CollegeBoard
connect to college success™

Research Report
No. 2006-8

The Impact of Course-Taking on Performance on SAT® Items with Higher-Level Mathematics Content

Hui Deng and Jennifer L. Kobrin

www.collegeboard.com

College Board Research Report No. 2006-8

The Impact of Course-Taking on Performance on SAT® Items with Higher-Level Mathematics Content

Hui Deng and Jennifer L. Kobrin

The College Board, New York, 2007

Hui Deng is an associate psychometrician at the College Board.

Jennifer L. Kobrin is a research scientist at the College Board.

The College Board: Connecting Students to College Success

The College Board is a not-for-profit membership association whose mission is to connect students to college success and opportunity. Founded in 1900, the association is composed of more than 5,200 schools, colleges, universities, and other educational organizations. Each year, the College Board serves seven million students and their parents, 23,000 high schools, and 3,500 colleges through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns.

For further information, visit www.collegeboard.com.

Additional copies of this report (item #060482015) may be obtained from College Board Publications, Box 886, New York, NY 10101-0886, 800 323-7155. The price is $15. Please include $4 for postage and handling.

Printed in the United States of America.

Acknowledgments

The research reported in this paper was collaborative in every respect, and the order of authorship is alphabetical. The authors would like to thank Rochelle Michel for her assistance preparing the data files for the analyses presented in this report, and Glenn Milewski and Robin O'Callaghan for their helpful comments on earlier versions of this report.

Contents

Introduction

The SAT Reasoning Test™ (SAT®) measures developed verbal and mathematical reasoning abilities related to successful performance in college. The SAT mathematics section measures reasoning in the areas of arithmetic, Algebra I, and geometry, and since March 2005, topics from third-year college-preparatory mathematics courses (henceforth referred to as Algebra II). According to the National Center for Education Statistics (NCES), in the year 2000 nearly 70 percent of all high school students finished Algebra II or the equivalent by the end of their junior year (U.S. Department of Education, 2002). The College Board reports that in 2005, 97 percent of the graduating high school seniors who took the SAT at least once completed three years of mathematics, with 69 percent completing four or more years. Most four-year colleges require three years of mathematics for admission (College Board, 2005b).

To better align the SAT mathematics section to classroom practice, new content that is typically taught in third-year mathematics classes was added to the test. Many of the new items are more difficult than the items testing reasoning skills in arithmetic, Algebra I, and geometry because the content involved in the new items is taught later in the curriculum. However, not all of the new items are more difficult than the old items. For example, an arithmetic question could have a higher level of difficulty than an Algebra II question because it might require more advanced reasoning skills to solve the problem. Therefore, even with the addition of the higher-level content, i.e., Algebra II, the overall difficulty of the test was kept at the same level as in past tests by including items at all levels of difficulty in all content areas. In addition, the statistical process of equating was used to adjust the scale of the new SAT to account for any slight differences in difficulty between the new and old forms.

Liu, Schuppan, and Walker (2004) explored whether the addition of items with more advanced mathematics content to the SAT would impact test-taker performance. They assembled and administered new test forms with some SAT I mathematics items replaced by SAT II Subject Test in Mathematics items that covered higher-level content. The replacement items had difficulty levels similar to those of the SAT I items that were removed. They found that the mere presence of the higher-level mathematics items did not make the test more difficult. That is, test-taker performance on each item was directly related to the difficulty level of that item.

Some SAT users have voiced concern that the new mathematics content will disadvantage some students, including those who do not take certain advanced mathematics courses in high school. Although the percentage of students taking algebra and geometry is similar across ethnic groups, there is a disparity in the percentage taking higher-level mathematics courses. For example, in 2005 only 14 percent of African American students took a calculus course in high school, compared to 28 percent of white students and 44 percent of Asian American students (College Board, 2005a).

Several researchers have found an association between the number of mathematics courses taken in high school and achievement (e.g., Laing, Engen, and Maxey, 1990; Maxey, Cargile, and Laing, 1987; Schmidt, 1983). A few studies have examined whether taking specific mathematics courses is associated with higher achievement. Pelavin and Kane (1990) reported that taking college-preparatory mathematics courses (algebra and geometry) is strongly associated with college enrollment and graduation. Rock and Pollack (1995) found that students who eventually took higher-level courses (Algebra II and geometry up through calculus) showed consistently greater gains on the National Education Longitudinal Study of 1988 (NELS:88) mathematics test between eighth and tenth grades and between tenth and twelfth grades than those who did not.

The College Board, ACT, and many other testing organizations annually report data illustrating that students who complete more years of core academic courses in high school and more rigorous courses have higher test scores. This data can be misinterpreted as implying a strong causal relationship when in fact many other factors such as student interest in the subject, student ability, and self-perceptions of ability may moderate these relationships (W. J. Camara, personal communication, 2003). After adjusting for background and academic variables, Morgan (1989) found that course work in mathematics, the natural sciences, and foreign languages still had a strong relationship with SAT mathematics (SAT-M) scores. The relationships were generally consistent across ethnic groups and income levels, but were stronger for students with higher GPAs.

Rock and Pollack (1995) reported that students who stop taking mathematics after basic or Algebra I–level courses are typically learning skills that improve their computational abilities but have little direct impact on their growth in more complex mathematical concepts and/or ability to successfully carry out complex problem-solving exercises. Rock and Pollack assembled clusters of NELS items marking five ascending points on the test score scale associated with increasingly higher levels of mathematical complexity. The items making up these clusters exemplified the skills required to answer successfully the typical item located at these points along the scale. They compared performance gains between the eighth- and tenth-grade levels and between the tenth- and twelfth-grade levels, and found that students who did not take advanced courses made greater gains on test items dealing with computational skills than they did on items testing higher-level skills. There was little growth in understanding of intermediate mathematical concepts and multistep problem-solving skills in the absence of advanced course work. In contrast, students taking advanced courses made larger gains on test items requiring conceptual understanding and problem-solving skills, but significant growth in these areas did not occur until students moved into the precalculus level of course work. It is noted that the sample of students in the NELS database differed substantially from the sample of college-bound students taking the SAT. In 2004, for example, 46 percent of SAT takers reported taking precalculus and 25 percent reported taking calculus, compared to 28 percent and 11 percent, respectively, in the NELS dataset.

This report describes the results of two studies designed to evaluate the impact of self-reported mathematics course-taking on performance on SAT mathematics questions measuring new content (Algebra II). Both studies analyzed data collected during the field trial of the new SAT. In Study 1, standardized mean differences (effect sizes) were computed between students taking or planning to take certain mathematics courses and those not taking such courses to show the impact of course-taking on performance on old and new SAT mathematics questions. In addition, correlation analyses were performed to determine the association between course-taking and performance on the old and new items. Study 2 was focused on more advanced courses than Algebra II, and differential item functioning (DIF) analyses were conducted using the simultaneous item bias test (SIBTEST) to explore whether items functioned similarly for students of equal ability with different course-taking patterns.

