the greater time condition. However, the conditions under which the test was administered and the sample that was used were not representative of the actual conditions and sample of the SAT. Findings of this study, therefore, may not generalize to the conditions or testing population of the SAT.

Other studies suggest that a competitive environment may be conducive to males' but detrimental to females' learning in mathematics (Eccles and Blumenfeld 1985; Peterson and Fennema 1985). The pressures of competitive performance may be similar to the pressures of a timed testing situation. It is possible that time constraints may interact with females' lower self-confidence in mathematics to inhibit their performance.

It appears, then, that gender differences in performance on standardized tests of mathematical ability are probably the result of a combination of background variables and differences in approaches to problem solving and learning in mathematics. Males may possess a combination of these characteristics that allows them to outperform females in competitive circumstances such as standardized tests of mathematical ability, but that does not generally give them an advantage on less-competitive measures such as grades in mathematics courses.

Most research that has examined differences in male and female performance on the SAT-M or on other standardized tests of quantitative ability has examined differences in the examinee population as a whole and not at specific levels of that population such as high performers (e.g., Clark and Grandy 1984; Doolittle and Cleary 1987; Dossey et al. 1988; O'Neill, Wild, and McPeek 1989). Although Benbow and Stanley (1980) have examined high performers, the population used in their investigations is highly specialized. All subjects are younger than the usual SAT population and may not yet have had the opportunity to study high school mathematics. The strategies they use to solve mathematics problems on the SAT may be different from those of mathematically talented high school students because the former have not yet had the opportunity to learn many of the algorithms.

OBJECTIVES

There were three major objectives of this study. The first was to identify item content and solution strategies that may contribute to the differential performance of high-scoring males and females on the SAT-M. As part of this process, a coding system was developed to identify items solved more efficiently by estimation or logical strategies than by using an algorithm. The coding system was applied to SAT-M items, and statistical analyses were performed to determine whether males perform better than females on items for which use of an estimation or logical strategy is required.

The second part of the study addressed the issue of differential speededness. In this part, patterns of omitted and not reached items were examined for males and females to determine whether females' poorer performance could be attributed to time limitations.

The third, and final, part of the study examined the verbal performance of high-scoring males and females to determine whether the relationship between mathematical and verbal scores is different for males and females in this group. In addition, this part of the study compared differences found between males and females scoring in the top 5 percent on the SAT-V with those found in the same group on the SAT-M. The purpose of this comparison was to determine whether similar patterns of differences between the top 5 percent and the total group of examinees on the SAT-M are found on the SAT-V.

ANALYSIS OF SAT-M ITEM TYPES

Method

Subjects

Subjects for this part of the study were high school juniors or seniors who considered English to be their first language and had taken the SAT at the most recent administration of each of three test forms. Since the number of females scoring in the top 5 percent of all examinees is about three times smaller than the number of males, it was decided that the lower cutoff score for subjects in this analysis would be based on the ninety-fifth percentile score for the female population for each administration. Examinees were selected if they scored at or above 650 on the SAT-M in May 1987 ($n = 18,683$ males and 7,577 females) or November 1987

($n = 24,984$ males and 12,157 females) and at or above 660 in May 1988 ($n = 19,145$ males and 7,650 females).

Item Classification

Forms from three of the most recently disclosed, largest volume administrations were selected for analysis: May 1987, November 1987, and May 1988. Items on the three forms were classified according to their primary content and the types of strategies that could be used in a successful solution. Each item received one numeric classification for content and one alphabetic classification for strategy. Categories for content were:

1. Arithmetic
2. Algebra
3. Geometry
4. A combination of algebra and geometry.

Categories for solution strategies were:

A Requires the use of an established (school-taught) algorithm

B Either an algorithm or a logical/estimation strategy can be used

C Requires the use of an algorithm, but also requires insight

D Can be solved only with a logical/estimation strategy and cannot be solved using a school-learned algorithm.

Two raters, one male and one female, worked independently to classify items. Both raters have worked extensively in developing items for the quantitative section of the SAT and have worked with other item classification systems in the past. After all the items in each form had been classified by raters independently, items receiving discrepant ratings were discussed, and agreement was reached on most items. Prior to discussion, only about 50 percent of the items received identical classifications from both raters. This was due to the tendency of one rater to approach the problems in a more algorithmic fashion than the other. However, after discussing the rationale for classifications, rater agreement increased to 99 percent.

Item Analysis: DIF Procedure

The Mantel-Haenszel Differential Item Functioning procedure (MH-DIF) was used to identify items on which males and females performed differently. In this procedure males and females are matched on total SAT-M scores and an odds-ratio, which compares the number of males and females answering each item either correctly or incorrectly, is calculated at each score level. The ratios are summed across score levels using weights that are a function of the frequencies at those score levels. The weighted average odds-ratio is then tested for significance using a chi-square goodness of fit test. For example, the number of females who got item x right is divided by the number who got that item wrong at score 600. This same ratio is calculated for males and is then divided by the ratio for females. This comparison is performed at each score level (200, 210 . . . 790, 800) for each item. In practice, the weighted average odds-ratio is transformed to a scale that is meaningful to test developers. The ratio can be transformed to a percent-correct difference scale or a delta scale.

Two types of DIF statistics were used to determine differences in item performance: the Mantel-Haenszel P-DIF (MH P-DIF) and the Mantel-Haenszel D-DIF (MH D-DIF). Both statistics are logistic transformations of the MH odds-ratio described above. If both statistics for any particular item were greater than the criterion, then the item was flagged as a differentially functioning item.

The P-DIF statistic is based on the percent correct (P+). Assuming a population is normally distributed, the placement of percent values along an ability continuum would be unevenly spaced, with percent values around the mean (50 percent) lying closer together than values at the extremes. As a result, the P-DIF statistic is most sensitive to differences in items of middle difficulty for which small differences on the ability continuum are equivalent to a larger area under the normal curve than differences at either end of the scale. The D-DIF statistic, on the other hand, is based on delta, which is a normal transformation of percent correct (much like a z value) with a mean of 13 and a standard deviation of ± 4. Because delta is a linear scale, the position of values along the ability continuum is not affected by the area under the normal curve; on the ability scale, then, the distance between two delta points around the mean is the same as the distance between two delta points at the extremes. D-DIF, therefore, is more sensitive than P-DIF to differences in items of high or low difficulty.

The examinees in this part of the study all scored at or above 650, which restricted the range of possible scores to 150 points as opposed to the standard 600-point range. This restriction of range in score was reflected in relatively small DIF values. Consequently, a relatively small criterion value was used to identify differentially functioning items. For the group being examined, the percent correct for each item was generally high. It was therefore decided that the P-DIF statistic (the more conservative indicator for easy items) should be the first criterion for flagging items because very small differences in performance on easy items could result in high DIF values of D-DIF. Items were flagged when the MH P-DIF was greater than .05. The second criterion used was a value of MH D-DIF at or above 0.5. This was used in conjunction with the P-DIF statistic in order to ensure that an equally conservative indicator was used for items of middle difficulty as well.

Results

Table 1 shows the frequency distribution of item types across all three test administrations. The most commonly found type of item was type A (algorithm), with C (algorithm + insight) and D (logic alone) following closely be-

Table 1. Frequency Distribution of Item Types across Three Test Administrations

Item Type	*Frequency*	*Percentage of Test*
A (algorithm)	58	32
B (algorithm or logic/ estimation)	25	14
C (algorithm + insight)	48	27
D (only logic/estimation)	42	23
Unclassified	7	4

Table 2. Mean P+ Values across Item Type and Test Administration

	Test Administration		
Type	*May 1987*	*Nov. 1987*	*May 1988*
A	.96	.94	.96
B	.94	.96	.91
C	.74	.76	.78
D	.84	.82	.87

Table 3. Distribution of Items Flagged for Total Group Analysis

	Favored Sex		
Type	*Males*	*Females*	*Total*
A	1	2	3
B	1	0	1
C	6	2	8
D	5	1	6
Total	13	5	18

hind. Relatively few items were classified B (algorithm *or* logic), and seven items failed to fall clearly into any of the four categories.

