DOCUMENT RESUME

ED 390 892 TM 024 173

AUTHOR Hale, Gordon A.; And Others
TITLE A Comparison of the Predictive Validity of the Current SAT and an Experimental Prototype. Research Report.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-92-32
PUB DATE Jun 92
NOTE 61p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *College Entrance Examinations; Colleges; Comparative Analysis; Higher Education; High Schools; *Mathematics Tests; *Predictive Validity; Sex Differences; *Test Construction; *Verbal Tests
IDENTIFIERS Composite Scores; Prototypes; Revision Processes; *Scholastic Aptitude Test; *Test Revision

ABSTRACT

As part of a large-scale project to remodel the Scholastic Aptitude Test (SAT), this study examined the predictive validity of a prototype revised SAT, which incorporated many of the important structural changes planned for the test. This prototype was compared to a form of the current SAT with regard to several validity-related issues. Nineteen colleges, with at least 280 participating students at each, provided study data. The results showed that the test revisions: (1) tended to increase predictive validity for the verbal score, the mathematical score, and the composite of verbal and mathematical scores; (2) slightly increased the incremental validity of the test over high school rank; and (3) produced a modest reduction in gender-related prediction differences for the verbal score. This evidence is regarded as preliminary, as the data were collected under experimental conditions with a limited sample of colleges, and the prototype used here was not identical in form to the remodeled SAT as it will be implemented operationally. Tentatively, however, the results were consistent with the goals of the overall project in that, where the present revisions had effects, they tended to be in the desired direction. An appendix presents analyses of the sections of the original and the prototype revision of the SAT. (Contains 10 tables and 14 references.) (Author/SLD)

RESEARCH REPORT

A COMPARISON OF THE PREDICTIVE VALIDITY OF THE CURRENT SAT AND AN EXPERIMENTAL PROTOTYPE

Gordon A. Hale
Brent Bridgeman
Charles Lewis
Judith M. Pollack
Ming-mei Wang

Educational Testing Service
Princeton, New Jersey
June 1992

A Comparison of the Predictive Validity of the Current SAT and an Experimental Prototype

Gordon A. Hale
Brent Bridgeman
Charles Lewis
Judith M. Pollack
and
Ming-mei Wang

Educational Testing Service

Princeton, NJ

Acknowledgments

The authors express their sincere appreciation to:

John Fremer, Steve Graff, Larry Hecht, and Larry Litten for their leadership of the New Possibilities Project, of which the present study was a part
Nancy Burton, Sydell Carlton, Linda Cook, Samuel Messick, Rick Morgan, Donald Rock, and Warren Willingham for providing general advice regarding the methodology of the study
Gloria Weiss for supervising the data collection
Elizabeth Burton, Anne Neeff, and Brian O'Reilly for making arrangements for data collection
Sheila Krolikowski and the Operations staff for preparing the data for analysis
Neil Dorans, Nancy Feryok, Ida Lawrence, and Samuel Livingston for conducting preliminary analyses of the test data
Lucient Chan and Laura McCamley for assisting in data analysis
James Braswell, Edward Curley, and Marilyn Sudlow for furnishing information about the development of the test
Donald Powers and Howard Wainer for reviewing drafts of this report
Eleanore DeYoung, Joanne Farr, and Ruth Yoder for providing secretarial assistance

The authors wish to express particular gratitude to the institutions listed below, whose participation made the study possible.

Augustana College, Rock Island, IL
Babson College, Wellesley, MA
Bridgewater State College, Bridgewater, MA
Clarkson University, Potsdam, NY
Colorado School of Mines, Golden, CO
Fisk University, Nashville, TN
Flagler College, Saint Augustine, FL
Indiana University, Bloomington, IN
Lincoln Memorial University, Harrogate, TN
Lock Haven University, Lock Haven, PA
Manchester College, North Manchester, IN
Mansfield University, Mansfield, PA
Marywood College, Scranton, PA
Merrimack College, North Andover, MA
North Georgia College, Dahlenoga, GA
Roanoke College, Salem, VA
Saint Leo College, Saint Leo, FL
Saint Peter's College, Jersey City, NJ
San Jose State University, San Jose, CA
Seton Hall University, South Orange, NJ
University of California, Riverside, CA
University of New Haven, West Haven, CT
Vanderbilt University, Nashville, TN
Villanova University, Villanova, PA
Westminster College, New Wilmington, PA
Whittier College, Whittier, CA
Williams College, Williamstown, MA

Abstract

As part of a large-scale project to remodel the Scholastic Aptitude Test (SAT), this study examined the predictive validity of a prototype revised SAT, which incorporated many of the important structural changes planned for the test. This prototype was compared to a form of the current SAT with regard to several validity-related issues. The results showed that the test revisions (a) tended to increase predictive validity for the verbal score, the mathematical score, and the composite of verbal and mathematical scores; (b) slightly increased the incremental validity of the test over high school rank; and (c) produced a modest reduction in gender-related prediction differences for the verbal score. Also, inclusion of scores from a new writing test tended to increase validity and to reduce gender-related prediction differences. This evidence is regarded as preliminary, as the data were collected under experimental conditions with a limited sample of colleges, and the prototype used here was not identical in form to the remodeled SAT as it will be implemented operationally. Tentatively, however, the results were consistent with the goals of the overall project in that, where the present revisions had effects, they tended to be in the desired direction.

Table of Contents

List of Tables

Introduction

Background

Since 1987, work has been under way to revise the Scholastic Aptitude Test (SAT). Conducted under the rubric "New Possibilities Project," this work has been directed toward making the SAT more educationally relevant to all segments of the test-taking population and to provide more meaningful information to colleges, high schools, counselors and students. Toward this end, both the verbal and mathematical sections of the SAT, now to be called "SAT-I: Reasoning Tests," are to be revised and expanded. In addition, a test of writing ability, including an essay component, is to be incorporated into the companion tests that have been known as the Achievement Tests, now to be called "SAT-II: Subject Tests". The plan for revision of the test is detailed in the College Board announcement "Background on the New SAT-I and SAT-II" (College Entrance Examination Board, 1990a). Some excerpts from that announcement follow.

> Inclusion of critical reading passages and questions [in the verbal section of SAT-I] will bring the test into closer alignment with current professional thinking about how reading ability develops and how it is best assessed. The new critical reading passages will be longer than the reading passages in the current SAT and will better allow assessment of the ability of students to evaluate and make judgments about points of view expressed in written passages, an important skill required in much college reading....Vocabulary knowledge will continue to be tested through the use of vocabulary-in-context questions....
>
> The mathematical component of SAT-I will include questions that require students to produce a response--not just to select a response from a set of multiple-choice alternatives. This new format (sometimes informally referred to as "grid-in" questions) will make up about 20 percent of the proposed new mathematical test. The rest of the test will consist of established problem-solving questions in five-choice and quantitative-comparison formats. The new test will emphasize the application of mathematical concepts and the interpretation of data.
>
> The writing test in SAT-II will consist of a combination of multiple-choice questions and a direct writing sample at each of five administrations per year. The new SAT writing test will replace the all-multiple-choice Test of Standard Written English (TSWE), which currently accompanies every SAT, and the English Composition Test (ECT), which currently is offered with a direct writing sample only once a year. The writing test will be offered each time the SAT-II series is administered. It will be designed to be useful in both admissions and placement, and in that regard will combine into a single test the functions of the current ECT and TSWE.

Purpose of the Present Study

As the SAT is revised, it is critical that the psychometric integrity of the test be maintained. An important issue in this regard concerns the predictive validity of the test, as typically measured by the test's ability to predict students' academic performance during their first year in college. The present study was conducted to provide initial information about the predictive validity of a revised SAT, in comparison with the current test, and to address certain key questions related to predictive validity. It must be stressed that the major objective of the New Possibilities Project, of which this study was a part, was not to improve the predictive validity of the test. Rather, the goal was to implement changes designed to enhance the SAT from a content-related standpoint, while seeking to ensure that the test's validity would not be compromised by doing so.

For this study, a prototype revised SAT was developed, which is herein labeled the New Possibilities Project test, or NPP test. The prototype contained the essential elements of the proposed new test, as it was envisioned at the time the present study was initiated, in the fall of 1989. It must be noted that, since the present study was conducted, further planning has led to some additional changes in the proposed format of the revised SAT. Notably, it has been decided to allow use of calculators in the mathematical section of the new SAT-I, whereas calculator use was not permitted in the present study. Also, the sentence-completion item type from the current SAT is now to be retained in SAT-I, although it was not included in the NPP test used here. And, due to procedural constraints, the timing of each test section in the current study was set at 30 minutes. By contrast, a key feature of SAT-I will be the extension of the time limits of the verbal and mathematical sections, to accommodate the changes in the test structure; and in the writing test, the time limits of the multiple-choice and essay sections are likely to be 40 and 20 minutes, respectively. For these reasons, the present study is regarded as an interim investigation and not a test of the validity of SAT-I and the writing test as they will ultimately be implemented in operational form. Nevertheless, the prototype NPP test used here incorporates the most substantial structural revisions that have been planned for the revised SAT. Thus, by examining the effects of these revisions, the study provides useful preliminary information bearing on the validity of the remodeled SAT.

Issues under Study

The most central issue under investigation was whether the proposed revisions substantially altered the test's predictive validity. This issue was addressed through comparison of the current SAT and the NPP test with regard to predictive validity, and two principal questions were posed. The first question concerned the separate area scores: Is there a substantial difference between the NPP and current versions of the test in predictive validity of either the verbal or the mathematical parts? The second question concerned composite scores: Is there a substantial difference

between versions of the test in predictive validity when prediction is based on the verbal and mathematical area scores taken in combination?

A second major issue concerned the incremental validity of each version of the SAT over high school record. Because the SAT is typically used along with high school record (among other information) in making admissions decisions, it is essential to determine the extent to which prediction is improved when the SAT is added to high school record, and whether the degree of improvement is noticeably greater for one version of the test than the other.

A third issue had to do with differential prediction by gender. Previous research has found that GPA tends to be underpredicted slightly by the SAT for women relative to men.[1] In the present study, the prediction difference was examined for both the current SAT and the NPP test, in order to determine whether, and to what extent, the revisions of the test affected the gender-related prediction difference.

A final issue concerned the role of the writing test. Although this test is to be administered separately as part of the new SAT-II, it was believed that the score on this test could be a useful source of information along with the verbal and mathematical scores in predicting college GPA. Thus, a question of interest was whether predictive validity would be improved by the inclusion of the writing score along with scores from the verbal and mathematical parts of the SAT in the prediction equation.

To address these issues, an experimental study was conducted in a group of colleges that volunteered to participate. Students entering these colleges were administered components of the NPP test or the current SAT in a special test administration conducted during the orientation period preceding the beginning of classes. Scores on these tests, as well as high school performance, were then related to the students' grade-point averages at the end of the year.

---

[1]Various hypotheses have been offered to account for this phenomenon, including the fact that there are gender differences in courses taken by the two groups (cf., Elliott and Strenta, 1988; Young, 1991). Specific explanations for the phenomenon were not pursued in the present study, as the main objective here was to compare versions of the test with respect to differential prediction by gender, with full GPA as the criterion.

Method

Sample of Colleges

The participating colleges were ones that expressed interest, in response to a general solicitation of four-year coeducational colleges and universities in the United States with freshmen enrollments of 280 or greater. (For simplicity, the term "colleges" will be used hereafter to refer to the participating institutions.) A minimum of 280 students was needed to ensure that sufficient numbers would be assigned to each of the 21 combinations of test components, or spirals (see Experimental Design). A total of 27 colleges agreed to participate and administered the test materials. Of those colleges, 19 were used in the final analyses; the eight other colleges were not included in the analyses due to insufficient numbers of participating students. The 19 colleges in the final sample had the characteristics indicated in Table 1, as listed in The College Handbook (College Entrance Examination Board, 1990b).