Data and Methods

The data used for both studies were obtained during the field trial of the new SAT and the new PSAT/NMSQT® (PSAT) conducted in March 2003. The purpose of the field trial was to evaluate the content, statistical, and timing specifications for the new SAT and PSAT as well as to determine whether scores on the new tests were comparable to scores on the old tests.[1] More than 45,000 students from 679 high schools participated in the field trial. These students were from both public and private schools across rural, suburban, and urban areas, and represented every geographical region in the United States. To ensure that the research was based on sufficient numbers of students from each major racial/ethnic subgroup, higher proportions of African American and Latino students were included in the field trial sample.

The field trial included two different data collection designs with a total of 23 test booklets. This study made use of only the subset of the field trial data that included pretest mathematics sections on seven different new SAT prototype forms. The pretest mathematics sections were designed with a large proportion of items representing the new mathematics content in order to get item statistics on a large number of new content items to build up the pool of these items for upcoming forms of the test. Each test booklet included either eight or nine sections.

The mathematics items in the pretest sections in the field trial were coded for new content. Subscores were created for items measuring the new content and for items measuring the old content. Appendix A displays the number of items that measure each new content area within each form. All students participating in the field trial answered a series of questions about the mathematics courses they took or anticipated taking in high school and the year(s) in which the courses were taken. Several "course pattern" variables were created, including total number of mathematics courses taken, highest level course taken, and highest level course planned. Table 1 shows the questions that the field trial participants answered regarding course-taking. Based on the relationship between a student's grade level and responses to these questions, separate variables were created to indicate whether each student actually took or planned to take each course. For example, if a student was in the eleventh grade at the time of the field trial, all of the courses that were indicated for grade 11 and earlier were assumed to have been taken, while it was assumed that all the courses indicated for grade 12 were planned.

Methods for Study 1

The purpose of Study 1 was to examine the descriptive statistics on old and new item performance for different course-taking patterns, to determine the impact of taking or planning to take each mathematics course, and to explore the association between course-taking and item performance. Table 2 shows the structure of the forms

Table 1

SAT Field Trial Question Format on Mathematics Course-Taking

Indicate which math course(s) you have taken or *plan to take* in each school year. If more than one in a year, check all that apply.

	8th or earlier	*9th*	*10th*	*11th*	*12th*
Algebra I	❑	❑	❑	❑	❑
Algebra II	❑	❑	❑	❑	❑
Geometry	❑	❑	❑	❑	❑
Trigonometry	❑	❑	❑	❑	❑
Precalculus	❑	❑	❑	❑	❑
Calculus	❑	❑	❑	❑	❑

Note: Additional courses were included in this question in the field trial, but only the courses listed here were used for the analyses in this study.

[1] See Liu and Fiegenbaum (2003) for a complete description of the new SAT field trial and its results.

Table 2

SAT Field Trial Test Forms Used in the Study

Form	*Position of First Mathematics Pretest Section*	*Position of Second Mathematics Pretest Section*	*Number of Students*
21	7	8	4,575
22	7	8	4,474
31	4	–	1,601
41	4	–	1,577
123*	8	–	2,112
133	2	7	1,515
143	3	6	1,493

*Students taking Form 123 were given "right-scoring" instructions.

that were used in the analyses for Study 1, and the number of students with valid data on each form. There were 17,347 cases used for the analyses in Study 1, but most of the results presented in this report are based only on eleventh-grade students, unless otherwise noted.[2]

The standardized mean differences (impact) were calculated for students who had taken or were taking each course compared to students *planning* to take the course (Took–Planned) and compared to students *not taking* the course (Took–Didn't Take). The standardized mean difference, or effect size (Cohen, 1988), was calculated as the raw mean difference divided by the pooled standard deviation for the two groups being compared. Cohen suggested classifying an effect size of .2 or lower as "small," .8 or higher as "large," and effect sizes in between as "medium." However, the practical significance of an effect size is a judgment call on the part of the researcher.

To supplement the standardized mean differences, correlation analyses were performed to explore the relationships between course-taking and performance on the old and new items. The highest course taken and highest course planned were recoded from 1 to 6, with Algebra I = 1, geometry = 2, Algebra II = 3, trigonometry = 4, precalculus = 5, and calculus = 6.

Methods for Study 2

Since the majority of the juniors in the field trial sample had already taken Algebra II by the time they participated in the field trial, the focus of Study 2 was to explore the impact of taking or planning to take more advanced courses (trigonometry, precalculus, calculus) on performance on items measuring the old versus the new content. DIF analyses were conducted using the SIBTEST (Stout, 1999) to explore whether each pretested item functioned similarly for students of equal ability with different course-taking patterns.

SIBTEST implements a nonparametric statistical method of assessing DIF based on Shealy and Stout's (1993) multidimensional model. It is presumed that the reference group (e.g., those taking advanced courses) and focal group (e.g., those not taking advanced courses) are given a set of items that measure an intended ability—the dominant ability. While most items measure the dominant ability, a few items may be influenced by some nuisance ability (or abilities) in addition to the dominant ability, and potentially cause DIF. The procedure to test DIF via SIBTEST involves a valid subtest that functions as the matching criterion and a studied subtest that contains potentially biased items. The valid subtest contains items known to be unbiased and is essentially unidimensional. The procedure detects DIF by comparing item responses of examinees in the reference and focal groups that are grouped based on their scores on the matching subtest (Roussos and Stout, 1996).

The performance of SIBTEST for detecting DIF has been evaluated through Monte Carlo simulations by various researchers. It has been found that the test statistic Beta-uni has good adherence to type-I error rate, and that the procedure can be applied to sample sizes as small as 100 (Shealy and Stout, 1993; Roussos and Stout, 1996a). Roussos and Stout proposed the following guidelines for classifying DIF on a single item using the Beta-uni statistic: given statistical rejection, the absolute value of Beta-uni < 0.059 indicates negligible or A-level DIF, the absolute value of Beta-uni ≥ 0.088 indicates large or C-level DIF, and intermediate values of Beta-uni indicate moderate or B-level DIF.

Three DIF comparisons were conducted on each field trial form. The first comparison examined whether each item functioned similarly for students who took one or more advanced mathematics courses compared to those who didn't take any advanced course (the Took–Didn't Take contrast). The second comparison examined whether the items functioned similarly for students who hadn't taken, yet planned to take advanced courses, compared to those who didn't take or plan to take advanced courses (the Planned–Didn't Plan contrast). And the third comparison examined whether each item functioned similarly for students who took one or more advanced mathematics courses compared to those who didn't take, yet planned to take advanced courses (the Took–Planned contrast). The research questions to be answered by each DIF comparison and the corresponding focal and reference groups involved in each comparison are summarized in Table 3.