When placement of the items in the test was examined, it was apparent that across the three forms A and B items were more frequently placed in the first part of each section (constituting from 53 percent to 70 percent of the items in the first half of each section), and C and D items were predominantly at the end (constituting from 20 percent to 36 percent of the items in the first half of each section). Item location in the SAT is determined by an item difficulty statistic that is calculated when an item is pretested (administered as an unscored section of a previous SAT). Each section in the SAT is arranged so that easy items are located at the beginning of the section and difficult items at the end. Consequently, one could conclude that for the total group of examinees at any given administration, items classified as C or D are generally more difficult than those classified as A or B.

This was the case for the subjects in the present study. Thirty-six percent of all C items and 27 percent of all D items had P+ values below .8 (less than 80 percent of the examinees answered correctly). In contrast, less than 10 percent of all A and B items had P+ values below .8. Table 2 presents the mean P+ values for each type of item. Across the three test forms, mean values of A and B item types are larger (indicating that a greater proportion of the sample answered them correctly) than item types C and D.

Two sets of DIF analyses were run. The first set matched examinees from all three administrations on their total SAT-M scores without controlling for mathematics background variables. The second set of analyses matched examinees on total scores as well as relevant background variables.

Results from the first set of analyses flagged a total of 18 items of the 180 items across the three administrations; five items each from May and November 1987, and eight items from May 1988. All of the flagged items demonstrated differences in male and female performance that were significant at the $p \leq .001$ level on the Mantel-Haenszel chi-square test.

Table 3 lists flagged items by type and by direction of sex difference. With few exceptions, on flagged items of type C or D, males outperformed matched females. Fourteen of the 18 items flagged were either C or D items, and 11 of the 14 favored males. There was little difference in performance on item types A and B, with one in each category favoring males and two type A items favoring females. To test the significance of these findings, a binomial test (a test similar to chi-square that can be performed on data with small cell sizes) was run on (1) items that could be solved using an algorithm (categories A and B) and (2) items that require insight or estimation (categories C and D). The difference in the number of insight or estimation items favoring males over females was significant ($p \leq .05$). No significant difference was found for items that could be solved using an algorithm.

Contrary to previous reports, there was no indication that item content or format influenced differential performance. The primary content of four of the five items on which females outperformed males was geometry (which has previously been shown to favor males). Items favoring males were evenly distributed across algebra and geometry. Finally, only 3 of the 18 flagged items were word problems (a format variable previously found to favor males).

The second set of analyses was run controlling for the number of years examinees had studied high school mathematics, the types of mathematics courses taken, and reported mathematics grades. Frequency distributions of each of these variables were examined across sex and score levels to determine whether there was any systematic variation. There was little male-female variation in mathematics background. The majority (about 80 percent) of both males and females in the sample had taken at least four years of high school math. When course-taking behavior was examined across all score levels, about half of each sex had studied two years or more of algebra, about 95 percent had studied less than two years of geometry, and about 92 percent had studied some trigonometry.

Calculus was the only mathematics course that appeared to have some relationship to high scoring for both males and females. Although the majority of examinees had taken calculus, as scores increased, so did the percentage of examinees who had taken calculus: about 72 percent at scores between 650 and 700; about 83 percent between 700 and 750; and about 91 percent between 750 and 800.

Across the three administrations, between 77 percent and 81 percent of examinees in the sample reported receiving mathematics grades of A or B. The majority of that group (60 percent) reported that they generally received A's.

On the basis of the above findings, the second set of DIF analyses was performed on examinees who had studied at least four years of high school mathematics, controlling for calculus (studied or not) and average math grade (A or B). Six analyses were performed for the following groups in each test administration:

1. All examinees who had calculus (CALC);
2. All examinees who did not have calculus (NOCALC);
3. Examinees who had calculus and reported an average mathematics grade of A (CALC-A);
4. Examinees who had calculus and reported an average mathematics grade of B (CALC-B);
5. Examinees without calculus who reported an average mathematics grade of A (NOCALC-A);
6. Examinees without calculus who reported an average mathematics grade of B (NOCALC-B).

A total of 36 items were flagged across the three administrations in the analyses by gender. These 36 fell into the following categories:

A Requires use of an established algorithm ($N = 7$)
B Algorithm *or* a logical/estimation strategy can be used ($N = 1$)
C Requires insight, but *also* requires use of an algorithm ($N = 18$)
D Can be solved *only* with a logical/estimation strategy ($N = 8$)
X Does not clearly fall into any of the four categories ($N = 2$)

Table 4 presents distributions of flagged items by item type and the four calculus by mathematics grade groupings. As the table illustrates, the fewest items were flagged for the CALC-A group (16 items across the three administrations). However, the ratio of items favoring males to those favoring females for this group was the highest of all groups at 3 to 1. Although a greater number of items were flagged for other groupings of examinees (23 to 27), the ratio of items favoring males to items favoring females was closer to 1 for these groups; 1.5 to 1 for CALC-B, and 1.33 to 1 and 1.3 to 1 respectively for NOCALC-A and NOCALC-B groups. Furthermore, the distribution of item types flagged for the CALC-A group is different from that of the other groups. Proportionally fewer type A items (which tend to favor females over males) were flagged for CALC-A than for the other groups.

Table 4. Distributions of Items Flagged for Males and Females in Groups Based on Course Taking

	Group									
	CALC-A		*CALC-B*		*NOCALC-A*		*NOCALC-B*		*Total*	
Type	*M*	*F*	*M*	*F*	*M*	*F*	*M*	*F*	*M*	*F*
A	1	1	2	3	1	3	2	2	6	9
B	1	0	1	0	0	0	1	0	3	0
C	6	2	7	5	6	4	5	6	24	17
D	4	1	6	2	5	1	5	2	20	6
X	0	0	0	1	0	1	0	0	0	2
Total	12	4	16	11	12	9	13	10	53	34

Two sets of statistical tests were run to determine whether any item type significantly favored one sex over the other. The first was a chi-square goodness of fit test run on the four categories in the item taxonomy. Item categories A and B and categories C and D were collapsed in order to have large enough cells to perform a chi-square test. The resulting collapsed categories were labeled "algorithm" (categories A and B) and "logical" (categories C and D). Only the largest grouping of examinees (CALC) showed any significant effect $X^2(1, N = 33) = 5.1$, $p \leq .05$ with more "logical" items favoring males. For smaller clusters of subjects (CALC-A or CALC-B), a binomial test was used due to small cell sizes. As in the more comprehensive grouping, only the largest group, CALC-A, showed a significantly greater number of "logical" items favoring males ($p \leq .05$).

A series of ANOVAs were run on the P-DIF values of all math items to determine whether there was a significant effect across all items by item type, calculus background, or reported mathematics grades. Because a large proportion of examinees correctly answered the majority of the mathematics items, the P+ values were generally high. Therefore, the P-DIF statistic was selected as the unit of analysis over the D-DIF statistic because it is the more conservative measure of DIF for items with high P+ values. Two ANOVA models were used; sex by calculus by item types, and sex by item type. The only significant effect was a main effect for item type ($F(5,1063) = 9.49$, $p \leq .0001$). All other main effects and interactions were not significant. Tukey's post hoc test showed significant differences between type A items and type C and D items. Type B items were not significantly different from the other three item types.

Table 5 presents mean P-DIF values for items. A mean of 0 signifies that neither sex is favored, whereas a negatively signed mean favors males and a positively signed mean favors females. An examination of the means in Table 5 reveals that females performed somewhat better than males on type A items, there is practically no difference between males and females on type B items, and males per-

Table 5. Mean Values of P-DIF for Item Types

Type	*Mean **	*Standard Deviation*
A**	.0047	±0 .02
B	−.0000	±0 .02
C**	−.0032	±0 .04
D**	−.0090	±0. 04

*Positive values indicate that females were favored and negative values indicate that males were favored.