To provide a general context for comparison, it is useful to note the characteristics of four-year colleges in general, as indicated in Annual Survey of Colleges, 1989-90: Summary Statistics (College Entrance Examination Board, 1989). These data are shown below under the heading "All Colleges." Although the freshman class data in this column are for the clas entering in 1988, these data are not expected to differ substantially from those of the 1989 class.

	Colleges in Present Sample	All Colleges
Mean freshman class size:	975	738
Mean percentage of minority students in freshman class:	12%	19%
Control: Private:	68%	55%
Public:	32%	45%
Percentage of colleges in each region:		
Middle States:	37%	23%
Midwestern:	16%	28%
New England:	21%	9%
Southern:	11%	20%
Southwestern:	0%	8%
Western:	16%	14%
Doctorate granting institutions:	21%	19%

Table 1

Characteristics of Colleges in Sample

	Fresh. class size	Percent minority freshmen	Type of institution	Region[a]
1.	6006	7	public university	Midwestern
2.	2448	52	private university	Western
3.	1597	10	private university	Middle States
4.	1046	9	public liberal arts college	New England
5.	835	4	public university	Middle States
6.	714	9	public university	Middle States
7.	680	2	private university	Middle States
8.	584	5	private liberal arts college	New England
9.	545	4	private liberal arts college	Midwestern
10.	523	24	private liberal arts college	New England
11.	494	3	public liberal arts college	Southern
12.	434	41	private liberal arts college	Middle States
13.	429	3	private liberal arts college	Southern
14.	402	7	private liberal arts college	Midwestern
15.	382	12	public engineering college	Western
16.	368	2	private liberal arts college	Middle States
17.	363	6	private business college	New England
18.	362	3	private liberal arts college	Middle States
19.	307	31	private liberal arts college	Western

[a]Regions are defined in The College Handbook (College Entrance Examination Board, 1990b).

Compared with four-year institutions in general, then, the present sample had a somewhat greater than average representation of the Middle States and New England regions, and a somewhat smaller than average representation of minority students. The sample was roughly comparable to the general population of four-year institutions with regard to percentages of private versus public institutions and doctorate versus nondoctorate granting institutions.

Concerning freshman class size, it is most useful to compare the distribution of freshman class enrollments for colleges in this study with that of U.S. colleges in general. Toward this end, the database for the 1988 edition of The College Handbook (College Entrance Examination Board, 1988), which contains information for 1987 freshman classes, was available to the authors. From this database, it was determined that 21% of four-year institutions had freshmen enrollments exceeding 1000 students, 19% between 501 and 1000, 17% between 300 and 500, and 42% below 300. The comparable percentages in the present sample were 21%, 31%, 47%, and 0%. In general, then, the present sample contained an overrepresentation of medium-sized colleges and an underrepresentation of small colleges, due to the necessity to exclude the latter for methodological reasons noted above.

With regard to type of institution, the Annual Survey of Colleges, 1989-90: Summary Statistics (College Entrance Examination Board, 1989) includes the category "private liberal arts colleges" and indicates that 15% of four-year institutions fall into this category; in contrast, 47% (9 out of 19) of the institutions in the present sample fell into this category. Thus, private liberal arts colleges were more heavily represented in the present sample than among four-year institutions in general.

For the present sample, admissions test requirements were: (a) SAT required: one institution, (b) either SAT or ACT required, but SAT preferred: eight institutions, (c) Either SAT or ACT required, no preference given to either: 10 institutions. Means of 25th and 75th percentiles of SAT scores (available for 15 colleges) were. SAT-Verbal scores: 417 - 534; SAT-Mathematical scores: 466 - 584. Admissions test information was not available for four-year colleges in general.

Sample of Students

Students were assembled in one or more testing rooms at orientation time. Participating students were those who took the test, upon being instructed that participation was not mandatory. This instruction was included in a letter from Educational Testing Service, handed to the students just prior to the test administration, thanking them for their assistance. It is not possible to determine exactly what percentage of students chose to participate, because data are not available on the number

of students present before the instructions were given. It is known, however, that a number of students opted not to participate.[2]

The sample consisted principally of freshmen, although one college reported having included transfer students as well. Of students in the final sample, 92% reported having graduated from high school in 1989 (i.e., in the spring prior to initiation of this study), 3% in 1988, 3% prior to 1988, and 2% did not respond.

Among students who participated, motivation to perform well could not, of course, be assured in an experimental testing situation such as this. In order to eliminate at least the most obvious cases of low motivation, the performance data were analyzed in an attempt to find students who were performing substantially below expectation. Students were identified who (a) marked their answers according to a clearly identifiable repetitive pattern, according to two judges--e.g., marking answers "abcde" for the first five questions and again for the next five questions, and so forth--(pattern responders were approximately 2/10 of 1% of the total number of examinees); (b) failed to respond in both of the subtests administered (3.8% of examinees), where failure to respond was defined as not answering at least three questions in a multiple-choice subtest, or failure to write an essay that addressed the topic (as judged by raters) in the essay subtest, and (c) failed to respond to any of the Grid-in mathematics items, for students administered this item type (2/10 of 1% of the total number of examinees, or 7/10 of 1% of those administered this item type; see definition and sample description below).

In addition, an outlier analysis was performed, designed to exclude students whose performance was substantially below that predicted by their scores on previously administered standardized tests (thus suggesting lack of motivation here). For this analysis, the data for the students participating in the present study were combined with the data for the approximately 4000 students participating in a companion study in which the same or similar test materials were administered under comparable testing conditions (Bridgeman, Hale, Lewis, Pollack, and Wang, 1992). The predictor was one of the following: (a) the student's most recent SAT score, obtained from Educational Testing Service data files or the participating college's records, (b) American College Test (ACT) score provided by the participating

---

[2]Before test materials were sent to the colleges, the 19 colleges in the final sample had indicated an expected total of 11865 test takers. Thus, the number of actual test takers, 7833, was 66% of the total number originally expected. Because the colleges' original expectations were gross approximations, based on enrollment estimates made in late summer, this percentage is not an actual participation rate, but it at least provides a rough approximation to a participation rate. Note that, in each institution that had a relatively large freshman class size, only a fraction of the class was enlisted for participation, for reasons of logistics. Usually these were students in attendance at selected orientation sessions, where such orientation sessions were offered at several different times.

college, (c) student's self-reported SAT score, or (d) student's self-reported ACT score. Four sets of regressions were computed, each using one of these four scores as a predictor variable. Within each set, separate regressions were computed for each of the 10 test sections as the dependent variable (see Test Sections, below).

Using the data from these regression analyses, the following procedure was employed to determine whether a given student should be eliminated from the sample. If the student had an SAT score on file, only the first of the above-mentioned four sets of regressions was used for that student; if not, the predictor was the first available of the other three: i.e., recorded ACT score, self-reported SAT score, and self-reported ACT score. In the applicable regression analysis, if the student's score was found to be two standard deviations or more below the predicted score for either of the two test sections administered to that student (see Experimental Design) the student was eliminated from the final sample. Those eliminated from the final sample for the present study comprised 6.1% of all examinees in this study.

The sample was further reduced by excluding the 4.8% of students for whom a valid first-year college grade-point average was not available. The final sample on which the data analyses were based consisted of 6,641 students, with the number per college ranging from 163 to 667.

According to the students' responses to a few background questions, the final sample consisted of (a) 54% women, (b) 11% minority students, (c) 51% with a probable major in the humanities or social sciences, 31% biological or physical sciences, and 18% "other" or "undecided"; (d) 54% ranking in the top fifth, 26% in the second fifth, and 18% in the middle fifth of their high school classes, and (e) students with a mean high school grade-point average of 2.9 on a 4-point scale.

Test Materials

For the form of the current SAT used here, test items were taken from Form 3HSAO2, originally administered as an operational test in March 1985. The two major scores on this test were the verbal and mathematical scores. For the NPP test, items were developed specially for the present study. In addition to the verbal and mathematical parts, the NPP test battery included the writing test, which consisted of a section with multiple-choice items and an essay section.

Test Sections. In all, the materials used in the study included 10 different test sections, each requiring 30 minutes to administer. The sections were constructed as follows. (The item types making up each test section are defined below.)

CV1: Current SAT, Verbal Section 1: 45 items
15 Antonym items, 10 Sentence Completion items, 10 Reading Comprehension items, 10 Analogy items

CV2: Current SAT, Verbal Section 2: 40 items
10 Antonym items, 5 Sentence Completion items, 15 Reading Comprehension items, 10 Analogy items

CM1: Current SAT, Mathematical Section 1: 25 items
25 five-choice mathematics items

CM2: Current SAT, Mathematical Section 2: 35 items
15 five-choice mathematics items, 20 Quantitative Comparison items

NV1: NPP test, Verbal Section 1: 25 items
25 Critical Reading items based on two passages

NV2: NPP test, Verbal Section 2: 35 items
13 Synonyms in Context items; 12 Critical Reading items based on one double passage; 10 Analogy items

NM1: NPP test, Mathematical Section 1: 33 items
15 five-choice items, 18 Quantitative Comparison items

NM2: NPP test, Mathematical Section 2: 22 items
12 four-choice Algebra Placement items, 10 Grid-in items
(For reasons explained below, all analyses were based on a score labeled NM2e, which excluded the Algebra Placement items.)

NW-MC: NPP test, Writing, multiple-choice section: 43 items
28 Usage items, 8 Sentence Correction items, 7 Revision-in-Context items

NW-ESS: NPP test, Writing, essay section
One essay question

The score for each multiple-choice test section was the number of items answered correctly minus 1/(K-1) times the number answered incorrectly, where K equals the number of response options. (One of the sections contained a group of four-choice items and a group of five-choice items; each group of items was formula scored separately, and the section score was the sum of scores for these two groups of items.) The score for the Grid-in portion was simply the number of items answered correctly. For the essay, scores ranged from 2 to 12 and were computed as the sum of scores assigned by two readers on a six-point scale.

Item Types

The item types used in the current SAT and the NPP test are described below. All multiple-choice item types contained five answer options, except for the Quantitative Comparison and Algebra Placement items, which contained four options.

Item types used in the current SAT verbal sections were as follows.

Antonyms: This item type is designed primarily to test the extent of the examinee's vocabulary. Each item consists of a word in capital letters followed by five lettered words or phrases. The examinee chooses the word or phrase that is most nearly opposite in meaning to the word in capital letters.

Analogies: Analogy items test the examinee's ability to see a relationship in a pair of words, to understand the ideas expressed in the relationship, and to recognize a similar or parallel relationship. Each item consists of a related pair of words or phrases, followed by five lettered pairs of words or phrases. The examinee selects the lettered pair that best expresses a relationship similar to that expressed in the original pair.

Sentence Completions:

These items test the examinee's ability to recognize relationships among parts of a sentence. Each item has a sentence with one or two words missing. Below the sentence, five words or pairs of words are given. The examinee must choose the word or set of words that best fits the meaning of the sentence as a whole.

Reading Comprehension:

These items test the examinee's ability to read and understand a text passage. The form of the current SAT used in the present study included passages ranging from approximately 230 to 480 words, which is typical of the SAT in its present form. Following each passage are several questions. For each question the examinee must choose the best answer from among five options.

Item types used in the current SAT mathematical sections were:

Five-choice items:

Within this item type, some items require application of numerical, graphic, spatial, symbolic, and logical techniques to familiar situations. Other items may require some original thinking. The questions involve concepts covered in one year of algebra plus some geometry. For each item, the examinee must select the correct answer from among five options.