Because the field trial forms involved different sets of items, the DIF analyses were conducted separately for each form. Form 123 contained a pretest section that was also used in Form 22, but Form 123 was

[2] The field trial was designed for eleventh-grade students, and there was a relatively small number of students participating from other grade levels.

Table 3

Description of DIF Analyses

Research Question	*Focal Group*	*Reference Group*
Are there items that perform differently for students who have taken courses higher than Algebra II, compared to those who have taken only up to Algebra II?	**Didn't Take** (Took Algebra II but no advanced course)	**Took** (Took Algebra II and one or more advanced courses of trigonometry, precalculus, and calculus)
Are there items that perform differently for students who planned to take courses higher than Algebra II but have not done so yet, compared to those who did not take nor plan to take advanced courses?	**Didn't Plan** (Took Algebra II, didn't take nor plan to take any advanced course)	**Planned** (Took Algebra II, didn't take but planned to take advanced course)
Are there items that perform differently for students who have taken courses higher than Algebra II, compared to those who planned to take advanced courses, but have not done so yet?	**Planned** (Took Algebra II, didn't take but planned to take advanced course)	**Took** (Took Algebra II and one or more advanced courses of trigonometry, precalculus, and calculus)

tested under right-scoring conditions while all the other forms were tested with formula scoring.[3] Due to this administrative difference, the same set of pretest items used for Form 22 and Form 123 were analyzed separately. Within each form, items were tested individually for DIF, and PSAT item responses for the same students were used to match ability. The use of PSAT scores as a matching criterion was based on the assumption that both the PSAT and SAT measure students' mathematical reasoning ability, not just the knowledge pertinent to specific content covered by the test. Although the PSAT may not reflect some of the new content introduced for the new SAT mathematics section, the two tests were considered to measure the same construct. The majority of the eleventh-grade participants in the field trial took the PSAT in 2002; therefore, the sample used for the DIF analyses included those eleventh-graders who had valid PSAT scores from 2002.

Results

Study 1

For the majority of eleventh-grade participants in the field trial, Algebra II was the highest level mathematics course taken (46 percent), followed by precalculus (29 percent), trigonometry (13 percent), geometry (7 percent), calculus (4 percent), and Algebra I (2 percent). The majority of participants planned to take calculus as the highest course in their senior year (41 percent); 36 percent planned to take precalculus, 15 percent planned to take trigonometry, and the remaining 8 percent planned to take Algebra II, geometry, or Algebra I as their highest course.

Figures 1 and 2 show a series of box plots of the proportion of old and new items correct by the highest mathematics course taken and highest course planned. The boxes extend from the lower hinge to the upper hinge, with a crossbar at the median. A whisker extends from each end of the box to the most extreme value that is not an outlier (Smith and Prentice, 1993). Each outlier is marked separately as circles and extreme cases are shown as asterisks and indicate cases with values greater than three box lengths. The box plots show that students performed better on the old content regardless of the highest mathematics course taken. There were many more outliers and extreme cases for students whose highest course was Algebra I, Geometry, or Algebra II.

Table 4a shows the mean proportion of old and new items correct for students indicating whether they took or planned to take each of the six mathematics courses. Table 4b shows the standardized mean difference for students taking each course compared to students *planning* to take the course (Took–Planned) and compared to students *not taking* the course (Took–Didn't Take). The standardized mean differences are also illustrated in Figure 3.

For both the old and new items, students who took each course scored higher than students who planned to take or didn't take the course. From Algebra I through trigonometry, students who didn't take the course scored higher than those planning to take the course; but for precalculus and calculus, those planning to take the course scored higher than those not taking the course.

[3] The traditional formula scoring on the SAT involves subtracting a fraction of a point for each incorrect answer to discourage test-takers from blind guessing. The right-scoring approach does not penalize students for guessing and bases the score only on the total number of correct answers.

Table 4a

Mean Proportion of Old and New Items Correct by Course-Taking Group

	N			*Old Items*			*New Items*		
Course	*Took*	*Planned*	*Didn't Take*	*Took*	*Planned*	*Didn't Take*	*Took*	*Planned*	*Didn't Take*
Algebra I	15,194	23	791	.47	.26	.44	.37	.17	.36
Geometry	14,808	166	1,034	.47	.23	.40	.37	.19	.32
Algebra II	13,898	758	1,352	.49	.28	.37	.38	.19	.28
Trigonometry	4,456	2,182	9,370	.58	.40	.43	.47	.29	.33
Precalculus	5,050	3,865	7,093	.64	.44	.35	.55	.32	.26
Calculus	561	4,198	11,249	.73	.63	.39	.66	.53	.29

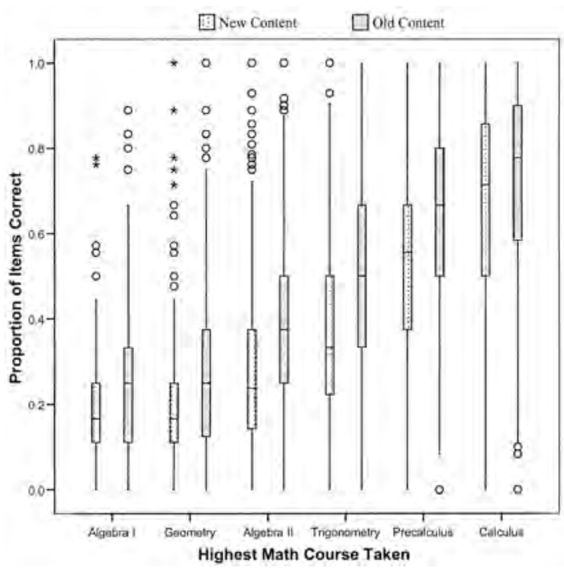

Figure 1. Box plots of proportion of old and new items correct by highest math course taken.

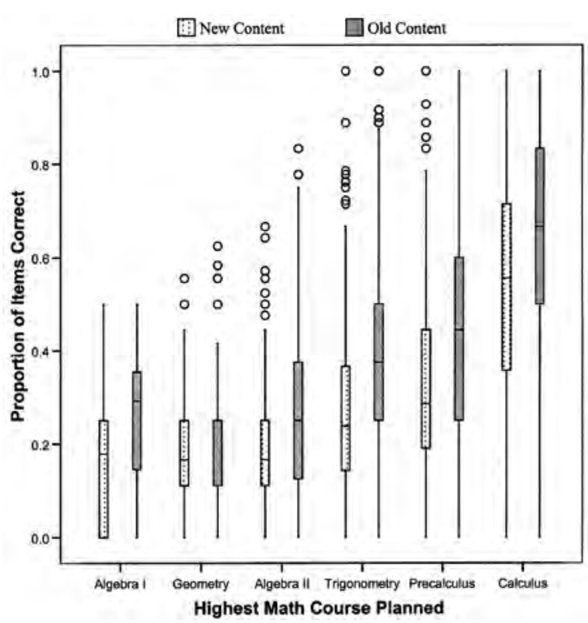

Figure 2. Box plots of proportion of old and new items correct by highest math course planned.

