DOCUMENT RESUME

ED 390 893 TM 024 175

AUTHOR	Bridgeman, Brent; And Others
TITLE	Placement Validity of a Prototype SAT with an Essay. Research Report.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-92-28
PUB DATE	May 92
NOTE	33p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*College Entrance Examinations; College Freshmen; *Essay Tests; Grades (Scholastic); High Schools; Mathematics Achievement; Multiple Choice Tests; Predictive Validity; *Student Placement; *Test Validity; Verbal Tests; Writing Tests
IDENTIFIERS	Composite Scores; Prototypes; *Scholastic Aptitude Test; *Test Revision

ABSTRACT

Grades in college freshman English composition courses were predicted from high school rank in class, multiple-choice writing scores, essays, current Scholastic Aptitude Test (SAT) Verbal scores, and scores from a revised version of the SAT-Verbal. Data were obtained from 21 English courses at 17 different colleges with some supplementary data provided by an additional college. In general, a writing composite score consisting of essay and multiple-choice writing scores appeared to outperform current or revised SAT-Verbal scores; validity coefficients were as high or higher for the writing composite score; and the underprediction of the grades of women students was reduced. The best predictions were obtained from the combination of high school rank with the writing composite score. Appendixes discuss the prediction of freshman mathematics grades 3 tables), provides lists of colleges in the sample, and describes test item types. (Contains 6 tables and 8 references.) (Author/SLD)

RESEARCH REPORT

PLACEMENT VALIDITY OF A PROTOTYPE SAT WITH AN ESSAY

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

Educational Testing Service
Princeton, New Jersey
May 1992

Placement Validity of a Prototype SAT with an Essay

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

May, 1992

Acknowledgments

The authors wish to express their gratitude to the students and staff at all of the participating institutions. In addition, we would like to thank all of our colleagues who assisted with this project, especially:

John Fremer, Steve Graff, Larry Hecht, and Larry Litten for their overall leadership of the New Possibilities Project;

Sydell Carlton, Linda Cook, Samuel Messick, Donald Rock, and Warren Willingham for their general advice;

James Braswell, Edward Curly, and Marilyn Sudlow for their assistance in explaining to us how the tests were developed;

Gloria Weiss for supervising the data collection;

Elizabeth Burton, Anne Neeff, and Brian O'Reilly for making arrangements for the data collection;

Sheila Krolikowski and the entire operations staff for processing the data sent from the various testing sites;

Nancy Feryok, Ida Lawrence, and Samuel Livingston for preliminary analyses of the test data;

Lucient Chan and Laura McCamley for assisting in the data analyses;

Nancy Burton, Donald Powers, and Howard Wainer for their reviews of earlier drafts of this report; and Sabrina Waller for secretarial assistance.

Abstract

Grades in college freshman English composition courses were predicted from high school rank in class, multiple-choice writing scores, essays, current SAT-Verbal scores, and scores from a revised version of SAT-Verbal. Data were obtained from 21 English courses at 17 different colleges with some supplementary data provided by an additional college. In general, a writing composite score consisting of essay and multiple-choice writing scores appeared to outperform current or revised SAT-Verbal scores; validity coefficients were as high or higher for the writing composite score, and the underprediction of the grades of women students was reduced. The best predictions were obtained from the combination of high school rank with the writing composite score.

Placement Validity of a Prototype SAT with an Essay

The primary purpose of the SAT has historically been, and continues to be, the selection of students for admission to colleges and universities. Although not designed as a placement instrument, the ready availability of SAT scores on student transcripts has led some institutions to use those scores for making course placement decisions. The two course sequences in which placement decisions must be made most frequently are English and mathematics.

The Test of Standard Written English (TSWE) was added to regular SAT administrations in 1974 for the specific purpose of aiding in English placement decisions, and its validity for placement purposes in conjunction with the SAT-Verbal (SAT-V) has been studied (Breland, Conlan, & Rogosa, 1976; Breland, 1977). In these studies, TSWE predicted freshman-year writing performance at least as well as other available measures including precourse writing samples, high school English grades, and high school rank in class. The TSWE was more useful for placing students into English classes than the SAT-V. Multiple regression analyses suggested that a score on a holistically-graded 20-minute writing sample could significantly improve predictions even after high school grades and TSWE scores were already taken into consideration.

The validity of SAT-Mathematical (SAT-M) scores for mathematics placement decisions was studied by Bridgeman and Wendler (1989). Although SAT-M might be useful as an initial screening test, other tests that were specifically designed for mathematics placement appeared to be superior to SAT-M for making mathematics placement decisions.

Since 1987, efforts have been under way to make modifications in the Admissions Testing Program (including SAT and Achievement Tests) "that would make the program more useful to students, schools, and colleges" (College Board, 1991). The purpose of the current study was to determine the probable impact of the proposed modifications on the placement validity of the revised SAT. Specifically, the current study was designed to focus primarily on how well the modified tests predicted success in introductory English composition and mathematics courses in comparison to the current SAT. This kind of course-specific predictive validity is a necessary but not sufficient criterion for an effective placement program. If, for example, potential failures were accurately identified and placed in special remedial classes, the system may seem to be working fine from the standpoint of the person teaching the regular class. But from the student point of view, the system would only be seen as successful if the remedial course effectively provided the skills needed to succeed in the regular course (or if the student could avoid taking the regular course, as often happens to students who fail to place into calculus courses).