Quantitative Comparisons:

This type of item requires the same general skills described for the five-choice items. However, the item format is different, emphasizing the concepts of equalities, inequalities and estimation. In a given item, two quantities are presented, one in each of two columns, and the examinee indicates which (if either) quantity is greater, or if the relationship cannot be determined.

The NPP test verbal sections included the following item types, along with the analogies item type described above.

Critical Reading:

These items focus on the abilities to interpret, synthesize, analyze, and evaluate the information in text passages. Also, some items test knowledge of vocabulary words appearing in the passages. The passages differ from those used in the current SAT in that (a) they are considerably longer (ranging from approximately 580 to 860 words in the forms used in this study), (b) they are designed to be more accessible and engaging to the reader, and (c) they have introductory material that places them in a broader context. Also, one of the "passages" in the NPP test is a double passage presenting two points of view, with certain questions requiring a comparison or contrast of the two passages. Each passage is followed by several questions, and for each question the examinee must choose the best answer from five options.

Synonyms in Context:

A short sentence is presented, in which either one word is underlined, or two words from different parts of the sentence are underlined. The examinee must choose, from five options, the word or pair of words that have the same meaning as the one(s) underlined.

The NPP test mathematical items, like those on the current SAT, emphasized problem solving in the domains of arithmetic, algebra, and geometry. The NPP test included the following item types along with the Quantitative Comparison item type described above.

Five-choice items:

These items are similar in nature to the five-choice items of the current SAT, but with two changes in format. First, to provide a format consistent with that of the Grid-in items described below, the right half of each page is left

blank and headed "Use this space for scratchwork"; in the current SAT, scratchwork may be done on any page, but no specific space is allocated for it. Second, the few reference formulas and definitions appearing above the five-choice items in the current SAT (e.g., formulas for the areas of a circle and triangle) are excluded here as they were assumed to be commonly known.

Grid-in Items:

In this item type, the examinee constructs a response rather than selecting an answer from among several options. Scratchwork space is provided as described above. The examinee writes, then grids in the response. The answer sheet allows the examinee to grid up to four characters--numerals 0 through 9, a decimal point, or a slash for representation of fractions.

Algebra Placement Items:

These items contain the routine types of problems students must solve in algebra classes, for which there is a standard algorithm. Their name derives from the fact that they assess the kinds of skills required for making placement decisions at the level of elementary college algebra. For each question, the examinee chooses the correct answer from among four options. (Because it subsequently became apparent that these items would not be included in the remodeled SAT, performance on these items was not included in scores used in the predictive validity analyses.)

In the writing test, multiple-choice portion, the items were designed to test the examinee's ability to recognize and use correct standard written English.

Usage: Each of these items consists of a sentence in which four short portions of the sentence are underlined, followed by a fifth underline, "No error." The examinee must identify the portion containing an error, if there is one.

Sentence Correction:

Each item consists of a sentence, part or all of which is underlined. Four possible revisions of the underlined portion appear below the sentence, along with a fifth option, which is a repetition of the underlined portion of the sentence. The examinee must choose the option that produces the most effective sentence.

Revision-in-Context:

> A text passage, described as the first draft of an essay, is followed by several questions. Some questions ask the examinee to select the best rephrasing of part of the passage. Other questions ask about the writer's composition strategies or the organization of the passage. The examinee must choose, from among five options, the answer that most effectively makes the intended meaning clear.

In the writing test, essay portion, the examinee was given the following prompt:

> In a well-organized composition, describe a situation in which it was necessary for people to find a balance between conflicting interests. Include the following in your composition:
>
> A description of the circumstances of the conflict
>
> A discussion of problems encountered by those attempting to reach a compromise or balance between the conflicting interests
>
> An explanation of what was learned by those involved
>
> Be sure that you support your discussion. You may wish to use an example or examples from your own experience, your observation of others, or your reading in history, literature, science, or current events.

Area Scores

The following area scores were derived, based on the combination of individual section scores defined above.

CV	Current SAT, verbal area (CV1 + CV2)
CM	Current SAT, mathematical area (CM1 + CM2)
NV	NPP test, verbal area (NV1 + NV2)
NMe	NPP test, mathematical area, excluding Algebra Placement items (NM1 + NM2e); to be used in all analyses of test combinations, as explained below
NW	NPP test, writing area, defined as the combination of the multiple-choice section (NW-MC) and the essay (NW-ESS), weighted 70%-30%, respectively (in this study, .70 times NW-MC divided by

its standard deviation plus .30 times NW-ESS divided by its standard deviation).[3]

Combinations of Test Components

Throughout the period of the New Possibilities Project, it has been assumed that an NPP test composite that includes the writing score is of interest in addition to the verbal/mathematical combination, under the assumption that the writing score could contribute valuable additional information. Thus, it was regarded as important to examine the validity of predictor combinations that included the writing score as well as the validity of scores from the verbal and mathematical sections of the test.

As mentioned above, 12 four-choice Algebra Placement items were included in one mathematical section of the NPP test administered to the students. However, it has since been decided that such items will not be included in the remodeled SAT that is ultimately implemented in operational test administrations, so these items were excluded from the principal data analyses in this study. Therefore, in all principal analyses here, the mathematical part of the NPP test is represented by the score labeled NMe, which is based on all mathematical items that were administered, excluding the Algebra Placement items. Also, in analyses involving individual test sections, section NM2 (the second NPP mathematical section) is replaced by subsection NM2e, which is based on all items administered in that section, excluding the Algebra Placement items.

In light of these considerations, the principal combinations of test components to be analyzed are as follows.

Current SAT:	CV and CM
NPP test without writing test:	NV and NMe
NPP test with writing test:	NV, NMe, and NW

Experimental Design

Because the time available did not permit administration of a full test to any given student, each student was administered two test sections, each requiring 30 minutes. Twenty-one different pairs of sections were spiralled randomly within a college: (a) six representing all possible pairs of the four sections of the current SAT, and (b) 15 representing all possible pairs of the six sections of the NPP test, including the two writing test sections. Although all 21 pairs were spiralled randomly, students administered parts of the current SAT are labeled the "current SAT sample"

---

[3]Weighting of the two parts closely mirrors the weighting of the multiple-choice and essay portions of the English Composition Test (ECT), at those administrations in which the essay has been included. The ECT, which has been offered among the Achievement Tests, is to be replaced by the new writing test.

in analyses presented below, and those administered parts of the NPP test are labeled the "NPP test sample". Given that the numbers of students per college in the final sample ranged from 163 to 667, the average number receiving each spiral ranged from about 8 (at the college with the smallest number of students) to about 32. The statistical procedures used, in effect, permitted estimation of the validity coefficients that would have been obtained for the current SAT and the NPP test, if it had been possible to administer a full test to each student.

Grade-Point Average and High School Rank

For each college, grade-point averages for the participating students were computed on a four-point scale, with A = 4, and F = 0. Every student who received grades for at least one term was included in the sample, with the grade-point average for each student based on the total number of courses taken by that student. Thus, for students who remained enrolled for the full year (98% of the sample), the grade-point average used in the analysis was that based on all courses taken during the year. For the remaining 2% of the students, who apparently dropped out of college after the first term, the grade-point average for the first term was used in the analysis.

Sixteen of the 19 colleges furnished data on students' high school ranks, which were converted to percentiles. For the other three colleges, students' self-reported high school ranks were used. (Analyses for the first 16 colleges had shown a high degree of correspondence between self-reported and college-reported high school ranks.) Students indicated their ranks according to the following categories: (a) top fifth: highest tenth, (b) top fifth: second tenth, (c) second fifth, (d) middle fifth, (e) fourth fifth, and (f) lowest fifth; each student was assigned a percentile representing the midpoint of the category selected.

Methods of Analysis

Within each college, the Expectation and Maximization (EM) algorithm (Little & Rubin, 1987) was used to obtain the Maximum Likelihood (ML) estimate of the variance-covariance matrix for the section scores in the current SAT, high school rank, and GPA from the incomplete test data (inherent in the design for test administration explained earlier). Likewise, an EM algorithm was employed to estimate the variance-covariance matrix for the section scores in the NPP test, high school rank, and GPA from the observed test scores and reported college performance.

A multivariate normal distribution for the test and college performance scores is assumed in the ML estimation procedure. In effect, the EM method estimates the variance-covariance matrix that would have been obtained if every section of the test (current or NPP) had been administered to every participating student in the college. The procedure implicitly makes an adjustment for the observed variances and covariances on the basis of partial information provided by all students having taken at least one part of the test and having had their GPA reported.

The EM method was employed in order to obtain more stable estimates of variances and covariances. This is an important concern in light of the small number of students within each college taking any given pair of test sections as dictated by the study design.

Essentially, all participating students in a college with a GPA available have contributed some information to the estimation of one of the variance-covariance matrices (current SAT or NPP test). Thus, for each college, the number of such students associated with each of these variance-covariance matrices will be identified as the total sample size for that matrix. Using this convention the total sample size across the 19 colleges for the current SAT is 1948, and for the NPP test, 4693.

All analyses and results described in the remainder of this report have as their basis the estimates obtained using the EM algorithm. This applies to all correlations, both those arising from variable-weights analyses and those from equal-weights analyses for predictive and incremental validity analyses (described below). It also applies to all means used in the analyses by gender.

Predictive Validity

Analyses involving test sections and areas. The predictive validity of a given component of the test (individual section or combination of sections) was obtained as its correlation with GPA. In the principal analyses, predictive validities were estimated for each of the area scores, CV, CM, NV, NMe, and NW. In supplementary analyses, predictive validities were also estimated for each section (i.e., CV1, CV2, NV1, NV2, NM1, NM2e, NW-MC, and NW-ES).

Variable-weights analyses for combinations of test scores. The principal set of analyses used variable weighting of scores to examine the predictive validity of each of three combinations of test scores: (a) CV and CM, (b) NV and NMe, and (c) NV, NMe, and NW. Empirical weights were derived to form composite scores, and predictive validities were obtained for each of the composite scores so defined. The weights were determined for each college with a linear multiple-regression model for predicting GPA. Estimates of the regression weights for each area test score were used to define the composite scores. The variable-weights analyses used the above-mentioned combinations of test components as predictors.

The regression analyses were performed based on the EM estimates of the variance-covariance matrices for the test area scores and the college performance scores within each college. Because the total number of participating students in each college tended to be small, an Empirical Bayes (EB) procedure was employed to obtain simultaneous estimates of the within-college regression coefficients (cf., Braun & Jones, 1985). In essence, the EB approach utilizes collateral information on the relationships between college performance and the test scores to arrive at more stable estimates of the regression weights for each college, particularly for colleges with small samples. This method effectively achieves a shrinkage (from the usual least-squares estimates) of the multiple correlations between the GPA and the test scores within each college, thus at least partially addressing the typical concern of shrinkage in cross-validation. (Empirical cross-validation was not possible with the type of data collection design used here, given the limited numbers of students available in each college.)