Table 4b

Standardized Mean Differences (Impact) of Course-Taking on Proportion of Old and New Items Correct

	Old Items		*New Items*	
Course	*Took–Planned*	*Took–Didn't Take*	*Took–Planned*	*Took–Didn't Take*
Algebra I	.80 (***)	.11 (.03-.19)	.86 (***)	.04 (-.02-.11)
Geometry	.94 (.69-1.27)	.31 (.19-.44)	.77 (.53-1.11)	.21 (.08-.39)
Algebra II	.82 (.56-1.03)	.47 (.32-.63)	.83 (.54-1.28)	.42 (.23-.65)
Trigonometry	.72 (.51-.87)	.61 (.48-.70)	.80 (.55-.97)	.61 (.36-.79)
Precalculus	.79 (.65-.90)	1.13 (.98-1.31)	.95 (.87-1.09)	1.24 (.95-1.56)
Calculus	.36 (.27-.77)	1.30 (1.16-1.51)	.58 (.44-.97)	1.61 (1.47-1.86)

Note: The minimum and maximum proportion of items correct across the seven forms are shown in parentheses under the means. There were not enough students indicating that they *planned to take* Algebra I to compute the standardized difference.

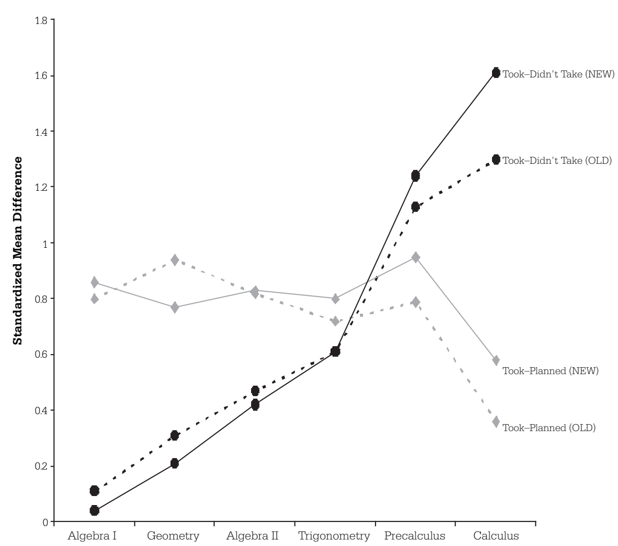

Figure 3. Standardized mean differences in proportion of old and new items correct by course-taking.

Table 5

Correlations of Number of Mathematics Courses and Highest Course Taken with Mean Proportion of Items Correct for Old and New Items

Form	*Number of Mathematics Courses Taken*		*Highest Course Taken*		*Highest Course Planned*	
	Old Items	*New Items*	*Old Items*	*New Items*	*Old Items*	*New Items*
21	.38	.41	.51	.59	.45	.51
22	.41	.43	.53	.59	.49	.51
31	.41	.42	.54	.60	.51	.53
41	.41	.38	.54	.55	.51	.48
123*	.41	.41	.54	.53	.51	.49
133	.40	.38	.53	.58	.40	.44
143	.40	.42	.52	.60	.45	.50

* Students taking Form 123 were given "right-scoring" instructions.

The standardized mean differences between students who took Algebra I, geometry, and Algebra II and students who didn't take those courses were higher on the old items than on the new items. The standardized mean difference between students taking trigonometry and those not taking the course were the same for the old and new items. However, the standardized mean difference between students who took precalculus and calculus and those not taking these courses were higher for the new items compared to the old items. There was considerable variability in the standardized mean differences across forms, as indicated by the range shown in the parentheses under each mean. The standardized mean differences between students taking a course and those planning to take the course were generally higher for the new items, with the exception of geometry.

Table 5 shows the Pearson correlations of number of mathematics courses taken, highest mathematics course taken, and highest mathematics course planned with the percentage of old and new items correct. The number of mathematics courses a student took correlated similarly with the percentage of both old and new items correct at about $r = .40$. The level of course-taking correlated slightly higher with performance on the new items compared to performance on the old items. The correlation of highest course taken and percentage of items correct ranged from .51 to .54 across forms for the old items and ranged from .53 to .60 across forms for the new items. The correlation of highest course planned and percentage of items correct was slightly lower, ranging from .40 to .51 across forms for the old items and ranging from .44 to .53 across forms for the new items.

Study 2

While Study 1 examined the impact of taking or planning to take each of the six different mathematics courses on old and new SAT mathematics item performance, Study 2 focused on the impact of taking or planning to take advanced mathematics courses (trigonometry, precalculus, and calculus) on old and new math item performance. The mean percentage of items correct and the standardized mean differences for students who took, didn't take, planned to take, or didn't plan to take one or more advanced mathematics courses on each field trial form is shown in Tables 6 and 7 for the old content and new content, respectively.

From Table 6, it is observed that for items measuring the old content, students who took one or more advanced courses scored higher than students who did not take any advanced course or just planned to do so, and students who planned to take one or more advanced courses scored higher than students who did not plan to take any advanced course. This pattern is consistently observed for all seven forms. Note that the Didn't Take group includes all the students who did not take advanced courses, including those who planned or didn't plan to take advanced courses, therefore it

Table 6

Mean Percentage of Items Correct by Course-Taking Group for Old Content

Form	*#Items*	*Mean Percent of Items Correct*												*Standardized Mean Differences*		
		Took			*Didn't Take*			*Planned*			*Didn't Plan*			*Took–Didn't Take*	*Planned–Didn't Plan*	*Took–Planned*
		N	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*			
21	8	1,232	0.56	0.23	1,096	0.34	0.20	741	0.36	0.21	355	0.30	0.18	0.98	0.34	0.86
22	12	1,213	0.67	0.23	999	0.46	0.21	674	0.49	0.21	352	0.41	0.20	0.95	0.38	0.82
31	9	433	0.66	0.22	372	0.42	0.21	247	0.47	0.21	125	0.34	0.20	1.10	0.62	0.91
42	10	392	0.64	0.21	360	0.40	0.22	229	0.43	0.21	131	0.36	0.21	1.10	0.36	0.98
123	9	610	0.66	0.25	470	0.42	0.22	298	0.45	0.21	172	0.37	0.22	1.03	0.33	0.92
133	12	391	0.53	0.21	371	0.35	0.19	239	0.38	0.19	132	0.28	0.18	0.92	0.54	0.74
143	8	366	0.60	0.24	350	0.37	0.20	249	0.38	0.20	101	0.34	0.18	1.07	0.20	0.99

Note: Didn't Take=Took Algebra II but no advanced course; Took=Took Algebra II and one or more advanced course; Didn't Plan=Took Algebra II, didn't take nor plan to take any advanced course; Planned=Took Algebra II, didn't take but plan to take an advanced course.