**Significantly different from all categories except B ($p<.001$).

formed somewhat better than females on type C and type D items.

Figures 1 through 3 display the P-DIF values for each item for the CALC group (examinees at or above the ninety-fifth percentile who have had four years of mathematics including calculus) and the total group of examinees. The mean standard error of the P-DIF statistic in the CALC group is 0.005 for the May 1987 and the May 1988 administrations. For November 1987 the standard error of the P-DIF is 0.004. For about half of the items in each form of the test, there is a difference of about two standard deviations between the P-DIF values of the total group and those of the CALC group. In general, the differences are closer to zero for the CALC group than they are for the total group.

Table 6 displays items exhibiting values of the P-DIF statistic that are at least two standard deviations more extreme in the CALC group than in the total group. We can see that almost all of the items are insight/estimation items (item type C or D) occurring at the end of the section (Section I contains 25 items, and Section II contains 35 items).

SAT-M Speededness

Method

Subjects

Subjects for this part of the study were from the top 5 percent of females and the top 5 percent of males for each of the three administrations. Selection was performed in this manner to avoid problems associated with using total score as an index of ability. If the test is differentially speeded for males and females, then total score is not a true indication of an examinee's ability. If, for example, the test is more speeded for females and not males, then matching females at score 700 with males at 700 would pair males with females who are actually of higher ability, because in order for a female to obtain that score she would either have to overcome the effect of speed or make fewer errors. By selecting subjects at or above the ninety-fifth percentile for their own sex and not from the ninety-fifth percentile of the total group, problems of relying on total score as an indicator of true ability are eliminated.

Table 6. Comparison of 95th Percentile and Total Group: Items with Two Standard Deviations' Difference in P-DIF

Administration	*Section**	*Item*	*Type*
May 1987	I	20	D
		21	D
		24	C
	II	32	D
Nov. 1987	I	18	D
		20	D
		21	D
	II	23	C
		24	C
		31	B
		32	C
		35	C
May 1988	I	23	B
	II	27	C
		32	C
		34	C
		35	C

*Section I contains 25 items, Section II contains 35 items.

Speededness Analysis

One of the standard measures of speededness used at Educational Testing Service is whether 80 percent of the examinees reached the last item of a section, and all examinees complete at least 75 percent of each section. Therefore, the percentage of males and females reaching the last item of each section on the SAT-M for the three administrations was examined as well as the percentage of each sex reaching item 15 in Section I (a 25-item section) and item 26 in Section II (a 35-item section). In addition to this standard indication of speededness, patterns of omitted and not reached items were examined to determine whether they vary by sex. Specifically, the ratio of omitted to not reached items was examined for each sex, as well as the percentage of each sex omitting or not reaching an item.

Results

The SAT-M is divided into two sections; one of 25 items and one of 35 items. Examinees are allowed 30 minutes to complete each section, which gives them less time per item in the 35-item section than in the 25-item section. Although in the actual test, these sections are randomly ordered, for simplicity's sake here, the shorter (25-item) section is labeled Section I and the longer (35-item) section, Section II. Further, it should be noted here that in scoring, all items except the last item are scored as right, wrong, omitted, or not reached. An item is scored "not reached" if all subse-

quent items are left blank and "omitted" if any subsequent answer is marked. The last item in each section, therefore, can never receive a score of "omitted" since there are no items following it. All blank responses to the last item in a section are scored "not reached."

Table 7 displays items with omit rates greater than 5 percent for males or females. Item type, frequencies of males and females, and the percentages of each sex omitting or not reaching the item are displayed. Table 7 also presents the ratio of omitted to not reached items for each sex. An examination of this table reveals that more than 80 percent of both male and female examinees completed section I on all three administrations examined.

In two out of the three administrations, however, fewer than 80 percent of the females reached the last item in Section II. In May 1987, only 75 percent of the females reached the last item in Section II and in May 1988, only 78.6 percent of the females reached this item. However, all of the females in both administrations reached item 51 (or 75 percent of the items) in Section II. It can be concluded, then, that standard measures of speededness show the test to be equally unspeeded for males and females.

A further examination of Table 7 reveals that although the omit/not reached ratio is generally larger for males (except on items where all males reached the item) the percentage of females omitting or not reaching any item is consistently greater than the percentage of males omitting or not reaching that same item. Across all three forms, there are no items where males omitted at a greater rate than females. Although the difference between the percentage of females and males omitting an item is generally less than 10 percentage points, on some items it is as high as 22 percent. Thirteen items show differences of 10 percent or greater. All except one of these items require insight or estimation and all of them are located in the second half of the section.

At this point it could be argued that these larger omit rates for females are due to lower ability, because their average total score is lower than that of the males. If an examinee's total score is considered an equally accurate measure for males' as for females' mathematical ability, then a comparison between groups that are not matched on score is not valid. An examination of the distractor analysis performed on the matched examinees used in the first part of this study also reveals a tendency across all three forms for females to omit to a greater extent than males. This difference is rarely greater than 5 percent and is never as high as 10 percent; however, it is generally positive (e.g., females are omitting more often than males). Of the 45 items showing differences in omit rates for the matched group, 39 hold positive values. Nine of the items have positive values indicating a difference of 5 percent or greater, and only six items have negative values. The nine items with 5 percent or more difference for the matched group are included in the items that showed 10 percent or greater difference in the unmatched group.

Clearly, then, females are omitting more items than males. Differences are more obvious in the unmatched sample. However the same pattern is found on the same items in the group that has been matched on total mathematics score. Further, the fact that most of the items with higher omit rates for females require insight or estimation may indicate that females are less willing to use estimation strategies than males. However, since the second half of each section is composed primarily of items of this type, the effects of item type and item location are confounded. On the basis of these data, it is not possible to determine whether females are omitting these items because they require insight or estimation, or because the females are running out of time.

COMPARISON OF SAT-M AND SAT-V

The final part of this study compared patterns in the P-DIF statistics for the total group and the group at or above the ninety-fifth percentile on the SAT-V to patterns found on the SAT-M. In addition, mean verbal scores of examinees at or above the ninety-fifth percentile for the total group on the SAT-M were examined.

Results

Figures 4 through 6 display the P-DIF values for the 85 SAT-verbal items for examinees at or above the ninety-fifth percentile and for the total group across the three administrations. The mean standard error of the P-DIF statistic in the ninety-fifth percentile group is 0.004 across all three administrations. As was found on the SAT-M, at least half the items show more than two standard deviations' difference between P-DIF values for the ninety-fifth percentile group and the total group. Again, like the SAT-M, on the majority of the items the P-DIF values for the group at or above the ninety-fifth percentile are closer to zero than those for the total group (e.g., there is generally less difference between males and females). A few items demonstrate higher P-DIF values for the ninety-fifth percentile group, and further study should be conducted to determine similarities among these items.

Figures 7 through 9 display the mean SAT-V score for examinees at each score level on the SAT-M. Females' scores on the SAT-V are consistently higher than males' scores by about 20 points. Differences are somewhat smaller at the highest score levels on the SAT-M.