Each composite score was defined in terms of the college-specific weights (the weights being the total N per college, as explained above), and the weighted average of the within-college correlations between GPA and the differentially weighted composite score is reported later as an estimate of its predictive validity.[4]

Equal-weights analyses for combinations of test scores. An additional set of analyses used equal weighting of test components in examining the predictive validity of each of the three combinations of test components mentioned above: (a) CV and CM, (b) NV and NMe, and (c) NV, NMe, and NW. For the current SAT, the area scores, CV and CM, were rescaled by dividing each by its standard deviation (SD) for the present sample and then weighted equally to form a composite score. Likewise, the area scores in the NPP test, NV, NMe, and NW, were each rescaled by the sample standard deviation and then equally weighted to form the two alternative composite scores under consideration (NV and NMe; NV, NMe, and NW). These composite scores are referred to as "equally-weighted" composites. The correlation of GPA with

[4]The term "weights" is used in two senses. "Variable weights" refers to the relative contribution of different test components in prediction. The terms "college-specific weights" and "weighted average" refer to computation of an overall statistic wherein each college's contribution to the statistic is based on the size of the sample from that college.

each of these composite scores was calculated for each college, and the weighted averages of these correlations across colleges were obtained, the weights being the total N per college (see explanation in the preceding section). These average correlations are reported below as the predictive validities. The equal-weights analyses were regarded as supplementary to those involving variable weights. For reasons discussed in the Results section, however, they provided a valuable alternative to variable weighting.

It should be noted that the predictive validity coefficients routinely reported for the SAT by the College Board Validity Study Service (cf., Ramist, 1984) are computed by applying standard multiple-regression techniques to data available for entire freshman classes. By contrast, use of the EM and EB techniques in computing predictive validities here was necessitated by the missing-data design and the small number of students per cell of the design in each college. Under the assumptions associated with the EM and EB techniques, the validity coefficients presented here should be of roughly the same magnitude as those that would be obtained if standard multiple-regression analyses were used, given comparable testing conditions. However, a more critical difference between the present situation and that of the typical Validity Study Service analysis lies in the nature of the testing conditions. The present analyses were based on tests administered under experimental conditions, whereas the Validity Study Service analyses are based on SAT scores from operational administrations. For this reason, it is important not to focus on the absolute magnitudes of the validities observed here or to compare them with those reported elsewhere. Of central importance here are the differences in validities observed for different test versions within the experimental situation created for this study.

Incremental Validity over High School Rank

To examine the incremental validity when a current or NPP test score is used in addition to high school rank to predict college performance, the simple validity of high school rank was estimated as the weighted average of its correlation with GPA within each college (again based on the EM estimates of their variances and covariances for each college).

For each of the equally-weighted composite scores noted earlier, an estimate of the predictive validity when it was used jointly with high school rank as a predictor of GPA was obtained. Optimal predictor weights for the composite score and high school rank were simultaneously estimated for all colleges using the EB regression procedure and applied to define a differentially weighted composite of high school rank and the composite test score (i.e., CV and CM; NV and NMe; NV, NMe, and NW). Within-college correlations of the resulting test score composite and high school rank with GPA were then calculated for each college, and the weighted average of these correlations (which are essentially equivalent to shrunken estimates of multiple correlations) across colleges was taken as an estimate of the joint predictive validity. The difference between the joint predictive validity and the simple validity of high school rank is reported as the incremental validity of the equally-weighted composite test score over high school rank.

The same procedure was employed to estimate the incremental validity in the variable-weights analysis. However, in place of the differentially-weighted composite score defined in the preceding section, the weights for the component test scores were re-estimated when the component scores were used jointly with high school rank as predictors in the EB regression procedure. The joint predictive validity of high school rank and the differentially-weighted composite of the test component scores was obtained as the weighted average of the resulting shrunken estimates of the within-college multiple correlations for high school rank and the test component scores in predicting GPA. Again, the difference between the joint predictive validity and the simple validity of high school rank is reported as the incremental validity of the particular combination of area test scores.

Analyses by Gender

To assess the extent to which the current SAT and the NPP test may under- or overpredict college performance of men versus women, similar EB regression analyses were employed to estimate the prediction difference for several likely uses of the test scores (test area scores used singly or jointly in the form of weighted composite scores) in predicting GPA.

As a rough approximation, the predictor weights in each analysis were assumed to be equal for the groups of men and women within each college (but not across colleges); adopting this assumption permitted unambiguous discussion of gender differences. Under this assumption, the estimated difference between the intercepts (regression constants) for the male and female prediction equations can be interpreted as the prediction difference (higher intercept for the women suggests underprediction of women's GPA when using a common regression equation). For each college, the EB approach provided a more stable estimate of such prediction differences, particularly for colleges with very small numbers of men and/or women, through the use of collateral information from similar colleges.

For the verbal, mathematical, and writing area scores and for the equally-weighted composite scores, the estimates of under- or overprediction were obtained from a regression model that included the single score of interest and an indicator variable (men = 1, women = 0) as independent predictors. For differentially weighted composite scores, the estimate of the male/female prediction difference was obtained from a regression model that included each of the component scores in the composite of interest and the same indicator variable as independent predictors. In either case, the EB estimate of the regression coefficient for the indicator variable is taken as an estimate of the prediction difference for men versus women.

Inferential Statistics

Analyses were performed to estimate the sampling error for each of several comparisons between validity coefficients that were of potential interest for test development and program policy. Similar analyses were applied to selected comparisons between incremental validities and between gender-related prediction differences. For this purpose, a "bootstrap" method (Efron, 1982) with 10,000 trials (of sample size 19) was employed to estimate confidence intervals for the difference observed in each specified comparison. For a given comparison, the procedure involves simulating a sampling distribution of observed sample differences for a hypothetical population of colleges similar to those participating in the study. In essence, each "bootstrap" trial draws a random sample of size N (here, 19) based on the empirically determined distribution of the statistics to be compared (e.g., predictive validities, incremental validities, gender-related prediction differences) employing a sampling-with-replacement procedure.

The comparisons of interest were organized into logically related groups, or families. Simultaneous confidence intervals for each family of comparisons were estimated from the bootstrap samples. To protect against inflated alpha level (and thus inaccurately narrow confidence intervals), a smaller nominal alpha level (equal to the desired family-wise alpha divided by the number of comparisons in a family) was used to obtain the confidence interval for each comparison. This approach applies the Bonferroni inequality to set an alpha level for an individual comparison and may result in conservative (wider) estimates of the actual simultaneous confidence intervals. The method is commonly known as the Bonferroni procedure for multiple comparisons (Miller, 1966; Rosenthal & Rubin, 1984).

As will be seen in the Results section, each family of comparisons being made in the present study contains two to three contrasts. Accordingly, a nominal alpha of .017 to .025 may be used following the Bonferroni method if an alpha of .05 is desired for a family. For simplicity and uniformity, however, an alpha of .01 was used in all reported comparisons; i.e., 99% confidence intervals (two-tailed) were reported for each of the comparisons being made.

Due to the complex statistical procedures employed in the study (which involve an EM algorithm for estimating sample statistics for incomplete data and an Empirical Bayes (EB) method for estimating validities and prediction differences), the bootstrap procedure described above takes into account sampling errors at the college level, but does not account for other sources of error at the student level. Furthermore, statistical errors due to the EM and EB procedures were also not examined. (Computational efforts for this work are extremely expensive and were beyond the resources available in this study). Nevertheless, the confidence intervals reported here provide a reasonable indication of the ranges within which the various contrasts might be expected to fall, at the 99% level of confidence, for the population of colleges similar to those sampled in the present study.

Results

Test Analyses

Prior to presentation of the major analyses, which involve predictive validity, it is useful to consider such characteristics of the test as reliability, item difficulty, item discrimination, and speededness. As part of a preliminary examination to assist further test development, analyses were conducted using students from the original 27 colleges, before screening was done to eliminate colleges with small numbers of students and to eliminate unmotivated students within colleges (except that students who failed to provide at least three responses per test section were excluded). These analyses were based on the students who were given both verbal sections, or both mathematical sections, of a given test version. Although these analyses did not use the exact same sample as the predictive validity analyses, they nevertheless provide useful background information for understanding the predictive validity data. Presented in detail in the Appendix, results of the test analyses are summarized here.

Reliability. Coefficient alpha was computed for each test section, and reliabilities for the whole verbal and whole mathematical area scores were estimated using the Angoff/Feldt procedure (Educational Testing Service, 1989). For the verbal sections, the coefficients were obtained by treating each item in the reading passages as a unit. H. Wainer (personal communication, June, 1990) has expressed concern that, with the long reading passages in the NPP test, coefficient alpha is inflated when individual items are used as the units of analysis.[5] Thus, although the reliability computed here was not substantially different for the current SAT verbal score (.92) and the NPP verbal score (.88), conclusions about the reliabilities of the two tests cannot be drawn without further study into the role of the number of items per passage in computation of reliability.

Reliabilities for the mathematical area were estimated to be .90 for the set of items in the current SAT and .88 for the set of items from the

[5]An alternative is to consider each reading-passage score as a larger unit and the remaining items as a unit; this would lead to an underestimate of the reliability because of the large disparity in the size of score units between passages and the discrete items. Another alternative is to regard each of the passage scores, the Synonyms in Context score, and the Analogies score as the score units. This may produce different estimates of reliabilities depending on the intra-cluster correlations among individual item scores within a score unit. Research is currently underway to determine the extent to which these methods produce estimates of reliability that differ from that obtained when using items as the units of analysis.

NPP test that were included in the validity analyses.[6] For the writing test the reliability of the multiple-choice section was .86, and the inter-reader reliability for the essay section was .81.

Item difficulty. The average and the variability of item difficulties for the verbal area were slightly lower for the NPP test than for the current SAT; the mean equated delta scores (with higher scores reflecting greater difficulty), were 11.4 and 10.6 for the current SAT and NPP test, respectively (SDs = 3.0 and 2.4). The average difficulty of the mathematical items was approximately the same for the current SAT (12.5) as for the NPP test excluding Algebra Placement items (12.6), SDs = 3.2 and 3.1, respectively. For the writing test, the mean delta was 11.1 (SD = 3.0) for the multiple-choice section; difficulty of the essay section cannot be evaluated in comparable terms.

Item discrimination. As measured by the biserial correlations of items with the appropriate area score (verbal or mathematical), the levels of item discrimination were slightly higher for the NPP test than the current SAT. The mean and SD, across items, of the biserial correlations were: current SAT verbal: .49 (SD = .11), NPP verbal: .53 (SD = .10); current SAT mathematical: .53 (SD = .10), NPP mathematical (excluding Algebra Placement items): .59 (SD = .13). For the writing test, multiple-choice section, the mean biserial correlation of items with total multiple-choice score was .50 (SD = .08).

Test speededness. Interpretation of test speededness was based on a combination of indices, as discussed in the Appendix. For the verbal part, the NPP test appeared to be slightly less speeded than the current SAT, whereas for the mathematical part, the NPP test appeared to be slightly more speeded than the current SAT. Because speededness is examined within a full section, speededness of the mathematical part excluding the Algebra Placement items cannot be determined. For the writing test, multiple-choice section, most indices indicated a relatively low degree of speededness.

In general, broad conclusions about the reliabilities of the two test versions are difficult to make, given the possibility of inflated estimates for the larger number of items per reading passage in the verbal area and given that the mathematical area was represented here by considerably fewer items for the NPP test than for the current SAT. Nevertheless, for purposes of interpreting the predictive validity data, presented below, it is worth noting that the reliability figures tended to be slightly lower for the set

---

[6]In the predictive validity analyses, the mathematical area of the NPP test was represented by score NMe--the score based on all mathematical items except the Algebra Placement items. Thus, comparisons between the current SAT and NPP mathematical scores involved different numbers of items for the two versions of the test. Adjustment for the number of items per test was not attempted, as such an adjustment would be extremely difficult to accomplish with the complex methods of analysis used in this study. The present reliability estimates for the NPP mathematical score as well as the validity estimates, therefore, are conservative.

of items analyzed for the NPP test than for those on the current SAT. If a validity difference in favor of the former set of items were found, then, it could not be attributed to differences in reliability.

