DOCUMENT RESUME

ED 381 554 TM 022 889

AUTHOR	Bridgeman, Brent; Morgan, Rick
TITLE	Relationships between Differential Performance on Multiple-Choice and Essay Sections of Selected AP Exams and Measures of Performance in High School and College. College Board Report No. 94-5.
INSTITUTION	College Board, New York, NY.; Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-94-41
PUB DATE	94
NOTE	16p.
AVAILABLE FROM	College Board Publications, Box 886, New York, NY 10101-0886 ($15).
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Academic Achievement; Advanced Placement; Advanced Placement Programs; Biology; College Entrance Examinations; College Students; English; *Essay Tests; High Achievement; Higher Education; High Schools; *High School Students; History; *Low Achievement; *Multiple Choice Tests; *Performance; Racial Differences; Sex Differences
IDENTIFIERS	*Advanced Placement Examinations (CEEB); *College Entrance Examination Board

ABSTRACT

Students with high scores (top third) on the essay portion of an Advanced Placement Examination (AP) (College Board) and low scores (bottom third) on the multiple-choice portion of the same examination were compared with students whose performance showed the opposite pattern. Across examinations in different subject areas (history, English, and biology) students who were relatively strong in the essay format and weak in the multiple-choice format were about as successful in their college courses as students who showed the opposite pattern, especially in courses where grades are not typically determined by multiple choice tests. Across several ethnic/racial groups, males tended to receive relatively high scores on the multiple-choice portion of the AP United States History Examination while females received higher scores on the essays than the multiple-choice questions. Because the population of students who take the AP Examinations is exceptionally able, generalizations to less able students are not warranted. Nine tables present study data. (Contains 14 references.) (Author/SLD)

College Board Report No. 94-5

Relationships Between Differential Performance on Multiple-Choice and Essay Sections of Selected AP® Exams and Measures of Performance in High School and College

BRENT BRIDGEMAN and RICK MORGAN

The College Board
Educational Excellence for *All* Students

College Board Report No. 94-5
ETS RR No. 94-41

Relationships Between Differential Performance on Multiple-Choice and Essay Sections of Selected AP® Exams and Measures of Performance in High School and College

BRENT BRIDGEMAN and RICK MORGAN

College Entrance Examination Board, New York, 1994

Brent Bridgeman is a senior research scientist at ETS.
Rick Morgan is a senior measurement statistician at ETS.

The College Board is a national nonprofit association that champions educational excellence for all students through the ongoing collaboration of more than 2,900 member schools, colleges, universities, education systems, and organizations. The Board promotes—by means of responsive forums, research, programs, and policy development—universal access to high standards of learning, equity of opportunity, and sufficient financial support so that every student is prepared for success in college and work.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101-0886. The price is $15.

Printed in the United States of America.

Contents

Abstract

Students with high scores (top third) on the essay portion of an Advanced Placement Examination and low scores (bottom third) on the multiple-choice portion of the same examination were compared with students whose performance showed the opposite pattern (top third on the multiple-choice questions and bottom third on the essay questions). Across examinations in different subject areas (history, English, and biology), students who were relatively strong in the essay format and weak in the multiple-choice format were about as successful in their college courses as students whose performance showed the opposite pattern, especially in those courses where grades are typically not determined by multiple-choice tests. Students who scored high on the multiple-choice portion and low on the essay portion performed relatively well on other multiple-choice tests, especially the verbal section of the SAT. Across several ethnic/racial groups, males tended to receive relatively high scores on the multiple-choice portion of the AP United States History Examination while females received higher scores on the essays than on the multiple-choice questions. Among females whose best language was not English, scores were substantially higher on the essay portion of the history examination; among males in this group, scores were slightly higher for the multiple-choice questions. Because the population of students who take Advanced Placement Examinations is exceptionally able, generalizations to less able populations are not warranted.

Introduction

Essay examinations and multiple-choice tests are both used to assess mastery of academic courses. Each question format has unique advantages as well as limitations. Multiple-choice tests provide an inexpensive means of assessing understanding of facts and concepts across a broad range of topics while essays assess organizational and productive skills in a more limited content domain. Because of measurement error due to subjective scoring and to relatively narrow content coverage, essay tests may be less reliable than multiple-choice tests in the same general subject area. But if the kinds of productive skills that only essay tests can assess are considered central to the definition of competence in a particular subject area, essay scores may be more valid indicators of competence than the more reliable multiple-choice scores.