Finally, in contrast to the findings in the comparison performed for the SAT-M, there is no apparent pattern in the location of the items that display greater differences for the group at or above the ninety-fifth percentile. On the SAT-M, differences at the beginning of each section decreased, and any increases in differences were found at the end of the section. On the SAT-V, however, items demonstrating greater values of P-DIF for the ninety-fifth percen-

Table 7. Items with Omit Rates Greater than 100

			Females					Males					Difference
Admin.	*Item*	*Type*	*Omit*	*%*	*NR*	*%*	*Ratio*	*Omit*	*%*	*NR*	*%*	*Ratio*	*%Omit*
May 1987													
	18	D	785	15	0	0	0	54	1	0	0	0	14*
	19	C	1087	21	1	0	1087	183	5	0	0	0	16
	21	D	1313	25	4	0	328	140	4	0	0	0	21*
	23	C	911	18	116	2	8	88	2	2	0	44	16*
	24	C	398	8	290	6	1	44	1	29	1	1	7
	25	D			941	18				187	5		
	31	D	230	5	0	0	0	52	1	0	0	0	4
	50	C	362	7	0	0	0	42	1	0	0	0	6
	51	A	252	5	0	0	0	14	0	0	0	0	5
	54	C	585	11	7	0	84	58	1	0	0	0	10
	55	D	277	5	15	0	19	48	1	1	0	48	4
	57	D	237	5	101	2	2	22	1	6	0	4	4
	58	C	914	18	302	6	3	116	3	17	0	7	15
	59	C	307	6	512	10	1	46	1	44	1	1	5
	60	C			1294	25		204	5				
Nov. 1987													
	12	C	414	5	0	0	0	63	1	0	0	0	4
	14	A	471	6	0	0	0	56	1	0	0	0	5
	17	C	627	8	0	0	0	45	1	0	0	0	7
	18	D	1836	23	1	0	1836	229	4	0	0	0	19*
	19	C	490	6	2	0	245	36	1	0	0	0	5
	20	D	1045	13	16	0	65	165	3	0	0	0	10*
	21	D	1537	19	58	1	27	222	4	1	0	222	15*
	22	A	404	5	119	2	3	39	1	2	0	20	4
	23	D	575	7	224	3	3	66	1	12	0	6	6
	24	C	732	9	543	7	1	169	3	57	1	3	6
	25	C			1114	14				167	3		
	50	C	1343	17	0	0	0	314	6	0	0	0	11
	51	C	670	8	0	0	0	72	1	0	0	0	7
	56	B	441	6	28	0	0	55	1	0	0	0	5
	57	C	904	11	88	1	10	123	2	4	1	31	9
	58	A	542	7	192	2	3	31	1	12	0	3	6
	59	X	392	5	446	6	1	115	2	52	1	2	3
	60	C			1275	16				162	3		
May 1988													
	19	C	317	6	0	0	0	22	0	0	0	0	6
	20	D	243	5	0	0	0	57	1	0	0	0	4
	21	X	290	6	7	0	41	72	1	0	0	0	5
	23	B	1587	30	172	3	9	420	8	9	0	47	22*
	24	C	409	8	278	5	2	126	2	24	0	5	6
	25	C			763	15				121	2		
	50	C	1166	22	0	0	0	397	7	0	0	0	15*
	52	C	251	5	0	0	0	65	1	0	0	0	4
	56	D	294	6	16	0	18	96	2	1	0	96	4
	57	C	656	12	86	2	8	328	6	17	0	19	6
	58	C	258	5	205	4	1	141	3	54	1	3	2
	59	C	1189	23	792	15	2	548	10	244	4	2	13*
	60				1180	22				471	9		

*Items that had 5% or greater difference in distractor analysis for sample matched on score.

tile group than for the total group are fairly evenly spread across the entire test.

DISCUSSION

A taxonomy was developed to classify mathematics items on three forms of the SAT on the basis of strategies that could be used in their solution. When differential item functioning procedures were performed on item data for examinees scoring at or above 650, significantly more items requiring insight or estimation were flagged in favor of males than items requiring the use of standard algorithms. When self-reported course taking in mathematics was controlled, similar differences favoring males were found for the best-prepared group (examinees who had four years or more of high school mathematics including some calculus).

One limitation of this taxonomy is that all categories are not mutually exclusive (e.g., Category B overlaps Categories A and D). However, it is necessary to include this type of category because there were a number of items on each test that could be solved using more than one strategy. The fact that these categories do overlap was borne out in the post hoc analyses; item type A was significantly different from item types C and D, but there was no significant difference between item type B and the other three item types.

Although only about 10 percent of the items showed differential performance in the predicted direction, and the size of the difference was fairly small, there is some support for the notion that differential performance among high scorers on the SAT-M may result at least partly from the use of different strategies by males and females. The taxonomy that was developed successfully identified broad categories of items where males generally outperformed females. However, it could be refined further to predict which items would show the largest differences and to account for items that unexpectedly favor one of the sexes (i.e., logic items where females outperform males).

One finding initially seems counterintuitive: Although fewer items were flagged for the CALC-A group, proportionally more items favored males in this group than in any other group. A number of explanations could be offered for this finding. One hypothesis might be that the criteria teachers use for awarding A's are different for males than for females. Another explanation could be that differences in problem-solving strategies are more pronounced in the more prepared groups.

Yet an alternative explanation, and the one favored here, is that proportionally fewer A items were flagged for the CALC-A group than for the other three groups. Consequently, CALC-A females lost some of the "advantage" that females in other groups had over males on items requiring application of standard algorithms. It appears, then, that for the most highly qualified group (CALC-A), the overwhelming majority of items differentially favoring males are items that require examinees to use some type of logical or estimation strategy.

The analysis of omitted items also supports this hypothesis. Items that females omitted to a much greater extent than males were almost exclusively insight/estimation items. This may be a result of females' lower confidence in their mathematical ability. Females may omit these items more frequently than males because they cannot solve them in the allotted time using a standard algorithm and they do not have enough confidence in their mathematical "intuition" to make an educated guess.

Further analyses are needed to clarify the underlying causes of this pattern. Protocol analyses would shed some light on whether these observed differences are due to cognitive processing differences or to other factors such as self-confidence or risk-taking behavior. Furthermore, because previous work shows that sex differences favoring males are smaller or nonexistent on other measures of mathematics ability such as course grades (Clark and Grandy 1984), it would be interesting to compare performance on some of these same items presented in both multiple-choice and free-response formats to determine whether this phenomenon is the consequence of item format.

In the larger context of mathematics education, these findings indicate that even some of the most highly prepared females have some difficulty tackling nonstandard mathematics problems that require insight or the logical application of basic mathematical concepts. Perhaps this is due to an interaction between the way mathematics is taught and girls' tendency to be more compliant than boys from an early age (Brophy 1985). Some girls who get A's in mathematics may be doing so by carefully following the teacher's instructions and perhaps some boys get A's because they do not always do as they are told. If mathematics is taught as a set of rules and formulas to be memorized with little or no connection between them, girls who get A's may be missing some of the higher-order connections that boys who get A's are making on their own by "figuring it out for themselves."

The analyses of the verbal data indicate that females who score in the top 5 percent on the SAT-M generally outperform their male counterparts on the SAT-V. Coupled with the data from the analyses of item types and omit rates, this appears to indicate that mathematically well-prepared females are at a disadvantage on the SAT-M, but not on the SAT-V. The comparison of the top group with the total group on the SAT-M and the SAT-V reveals that there are generally fewer items that demonstrate differential functioning by sex in the top group. However, on the SAT-M, but not on the SAT-V, there appears to be a pattern of decreasing differences on items placed at the beginning of each section and increasing differences on items at the end.

It is suggested here that this pattern on the SAT-M may be the result of differences in solution strategies and self-confidence in mathematics combined with the effects of time limitations. Females, more than males, may tend to

use standard algorithms over estimation strategies. On the more difficult items, this strategy would be time-consuming, and the examinee would probably first attempt items that obviously lend themselves to this type of solution, skipping items that do not. Time limitations may not allow such an examinee to go back and try items that were omitted, and lowered confidence may preclude estimating answers on items that cannot be solved with an algorithm.

APPENDIX A. ITEM CLASSIFICATION OF SAT-M SEX DIFFERENCES STUDY

Content Categories

1. Items that are primarily arithmetic problems, or "word" arithmetic problems.
2. Items that are primarily algebra problems. Questions in which the candidate uses established algebraic theory, formulae, solving equations, etc. to obtain the correct answer.
3. Items that are primarily geometry problems. Questions in which candidates use established geometric formulas, proofs, etc., to solve the problem.
4. Items that require use of both algebra and geometry.

Strategy Categories

A. Items for which the examinee *must* use an established algorithm, including items that require substitution or following directions.

B. Items for which examinees may use *either* a logical/estimation problem-solving strategy *or* an established algorithm.