The two versions of the test differed somewhat with regard to item difficulty, item discrimination, and speededness. These differences may simply reflect characteristics of the particular items appearing in the test forms used in this study. It cannot be determined whether these differences affected the predictive validity results to be presented below. However, these results serve to stress an important point: because the validity data are based on comparison of only one form of the current SAT and only one prototype form of the NPP test, conclusions drawn from the present data should not be overgeneralized. Firmer conclusions about the validity of the revised SAT must await the collection of data based on several test forms with larger and more representative samples.

Predictive Validity Analyses

The predictive validity for a given test component was the mean across colleges (weighted by the N per college) of the correlation of that test component with GPA. Before the predictive validity results are presented, it is useful first to examine relationships among the predictors. Table 2 presents the weighted means of the individual college correlations, derived from the EM-estimated variance-covariance matrix. Note that the mathematical score for the NPP test here, as in all other analyses to be presented, is that based on all mathematical items except the Algebra Placement items (i.e., score NMe).

Among notable aspects of these data, the relation between verbal and mathematical scores was approximately the same for the two versions of the test. (Disattenuated correlations were also approximately the same--.54 for the current SAT and .53 for the NPP test.) Further research now underway will provide a more thorough test of the issue whether proposed revisions in the SAT affect the degree of interdependence of verbal and mathematical scores.

The relation between the writing score and the NPP verbal score was relatively high, suggesting a greater degree of similarity in the abilities reflected in these two scores than in those reflected in the verbal and mathematical scores. High school rank appeared to be somewhat more highly correlated with scores on the NPP test than on the current SAT. These data will be considered further below in connection with interpretation of the predictive validity results.

Table 2

Correlations among Predictors

	Current SAT		NPP test		
	V	M	V	M[a]	W
Verbal score (V)					
Math. score (M)[a]	.49		.47		
Writing score (W)			.70	.42	
H. S. rank (R)	.23	.26	.29	.33	.34

[a]Mathematical score for NPP test here and in subsequent analyses and tables is that based on all mathematical items except the Algebra Placement items (i.e., score NMe).

Table 3 presents the predictive validity coefficients for the current SAT and the NPP test. As stressed above, the absolute levels of coefficients presented here are not of primary interest, as they may not compare with the levels that are typically observed when the SAT is administered under operational testing conditions. More important are the comparisons among coefficients, which are assumed to have been equally affected by the reduction in motivation associated with the use of experimental testing conditions. Analyses to be presented below will examine critical comparisons among coefficients.

Certain comparisons between pairs of validity coefficients in Table 3 bear on the issues under study and are thus of particular interest. These comparisons were singled out for examination and, using procedures described above, confidence intervals were computed for each of these comparisons. Three families of comparisons were of interest, which are presented in Table 4. For each comparison, a 99% confidence interval for the difference between validities was computed, indicating the boundaries within which the true population difference would be expected to fall, with a probability of .99 over repeated sampling. It is important to note that the term "population" as used in the discussion to follow refers to the particular kinds of colleges and students that participated in this study and the specific set of circumstances under which the study was conducted.

Consider the first family of comparisons presented in Table 4, the differences between the predictive validities for the NPP and current SAT area scores. The estimated difference for the verbal area was .049 in favor of the NPP test. The 99% confidence interval for the true population difference extended from .001 to .133. This indicates that the true difference in validities for the two verbal scores may be essentially zero or it may be quite large; the data from this study do not allow either extreme to be ruled out. The interval does, however, essentially rule out the possibility (at the 99% confidence level) that the current SAT verbal score has a distinctly higher predictive validity than that of the NPP verbal score, and this is the major conclusion that should be drawn regarding this comparison.

Table 3

Predictive Validity Coefficients for Current SAT and NPP Test

Predictor	Current SAT	NPP test
Area scores		
Verbal	.30	.35
Mathematical	.28	.35
Writing		.37
Test composites--variable-weights analyses		
Verbal and mathematical	.35	.42
Verbal, mathematical, and writing		.47
Test composites--equal-weights analyses		
Verbal and mathematical	.33	.41
Verbal, mathematical, and writing		.43

Note: The criterion in each case was first-year GPA.

Table 4

Target Comparisons for Predictive Validity Coefficients in Table 3

Comparison	Diff. bet. validities	Confidence interval Lower bound	Upper bound
Area scores			
NPP verbal vs. current SAT verbal (NV-CV)	.049	.001	.133
NPP math. vs. current SAT math. (NMe-CM)	.077	-.009	.151
Test composites--variable-weights analyses			
Verbal-math. composite for NPP test vs. current SAT (NV/NMe - CV/CM)	.071	.017	.127
Change due to addition of writing score to NPP verbal-math. composite (NV/NMe/NW - NV/NMe)	.047[a]	.029	.091
Verbal-math.-writing composite for NPP test vs. verbal-math. composite for current SAT (NV/NMe/NW - CV/CM)	.118	.061	.195
Test composites--equal-weights analyses			
Verbal-math. composite for NPP test vs. current SAT (NV/NMe - CV/CM)	.080	.019	.143
Change due to addition of writing score to NPP verbal-math. composite (NV/NMe/NW - NV/NMe)	.017[a]	-.005	.048
Verbal-math.-writing composite for NPP test vs. verbal-math. composite for current SAT (NV/NMe/NW - CV/CM)	.097	.037	.170

[a]This difference reflects changes due to addition of a predictor, rather than comparisons between independent predictor sets, and is thus not on the same scale as the differences immediately above and below it in the table.

Turning to the comparison for the mathematical area, the point estimate of the difference in predictive validities was .077, again in favor of the NPP test. Although the estimated difference for the mathematical area was larger than for the verbal, the interval for the mathematical area was also larger, from -.009 to .151, reflecting greater sampling variability from college to college. The basic conclusions from both comparisons in this family--i.e., those involving the verbal and mathematical scores--is the same: essentially, the predictive validities for NPP area scores are at least as great as those for the corresponding area scores of the current SAT. (There is an overall level of confidence of at least 98% associated with this statement, for this family of two comparisons.) The true population differences might be relatively small or relatively large. However, the confidence intervals observed here suggest that the data most likely represent modest differences in favor of the NPP test.[7]

These validity differences in favor of the NPP test cannot be attributed to reliability differences, since the reliabilities were not higher for the set of NPP test items examined here than for those in the current SAT. As discussed above, conclusions about effects of the test revisions on reliability cannot be drawn from this study. But it is reasonable to conclude that differences in reliability are not responsible for the observed differences in predictive validity.

The second family of comparisons shown in Table 4 compares the predictive validities of three predictor sets, current SAT verbal and mathematical, NPP verbal and mathematical, and NPP verbal, mathematical, and writing scores. With a 99% confidence interval per comparison, the overall confidence level for this family of three comparisons is 97%. As described in the Method section, these validities are averages across colleges of the result of estimating multiple regressions within each college using an Empirical Bayes estimation procedure. The label "variable weights" is used to indicate that the relative weights assigned to different predictors may vary from college to college in this analysis. The confidence intervals for the three differences confirm the ordering given by the estimates; i.e., the population predictive validity for the current SAT is at least slightly less than that for the NPP verbal and mathematical combination which, in turn, is less than that for the NPP verbal, mathematical and writing scores taken together. As a result, the NPP verbal, mathematical and writing combination appears to have a population predictive validity that is substantially higher (by at least .061 and as much as .195) than that for the current SAT verbal and mathematical combination.

---

[7]The intervals in this family illustrate an important advantage of working with confidence intervals instead of relying on conventional significance testing. Viewed as significance tests, the current results would show a "significant" difference for verbal scores but not for mathematical scores, since the population difference of zero falls outside the first interval but inside the second. This fact, however, obscures the essential similarity of the two intervals and the corresponding similarity of the conclusions that should be drawn.

Note the relative precision in estimating the change due to the addition of the writing score to the NPP verbal and mathematical composite. This is primarily a result of one of these sets being a subset of the other. In the population, as well as in any sample, the multiple correlation for the larger set cannot be smaller than that for the subset. The interval here indicates that at least some gain in predictive validity can be obtained when NPP writing is used together with NPP verbal and mathematical scores.

The third family of comparisons presented in Table 4 is very similar to the second. The important difference, as described in the Method section, is that the relative weights used for different predictors were determined a priori (based on their standard deviations) and were not allowed to vary from college to college, hence the label "equal weights." The main difference between the results obtained for these analyses and the variable-weights analyses is that the improvement in predictive validity resulting from the inclusion of NPP writing along with NPP verbal and mathematical scores is no longer so clear. Here, since population multiple correlations are no longer being estimated, it is theoretically possible for the true validity of the composite to decrease when writing is added. There is no evidence that this has in fact occurred to more than a trivial degree (cf., the lower bound of -.005) and there may have been an increase of as much as .048.

It is useful to summarize the predictive validity findings in relation to two fundamental issues under study here. One issue was whether the revisions in the basic components of the test--the verbal and mathematical areas--affected the test's predictive validity. The findings for this issue were relatively straightforward. Predictive validity tended to be higher for the NPP test than the current SAT, and this was true for the verbal scores, the mathematical scores, and the combination of verbal and mathematical scores.

The other fundamental issue concerned the potential increase in validity of adding the writing score to the NPP verbal-mathematical composite. Conclusions in this case are complicated by the fact that a moderate increase was observed in the variable-weights analysis, whereas only a slight increase was observed in the equal-weights analysis. In interpreting these results it is necessary to consider the kinds of information obtained in these two types of analyses. The variable-weights analysis allowed the relative weights of verbal, mathematical, and writing scores to take on values for each of the 19 colleges that maximized prediction of GPA per college. Had it been possible to cross-validate, using the same weights, marked shrinkage in the validity coefficient might have occurred with the small samples per college used in this study (even though the Empirical Bayes procedure adjusts somewhat for this problem). Furthermore, in attaining maximum prediction for the combination of verbal, mathematical, and writing scores, several weights observed here were negative; specifically, for 8 of the 19 colleges, the weight for either the verbal or writing score was negative, which is not unusual whenever two predictors are strongly related, as were the NPP verbal and writing scores. In these respects, then, the variable-weights analysis as used here deviates

from the kind of situation that would exist if a college used these scores to predict students' GPA, and may overestimate the predictive validity (and the increase in validity provided by the writing score) that would be observed in actual practice.

The equal-weights analysis provides an alternative to variable weighting that eliminates the problem of validity shrinkage due to cross-validation and the problem of encountering negative weights. However, it does not necessarily represent the way colleges might choose to employ the various scores in predicting GPA. A given college may prefer to assign differential weights to these scores, as determined by the nature of its curriculum or other considerations, in order to improve prediction. The equal-weights analysis, therefore, may underestimate predictive validity (and the increase in validity provided by the writing score), given that it was not based on empirical determination of the most suitable weights and did not involve the use of weights tailored to the needs of each individual college.

It seems reasonable to hypothesize that, in actual practice, the improvement in predictive validity due to addition of the writing score would lie somewhere between the moderate increase observed in the present variable-weights analysis and the slight increase observed in the equal-weights analysis. Tentatively, then, it appears that adding the writing score to the verbal and mathematical scores may have at least some beneficial effect on prediction of students' first-year GPA, albeit relatively modest in magnitude.

Although the validity coefficients of primary interest were those discussed above for area scores and composites, validity coefficients for individual test sections are presented in Table 5. These coefficients are based on smaller numbers of items, so comparisons among them would be subject to considerable error and have not been attempted. They are presented here to aid in interpretation of the validity coefficients for the area scores and composites and will be considered further in the Discussion section.