Table 7

Mean Percentage of Items Correct by Course-Taking Group for New Content

Form	#Items	Mean Percent of Items Correct												Standardized Mean Differences		
		Took			Didn't Take			Planned			Didn't Plan			Took–Didn't Take	Planned–Didn't Plan	Took–Planned
		N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD			
21	18	1,232	0.49	0.21	1,096	0.28	0.15	741	0.30	0.15	355	0.24	0.13	1.15	0.39	1.02
22	14	1,213	0.56	0.23	999	0.32	0.18	647	0.35	0.18	352	0.27	0.15	1.15	0.49	0.98
31	9	433	0.62	0.25	372	0.34	0.21	247	0.38	0.22	125	0.26	0.17	1.21	0.59	1.01
41	8	392	0.46	0.22	360	0.21	0.17	229	0.23	0.17	131	0.19	0.16	1.27	0.26	1.16
123	9	610	0.55	0.24	470	0.33	0.20	298	0.35	0.21	172	0.28	0.19	0.96	0.35	0.83
133	18	391	0.35	0.20	371	0.17	0.11	239	0.18	0.12	132	0.16	0.10	1.15	0.16	1.04
143	21	366	0.50	0.20	350	0.27	0.13	249	0.28	0.14	101	0.25	0.10	1.31	0.21	1.20

Note: Didn't Take=Took Algebra II but no advanced course; Took=Took Algebra II and one or more advanced course; Didn't Plan=Took Algebra II, didn't take nor plan to take any advanced course; Planned=Took Algebra II, didn't take but plan to take an advanced course.

is the aggregate of students in the Planned and Didn't Plan groups. The partitioning of Didn't Take into the Planned and Didn't Plan groups was to examine whether students planning to take advanced courses perform differently than those who did not plan to take any further advanced courses via the Planned–Didn't Plan contrast. It is observed that for all seven forms, the Took–Didn't Take contrast has the largest standardized mean differences, followed by the Took–Planned contrast, and the Planned–Didn't Plan contrast has the lowest standardized mean differences. This suggests that although both taking and planning to take advanced courses are associated with higher performance on items measuring the old content, students benefit more from the actual course-taking than planning.

Table 7 shows the average percentage of items correct by course-taking group for items measuring the new content. The results exhibit the same patterns as observed for the old content. Across all the seven forms, students who took one or more advanced courses scored higher than students who did not take or just planned to take such courses, and students who planned to take advanced courses scored higher than those who did not plan to take courses. The standardized mean differences between students who took advanced courses and those who didn't take any course were the highest, and the standardized mean differences between students who planned to take advanced courses and those who didn't plan to take these courses were the lowest. Again, this indicates that actual course-taking instead of planning was associated with higher performance on items measuring the new content. In addition, when comparing results from Tables 6 and 7 for the same course-taking group within each form, it was observed that students consistently scored higher on items measuring the old content than on items measuring the new content. For the Took–Didn't Take and Took–Planned contrasts, the standardized mean differences were higher for items of new content than for items of old content for each form, which suggests that items measuring the new content were more sensitive to the effects of taking advanced math courses than items that measure the old content. For the Planned–Didn't Plan contrast, three forms had standardized differences that were smaller for the new items than for the old items.

The effects of course-taking on individual item performance, after controlling for students' mathematical reasoning ability using PSAT scores, are summarized in Tables 8 and 9. Table 8 shows the number and percentage of items identified with C DIF for the three course-taking contrasts for each form, separately for items measuring the old and new content. Some general patterns can be observed. Within each form and for both the old and new content, a higher percentage of items were flagged with C DIF for comparisons involving taking versus not taking (Took–Didn't Take), or taking versus just planning to take advanced courses (Took–Planned), than comparisons that involved planning versus not planning to take advanced courses (Planned–Didn't Plan). In fact, items that showed C DIF in the Took–Didn't Take contrast tended to show C DIF in the Took–Planned contrast as well.

Out of the 165 items that were pretested in the seven forms, 42 items (26 percent) were flagged with C DIF in the Took–Didn't Take contrasts, 14 items (9 percent) were flagged in the Planned–Didn't Plan contrasts, and 40 items (24 percent) were flagged in the Took–Planned contrasts. All but four of the items identified as having C DIF were found to favor the reference group. Three new items and one old item from Form 143 that were flagged in the Planned–Didn't Plan contrast favored the focal group. Higher percentages of new items were flagged with C DIF compared to the old items, especially for the Took–Didn't Take and Took–Planned contrasts. Across the seven forms, for the Took–Didn't Take contrast, the percentage of items flagged for C DIF ranged from 0 to 30 percent for the old content and ranged from 11 to 50 percent for the new content. Similarly, for the Took–Planned contrast, the percentages ranged from 0 to 25 percent

Table 8

Number and Percent of C DIF Items for Different Comparisons by Form

	All Items							*Old Items*							*New Items*						
	Total #	*Took–Didn't Take*		*Planned–Didn't Plan*		*Took–Planned*		*Total #*	*Took–Didn't Take*		*Planned–Didn't Plan*		*Took–Planned*		*Total #*	*Took–Didn't Take*		*Planned–Didn't Plan*		*Took–Planned*	
Form		*N*	*%*	*N*	*%*	*N*	*%*		*N*	*%*	*N*	*%*	*N*	*%*		*N*	*%*	*N*	*%*	*N*	*%*
21	26	7	26.9	1	3.8	7	26.9	8	1	12.5	0	0.0	1	12.5	18	6	33.3	1	5.6	6	33.3
22	26	4	15.4	1	3.8	3	11.5	12	0	0.0	0	0.0	0	0.0	14	4	28.6	1	7.1	3	21.4
31	18	5	27.8	3	16.7	5	27.8	9	2	22.2	1	11.1	2	22.2	9	3	33.3	2	22.2	3	33.3
41	18	7	38.9	2	11.1	6	33.3	10	3	30.0	2	20.0	2	20.0	8	4	50.0	0	0.0	4	50.0
123	18	2	11.1	1	5.6	1	5.6	9	1	11.1	0	0.0	1	11.1	9	1	11.1	1	11.1	0	0.0
133	30	6	20.0	2	6.7	5	16.7	12	2	16.7	2	16.7	1	8.3	18	4	22.2	0	0.0	4	22.2
143	29	11	37.9	4	13.8	13	44.8	8	1	12.5	1	12.5	2	25.0	21	10	47.6	3	14.3	11	52.4
Total	165	42	25.5	14	8.5	40	24.2	68	10	14.7	6	8.8	9	13.2	97	32	33.0	8	8.2	31	32.0

Note: Didn't Take=Took Algebra II but no advanced course; Took=Took Algebra II and one or more advanced course; Didn't Plan=Took Algebra II, didn't take nor plan to take any advanced course; Planned=Took Algebra II, didn't take but plan to take an advanced course.

for the old content and ranged from 0 to 52 percent for the new content. The percentage of C DIF items for the Planned–Didn't Plan contrast was similar for the old and new content, ranging from 0 to 20 percent for the old content and 0 to 22 percent for the new content.