For this study a prototype of the revised SAT was developed, which is herein labeled the New Possibilities Project test, or NPP test. This 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.

The plan for revision of the test is detailed in the College Board announcement "Background on the New SAT-I and SAT-II" (College Board, 1991). (SAT-I refers to the verbal and mathematical reasoning tests; SAT-II refers to the subject tests [formerly called ATP Achievement Tests].) 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 referred to as "student produced response") 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.

A major change in the mathematical section (SAT-M) that was contemplated when this study was designed in 1989 was the addition of algebra achievement items and an algebra subscore that might make the test more useful for course placement decisions. However, this suggestion was strongly criticized by several committees that reviewed these preliminary plans. In the words of the chair of the College Board's Mathematical Science Advisory Committee:

> We do not...believe that it is possible to fit an adequate college placement test in mathematics within the framework of a new SAT without seriously weakening the SAT itself. The Board has developed other instruments that can be used for such placement. It should not attempt to load down the SAT with more functions than it can reasonably bear. The proposed subscore risks damaging both the reputation and effectiveness of the Board's most widely used test.

Because of these concerns, the idea of an algebra placement subscore for SAT-I was dropped.[1] Therefore, results from the portion of this study dealing with mathematics placement are now of very limited interest; they are briefly summarized in Appendix A.

Changes in plans since this study was designed in 1989 also have a potential impact on the interpretation of the English placement results. The sentence-completion item type from the current SAT is now to be retained, although it was not included in the NPP test 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 to accommodate the changes in the test structure while ensuring non-speededness of the test; about 75 minutes will be allowed for each of the SAT-I tests. In the Writing Test, the multiple-choice portion and essay are tentatively scheduled for 40 minutes and 20 minutes, respectively. For these reasons, the present study is regarded as an interim investigation and not an exact indicator of the placement 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.

Method

Sample of Colleges

Two separate samples were used in the analyses. Some analyses were conducted on the same sample of colleges used in the predictive validity analyses (Hale, Bridgeman, Lewis, Pollack, & Wang, 1992); this sample is referred to as Sample A. A supplementary sample (Sample B) was obtained in which the tests were administered at both the beginning and the end of the semester in order to evaluate instructional sensitivity (gain from pretest to posttest) and concurrent correlations with course grades.

Sample A. Sample recruitment and test administration procedures are fully documented in Hale et al. (1992). Briefly, 27 colleges[2] responded to a general solicitation of four-year coeducational colleges in the United States with freshman enrollments of 280 or greater. They administered a form of the prototype test during freshman orientation. Eight colleges were eliminated because they did not return enough tests for analysis, resulting in a final sample of 19 colleges for this part of the predictive validity study.

---

[1]Note that the objections to incorporating an algebra placement score in SAT-I do not apply to SAT. Efforts are continuing to enhance the usefulness of current (and possibly new) mathematics achievement tests for placement.

[2]The term "college" is used in this report to refer to both colleges and universities.

The criterion for the predictive validity study was end-of-year GPA, but the placement study required grades in freshman English composition courses. Remedial and advanced English courses were eliminated as were courses that were not clearly focused on composition skills. Sufficient numbers of students for meaningful analysis were eventually obtained from 13 colleges (see Appendix B). One college had two distinct freshman composition courses at the same ability level, so the final sample consisted of 14 courses.

Sample B. This sample resulted from a separate solicitation and did not overlap with Sample A. Colleges had to agree to test in English classes at the beginning of the semester and again at the end of the semester. Of the six colleges that initially agreed to participate, usable data were obtained from four; two were Midwestern state universities, one was a Midwestern private college, and one was a New England private university (see Appendix B). One of the state universities tested in both remedial and regular classes, and the other large state university sent data from three distinct regular courses (each course had a writing component, but the emphasis in the readings differed across courses). Thus, the final sample consisted of 7 courses.

Sample of Students

All participants in both samples were unpaid volunteers who were instructed that participation was not required. Because the students understood that the scores were to be used only for research purposes, there were no extrinsic rewards motivating student performance. No procedure could correct for this lack of strong extrinsic motivation, but an attempt was made to eliminate from both samples students who were clearly exerting very little effort. 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, (b) failed to respond in both of the subtests administered, where failure to respond was defined as not answering at least three questions in the multiple-choice subtests, or failure to write an on-topic essay (as judged by raters) in the essay subtest, and (c) failed to respond to any of the "grid-in" mathematics items, for students administered that item type. In addition, an outlier analysis was performed, designed to exclude students whose performance was substantially below (two standard deviations or more) that predicted by their scores on previously administered standardized tests. Applying all of these procedures reduced Sample A by about 6% with a slightly smaller reduction in Sample B.