Table 5

Validity Coefficients for Individual Test Sections

Current SAT:	
Verbal Section 1 (CV1)	.28
Verbal Section 1 (CV2)	.29
Math. Section 1 (CM1)	.23
Math. Section 2 (CM2)	.28
NPP test:	
Verbal Section 1 (NV1)	.29
Verbal Section 2 (NV2)	.35
Math. Section 1 (NM1)	.35
Math. Section 2, excl. Alg. Pl. items (NM2e)	.28
Writing, multiple-choice section (NW-MC)	.36
Writing, essay (NW-ESS)	.24

Incremental Validity over High School Rank

Table 6 presents the data bearing on the incremental validity of the various test composites relative to high school rank.[8] Before considering the incremental validities, it is useful to compare the two subsamples--the current SAT sample and the NPP test sample--with respect to predictive validity of high school rank. (Although the essentially random assignment should have resulted in roughly equivalent subsamples, it is important to determine whether this is the case.) In testing the difference, a liberal alpha level of .10 was used, or a 90% confidence interval. The difference was .015 and the bounds of the confidence interval were found to be -.044 and .020, indicating that the difference between these two subsamples was negligible.

For the data in Table 6, the principal comparisons were those involving the incremental validities, which indicate the degree to which various test composites, taken together with high school rank, add to the predictive validity observed for high school rank alone. These comparisons are shown in Table 7.

For the verbal/mathematical composite, the difference between the NPP test and current SAT in incremental validity was slight in the variable-weights analysis (falling between -.017 and .041) and modest in the equal-weights analysis (-.006 to .047). Thus, the increment in validity due to the addition of the verbal/mathematical composite to high school rank tended to be more pronounced for the NPP test than the current SAT, although not markedly so.

In the variable-weights analysis, the incremental validity of the NPP test composite was moderately greater when it included the writing score than when it did not. However, in the equal-weights analysis, the effect due to addition of the writing score was negligible. Thus, considering the results of both analyses together, the implication is that, at best, inclusion of the writing score among the predictors produces a small increase in incremental validity. In both the variable-weights analysis and the equal-weights analysis the NPP composite consisting of three scores, verbal, mathematical and writing, yielded a modestly greater incremental validity than the two-part current SAT.

---

[8]The College Board Validity Study Service creates a 20- to 80-point scale for class rank. The transformation stretches the top of the percentile rank scale to allow greater discrimination in the upper ranges. Because simple percentile was used here, rather than the transformed scores, the observed correlations between high school rank and first-year GPA tend to be slightly lower than those typically observed in Validity Study Service analyses. Nevertheless, it seems reasonable to assume that correlations involving the current SAT and the NPP test will be affected to approximately the same degree; thus, conclusions regarding the comparison between tests are expected to be roughly the same, whichever scale is used for high school rank.

Table 6

Incremental Validity of Current SAT and NPP Test over High School Rank

Predictor	Current SAT		NPP test	
	Pred. validity	Increm. validity	Pred. validity	Increm. validity
Variable-weights analyses				
High school rank[a]	.45		.43	
H. S. rank and verbal/math. composite	.53	.08	.53	.09[b]
H. S. rank and verbal/math./writing composite			.55	.12
Equal-weights analyses (i.e., equal weights for test composites)				
High school rank[a]	.45		.43	
H. S. rank and verbal/math. composite	.51	.06	.52	.09
H. S. rank and verbal/math./writing composite			.52	.09

[a]Predictive validity of high school rank presented separately for the students given parts of the current SAT ("current SAT sample") and for students given parts of the NPP test ("NPP test sample").

[b]Each incremental validity is based on the predictive validity for the indicated combination of predictors minus the predictive validity for rank alone, rounded to two digits. In the case noted, the apparent discrepancy between the incremental validity and the difference between predictive validities shown is due to rounding.

Table 7

Target Comparisons for Incremental Validities in Table 6

Comparison	Diff. bet. validities	Confidence interval Lower bound	Upper bound
Variable-weights analyses			
Incremental validity of:			
Verbal-math. composite for NPP test vs. current SAT (NV/NMe - CV/CM)	.014	-.017	.041
Verbal-math.-writing composite vs. verbal-math. composite, NPP test (NV/NMe/NW - NV/NMe)	.023[a]	.010	.039
Verbal-math.-writing composite for NPP test vs. verbal-math. composite for current SAT (NV/NMe/NW - CV/CM)	.037	.000	.071
Equal-weights analyses (i.e., equal weights for test composites)			
Incremental validity of:			
Verbal-math. composite for NPP test vs. current SAT (NV/NMe - CV/CM)	.021	-.006	.047
Verbal-math.-writing composite vs. verbal-math. composite, NPP test (NV/NMe/NW - NV/NMe)	.004[a]	-.010	.016
Verbal-math.-writing composite for NPP test vs. verbal-math. composite for current SAT (NV/NMe/NW - CV/CM)	.026	-.005	.054

[a]This difference reflects changes due to addition of a predictor, rather than comparisons between independent predictor sets, and is thus not on the same scale as the differences immediately above and below it in the table.

Under- Versus Overprediction by Gender

When gender is included as an independent variable in the regression equation, the regression weight for gender reflects the degree of prediction difference between men and women, expressed in GPA units. Presented in Table 8 under the heading "mean prediction difference," positive figures show that GPA is underpredicted for women relative to men. For example, the observed prediction difference of +.20 for the current SAT verbal area indicates that the women's GPA was .20 points higher than that of the men with the same verbal score. Mean scores to be presented subsequently (see Table 10) show that the prediction differences were generally due to lower test scores, but higher GPA, for women than men. Table 9 presents the principal comparisons for the data in Table 8. Analogous to the comparisons among validity coefficients presented in previous tables, the comparisons here are between mean prediction differences for various test parts or composites.

The first results of interest are those involving the verbal, mathematical, and (for the NPP test) writing area scores. The mean prediction differences shown in Table 8 were all positive, indicating underprediction of GPA for women relative to men, especially for the mathematical score. Comparisons between the NPP test and current SAT, shown in Table 9, indicated a modest difference between versions of the test for the verbal area (falling between -.192 and -.001), reflecting a reduction in degree of underprediction for women with the test revisions. For the mathematical area, on the other hand, the mean prediction difference was roughly the same for both versions of the test, indicating that the test revisions had a negligible effect on the degree of underprediction for women in the case of the mathematical score.

When the verbal and mathematical scores were taken in combination, the difference between the NPP test and current SAT was small, both in the variable-weights and equal-weights analyses. However, addition of the writing score to the verbal/mathematical composite of the NPP test tended to reduce the prediction difference. As a consequence, comparison of the NPP verbal/mathematical/writing composite versus the current SAT verbal/mathematical composite showed a modestly lower prediction difference in the former case than the latter. Thus, when the writing score was included in the NPP test composite, the NPP test yielded a lower gender-related prediction difference than did the current SAT, and this was true both when the three area scores were given equal weights and when they were allowed to take on variable weights.

Table 8

Under- Versus Overprediction of GPA by Gender

Predictor	Mean prediction difference: Current SAT	Mean prediction difference: NPP test
Area scores		
Verbal score	.20	.10
Mathematical score	.26	.25
Writing score		.07
Test composites--variable-weights analyses		
Verbal/mathematical	.25	.20
Verbal/mathematical/writing		.16
Test composites--equal-weights analyses		
Verbal/mathematical	.25	.18
Verbal/mathematical/writing		.13
High school rank and variably weighted test composites		
Rank alone[a]	.08	.06
Rank and verbal/math. composite	.15	.12
Rank and verbal/math./writing composite		.09
High school rank and equally weighted test composites		
Rank and verbal/math. composite	.15	.12
Rank and verbal/math./writing composite		.08

[a]Prediction difference based on high school rank, presented separately for students given parts of the current SAT ("current SAT sample") and for students given parts of the NPP test ("NPP test sample")

Note: Positive numbers indicate underprediction for women.

Table 9

Target Comparisons for Prediction Differences in Table 8

Comparison	Diff. bet. pred. diffs.	Confidence interval Lower bound	Confidence interval Upper bound
Area scores			
NPP verbal vs. current SAT verbal (NV-CV)	-.098	-.192	-.001
NPP math. vs. current SAT math. (NM-CM)	-.010	-.104	.060
Test composites--variable-weights analyses			
Verbal-math. composite for NPP test vs. current SAT (NV/NMe - CV/CM)	-.048	-.141	.037
Change due to addition of writing score to NPP verbal-math. composite (NV/NMe/NW - NV/NMe)	-.037[a]	-.063	-.013
Verbal-math.-writing composite for NPP test vs. verbal-math. composite for current SAT (NV/NMe/NW - CV/CM)	-.086	-.174	-.006
Test composites--equal-weights analyses			
Verbal-math. composite for NPP test vs. current SAT (NV/NMe - CV/CM)	-.064	-.157	.019
Change due to addition of writing score to NPP verbal-math. composite (NV/NMe/NW - NV/NMe)	-.050[a]	-.061	-.038
Verbal-math.-writing composite for NPP test vs. verbal-math. composite for current SAT (NV/NMe/NW - CV/CM)	-.114	-.206	-.031

[a]This difference reflects changes due to addition of a predictor, rather than comparisons between independent predictor sets, and is thus not on the same scale as the differences immediately above and below it in the table.

Table 9 (continued)

Comparison	Diff. bet. pred. diffs.	Confidence interval Lower bound	Upper bound
High school rank and variably weighted test composites			
Combination rank and:			
Verbal-math. composite for NPP test vs. current SAT (NV/NMe - CV/CM)	-.026	-.112	.032
Verbal-math.-writing composite vs. verbal-math. composite, NPP test (NV/NMe/NW - NV/NMe)	-.033[a]	-.048	-.004
Verbal-math.-writing composite for NPP test vs. verbal-math. composite for current SAT (NV/NMe/NW - CV/CM)	-.058	-.134	.006
High school rank and equally weighted test composites			
Combination rank and:			
Verbal-math. composite for NPP test vs. current SAT (NV/NMe - CV/CM)	-.033	-.113	.023
Verbal-math.-writing composite vs. verbal-math. composite, NPP test (NV/NMe/NW - NV/NMe)	-.036[a]	-.044	-.025
Verbal-math.-writing composite for NPP test vs. verbal-math. composite for current SAT (NV/NMe/NW - CV/CM)	-.069	-.146	-.012

[a]This difference reflects changes due to addition of a predictor, rather than comparisons between independent predictor sets, and is thus not on the same scale as the differences immediately above and below it in the table.

For the analyses including high school rank, a separate test was first performed to compare the current SAT and NPP test with respect to the prediction difference based on prediction from high school rank alone. This difference proved to be small; the difference was equal to -.024 and the bounds of the 90% confidence interval were -.071 and .008. Thus, the current SAT sample and the NPP test sample appeared not to have differed substantially with respect to gender-related prediction differences (although the confidence interval suggested a relatively wide range of potential values for the difference). Observed differences in results for the two test versions, then, are believed not to be attributable to differences in nature of the two subsamples.

The combination of rank and test composites yielded gender-related prediction differences, as did the test composites alone. Results of the principal comparisons were largely consistent with those observed in the comparisons in which rank was not included among the predictors. The degree of underprediction differed only slightly for the NPP test and the current SAT, using the verbal/mathematical composite score along with high school rank as predictors. But addition of the writing score to the NPP test composite, together with high school rank, did tend to reduce the degree of underprediction for women on the NPP test. As a result, when the NPP test was represented by the three parts, verbal, mathematical and writing, the combination rank and test composite yielded a modestly lower prediction difference for the NPP test than the current SAT.