Table 9 shows the number and percentage of items flagged with B DIF for each contrast. The results for the B DIF items did not reveal the same patterns observed for the C DIF results. In general, there were fewer items flagged with B DIF than with C DIF for the Took–Didn't Take and Took–Planned contrasts, especially for the items measuring the new content. Out of the 165 items pretested in the seven forms, 15 items (9 percent) were flagged with B DIF in the Took–Didn't Take contrasts, 18 items (11 percent) were flagged in the Planned–Didn't Plan contrasts, and 16 items (10 percent) were flagged in the Took–Planned contrasts. Taken together, the results in Tables 9 and 10 consistently show that the items pretested in the field trial to a large extent favored students who took one or more advanced mathematics courses compared to those who did not take any advanced courses. Planning to take advanced courses had some impact on students' performance, but this impact was far less substantial than the effect of actual course-taking. Furthermore, the effect of course-taking was more significant for items that measure the new content than for items that measure the old content.

To examine the possible content effects on differential item functioning, the items that were flagged with B or C DIF in the various contrasts were grouped by the categories of new content, and Table 10 shows the number and percentage of DIF items from each content category. The percentage of DIF items by content is also depicted in Figure 4. Note that many items covered more than one content category, and the total number of items across content add up to 140, instead of 97, which is the total number of distinct items across forms that measure the new content.

The three subcontent categories of Algebra II, namely AV (absolute value), VA (direct and inverse variation), and FM (functions as models), had a small or zero percentage of DIF items in all contrasts. The other eight content categories, however, had a percentage of DIF items higher than 33 percent in the Took–Didn't Take and Took–Planned contrasts. Specifically, the percentages of items flagged for DIF ranged from 33 percent for DR

Table 9

Number and Percent of B DIF Items for Different Comparisons by Form

	All Items							*Old Items*							*New Items*						
	Total #	*Took–Didn't Take*		*Planned–Didn't Plan*		*Took–Planned*		*Total #*	*Took–Didn't Take*		*Planned–Didn't Plan*		*Took–Planned*		*Total #*	*Took–Didn't Take*		*Planned–Didn't Plan*		*Took–Planned*	
Form		*N*	*%*	*N*	*%*	*N*	*%*		*N*	*%*	*N*	*%*	*N*	*%*		*N*	*%*	*N*	*%*	*N*	*%*
21	26	0	0.0	1	3.8	0	0.0	8	0	0.0	1	12.5	0	0.0	18	0	0.0	0	0.0	0	0.0
22	26	0	0.0	3	11.5	2	7.7	12	0	0.0	1	8.3	1	8.3	14	0	0.0	2	14.3	1	7.1
31	18	2	11.1	4	22.2	0	0.0	9	0	0.0	2	22.2	0	0.0	9	2	22.2	2	22.2	0	0.0
41	18	2	11.1	4	22.2	1	5.6	10	0	0.0	2	20.0	0	0.0	8	2	25.0	2	25.0	1	12.5
123	18	4	22.2	1	5.6	4	22.2	9	3	33.3	1	11.1	3	33.3	9	1	11.1	0	0.0	1	11.1
133	30	4	13.3	3	10.0	8	26.7	12	1	8.3	1	8.3	4	33.3	18	3	16.7	2	11.1	4	22.2
143	29	3	10.3	2	6.9	1	3.4	8	1	12.5	2	25.0	0	0.0	21	2	9.5	0	0.0	1	4.8
Total	165	15	9.1	18	10.9	16	9.7	68	5	7.4	10	14.7	8	11.8	97	10	10.3	8	8.2	8	8.2

Note: Didn't Take=Took Algebra II but no advanced course; Took=Took Algebra II and one or more advanced course; Didn't Plan=Took Algebra II, didn't take nor plan to take any advanced course; Planned=Took Algebra II, didn't take but plan to take an advanced course.

Table 10

Number and Percent of DIF Items for Different Comparisons by New Content

Content	Total #	B DIF						C DIF						B and C DIF					
		Took–Didn't Take		*Planned–Didn't Plan*		*Took–Planned*		*Took–Didn't Take*		*Planned–Didn't Plan*		*Took–Planned*		*Took–Didn't Take*		*Planned–Didn't Plan*		*Took–Planned*	
		N	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*
AV	14	0	0.0	1	7.1	0	0.0	1	7.1	0	0.0	1	7.1	1	7.1	1	7.1	1	7.1
EX	10	1	10.0	0	0.0	1	10.0	3	30.0	1	10.0	3	30.0	4	40.0	1	10.0	4	40.0
VA	6	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
FN	31	5	16.1	1	3.2	4	12.9	15	48.4	5	16.1	15	48.4	20	64.5	6	19.4	19	61.3
DR	15	1	6.7	1	6.7	1	6.7	4	26.7	3	20.0	4	26.7	5	33.3	4	26.7	5	33.3
FM	5	0	0.0	1	20.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	1	20.0	0	0.0
LE	20	3	15.0	3	15.0	2	10.0	8	40.0	1	5.0	7	35.0	11	55.0	4	20.0	9	45.0
QE	16	2	12.5	2	12.5	1	6.3	8	50.0	1	6.3	8	50.0	10	62.5	3	18.8	9	56.3
SP	9	1	11.1	2	22.2	1	11.1	3	33.3	0	0.0	2	22.2	4	44.4	2	22.2	3	33.3
QG	10	2	20.0	0	0.0	1	10.0	7	70.0	0	0.0	8	80.0	9	90.0	0	0.0	9	90.0
TR	4	0	0.0	0	0.0	0	0.0	3	75.0	2	50.0	3	75.0	3	75.0	2	50.0	3	75.0
Total	140	15	10.7	11	7.9	11	7.9	52	37.1	13	9.3	51	36.4	67	47.9	24	17.1	62	44.3

Note: AV=absolute value; EX=exponents (negative and rational); VA=direct and inverse variation; FN =function notation; DR=concepts of domain and range; FM=functions as models; LE=graphs and equations of linear functions; QE=graphs and equations of quadratic functions; SP=slopes of partial and perpendicular lines; QG=qualitative behavior of graphs and functions; TR=transformations and their effects on graphs of functions.

(concept of domain and range) to 90 percent for QG (qualitative behavior of graphs and functions) for both the Took–Didn't Take and Took–Planned contrasts. The percentages of items showing DIF in the Planned–Didn't Plan contrasts were generally smaller.