Verbal and Writing Test Sections

Contents of the verbal and writing sections are summarized in Appendix C. The sections administered to Sample A were as follows:

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

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

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

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

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

NW-ESS: NPP test, Writing, Essay section (1 essay)

The score for each multiple-choice test section was a formula score that included a correction for random guessing. Essays were read by two independent readers each of whom assigned a score based on a one to six holistic scale; ratings from the two readers were added yielding a 2 to 12 score range. The same test forms and scoring procedures were used in Sample B, except that parallel versions of both the multiple-choice and essay writing tests were developed and administered along with the original versions.

Experimental Design

Sample A. Because the time available did not permit administration of a full test (verbal, math, and writing) to any individual student, each student was administered two test subsections, each requiring 30 minutes. Twenty-one different pairs of subsections were spiralled randomly within a college, six representing all possible pairs of subsections from the current SAT (including two mathematics sections that were not relevant for this report), and 15 representing all possible pairs of subsections from the revised test (including two mathematics sections) plus subsections of the writing test. It was then possible to use the statistical procedures described below to estimate the validity of various combinations of test sections even though no student actually took more than two 30-minute sections.

Sample B. The spiral plan varied from college to college depending on the amount of time that the particular college could allocate to the testing. Some students took a verbal section, a multiple-choice writing section, and an essay, others took only two out of the three. The parallel forms of the writing test were spiralled together so that some students received form 1 and others received form 2 during the fall pretest; at the end-of-semester testing each student was assigned to the form that she or he did not take at the pretest. Parallel forms of the verbal test were not available; half of the

NPP sample took section NV1 as a pretest and section NV2 as a posttest while the other half took NV2 as the pretest and NV1 as the posttest. Because no one took both NV1 and NV2 at the same time, statistics for the full verbal test could not be estimated in this sample. Although they are not strictly parallel, both verbal forms were intended to measure the same underlying verbal ability dimension. For the students who took the current verbal tests, CV1 and CV2 were similarly treated as pretest and posttest. Analyses were conducted separately for form 1 and form 2 tests, and the results of these analyses were averaged for final presentation in the tables.

Area Scores. Except as noted for the verbal scores in Sample B, the design permitted estimates of validity for combined scores as follows:

CV	Current SAT, Verbal area (CV1 + CV2)--only in Sample A
NV	NPP test, Verbal area (NV1 + NV2)--only in Sample A
NW	NPP test, Writing area, defined as the combination of NW-MC and NW-ESS, weighted 70%-30%, respectively. This weighting scheme was selected because it parallels the weighting in the English Composition Test with Essay.

Analysis Procedures

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

A multivariate normal distribution for the test and college performance scores is assumed in the maximum likelihood 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 English grade reported.

The EM method was employed in order to obtain more stable estimates of correlations. This is an important concern in light of the small number of students within each course taking any given pair of test sections as dictated by the study design. Essentially, all participating students in a course with an English grade available have contributed some information to the estimation of one of the correlation matrices (current SAT or NPP test). Thus, for each course, the number of such students associated with each of these correlation matrices will be identified as the total sample size for that matrix. Each within-course matrix was weighted by the n for that course and then averaged across courses to produce the correlations reported in the tables.

Across the 14 courses in Sample A, the n was 1189 for the current SAT sample and 2837 for the NPP sample. Across the 7 courses in Sample B, the ns were 767 and 1318 for the current SAT and NPP samples respectively.

For analyses that estimated the predictive validity of high school rank in combination with a test score, an Empirical Bayes (EB) procedure (Braun & Jones, 1985) was employed to obtain simultaneous estimates of the within-college regression coefficients. In essence, the EB approach uses 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 English grades and the test scores within each college, thus at least partially addressing the typical concern of shrinkage in cross-validation. The weighted average of the within-college correlations between English grade and the college-specific weighted composite of rank and test score is reported as an estimate of the predictive validity of the rank and score combination.

Confidence intervals for the differences in correlations (and gender-related prediction differences) were estimated by a bootstrap method (Efron, 1982) with 10,000 trials of sample size 14 for each trial. For a given comparison, the procedure simulates a sampling distribution of observed sample differences for a hypothetical population of courses similar to those participating in this study. Each bootstrap trial drew a random sample of size 14 employing a sampling with replacement procedure. Because of the relatively small number of courses (7) in Sample B, these procedures were used only in Sample A.

Results and Discussion

Correlations of the various predictors with grades in freshman English composition courses for the 14 courses in Sample A and the 7 courses in Sample B are presented in Table 1. Because high school rank was defined the same way in both the current SAT and NPP samples, the small difference in Sample A between the validity coefficients for rank (.32 vs. .30) simply represents sampling variation. In both samples, the writing composite score (analogous to SAT-II: Writing) was a better predictor of English grades than the NPP verbal score. The predictive validity for the writing composite was higher than the validity for the current verbal score in 9 out of the 14 courses (7 out of the 8 largest courses), but the considerable variability across courses resulted in a relatively broad 99% confidence interval around the mean difference of .09 (see Table 2). Although the slightly negative lower bound of the confidence interval suggests that actual differences in the population of courses similar to those sampled here could slightly favor the current SAT verbal test, superior predictions from the writing composite score are much more likely. This conclusion is strengthened by the replication of the key Sample A results in Sample B. Although each verbal score in Sample B was