The Advanced Placement (AP) Program of the College Board provides an ideal testing ground for comparing performance on multiple-choice and essay tests. Every year thousands of high school students complete college-level courses and then take AP Examinations to demonstrate their mastery of the course content. The three-hour examinations typically contain both multiple-choice and free-response (including essay) sections. The score on the essay portion of the test is based on at least two essays and each essay is scored by a different reader. Readers are high school and college teachers who are content specialists in the particular examination that they are grading. Scores on the essay and multiple-choice sections are combined to form a grade on a 1 to 5 scale. These grades are the only scores reported to students or colleges.

Correlations between multiple-choice and essay scores on AP Examinations are typically moderately high (College Board 1988, 53). Most students who do well with one format also do well with the other. But there are exceptions. Some students appear to perform better on essay tests and less well on multiple-choice tests, or vice versa. Although strong performance in both question formats has been shown to be predictive of success in college courses (Bridgeman and Lewis 1994), it is unclear whether students who are relatively strong on essays and weak on multiple-choice questions are more likely to succeed academically than students whose performance reflects the reverse pattern. Understanding these relationships may be useful not only for designing better assessment instruments but also for making optimal placement decisions. Thus a major purpose of the current study was to determine whether students with relatively high multiple-choice scores and low essay scores on AP Examinations were generally more successful in other testing situations and in college courses than students exhibiting the opposite pattern.

In several different AP subject areas, essay assessments have produced smaller gender differences in scores than multiple-choice tests (Mazzeo, Schmitt, and Bleistein 1993). Evidence from studies of other large-scale assessments has confirmed these findings (Murphy 1982; Beil and Hay 1987; Bolger and Kellaghan 1990). Nevertheless, gender differences remained even after correcting for the differential reliability of the two types of question and after removing items from the multiple-choice test on which men did particularly well. These differences were especially striking on the AP U. S. History Examination, in which estimated true score means for males and females were essentially equivalent on the essays (a difference of less than .02 in standard deviation units), but the mean for males was more than .3 standard deviation units higher than the mean for females on the multiple-choice portion of the test. Breland (1991) examined construct-irrelevant factors such as

handwriting to explain the relatively high scores females receive on essay tests, but concluded that males and females were nearly equal in the actual historical knowledge demonstrated in their essays as evaluated by specific facts included and errors avoided. Furthermore, Bridgeman and Lewis (1994) demonstrated that the performance of males and females in college history courses was essentially equal despite the advantage males enjoyed on the multiple-choice AP questions. Although the reasons for these gender differences are not yet known, identification of similar effects among specific ethnic/racial groups or among students whose best language is not English may provide some clues. Therefore another purpose of the current study was to determine whether examinees from such groups perform relatively better on questions in a multiple-choice or in an essay format.

Method

Data Sources

Three data files were used. One file was the same as that previously used by Bridgeman and Lewis (1994). The 38 colleges in this data base included both public and private institutions that use the SAT as part of the admission process. The file contained scores from selected AP examinations, SAT scores, and scores on the Test of Standard Written English (TSWE). TSWE is a multiple-choice test on the conventions of grammar and usage in written English. In addition, this file contained responses to the Student Descriptive Questionnaire (SDQ), which is completed when students register to take the SAT (typically near the end of the junior year or during the first few months of the senior year in high school) and asks for self-reported high school grade-point average (HSGPA) as well as grade-point averages in selected subject areas. Finally, this file included grades earned in college courses for students who had entered in the fall of 1985. The data base contained grades in individual courses as well as summary averages that grouped course grades in related fields. For example, the history average represented the grade-point average of a student in all the history courses that student took during the freshman year. Some colleges provided grades on a 5-point, A to F scale while others used a 13-point scale that included plus and minus indicators for all grades except F's. All grades were recoded on a 13-point scale (F = 0, D- = .7, D = 1,..., A+ = 4.3).

The data base included AP Examination scores from 1984 and 1985. Although most AP Examinations are taken at the end of the senior year in high school (e.g., spring 1985 for students who began college in fall 1985), a notable exception is the AP U.S. History Examination, which is typically (but not exclusively) taken at the end of the junior year, because most students take U.S. history as an eleventh-grade course. Of the 53,859 students in the original data base, AP scores were located for 7,626 students (about 14 percent). Of these 7,626 students, 6,243 had taken one AP Examination, 1,237 had taken two AP Examinations, and the remaining 146 had taken three or more AP Examinations.