Discussion

The College Board has responded to the trends of mathematics course-taking in the nation's high schools by adding more advanced content to the SAT, but does the addition of this new content increase the SAT's "curriculum sensitivity"? That is, is the SAT now more closely linked to high school curriculum because the number or level of courses taken is related to performance on the test? If the SAT is not related to curriculum, then we should see little if any difference in the scores of students who take more rigorous courses or advance further in the curriculum than other students. On the other hand, if the SAT is curriculum relevant, then we should see higher scores among students who take more rigorous and more core courses in high school (Camara, 2003).

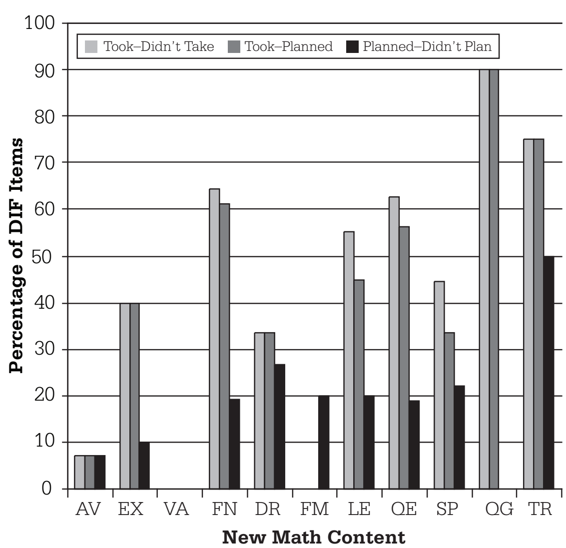

Figure 4. Percentage of B and C DIF items by new math content.

This study was designed to address this question, and the results suggest that course-taking is indeed more strongly related to performance on the new items than on the old items. The results from Study 1 indicated that students who took each course scored higher on both the old and new items than students who planned to take or didn't take the course, and the level of courses students took or planned to take correlated slightly higher with performance on the new items than with performance on the old items. The results from Study 2 indicated that students who took one or more advanced courses scored higher than those who didn't take any advanced course or just planned to do so. There were a larger percentage of items identified with DIF in the comparisons involving taking versus not taking, or taking versus just planning to take advanced courses, than in the comparisons that involved planning versus not planning to take advanced courses. This suggests that actually taking a course rather than just planning to take a course was the main factor contributing to DIF. In addition, a higher percentage of the items that measured the new content were flagged with DIF compared to the items that measured the old content. Taken together, the results from both studies suggest that students' experience taking or planning to take more rigorous mathematics courses benefit their performance on the math items, and the addition of new content to the SAT mathematics section potentially increased its curriculum sensitivity.

However, there are a myriad of methodological issues that precluded getting a definitive answer to this question. The most salient methodological challenge is self-selection. Because students self-select into courses, it is impossible to completely disentangle the effect of course-taking from the host of other factors that influence performance on the SAT. Students who take higher-level mathematics courses may be more interested in mathematics and more motivated to achieve than students who don't take these courses, and these interest and motivation factors are also related to high scores on the SAT. This was the case before the new content was added, and will remain the case in the future.

Another caveat to keep in mind is the nature of the sample of students that participated in the SAT field trial. The field trial sample is not completely representative of the population of students who take the SAT. The field trial was not administered under the same operational conditions as the actual SAT, and there was not the same motivation of students to perform their best. Future research should employ operational SAT data to mitigate these motivational effects. Currently the SAT Questionnaire does not include enough detail on course-taking patterns to allow a full replication of the analyses described in this report. However, there are plans to revise the SAT Questionnaire so that detailed information on course-taking will be collected in the near future. Furthermore, the pretest mathematics sections from the field trial that were the foci of the analyses were specifically designed with a large proportion of items covering the new mathematics content. These sections did not mirror an actual SAT mathematics section, which has only a few items with the new content. One cannot rule out the possibility that performance on both the old and new items in the field trial was subjected to contextual effects of a test heavily laden with new content.

Since few of the students in the field trial sample did not take nor plan to take advanced mathematics courses, the DIF comparisons involving the Didn't Plan group involved relatively small samples. In several cases the sample sizes were only slightly more than 100. The use of the PSAT as a matching variable for the DIF analyses presented some limitations. Different course-taking groups had consistently different PSAT score distributions across forms. Some score points on the low end of the scale were not represented in the distributions for the Took group, and an entire portion at the high end of the scale was not represented in the distributions for the Didn't Take, Didn't Plan, and Planned groups. Appendix B displays the PSAT score distributions for various course-taking groups for Form 21, and other forms share the same distributional patterns. Except for Form 31, the Didn't Take, Didn't Plan, and Planned groups had no PSAT scores falling in the range of 71–80. The inadequate and/or inconsistent scale coverage for the matching criterion in the reference and focal group distributions caused missing data for the SIBTEST procedure, and may have affected DIF estimation results to a certain degree.

Given these limitations, the findings should be interpreted with caution, and future studies are needed to gain a more definite understanding of the effect of taking advanced courses on students' mathematics performance. Replicating the DIF analyses with operational SAT data would allow larger sample sizes for the focal groups, and the distributions of the criterion scores can be expected to better cover the entire range of the score scale for various course-taking groups. In addition, it is recommended that differential bundle functioning analyses (DBF) be conducted using the operational data, with various mathematics content areas specified as item bundles, and especially those identified with a large percentage of DIF items in this study.

Finally, this study did not examine differences in the cognitive attributes of SAT items, and how performance on the old and new items differed by cognitive dimensions. There is a lot of variation among SAT items in the content and cognitive processes that are assessed. Future research may look at the cognitive dimensions tapped by the old and new SAT items to determine if the Rock and Pollack (1995) findings are replicated. That is, do the new items require a higher level of cognitive skill than the old items, and does taking certain courses give an advantage on items tapping such higher-level cognitive skills? Future research should also compare performance on the old and new items equalizing or holding constant cognitive demand and difficulty of the items.

References

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

College Board. (2005a). *Graph Set 5: Increase in the study of advanced math*. New York: The College Board. Retrieved December 15, 2006, from http://www.collegeboard.com/prod_downloads/about/news_info/cbsenior/yr2005/graph5-advanced-math.pdf.

College Board. (2005b). *New SAT for the press: Frequently asked questions*. Retrieved December 15, 2006, from http://www.collegeboard.com/about/news_info/sat/faqs.html.

Laing, J., Engen, H., & Maxey, J. (1990). Coursework patterns, ACT assessment scores, and related data. *College and University*, *65* (4), 335–44.

Liu, J., & Feigenbaum, M.D. (2003). *Prototype analysis of spring 2003 new SAT field trial* (ETS Statistical Report SR-2003-69). Princeton, NJ: Educational Testing Service.