Table 1

Correlations with English Grades Averaged Over 14 Courses in Sample A (and 7 Courses in Sample B)

Predictor	Current SAT	NPP
High School Rank	.32 (.34)	.30 (.34)
Verbal	.23 (.23)	.19 (.22)
M-C writing		.29 (.25)
Essay		.22 (.27)
Writing Composite (M-C and Essay)[a]		.32 (.30)
Verbal and Writing Composite[a]		.29 (.29)
Rank plus Verbal[b]	.38 (.39)	.34 (.38)
Rank plus Writing Composite[b]		.40 (.42)

Note. - Sample B correlations are in parentheses
Sample A n for Current SAT sample is 1189; n for NPP sample is 2837
Sample B n for Current SAT sample is 605; n for NPP sample is 1033
[a]equally weighted composite after rescaling by standard deviation
[b]combined with regression equation; different weights for each course

Table 2

Confidence Intervals for Selected Validity Differences

Scores	Validity Difference	99% Confidence interval lower bound	99% Confidence interval upper bound
Current Verbal-NPP Verbal	.04	-.09	.15
Writing Composite-NPP Verbal	.13	.02	.25
Writing Composite-Current Verbal	.09	-.04	.23
(Rank and Writing Composite)-(Rank and NPP Verbal)	.06	-.01	.12
(Rank and Writing Composite)-(Rank and Current Verbal)	.02	-.08	.13

derived from the administration of a single 30-minute section, the 30-minute verbal sections in Sample B predicted just as well as the 60-minute verbal sections in Sample A.

Because regression analyses (described below) suggested that the creation of a writing composite score by weighting the standardized NPP multiple-choice writing score by .7 and the essay by .3 might be less than optimal, an alternative equal weighting of the multiple-choice and essay components was tried for the 14 courses in Sample A. The weighted mean correlations for the .7-.3 weights and equal weights were nearly identical (.32 and .31 respectively). Validities for the .7-.3 weights were higher in half of the courses with higher validities for the equal weights in the other half. Using whichever of these two weighting schemes was best for a particular course and averaging these "best" validities resulted in an average correlation of .33 across courses. This "best" writing composite was a better predictor than the current SAT verbal score in 11 out of the 14 courses; the 95% confidence interval for the .10 difference between this "best" composite and the current SAT verbal score was from .01 to .21.

In both samples, the thirty minute essay by itself was about as good a predictor of first-year English grades as either the current or NPP verbal tests. The combination of the multiple-choice and essay writing scores to form the writing composite created a predictor that appeared to be superior to either score by itself. However, adding the verbal score to the writing composite to form an equally weighted combination of the these two scores did not improve the prediction of English grades. Indeed, in 10 out of the 14 courses in Sample A, the correlation for the writing composite by itself was higher than the correlation for the combination of the verbal and writing scores.

The combination of high school rank and the writing composite score in the EB regression analysis yielded a modest improvement in validity over the writing composite by itself. The combination of rank with the current SAT verbal score yielded a substantial increase in validity over the verbal score by itself. As indicated in Table 2, the combination of rank plus writing composite was not significantly higher than the combination of rank plus current SAT verbal score or rank plus NPP verbal score[3].

Results in a Motivated Sample

The above results must be interpreted cautiously because the tests were administered under low-motivation conditions. The students knew that the scores would be used only for research purposes. In order to determine whether similar results would be obtained for motivated students, an additional set of data was obtained from a state university in the Northeast. Although scores on the NPP writing test were not available, surrogates for the two components of the writing score (a multiple-choice writing test and an

[3]With a less conservative 95% confidence interval this latter difference was significant (95% interval .01 to .10).

holistically-scored essay) were obtained from the college's files. The multiple-choice writing test was the TSWE that students took at the same time as the SAT; this score is reported to colleges along with SAT scores. Two of the three item types on the multiple-choice NPP writing test also appear on the TSWE. The 20-minute essay was part of a basic skills battery taken by all incoming freshmen. As with the NPP essay, these essays were independently graded by two readers. Students were motivated to do well on this essay in order to avoid placement into remedial English. Information on high school rank and regular SAT scores were also in the college files.

The mean SAT-Verbal score of the 777 freshman English students was 500 with a standard deviation of 64. Correlations of the various scores with English grades are presented in Table 3. Multiple correlations based on

Table 3

Predictors of Regular English Grades at a State University in the Northeast

Predictor	r
High School Rank	.27
SAT Verbal	.24
TSWE	.26
Essay	.21
Verbal + Essay	.28
TSWE + Essay	.30
Verbal + TSWE + Essay	.31
Rank + Verbal	.32
Rank + Verbal + Essay	.34
Rank + TSWE + Essay	.35

$n = 777$

various score combinations are also presented. For all of the multiple correlations presented, the beta weights for each score were statistically significant ($p < .001$). In a model that combined rank, verbal, TSWE, and essay ($R = .36$), the weight on the verbal score was not significant ($p = .07$), thus replicating the finding in Samples A and B that the verbal score is superfluous when the other scores are available. In general, this analysis of motivated students replicated the key findings in Samples A and B (e.g., the

writing composite score was superior to the verbal score for prediction of English grades) and permits added confidence in the validity of the results in Samples A and B.