Table 10 presents mean test scores and GPAs by gender, to help illustrate the prediction differences. The means in Table 10 are estimated by the EM algorithm, to account for the fact that, in most cases, an individual student was not administered all test components contributing to a given mean. (The EM adjusted means for individual test sections should be generally quite close to the raw means, however.) The test composites shown are those based on equal weighting of the component scores. The scores for the verbal and mathematical areas are based on the number of correct responses. In contrast, the score for each test composite and the score for the writing test (a composite of a multiple-choice part and an essay part) was derived by dividing each part of the composite by its standard deviation and taking the sum. For this reason, the scales varied widely for the different scores shown.

The most notable results in Table 10 are the differences between men and women, expressed in standard deviation units. The differences are generally positive in the case of test performance, but negative in the case of GPA--hence, the underprediction of GPA for women relative to men. Other aspects of the results generally parallel the prediction differences presented above; notably, the gender-related prediction difference for the verbal area was less for the NPP test than the current SAT (and, in fact, was zero for the NPP test), and the prediction difference was most pronounced for the mathematical area.

Table 10

Mean (EM Adjusted) Scores by Gender
and Pooled Standard Deviation

Test area or composite[a]	Mean Men	Mean Women	Pooled SD[b]	Gender difference in SD units
Current SAT sample				
Verbal (V)	40.98	38.40	12.66	.20
Mathematical (M)	29.71	23.51	9.77	.63
V/M composite	4.92	4.27	1.35	.48
GPA	2.57	2.75	.68	-.26
NPP test sample				
Verbal (V)	31.76	31.75	10.50	.00
Mathematical (M)	21.98	17.31	7.52	.62
Writing (W)	2.89	3.02	.75	-.17
V/M composite	4.73	4.24	1.36	.36
V/M/W composite	8.18	7.85	2.07	.16
GPA	2.62	2.71	.66	-.14

[a]Test composites are based on equal weighting of parts indicated.

[b]Pooled SD is based on pooled within-college variances (EM estimates).

Note: The total Ns for men and women, respectively, are 904 and 1044 for the current SAT sample, and 2134 and 2559 for the NPP test sample.

Discussion

This study was designed to provide an initial view of how planned revisions in the SAT might affect its predictive validity. Conducted during the intermediate stages of the New Possibilities Project, when additional revisions were still under consideration, the study was intended purely as a formative investigation. Nevertheless, the prototype revised SAT used here, or NPP test, incorporated many of the most important structural revisions planned for the test, and this study provided useful preliminary information about the overall effects of these revisions on the test's predictive validity. Given that the research employed experimental testing conditions, generalizations from the results are necessarily limited. Nevertheless, because the current SAT and NPP test were both presented under these same experimental conditions, tentative conclusions could be drawn about their relative validities, if not their absolute validities.

Predictive Validity of NPP Test Versus Current SAT

Perhaps the results of greatest interest were that the validity coefficients were modestly higher for the verbal score, the mathematical score, and the verbal-mathematical combination, of the NPP test than the current SAT. Tentatively, then, it appears that the revisions implemented here had a favorable effect on predictive validity and, as discussed above, the observed differences do not appear to be attributable to test differences in reliability.

The two test versions differed in many respects, and it is only possible to speculate as to the factors that played the most substantial role in these results. Regarding the verbal area, the greater validity of the NPP test appeared to be due primarily to the second section of the test. This section contained several item types, including the new critical reading items and synonyms in context items, along with the analogies item type from the current SAT. It might be posited that validity is higher when a variety of item types is used than when a single item type is used, and that the particular mix of verbal item types used in the NPP test was conducive to higher validity than that used in the current SAT. The critical reading items in the NPP test, for example, tested reading comprehension in a relatively broad context and, among other skills, called for evaluation about points of view expressed. Also, unlike the antonym items of the current SAT, the synonyms in context items tested vocabulary in a format that simulated the way in which lexical knowledge is applied in actual practice.

In the mathematical area, the fact that there was a reduced number of items underlying the NPP mathematical score (thus rendering the validity estimate for this score conservative) makes the difference between tests particularly noteworthy. At the same time, the reduction in items makes interpretation of the results somewhat difficult. In particular, it is difficult to speculate on the role of the Grid-in items in NPP Mathematical Section 2 without evidence based on more than the 10 items of this type that

were used here. It appears that NPP Mathematical Section 1 played some role in the higher validity for the NPP than the SAT mathematical score. A key distinguishing characteristic of this section was the format change, in which the five-choice items were accompanied by space for calculations along with instructions that essentially encouraged its use. The effects of this factor are currently being studied further. Whatever the basis for test differences in the mathematical area, however, it is notable that predictive validity for this area, like the verbal area, tended to be increased with the revisions in the test.

In general, interpretation of the differences between the NPP test and the current SAT must also take account of the fact that the two tests differed somewhat in mean item difficulty, item discrimination, and speededness. Whether these differences helped contribute to the difference in validity cannot be determined from the present data. At the least, these results serve as a reminder that generalizations about the two versions of the test--the NPP test and the current SAT--are necessarily limited when based on only one form of each.

Incremental Validity over High School Rank

As expected, addition of test scores to high school rank as predictors, both for the NPP test and the current SAT, led to higher validity coefficients than did use of high school rank as a single predictor. These incremental validities, consistent with those commonly observed in data from the College Board Validity Study Service (cf., Ramist, 1984), reflect the way in which the SAT is typically used in the admissions process--as a predictor along with high school performance.

Unlike the findings for predictive validity of the tests alone, the incremental validity data showed only a slight to modest difference between the NPP test and the current SAT. Thus, the revisions did not substantially alter the test's contribution to prediction of first-year college performance when the predictors included high school rank together with the test scores. The difference in results for predictive and incremental validity may have to do with the nature of the test revisions. Perhaps the additional skills tapped by the revised test overlapped enough with skills underlying successful high school performance to result in only a small change in incremental validity. Relevant to this point is the finding that high school rank was somewhat more highly correlated with scores on the NPP test than the current SAT and, therefore, the NPP test might be expected to exhibit less incremental validity than the current SAT. It remains for further, more analytic research to identify characteristics of the NPP items that may have contributed to these results.

More important for the present purposes is the overall picture that emerges from the comparisons between the current SAT and the NPP test. For incremental as well as predictive validity, the observed coefficients were found to be no lower for the NPP test than for the current SAT and, where different, they were in the direction of higher values for the NPP test. This point is important to stress. As mentioned in the Introduction, it was

not an objective of the New Possibilities Project to increase the test's validity. Rather, the goal was to implement the test enhancements discussed above, which were chosen for content-related reasons, while seeking to ensure that the test's validity would not be reduced by doing so. In this respect, a key objective of the overall project appears to have been met, in that the level of validity was maintained in the face of revisions in the test's structure.

Gender-related Prediction Differences

As has been observed in other validity studies, first-year college GPA tended to be underpredicted slightly by the SAT for women relative to men. Among hypotheses that have been offered to account for this phenomenon is that men and women take different types of courses, thus making the criterion measure, GPA, somewhat different for the two gender groups. This hypothesis is supported in research using a criterion that adjusts for different courses taken (e.g., Elliott & Strenta; 1988; Young, 1991). It was beyond the scope of this study to examine the basis for the underprediction effect or to adjust the criterion according to courses taken. Rather, the objective here was to determine whether the gender-related prediction difference, using first-year GPA as the criterion, would be altered by the proposed revisions in the SAT.

The test revisions produced a modest reduction in the gender-related prediction difference for the verbal area, but a negligible change for the mathematical area. As a consequence, the effect for the composite of verbal and mathematical scores (as well as that for the combination of rank and verbal/mathematical composite scores) was small.

Apparently, the kinds of verbal items used on the NPP test led to test performance for women that was more in line with that of the men, relative to GPA, when compared with the kinds of items used on the current SAT. Whether this was due primarily to any one of the verbal item types used in the NPP test, or to the mixture of verbal item types, cannot be determined from these data. Whatever the reason for these results, however, the important point is that the gender-related prediction difference, where changed, tended to be reduced. Judging from research cited above, it is not realistic to expect that the prediction difference can be totally eliminated, as long as men's and women's GPAs are based on different sets of courses. Nevertheless, to the extent that it is desirable to reduce the prediction difference (or at least that part of the difference not due to differential course selection), the present revisions in the verbal area had a favorable effect.

Effects Involving the Writing Test Scores

Although not part of the SAT-I examination, it was believed that colleges might find the writing score to be a useful supplement to the verbal and mathematical scores. In anticipation of this possibility, analyses reported above examined the role of the writing scores in

predictive validity and incremental validity, and in gender-related prediction differences.

Inclusion of the writing score along with the verbal and mathematical scores appeared to improve predictive validity over that observed when the latter scores alone served as predictors. Interpretation of this result relates to the unique contribution of the writing test score. By assessing knowledge of appropriate English written expression, and by testing productive writing ability, the writing test provides useful information not already furnished by the verbal and mathematical scores. And because the kinds of knowledge and abilities measured by the writing test are ones that are important for effective communication in college, the test should aid in predicting college success.

The effect of the writing score on incremental validity, on the other hand, was relatively small. That the effect was less pronounced for incremental than predictive validity may partly reflect the fact that there was less "room for improvement" in the case of incremental validity (e.g., in the variable-weights analyses the incremental validity of the verbal and mathematical score composite was .53, whereas its predictive validity was .42.) Also, performance on the writing test may involve skills and knowledge similar enough to those involved in high school performance that the test does not make a substantial contribution beyond prediction from high school rank alone.

Addition of the writing score to the predictors--both with and without high school rank among the predictor set--had the effect of reducing the degree of underprediction for women. This effect undoubtedly relates to the fact that gender-related underprediction is less pronounced in the case of verbally oriented skills than quantitatively oriented skills, and the writing test taps abilities that fall generally within the verbal domain. In effect, addition of the writing score results in a set of predictors that provides a broader sampling of skills and knowledge of a verbal nature, and might thus be expected to benefit women more than men. To the extent that it is desirable to reduce the gender-related prediction difference, then, the writing score may be a useful addition to the variables employed in predicting students' college performance.

General Conclusions

This study was intended as a preliminary investigation, and generalization from the results is necessarily limited in several respects. The fact that the study used only one form of each version of the test restricts the degree to which generalizations can be drawn. Further, because the study was conducted under experimental conditions that presumably reduced examinee motivation and introduced some degree of self-selection, it is difficult to say whether the revisions would have a similar effect on the validity of the test when administered in an operational testing situation. Also, one cannot be certain that the results obtained with the present sample of institutions would necessarily apply to the broad range of colleges and universities that use SAT scores. Thus, it remains

for confirmatory research to determine the predictive validity of the final version of the revised SAT under operational testing conditions. Despite these qualifications, however, it is reasonable to draw at least tentative overall conclusions about the effects of the test revisions implemented in this study.

In evaluating the results, it is important to bear in mind the general objectives of the New Possibilities project to remodel the SAT. The revisions in the test were determined on the basis of conceptual considerations, as outlined in the Introduction. Although psychometric considerations also played a role, it was not an objective of the project to increase the test's predictive validity. Rather, the purpose of the project was to implement the proposed enhancements in the test, while being sure to maintain the psychometric integrity of the test.

The project's objectives appear to have been met, at least with regard to effects of the present test revisions upon the aspects of validity examined in this study. The revisions did not reduce the test's predictive validity or incremental validity, nor did they adversely affect the degree of underprediction for women relative to men. To the contrary, where differences were observed, they were in the direction of increased validity and decreased gender-related prediction differences.

The need for confirmatory research has been stressed, and one set of validity-related issues that remains to be studied concerns differences among ethnic groups--issues that could not be addressed with the modest sample in the present study. Among next steps in research, then, it is important to examine issues of equity for key subgroups, as well as to confirm the overall validity of the remodeled SAT as it is implemented in an operational context.