Liu, J., Schuppan, F., & Walker, M.E. (2004, April). *Evaluating SAT II: Mathematics IC items in the SAT I population*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Maxey, J., Cargile, S., & Laing, J. (1987). Three measures of academic achievement and their association with performance on the ACT assessment. *Journal of the National Council of Educational Opportunity Associations*, *2*, 6–10.

Morgan, R. (1989). *An examination of the relationships of academic course work with admissions test performance* (College Board Research Report No. 89-6). New York: The College Board.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement*, *16*, 159–76.

Pelavin, S.H., & Kane, M. (1990). *Changing the odds: Factors increasing access to college*. New York: The College Board.

Rock, D.A., & Pollack, J.M. (1995). *Mathematics course-taking and gains in mathematics achievement*. National Center for Education Statistics (NCES Statistics in Brief, No. 95-714). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Roussos, L.A., & Stout, W.F. (1996a). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*, 355–71.

Roussos, L.A., & Stout, W.F. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, *33*, 215–30.

Schmidt, W.H. (1983). High school course-taking: Its relationship to achievement. *Journal of Curriculum Studies*, *15*, 311–22.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, *58*, 159–94.

Smith, A.F., & Prentice, D.A. (1993). Exploratory data analysis. In G. Keren and C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Statistical issues* (pp. 349–90). Mahwah, NJ: Lawrence Erlbaum Associates.

Stout, W., & Roussos, L.A. (1999). *Dimensionality-Based DIF/DBF Package* [Computer program]. William Stout Institute for Measurement: University of Illinois.

U.S. Department of Education. (2002). *High school and beyond longitudinal study of 1980 sophomores: High school transcript study (HS&B-So:80/82);* and *The high school transcript study (HSTS)*. Available from National Center for Education Statistics Web site, http://nces.ed.gov/programs/digest/d04/tables/dt04_136.asp.

Appendix A

Table A1

Number of Items for Each New Content Category by Form

Content	*Classification*	*Description*	*Form 21*	*Form 22*	*Form 31*	*Form 41*	*Form 123*	*Form 133*	*Form 143*	*Total*
Algebra and Functions	AV	Absolute value	3	2	1	1	1	3	3	14
	EX	Exponents (negative and rational)	3	1	0	2	0	2	2	10
	VA	Direct and inverse variation	3	1	0	0	1	0	1	6
	FN	Function notation	4	6	3	2	4	6	6	31
	DR	Concepts of domain and range	5	3	1	0	2	2	2	15
	FM	Functions as models	1	0	1	0	0	1	2	5
	LE	Graphs and equations of linear functions	4	3	2	2	3	3	3	20
	QE	Graphs and equations of quadratic functions	1	3	2	2	2	3	3	16
Geometry	SP	Slopes of partial and perpendicular lines	1	1	1	2	1	1	2	9
	QG	Qualitative behavior of graphs and functions	2	1	2	0	0	1	4	10
	TR	Transformations and their effects on graphs of functions	1	0	0	0	0	0	3	4

Appendix B

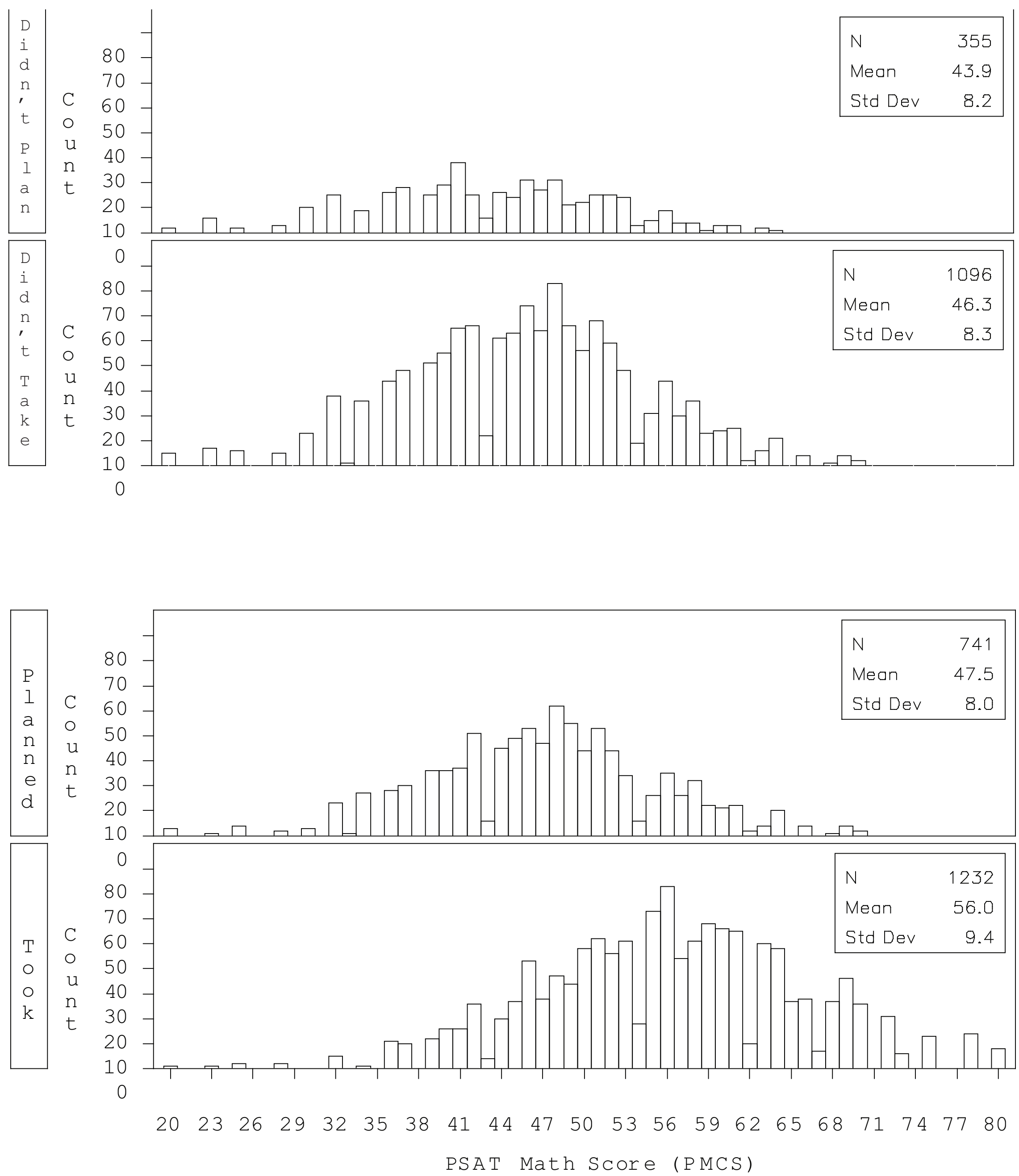

Figure B1. Distributions of PSAT scores by course-taking group for form 21.

CollegeBoard
connect to college success™