Weights for multiple-choice and essay scores. In the Northeastern state university, the standardized regression weights were .22 for the TSWE and .15 for the essay, suggesting that forming a writing composite by weighting the standardized NPP multiple-choice scores by .7 and the essay by .3 may undervalue the essay in some courses. Regression analyses of the two large courses in Sample B tended to support this finding. Analyses were run separately for the form 1 and form 2 scores. For the form 1 scores, the standardized regression weights were .28 for the essay and .06 for the multiple-choice scores at one college (n = 117) and at the other college (n = 200) the weights were .18 and .21 for the essay and multiple-choice scores respectively. For the form 2 scores, the weights were .33 (essay) and .15 (multiple-choice) at the first college (n = 105), and .12 (essay) and .35 (multiple-choice) at the other college (n = 208). Thus, in two of the four Sample B comparisons, the weight on the essay was substantially higher than the weight on the multiple-choice scores, in one comparison they were about equal, and in only one out of the four comparisons was the weight on the multiple-choice score substantially above the weight on the essay. However, as already noted, in Sample A .70-.30 weighting was superior to equal weighting in exactly half of the courses. These differences may reflect inevitable course to course fluctuations, but it may be possible in future research to identify college or course characteristics that are related to the optimal weights for each course. As data accumulates from the operational administration of SAT-II: Writing (when students are highly motivated to perform well on all sections of the test), the issue of the optimal weighting of the component scores should be revisited.

Gender Differences

Gender was included as an independent variable in the regression equations to reflect the degree of prediction difference between men and women expressed in grade point units. These differences, as computed in Sample A, are presented in Table 4. The .25 difference found for both the current and NPP verbal tests indicates that the grades of women were .25 grade points higher than the grades of men who had the same verbal score. The gender difference for the writing composite (.16) was .09 smaller than for the current verbal score (99% confidence interval .02 to .14). The prediction difference for the rank and writing composite combination was only .12.

Note that the values in the table indicate the difference, in grade point units, between the regression lines for men and women. However, the common practice is to use a single regression line for men and women. Assuming an approximately equal number of men and women, the underprediction of women's grades would then be half of the values indicated in Table 4. Thus, for example, if prediction were based on the writing score, women would be expected to receive English grades that were .08 higher than predicted from the single prediction equation (e.g., on average, women predicted to receive a 3.00 would receive a 3.08).

Table 4

Underprediction of Women's English Grades

Predictor	Mean Prediction Difference Current SAT	NPP
Verbal	.25	.25
Writing Composite		.16
Rank and Verbal	.15	.15
Rank and Writing Composite		.12

Grade by Score Crosstabulation

Correlations provide a convenient means of comparing the predictive efficiency of several different measures, but they are of little use for making practical decisions on where cut scores should be set. For this purpose, a crosstabulation of grades and scores may be useful. Table 5 presents such a crosstabulation for the regular English composition course at

Table 5

Crosstabluation of Writing Composite Scores and English Grades for the Regular Course at a Midwestern University from Sample B

		English Grades					
		F,D	D+,C	C+	B	B+,A	
Composite Scores	36+	3	4	6	13	18	49
	31-35	3	7	12	10	22	54
	26-30	17	19	19	22	23	100
	21-25	11	17	14	14	11	67
	0-20	23	24	19	14	8	88
		57	71	70	73	82	

a state university in the Midwest. The table clearly illustrates the imprecision of prediction from a correlation of .31 (the full range of grades is represented in each score category) while at the same time indicating the potential usefulness of such a correlation for making placement decisions.

For example, setting a cut off score of 21 for placement into the regular class would remove 40% of the unsuccessful students (those getting Ds or Fs) while keeping out fewer than 10% of the students who ultimately received B+s or As; a cut score of 26 would remove 60% of the unsuccessful students while eliminating 23% of the B+ or better students. This crosstabulation is provided only as an example; any college considering a placement program should conduct its own study that is consistent with local grading practices and student characteristics. The social and economic benefits of removing the potential failures from the regular class then must be weighed against the expenses of remedial instruction and the likelihood that the extra instruction will be successful.

Gain Analyses

In general, a test that is used to exempt students from college course requirements should be sensitive to instructional gain (Willingham, 1974). That is, scores should be higher at the end of instruction than they were at the beginning of instruction. Tests that are most closely tied to what is actually taught (and learned) should demonstrate the largest gains.