References

Braun, H. I., & Jones, D. H. (1985). Use of empirical Bayes methods in the study of the validity of academic predictions of graduate school performance. (Research Report No. 84-34). Princeton, NJ: Educational Testing Service.

Bridgeman, B., Hale, G. A., Lewis, C., Pollack, J., & Wang, M. (1992). Placement validity of a prototype SAT with an essay. Unpublished manuscript, Educational Testing Service, Princeton, NJ.

College Entrance Examination Board. (1988). The college handbook, 1988-89. New York: Author.

College Entrance Examination Board. (1989). Annual survey of colleges, 1989-90: Summary statistics. New York: Author.

College Entrance Examination Board. (1990a, October). Background on the new SAT-I and SAT-II. Unpublished report. New York: Author.

College Entrance Examination Board. (1990b). The college handbook, 1991. New York: Author.

Educational Testing Service. (1989). A reader's guide to Scholastic Aptitude Test (SAT) test analysis report. (Statistical Report No. SR-89-61). Princeton, NJ: Author.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. Philadelphia: Society for Industrial and Applied Mathematics.

Elliott, R., & Strenta, C. A. (1988). Effects of improving the reliability of the GPA on prediction generally and on comparative prediction for gender and race particularly. Journal of Educational Measurement, 25, 333-347.

Little. R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data. New York: Wiley.

Miller, R. G. (1966). Simultaneous statistical inference. New York: McGraw-Hill.

Ramist, L. (1984). Predictive validity of the ATP tests. In T. F. Donlon (Ed.), The College Board technical handbook for the Scholastic Aptitude and Achievement Tests (pp. 141-170). New York: College Entrance Examination Board.

Rosenthal, R., & Rubin, D. (1984). Multiple contrasts and ordered Bonferroni procedures. Journal of Educational Psychology, 76, 1028-1034.

Young, J. W. (1991). Gender bias in predicting college academic performance: A new approach using item response theory. Journal of Educational Measurement, 28, 37-47.

Appendix

Analyses were conducted to determine reliability, item difficulties, item discrimination, and speededness of the sections of the current SAT and NPP test. These analyses were conducted on students in all the original 27 colleges, prior to exclusion of colleges with small numbers of students and prior to screening of the sample of students within colleges. Unlike the predictive validity analyses, in which data were computed separately per college and averaged across colleges, these analyses were conducted by combining students across colleges. In analyses for a given test section, students who responded to fewer than three items in that section were excluded. Data for the analyses involving verbal and mathematical portions of the test were taken from students who were given both verbal sections, or both mathematical sections, of a given test version.

Verbal Sections

In the sample used here, analyses were based on 380 students given both verbal sections of the current SAT and 390 students given both verbal sections of the NPP test.

Reliability. Coefficient alpha reliabilities were computed for individual test sections, and reliabilities of the scores for sections combined were estimated using the Angoff/Feldt procedure (Educational Testing Service, 1989). The results were as follows.

Current SAT:
- Verbal Section 1 (CV1): .85
- Verbal Section 2 (CV2): .85
- All SAT verbal items: .92

NPP test:
- Verbal Section 1 (NV1): .83
- Verbal Section 2 (NV2): .85
- All NPP verbal items: .88

These reliabilities appeared to be reasonably high and to be within the same general range for the current SAT and NPP test. However, given the possibility that the greater number of items per reading passage in the NPP test than the current SAT may have inflated the coefficient alpha reliability (H. Wainer, personal communication, June 1990), conclusions cannot be drawn without further study into the most appropriate method of computing reliability. (See footnote 5, above.)

Item difficulty. Item difficulty was represented by the delta statistic, with higher delta scores indicating greater item difficulty. Delta equating was done using the sample of students who took both verbal sections of the current SAT (CV1 and CV2). A set of 60 items in SAT-V were used to obtain the transformation parameters for placing the SAT difficulty

estimates from this fall 1989 test administration on the SAT equated delta scale. This block of common items was selected to reflect a rectangular distribution of item difficulty values, excluding items at the beginning and end of sections in SAT-V. The transformation parameters from the delta equating were then applied to the observed deltas obtained on the sample of students who took both verbal sections of the NPP test (NV1 and NV2).

For the current SAT, the means (and SDs in parentheses) of the equated deltas, computed across items, were as follows.

Current SAT:
CV1: 11.4 (3.2)
CV2: 11.4 (2.8)
All SAT verbal items: 11.4 (3.0)

NPP test:
NV1: 10.7 (1.8)
NV2: 10.5 (2.7)
All NPP verbal items: 10.6 (2.4)

Apparently, then, the NPP test sections were somewhat easier and less variable than the current SAT sections.

Item discrimination. Item discrimination for a given item is typically represented by the r-biserial correlation with the total test of which that item is a part. The r-biserial correlations were computed using scores on the total test (i.e., both sections of either SAT verbal or NPP verbal) as the criterion.

The means of the r-biserials (and SDs in parentheses), computed across items, were:

Current SAT:
CV1: .48 (.13)
CV2: .51 (.08)
All SAT verbal items: .49 (.11)

NPP test:
NV1: .53 (.10)
NV2: .52 (.10)
All NPP verbal items: .53 (.10)

It appears that the r-biserials for the NPP test sections were slightly higher on average than the r-biserials for the current SAT sections.

Speededness

Speededness of a test section is evaluated by a combination of statistics. Among the criteria that are used to define a section of the SAT as nonspeeded are: (a) essentially all students complete 75% of the items, and (b) 80% of the examinees complete essentially all of the items. Some

subjective judgment must also be applied (for example, if the last item is so difficult that few examinees attempt it), so that these are not regarded as hard and fast rules. One other useful statistic is the ratio of not-reached item variance to score variance; generally, a lower ratio indicates a less speeded test section.

Speededness statistics for the current SAT verbal sections were as shown in the table below:

	CV1	CV2	NV1	NV2
Percent completing 75% of section	98.9	99.1	99.1	98.6
Number of items reached by 80% of examinees	44	40	25	34
Total number of items	45	40	25	35
Ratio of not-reached variance to score variance	.12	.08	.06	.08

According to these data, the first NPP verbal section (NV1) was slightly less speeded than the two sections of the current SAT, while the second NPP verbal section (NV2) was roughly comparable in speededness to the current SAT sections.

Mathematical Sections

For the mathematical area, data analyses were based on the 405 students given both mathematical sections of the current SAT and 355 students given both mathematical sections of the NPP test. Because the main analyses of the study involved all NPP mathematical items except the Algebra Placement items, data are presented here for that reduced set of items as well as for the full sections.

Reliability. Coefficient alpha reliabilities for the mathematical sections were:

Current SAT:
- Mathematical Section 1 (CM1): .82
- Mathematical Section 2 (CM2): .85
- All SAT mathematical items: .90

NPP test:
- Mathematical Section 1 (NM1): .85
- Full Mathematical Section 2 (NM2): .84
- Math. Section 2 excluding Algebra Placement items (NM2e): .75
- Estimated reliability for combination NM1 + NM2e (NMe): .88

The estimated reliability for score NMe was derived as follows. The reliability of the combination NM1 + NM2, using the Angoff/Feldt procedure, was estimated to be .90. The Spearman-Brown formula was then applied, in order to estimate the reliability of the set of all mathematical items excluding the 12 Algebra Placement items.

Item difficulty. Delta equating was done using the sample of students who took both mathematical sections of the current SAT (CM1 and CM2). All 60 items in SAT-M were used to obtain the transformation parameters for placing the SAT difficulty estimates from this fall 1989 test administration on the SAT equated delta scale. These parameters were then applied to the observed deltas obtained on the sample of students who took the two mathematical sections of the NPP test, NM1 and NM2.

The means (and SDs in parentheses) of the equated deltas, computed across items, were as follows.

Current SAT:
- CM1: 12.5 (3.4)
- CM2: 12.4 (3.0)
- All SAT math. items: 12.5 (3.2)

NPP test:
- NM1: 12.2 (3.1)
- NM2e: 14.0 (2.5)
- All NPP math. items, excl. Alg. Place. items (NMe): 12.6 (3.1)

Apparently, then, the mean difficulty of the NPP mathematical items was comparable to that of the current SAT sections.

Item discrimination. The r-biserial correlations were computed using scores on the total test (i. ., both mathematical sections of either the current SAT or the NPP test) as the criterion.

The means of the r-biserials (and SDs in parentheses), computed across items, were:

Current SAT:
- CM1: .55 (.07)
- CM2: .51 (.11)
- All SAT math. items: .53 (.10)

NPP test:
- NM1: .56 (.12)
- NM2e: .67 (.08)
- All NPP math. items, excl. Alg. Place. items (NMe): .59 (.13)

The r-biserials appear to be higher for the NPP test than for the current SAT, apparently because of the generally higher r-biserial for the Grid-in items (NM2e) than for the others.

Speededness. Speededness statistics for the current SAT mathematical sections were as shown in the table below. (Data are presented for the full section NM2, rather than for the subset of items labeled NM2e, since a single time limit applies to all items in section NM2, and speededness cannot be evaluated for a subset of items therein.)

	CM1	CM2	NM1	NM2
Percent completing 75% of section	97.8	98.3	98.3	95.8
Number of items reached by 80% of examinees	24	33	30	21
Total number of items	25	35	33	22
Ratio of not-reached variance to score variance	.10	.12	.13	.14

It appears that the NPP sections were slightly more speeded than the current SAT sections; however, it is impossible to say how the two tests would have compared in speededness with the Algebra Placement items excluded.

Writing Test

Statistics similar to those for the verbal and mathematical sections could be computed for the multiple-choice section, but not the essay section, of the NPP writing test. These analyses were based on all 2,033 students administered the multiple-choice writing section.

Reliability. Coefficient alpha reliability for the multiple-choice section of the NPP writing test, computed for all students taking this section, was .86. Although a direct measure of internal consistency is not available for the essay, the inter-reader reliability (based on the 1853 students taking the essay) was .81, indicating a relatively high degree of scoring consistency.

Item difficulty. For the multiple-choice section of the NPP writing test, delta equating was based on 15 items. These were Usage items that had been administered in an earlier prototype test to several thousand high school juniors, and these items had been delta-equated to the SAT scale on the basis of data from that administration. The mean delta for all 43 items in this section, computed for all students taking this section, was found to be 11.1 (SD = 3.0).

Item discrimination. The mean r-biserial correlation between items and total score on the multiple-choice writing section was found to be .50 (SD = .08).

Speededness. Speededness statistics for the multiple-choice section of the NPP writing test were as shown below.

Percent completing 75% of section	98.2
Number of items reached by 80% of examinees	43
Total number of items	43
Ratio of not-reached variance to score variance	.13

Most of these figures suggest a low degree of speededness. Nevertheless, the ratio index was higher than that of the NPP verbal sections (.06 and .08, respectively), suggesting that firm conclusions about the relative speededness of this section cannot be drawn.

General Conclusions

The coefficient alpha reliabilities for the test sections were reasonably high. Comparison between tests is rendered difficult, however, by the possibility that use of comparatively large sets of reading items may have produced inflated reliability estimates for the NPP verbal area, and by the fact that relatively few items contributed to the mathematical score (NMe) in the NPP test. There were some differences between tests in item difficulty, item discrimination, and speededness. These differences may reflect characteristics of the particular items that appeared in the forms of the tests used here. Because only one form of each version of the test was used, general conclusions about characteristics of the current SAT versus the proposed revision cannot be drawn without further research using additional forms of each.