The second data file contained SAT scores and grades in specific freshman courses from a campus of the University of California. Students in this file, who began college in 1989, were matched with AP files from 1987, 1988, and 1989. The analyses focused on the grades of these students in regular English courses who had taken the AP English Literature and Composition Examination.

The third data file was created by merging files from the Advanced Placement Program with files from the Admissions Testing Program, thus linking AP scores (multiple-choice and free-response), SAT scores (verbal and mathematical), scores on the English Composition Achievement Test (ECT), scores on the Test of Standard Written English (TSWE), and responses to the SDQ. The complete merged file contained large samples of students with AP scores, SAT scores, and SDQ scores (e.g., 58,596 AP U.S. History scores were matched to the SAT file), but it lacked the information on college grades available in the other files.

Descriptions of AP Examinations

The major focus of the study was on the AP Examinations in U.S. History and European History, with some consideration of the AP Examinations in Biology and in English Literature and Composition. These examinations were selected because they were taken by large numbers of students and showed relatively low correlations between the multiple-choice and essay sections.

AP U.S. History

Data from three different administrations of the U.S. History Examination were used (1984, 1985, and 1989). Prior to 1989 the examination was referred to as the AP American History Examination, but the format has remained consistent over the years.

The multiple-choice portion of this examination consisted of 100 items with five answer options for each item. Examinees were allowed 75 minutes to answer the questions. This section was formula scored (the score is the number of questions right minus one-quarter the

number wrong), with negative formula scores converted to 0. The free-response portion of the test consisted of two essays. In the first, examinees were provided with a set of documents and asked to construct an argument based on them. In order to receive an above-average score, candidates had to make reference to historical facts that were not directly discussed in the documents provided. In the second essay, examinees were asked to respond to one of five thematic history questions that were presented. An attempt was made to assess all five essay options on the same scale. Comparability of topics was monitored, but no statistical adjustments were made in the scores.

Each of the two essays was scored by a different reader using a 0 to 15 scale; thus, essay scores could range from 0 to 30. The composite AP score was arrived at by multiplying the multiple-choice formula score by .9, multiplying the essay score by 3, and summing the two weighted scores. Thus the two sections were given nominally equal weight in the composite score (each could contribute a maximum of 90 points). But because the standard deviation of the multiple-choice section was slightly larger (14.8 versus 12.7 in 1984), the multiple-choice section actually had slightly greater weight in the determination of the composite score. The chief reader and ETS professional staff then transformed the composite score into the 1 to 5 grading scale that was reported to colleges. This transformation was based to a large extent on an equating of the multiple-choice scores on a given examination form with the multiple-choice scores on an earlier form through a set of items common to both.

Reliability of the multiple-choice scores, as estimated by KR-20, was .90 in 1984 (Eignor, Flesher, and McClean 1984), .89 in 1985 (Livingston, McClean, and Flesher 1985), and .90 in 1989 (Bleistein, Damiano, and Flesher 1989). The coefficient alpha reliability of the essay scores was based on the correlation between the data-based essay question and the essay selected from five choices. These two types of essays were probably not essentially tau equivalent, so coefficient alpha was likely to underestimate the parallel form reliability. Because the two essays were read by different readers, this estimate included both differences among readers and differences among topics as sources of unreliability. The alpha reliability was .54 in 1984 and 1985, and .50 in 1989; reader reliability alone was about .79. Correlation between the multiple-choice and essay sections was .48 in 1984, .53 in 1985, and .51 in 1989.

AP European History

The general format and scoring rules for this examination were nearly identical to the AP U.S. History Examination except that candidates were not expected to use outside knowledge in answering the document-based essay question. The KR-20 reliability of the multiple-choice score was .91 and the coefficient alpha reliability of the essay score was .44. The correlation between the two sections was .50 (Mazzeo and Flesher 1985a).