C. Items for which solution *requires* the use of an algorithm, but *also* requires insight.

D. Items for which the problem can be solved only by using a logical/estimation problem-solving strategy based on general mathematical principles, and *cannot* be solved by using an established algorithm taught in schools. Items solved by inspection or insight are also included in this category.

REFERENCES

Armstrong, J. M. 1981. Achievement and participation of women in mathematics: Results of two national surveys. *Journal of research in mathematics education* 12(5): 356–372.

Armstrong, J. M. 1985. A national assessment of participation and achievement of women in mathematics. In *Women and mathematics: Balancing the equation,* ed. by S. F. Chipman, L. R. Brush, and D. M. Wilson, 59–94. Hillsdale, N.J.: Lawrence Erlbaum Associates.

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

Benbow, C. P., and J. C. Stanley. 1982. Consequences in high school and college of sex differences in mathematical reasoning ability: A longitudinal perspective. *American Educational Research Journal* 19(4): 598–622.

Brophy, J. 1985. Interactions of male and female students with male and female teachers. In *Gender influences in classroom interaction,* ed. by L. C. Wilkinson and C. B. Marrett. Orlando, Fla.: Academic Press.

Clark, M. J. and J. Grandy. 1984. *Sex differences in the academic performance of Scholastic Aptitude Test takers.* (College Board Report No. 84–8). New York: College Entrance Examination Board.

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

Dorans, N. J., and S. A. Livingston. 1987. Male-female difference in SAT-verbal ability among students of high SAT-mathematical ability. *Journal of Educational Measurement* 24(1): 65–71.

Dossey, J. A., et al. 1988. *The mathematics report card: Are we measuring up? Trends and achievement based on the 1986 National Assessment.* Princeton, N.J.: The Nation's Report Card, NAEP, Educational Testing Service.

Eccles, J. S., and P. Blumenfeld. 1985. Classroom experiences and student gender: Are there differences and do they matter? In *Gender influences in classroom interaction,* ed. by L. C. Wilkinson and C. B. Marrett, pp. 79–114. New York: Academic Press.

Evans, F. R. 1980. *A study of the relationships among speed and power aptitude test scores, and ethnic identity* (ETS Research Report No. 80–81, No. 2). Princeton, N.J.: Educational Testing Service.

Fennema, E. 1974. Mathematics learning and the sexes: A review. *Journal for Research in Mathematics Education* 5: 126–129.

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

Fox, L. H., L. Brody, and D. Tobin. 1985. The impact of early intervention programs upon course-taking and attitudes in high school. In *Women and mathematics: Balancing the equation,* ed. by S. F. Chipman, L. R. Brush, and D. M. Wilson, pp. 249–274. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Grieb, A., and J. Easley. 1984. A primary school impediment to mathematical equity: Case studies in rule-dependent socialization. In *Advances in motivation and achievement: women in science,* ed. by M. W. Steinkamp and M. L. Maehr, pp. 317–362. Greenwich, Conn.: JAI Press.

Hudson, L. 1986. Item-level analysis of sex differences in mathematics achievement test performance. *Dissertation Abstracts International* 47(2), Order no. DA8607283.

Kimball, M. M. 1989. A new perspective on women's math achievement. *Psychological Bulletin* 105(2): 198–214.

Leder, G. C. 1987. Teacher-student interaction: A case study. *Educational Studies in Mathematics* 18: 255–271.

Leinhardt, G., A. Seewald, and M. Engle. 1979. Learning what's taught: Sex differences in instruction. *Journal of Educational Psychology* 71: 432–439.

Maccoby, E. E., and C. N. Jacklin. 1974. *The psychology of sex differences.* Stanford, Calif.: Stanford University Press.

Marshall, S. P. 1984. Sex differences in children's mathematics

achievement: Solving computations and story problems. *Journal of Educational Psychology* 76: 195–204.

McPeek, W. M. and C. L. Wild. 1987. *Characteristics of quantitative items that function differently for men and women.* Paper presented at the annual meeting of the American Psychological Association, New York.

O'Neill, K., C. L. Wild, and W. M. McPeek. 1989. *Gender-related differential item performance on graduate admissions tests.* Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Peterson, P. L., and E. Fennema. 1985. Effective teaching, student engagement in classroom activities, and sex-related differences in learning mathematics. *American Educational Research Journal* 22: 309–336.

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

Senk, S., and Z. Usiskin. 1983. Geometry proof writing: A new view of sex differences in mathematics ability. *American Journal of Education* 91: 187–201.

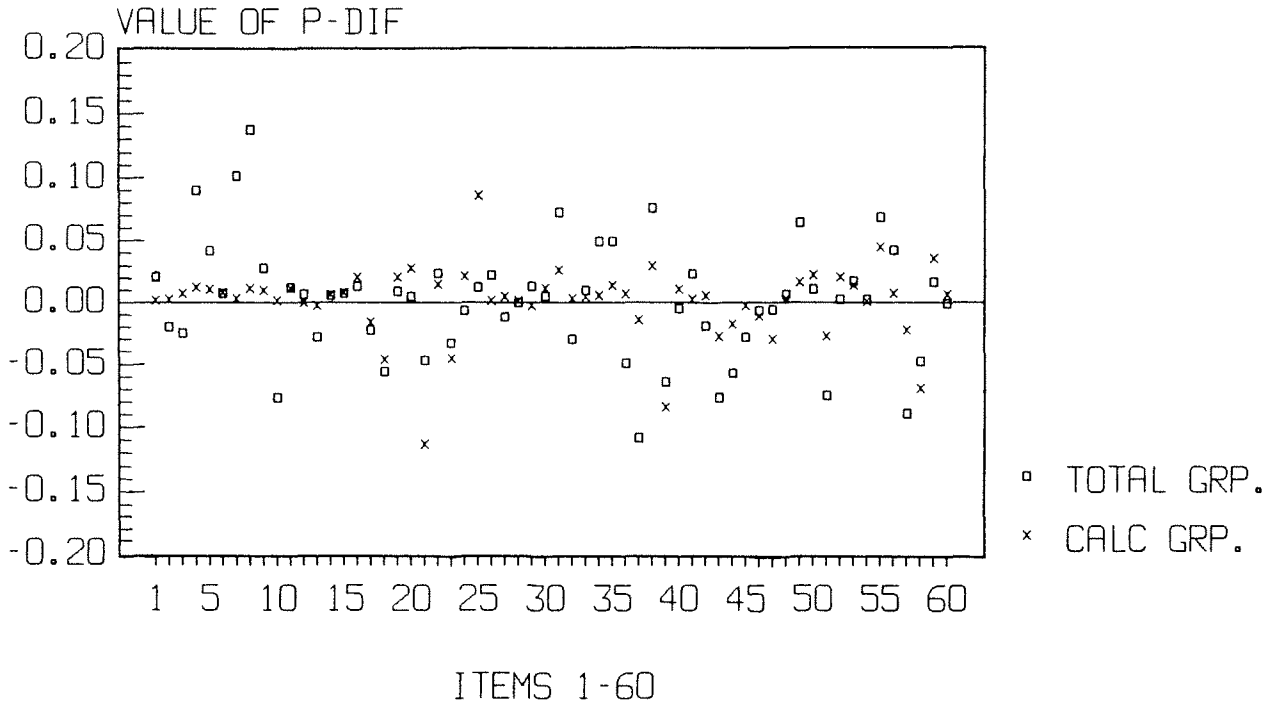

Figure 1. MH P-DIF for CALC and total groups for SAT-M May 1987 administration.