Gain analyses from the beginning of the semester pretest to the end of semester posttest were performed on the 7 courses in Sample B. Equipercentile and linear equating procedures were used on the pretest scores to equate scores on the two forms. (Both procedures yielded nearly equivalent results and the linear equating was used.) Because of the counterbalanced design (with half of the sample taking each form at pretest and the opposite form at posttest) small equating errors would not bias the results; the equating merely made it possible to express the gain for each measure as a single number. Mean gains with their associated 99% confidence intervals are presented in Table 6. The weighted average gain was negative for all four measures and was statistically significant for three of them. One possible conclusion is that English classes harm students' verbal and writing skills as defined by these measures. This conclusion does not seem very likely, and anecdotal evidence suggests an alternative explanation. A number of students wrote essays complaining about having to take a test that did not count for anything while they were preparing to take their final examinations that they cared about very much. Although these off-topic essays were not scored and these students were removed from the sample, they may have reflected a pervasive attitude. The most reasonable conclusion may be that college freshmen do not try as hard on the second test that does not count as they did on the first test, especially if the first test was while they were relatively intimidated beginning students and the second test was in the middle of finals. Because of these serious doubts about the validity of the posttest scores, plans to use the posttest essays as alternative criterion measures were dropped.

Table 6

English Gain Scores

College	CV n	CV gain	CV SD	CV 99% Interval low	CV 99% Interval high	NV n	NV gain	NV SD	NV 99% Interval low	NV 99% Interval high	MC Writing n	MC Writing gain	MC Writing SD	MC Writing 99% Interval low	MC Writing 99% Interval high	Essay n	Essay gain	Essay SD	Essay 99% Interval low	Essay 99% Interval high
1A	197	-0.382	5.61	-1.41	0.65	210	-0.236	4.59	-1.05	0.58	375	0.203	[illegible]99	-0.60	1.00	353	-0.130	2.22	-0.43	0.17
1B	100	-0.656	6.55	-2.39	1.07	108	-1.268	4.73	-2.46	-0.08	174	-2.596	6.43	-3.85	-1.34	119	-1.008	2.11	-1.51	-0.50
2	71	-0.019	5.55	-1.76	1.73	141	-0.542	4.36	-1.51	0.42	184	-0.545	5.66	-1.62	0.532	112	0.063	2.02	-0.44	0.56
3	15	-2.004	2.58	-3.99	-0.02	31	-0.318	4.68	-2.63	1.99	39	1.277	5.13	-0.95	3.50	29	-0.35	1.97	-0.97	1.04
4A	10	1.617	5.64	-4.18	7.41	17	0.068	3.88	-2.68	2.82	31	0.704	6.72	-2.62	4.02	18	-0.833	1.82	-2.07	0.40
4B	29	-2.392	5.08	-5.00	0.21	55	0.224	3.55	-1.03	1.48	80	-0.170	4.86	-1.60	1.26	54	-1.056	1.96	-1.76	-0.35
4C	20	1.872	3.40	-0.30	4.05	45	-7.17	3.93	-2.29	0.86	60	-1.444	6.00	-3.55	0.57	38	-0.895	1.90	-1.73	-0.06
Weighted average	442	-0.13	0.27*	-1.12	0.27	607	-0.48	0.18*	-0.94	-0.02	943	-0.54	0.19*	-1.04	-0.04	723	-0.36	0.08*	-0.57	-0.16

*Standard error of the weighted average.

Conclusions

At least for colleges that are similar to those in this study, it would appear that English placement decisions could be made more effectively with the new writing composite score than they could with the current or NPP verbal scores. In Sample A, the average validity coefficient for the prediction of English course grades was .09 higher with the writing composite score than with the current SAT verbal score. Furthermore, the replication of this apparent superiority in Sample B (albeit with a single verbal section) and in the motivated sample at the Northeastern university permits some confidence in its reality. This result is consistent with previous findings (Breland 1977) on the relative efficiency of a writing score that includes multiple-choice questions and an essay for the prediction of English grades. A further advantage of the writing composite score for English placement is that the underprediction of women's grades was about half as large as it was when the verbal score was used as a predictor. Thus, even if the validity coefficients were identical, the writing composite would be the preferred predictor.

The essay score by itself was about as good a predictor as the verbal score, but not as good as when it was combined with the multiple-choice writing score. The addition of a verbal score to the writing score did not improve prediction. The combination of high school rank and the writing composite score provided the best predictions.

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 84-34. Princeton, NJ: Educational Testing Service.

Breland, H.M. (1977). A study of college English placement and the Test of Standard Written English. College Entrance Examination Board Research and Development Reports (RDR 76-77, No. 4).

Breland, H.M., Conlan, G.C., & Rogosa, D. (1976). A preliminary study of the Test of Standard Written English. Princeton, NJ: Educational Testing Service.

Bridgeman, B., & Wendler, C. (1989). Prediction of grades in college mathematics courses as a component of the placement validity of SAT-Mathematics scores. College Board Report No. 89-9. New York: College Entrance Examination Board.

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

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. CBMS-NSF regional conference series in applied mathematics (No. 38), Philadelphia: Society for Industrial and Applied Mathematics.

Hale, G.A., Bridgeman, B., Lewis, C., Pollack, J., & Wang, M. (1992). A Comparison of the Predictive Validity of the Current SAT and an Experimental Prototype. (ETS RR) Princeton, NJ: Educational Testing Service.