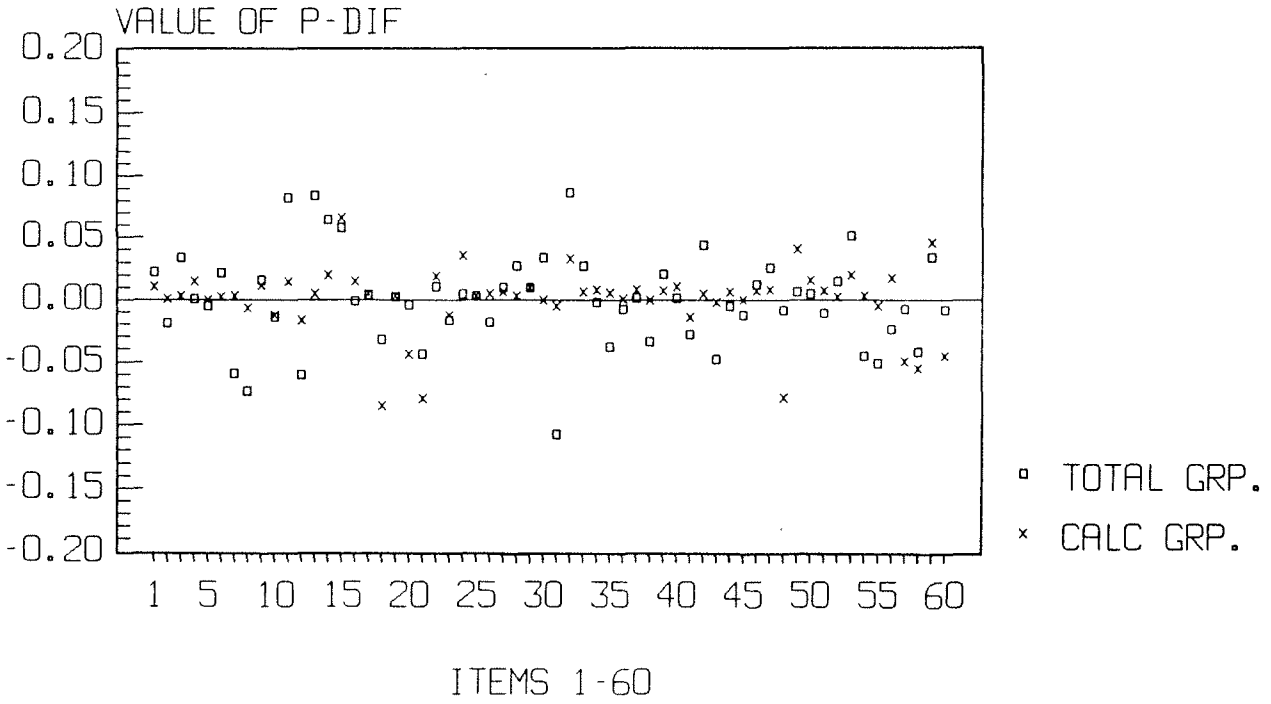

Figure 2. MH P-DIF for CALC and total groups for SAT-M November 1987 administration.

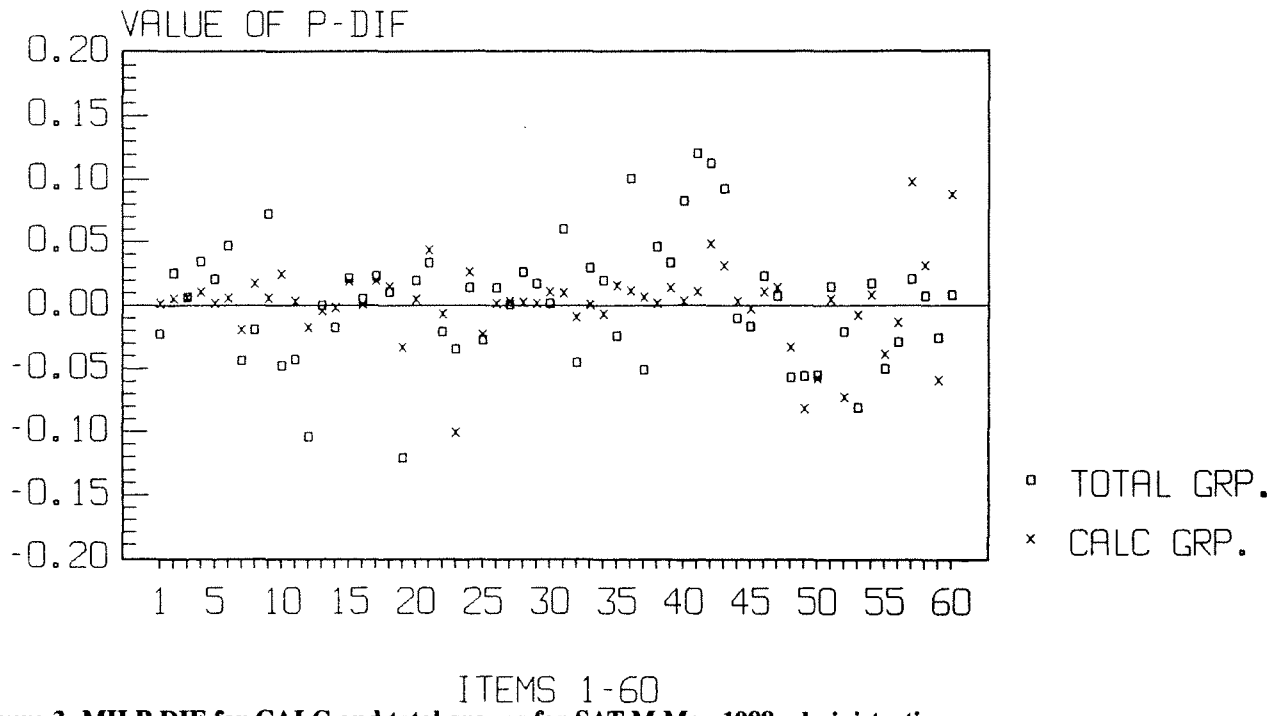

Figure 3. MH P-DIF for CALC and total groups for SAT-M May 1988 administration.

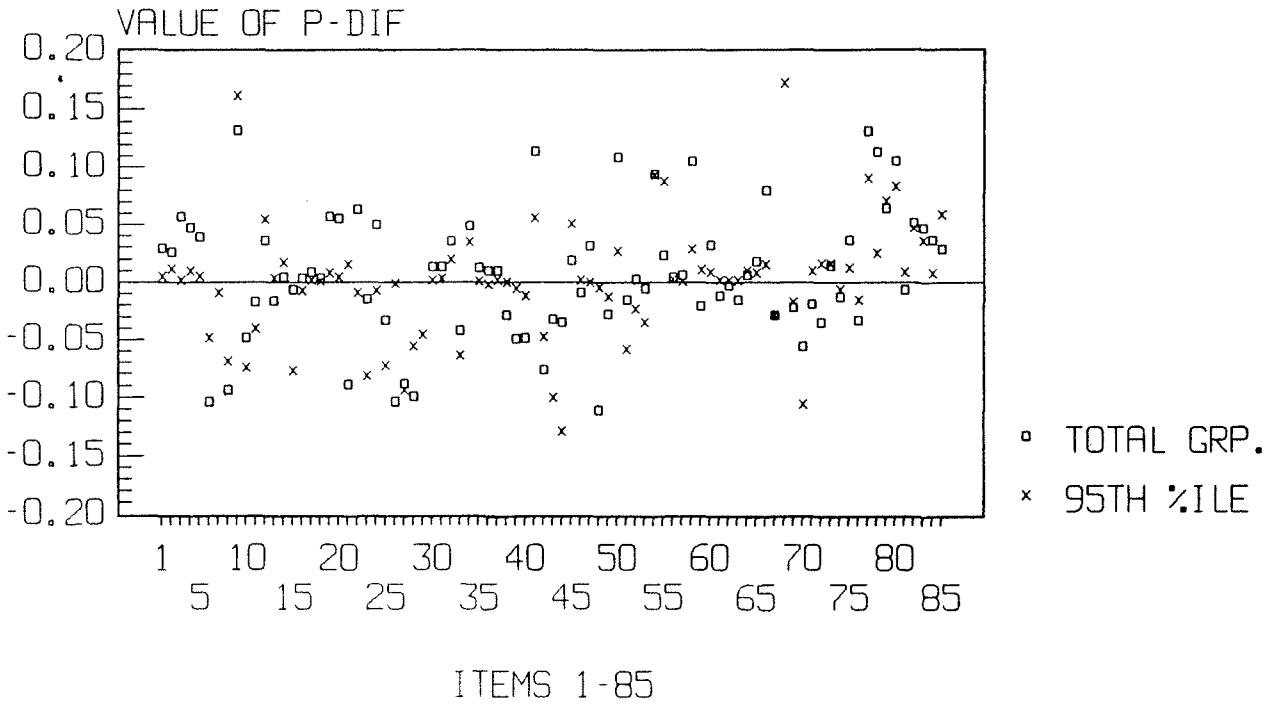

Figure 4. MH P-DIF 95th percentile and total groups for SAT-V May 1987 administration.