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

Appendix A

Prediction of Freshman Mathematics Grades

As with the English grade predictions, two different samples were used to predict freshman mathematics grades. Sample A was the same as for the English placement analyses except that freshman mathematics courses rather than English courses were sampled. Because many fewer students enroll in mathematics than in English, sufficient numbers for analysis were available from only four algebra courses and two calculus courses.

Because of the limitations of Sample A for the study of mathematics placement validity, another sample of colleges was recruited (designated Sample C to distinguish it from Sample B, the English placement sample). Colleges in Sample C administered the one-hour mathematics tests in mathematics classrooms at the beginning of the semester and parallel versions of the tests at the end of the semester. One of the six colleges that conducted the testing returned valid grades on only 13 students spread over three courses, so that college was dropped from the analysis. Four of the five remaining colleges were relatively unselective state colleges while the fifth was a moderately selective private college. All five colleges had at least one course below the level of introductory calculus. These included courses in intermediate algebra, college algebra, and trigonometry; for convenience, the nine courses below calculus were all labeled as "algebra" courses. Only three colleges in this sample sent grades for calculus courses. All courses were designated by the colleges as freshman-level courses (or non-credit remedial courses), but some of the courses included a small number of non-freshman students.

Test Sections

Test sections (30 minutes each) for both samples are as follows:

CM1: Current SAT, mathematical section 1 (25 items)
25 five-choice regular mathematics items

CM2: Current SAT, mathematical section 2 (35 items)
15 five-choice regular mathematics and 20 four-choice quantitative comparisons

NM1: NPP test, mathematical section 1 (33 items)
15 five-choice regular mathematic items and 18 four-choice quantitative comparisons

NM2: NPP test, mathematical section 2 (22 items)
12 four-choice algebra placement items and 10 grid-in items

Note that the item types in NM1 are identical to those in CM2, but there are two fewer quantitative comparisons items in NM1. Although this change was made to make the NPP form less speeded, in fact it was more speeded. About 73 percent of the sample completed CM2 but only 63 percent completed NM1. An additional difference between CM2 and NM1 was in the way that they were

printed in the test book. In CM2 the 35 items were squeezed onto four pages; the 33 items in NM1 were spread over seven pages, and each page included a separate column for scratch work. Further research on the possible effects of this difference in format is currently in progress.

The algebra placement items in NM2 emphasized the routine application of algebraic formulas. The grid-in items were similar in content to the five-choice regular mathematics items on the current SAT except that no answer choices were provided. Instead, students had to grid their numerical answers on an answer sheet that allowed them to grid up to four characters--numerals 0 through 9, a decimal point, and a slash for fractions.

In Sample C, parallel versions of each of the above forms were administered (CM3 parallel to CM1, CM4 parallel to CM2, NM3 parallel to NM1, and NM4 parallel to NM2). At the pretest, students were randomly given one of four test forms; each form consisted of two CM sections (either CM1 and CM2 or CM3 and CM4) or two NM sections (NM1 and NM2 or NM3 and NM4). The parallel forms were administered at the end-of-semester posttest so that if a student took one form at pretest he or she would take the other form at posttest.

Methods

Screening of students who showed relatively low motivation was conducted as previously described for the English placement study. Similarly, the previously described EM algorithm was used to estimate correlations in Sample A. The EM algorithm was not used in Sample C because all students took both sections of the mathematics test. Thus, the ns for Sample C represent the number of people who actually took the specified form, but in Sample A the ns represent the number included in the EM calculations (i.e., the number of students in all spirals, not just the spirals with the target mathematics sections). In Sample A, a given mathematics section was included in 1/2 of the current SAT spirals and 1/3 of the NPP spirals. Thus, for example, only about 86 students in Sample A took CM1 and had an algebra grade reported although the EM calculation is based on the 171 students who took any current test and had a grade reported.

All correlational analyses were performed within courses and the results were then weighted by the n for the course and averaged across courses. In Sample C, where parallel forms were given to different random samples of students, correlations of these forms with course grades were first averaged over forms within courses and then weighted by n and averaged over courses.

Results

Mean correlations for the 9 algebra courses in Sample C and the 4 algebra courses in Sample A are presented in Table A-1. In Sample C, predictions from the current and NPP tests appear to be quite similar. In Sample A, the NPP tests appear to be superior, but note that the biggest difference is between the two sections that contain the same item types (CM2 and NM1). When the three remedial-level courses were excluded from Sample C a similar, though smaller, difference between CM2 and NM1 emerged (rs of .30 and .36 respectively).

Table A-1

Correlation of Mathematics Test Scores and Algebra Grades Averaged over Courses

Sample/n	Test	r	Test	r
9 courses in Sample C Current n=416 NPP n=388	CM1 (or 3)	.34	NM1 (or 3)	.32
	CM2 (or 4)	.32	NM2 (or 4)	.35
	CM Total	.37	NM Total	.37
		[.28]		[.27]
		[.45]		[.45]
4 courses in Sample A Current n=171 NPP n=401	CM1	.26	NM1	.30
	CM2	.12	NM2	.23
	CM Total	.21	NM Total	.30

*Numbers in brackets represents the 95% confidence interval for the Sample C totals

Table A-2

Correlation of Mathematics Test Scores and Calculus Grades Averaged over Courses

Sample/n	Test	r	Test	r
3 courses in Sample C Current n=192 NPP n=182	CM1 (or 3)	.35	NM1 (or 3)	.41
	CM2 (or 4)	.37	NM2 (or 4)	.37
	CM Total	.39	NM Total	.43
		[.26]		[.30]
		[.50]		[.54]
2 courses in Sample A Current n=134 NPP n=318	CM1	.18	NM1	.43
	CM2	.20	NM2	.27
	CM Total	.21	NM Total	.42

*Numbers in brackets represents the 95% confidence interval for the Sample C totals

As shown in Table A-2, results for the calculus courses were similar to the results for the algebra courses with essentially no differences in Sample C; the highest correlations were associated with NM1 in both samples. A study has been proposed to test whether the spread-out format and encouragement of scratch work of the kind found in NM1 could contribute to a validity difference when the actual item content is held constant. (In the current study CM2 and NM1 had the same types of items, but not identical items.)

NM2 (and its parallel version NM4) were composed of 12 four-choice algebra placement items and 10 grid-in items. Validities for these item types in algebra and calculus courses are presented in Table A-3. Because these test sections were so short, direct comparisons with the validities in Table A-2 are unwarranted. However, the reliability of the grid-in items was higher than might be expected for a ten-item test because of the elimination of random guessing. The reliability of the grid-in items was estimated at .79 compared to .85 for the 32 items in NM1 and .73 for the 12 algebra achievement items. Corrected for unreliability, the grid-in correlations in Sample C go up to .26 for the algebra courses and .33 for the calculus courses, which may be compared to corrected correlations of .38 and .39 for NM1.[4] Thus, the grid-in items appear to be neither substantially better nor substantially worse predictors than the current item types. However, the low correlations for the small set of courses in Sample A suggest that further study of the validity of these items is warranted. Future studies should also unconfound item format and item difficulty; in the current study the grid-in items were all of above average difficulty.

Consistent with previous studies (e.g., Bridgeman & Wendler, 1989), the algebra placement items were relatively good predictors of performance in freshman mathematics courses. In both samples, the correlations based on just 12 four-choice items were as high as for the entire 60 items on the current SAT-M. They were as good at predicting calculus grades as they were at predicting algebra grades. Thus, these results are consistent with the notion that a specialized placement test (perhaps as part of an expanded offering of mathematics placement tests in SAT-II: Subject tests) is preferable to a mathematical reasoning test for placing students into mathematics courses.

Attempts to assess gain over the semester were abandoned because of the same concerns about the validity of the end-of-semester scores that were noted in the English placement analyses. Although gains were positive for all tests, the only confidence interval that did not include 0 was for NM1 where the mean gain was slightly less than one raw score point.

---

[4]Reliability was estimated across courses while validities were estimated within course and then averaged across courses. Within-course reliabilities would probably be lower, and corrected validities are therefore probably underestimates. Nevertheless, the relative ordering of the validities should be correct.

Table A-3

Correlation of Algebra Achievement and Grid-in Scores with grades

Sample	Course	N	r Course grade with Algebra Achievement	r Course grade with Grid-in
C	Algebra	388	.37	.23
A	Algebra	401	.28	.04
C	Calculus	182	.35	.29
A	Calculus	318	.29	.17

Appendix B

Colleges in Sample A

College	State
Augustana College	IL
Babson College	MA
Bridgewater College	MA
Lock Haven University	PA
Manchester College	IN
Mansfield University	PA
Merrimack College	MA
North Georgia College	GA
Roanoke College	VA
San Jose State University	CA
Villanova University (2)	PA
Williams College	MA
Whittier College	CA

Note.--Number in parentheses is the number of different courses sampled

Colleges in Sample B

College	State
University of Hartford	CN
Indiana State University (2)	IN
Michigan State University (3)	MI
Saint Olaf College	MN

Note.--Number in parentheses is the number of different courses sampled

Appendix C

Verbal and Writing Item Types

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 ability to read and understand a text passage. The form of the current SAT used in the present study included passages ranging from 200 to 450 words, which is typical of the SAT in its present form. Following each passage, there are several questions. For each question the examinee must choose the best answer from among five options.

The NPP 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 they are considerably longer (ranging up to 800 words) and are designed to be more accessible and engaging to the reader. 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. 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.

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 of the sentence 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 presented and followed by several questions. Some questions ask the examinees to rephrase 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, form 1, examinees were given the following assignment:

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.

The form 2 essay topic (administered to half of the students in Sample B) was as follows:

(First, the student was asked to read carefully a 187-word letter written by Albert Einstein.)

Write a well-organized essay that relates to the passage in the following ways:

First, explain the main points in Einstein's letter that lead him to the conclusion that "it is better to be an appreciative spectator than a floodlit actor."

Then, to illustrate how you may agree or disagree with Einstein's statement, choose a specific individual who has experienced local or widespread fame. This individual can be you, someone you know, or someone you have read about. Discuss in detail how being "a floodlit actor" has affected the individual you describe.