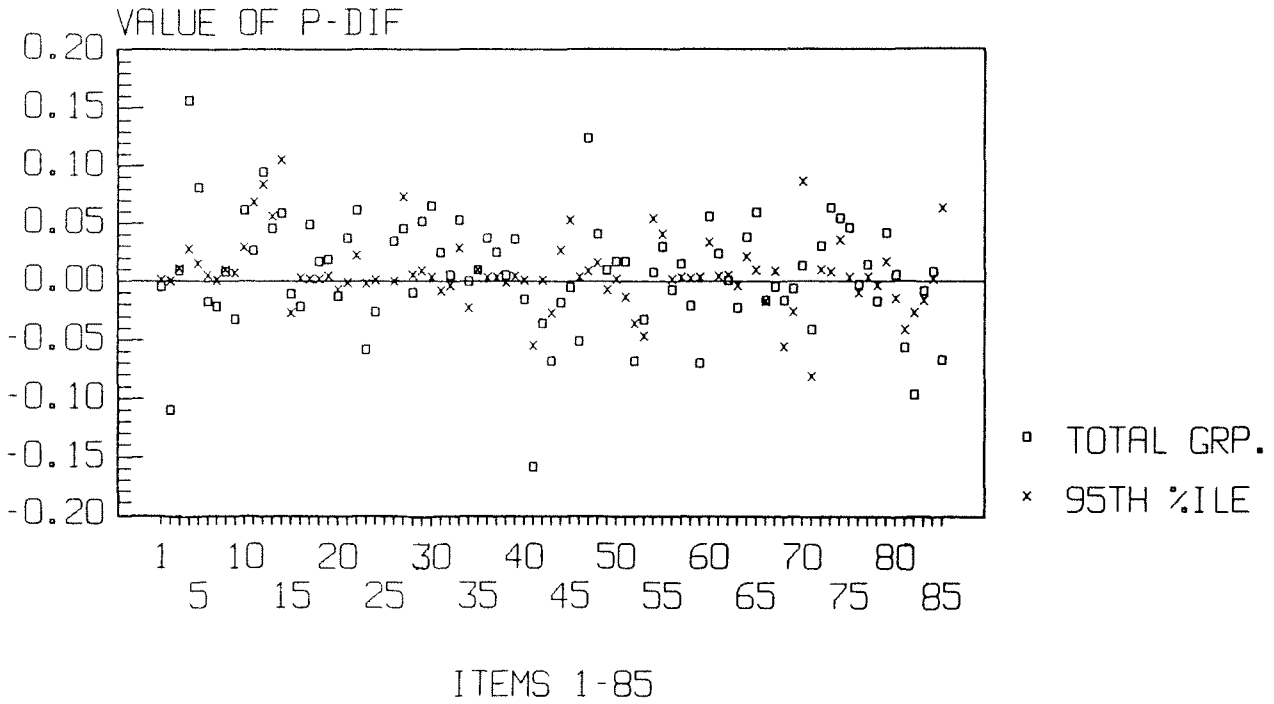

Figure 5. MH P-DIF 95th percentile and total groups for SAT-V November 1987 administration.

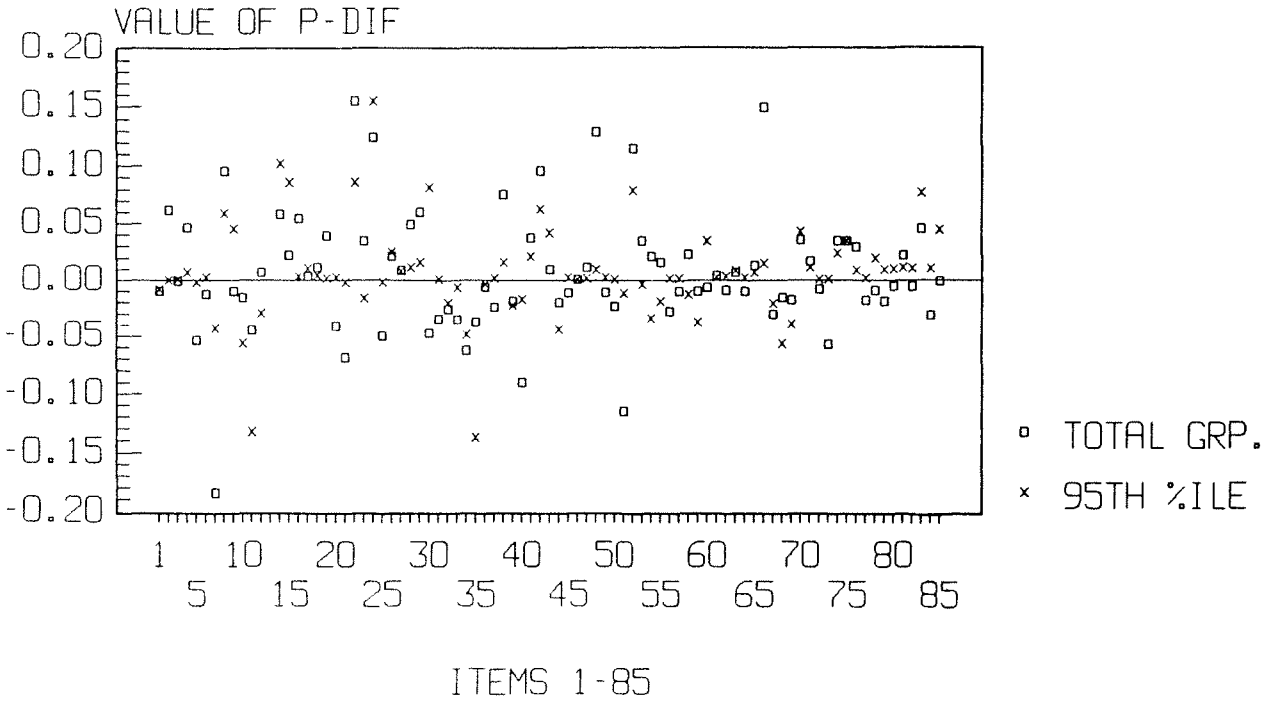

Figure 6. MH P-DIF 95th percentile and total groups for SAT-V May 1988 administration.

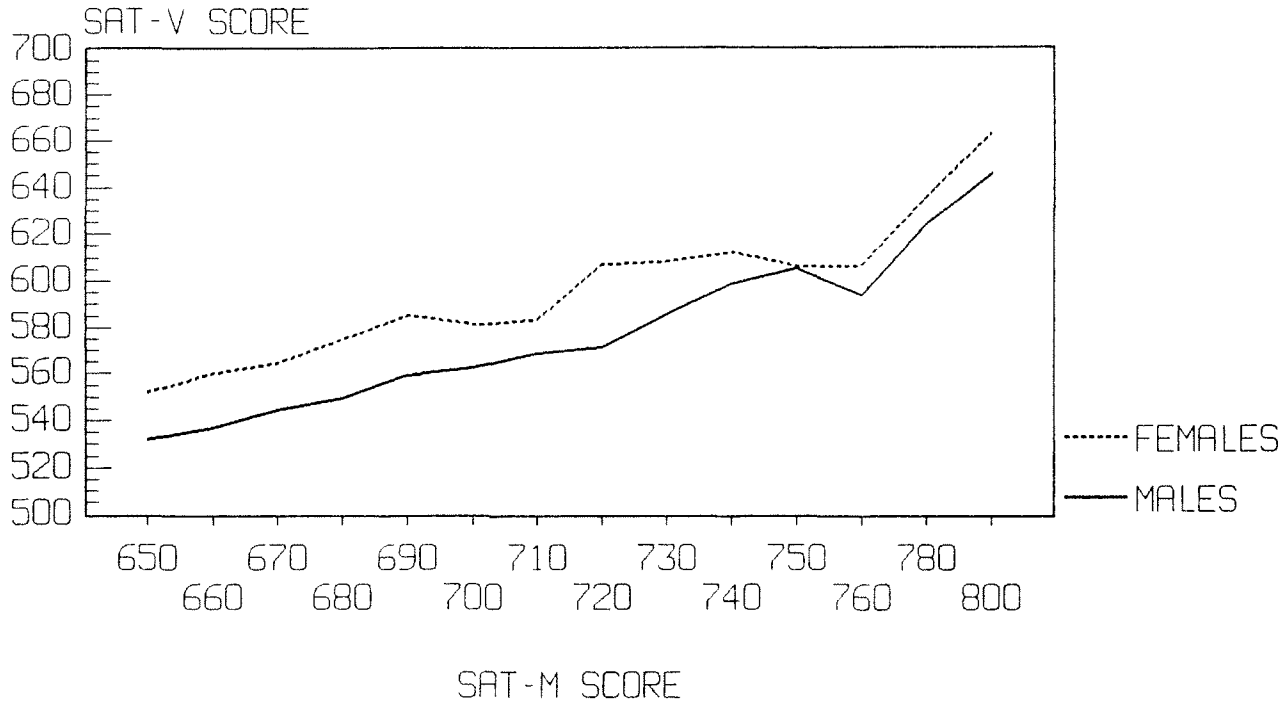

Figure 7. Mean SAT-V by SAT-M CALC group for May 1987 administration.

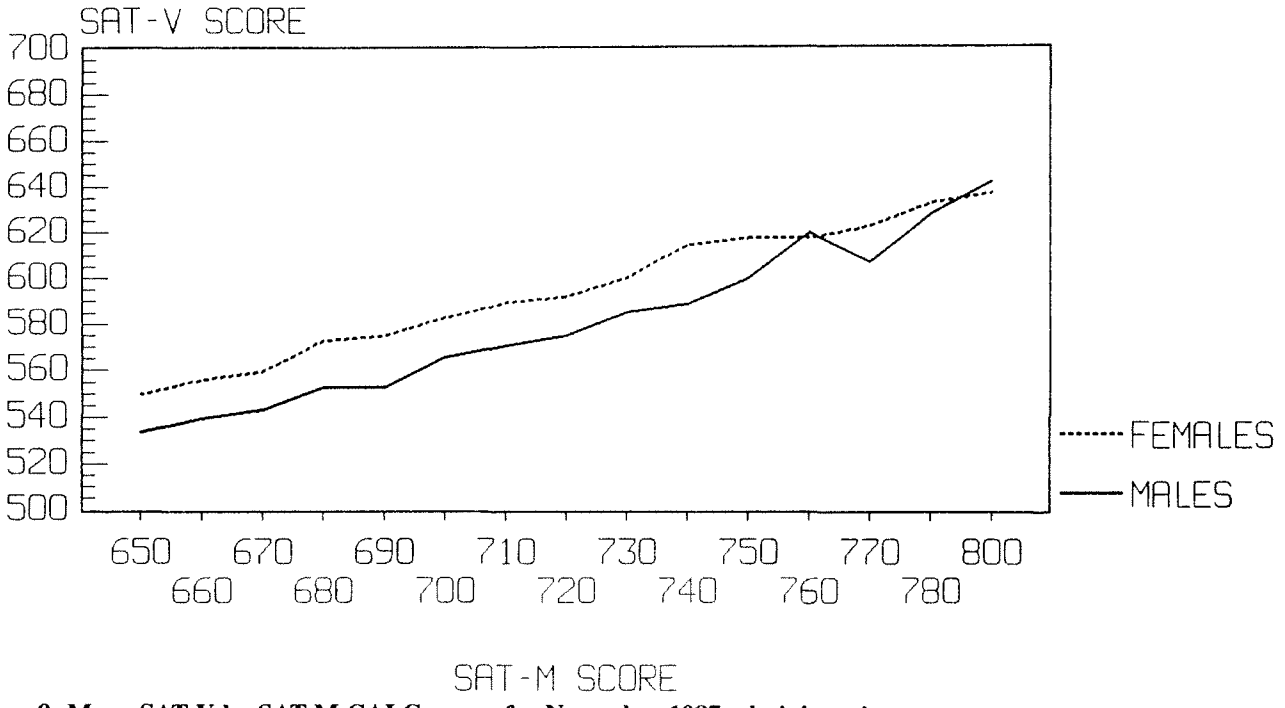

Figure 8. Mean SAT-V by SAT-M CALC group for November 1987 administration.

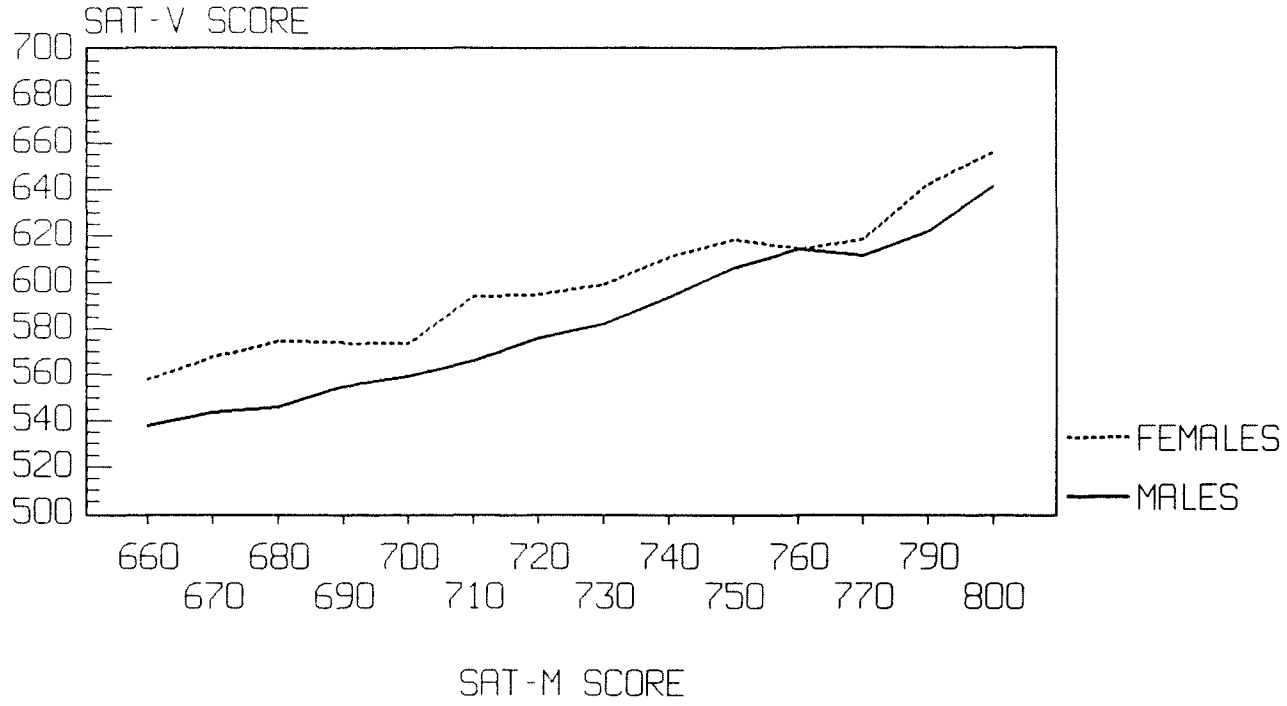

Figure 9. Mean SAT-V by SAT-M CALC group for May 1988 administration.