AP Biology

In 1985, the 90-minute, multiple-choice portion of this examination consisted of 120 five-option items that were formula scored. Three topics were assessed with 40 items on each topic: (A) Cellular and Molecular, (B) Organismal, and (C) Populational. On the 75-minute essay section there were three pairs of questions, one pair on each of the above topics. The candidate was instructed to choose one question from each pair. As with the history essays, an effort was made to use a common scoring scale, but no statistical adjustments were made. Each of the three essays was graded on a 0 to 15 scale. Multiple-choice scores were multiplied by .625 and essay scores were multiplied by 1.667 so that the two portions of the examination made nominally equal contributions to the total possible score of 150. Reliability of the multiple-choice section was .93, while the coefficient alpha reliability of the essay section was .66 (Mazzeo and Flesher 1985b). Reader reliability alone was about .85. The correlation of essay and multiple-choice scores was .73.

AP English Literature and Composition

The 60-minute, multiple-choice section of this examination consisted of 52 five-option items that were formula scored. The 120-minute essay section consisted of three essays, each graded by a different reader on a 9-point scale. Reliability estimates were .85 for the multiple-choice items and .58 for the essay section (Chiu, Maneckshana, and Flesher 1989). The correlation between the multiple-choice and essay sections was .49.

Analyses of Files with Course Grades

For all students in the 38-college sample with scores on the 1984 AP American History Examination, the essay scores and the multiple-choice scores were arranged in order from high to low, separately for each college. The high essay/low multiple-choice group included those students who scored in the top one-third on the essay section and the bottom one-third on the multiple-choice section. Similarly, the high multiple-choice/low essay group included those students who scored in the top one-third on the multiple-choice section and the bottom

one-third on the essay section.[1]

Because the essay scores contain more measurement error than the multiple-choice scores, the group definitions are not as symmetrical as they appear to be. If scores with no measurement error were available, the students in the top third of the multiple-choice score distribution would generally be those in the top third of the observed score distribution. However, the composition of the top third group for the essays would change substantially. The procedure adopted in this study makes sense as a means of contrasting a group of students that is relatively strong on essays with a group that is relatively strong on multiple-choice items, but it would be incorrect to imply that students in the high essay group are exactly as extreme on essay performance as students in the high multiple-choice group are extreme on multiple-choice performance.

Within each college, the difference in the overall freshman grade-point average (FGPA) between the high essay and high multiple-choice groups was determined and weighted by the number of students in the combined groups. The weighted average of these FGPA differences across colleges was computed. This procedure was repeated for three more specific grade-point averages (social sciences/humanities, English, and history), and for the following four additional scores: HSGPA, SAT-Verbal (SAT-V), SAT-Mathematical (SAT-M), and TSWE. The entire procedure was repeated for AP scores on each of the following AP Examinations: 1985 American History, European History, and Biology. For AP Biology, a math/science grade-point average was used instead of the history grade-point average. A combined history high essay/low multiple-choice group was created including all students in the high essay/low multiple-choice group for whom data were available in the 1984 AP American History, or 1985 AP American History, or 1985 AP European History Examination file; a combined history high multiple-choice/low essay group was created in the same manner. For comparison, two additional groups were created including students who scored (1) in the top third on both the essay and multiple-choice sections (high on both) and (2) in the bottom third on both (low on both). To permit analysis of gender differences, all the above groups were broken down by gender, except for students taking the AP Biology Examination, where small sample sizes prohibited meaningful within-gender analyses.

Analyses of the University of California campus file used these same procedures for identifying high- and low-scoring groups among students enrolled in the regular freshman English course. Because students with AP grades of 4 or 5 on the AP English Literature and Composition Examination could be exempted from this course, the groups included primarily students with AP grades from 1 to 3. The large number of students in this course who had taken the AP Examination (694) permitted additional cross-tabulations of grades by high essay and high multiple-choice groups.

Analyses of Files without Course Grades

Once again, high essay and high multiple-choice groups were created including students scoring in the top third on the essays and in the bottom third on the multiple-choice items, and vice versa. Means on a number of variables were compared with the performance of these two groups on the AP U.S. History Examination and the AP English Literature and Composition Examination.

In order to estimate the relative strength of the performance of ethnic/racial and gender groups in the two question formats, analyses were run that included all the students who had taken the examination, not just those in the top and bottom third groups. Standard scores (mean of 0 and standard deviation of 1) were generated separately for the essay and multiple-choice scores on the AP U.S. History Examination. The low reliability of the essay scores compared to the multiple-choice scores would attenuate group differences more on the essays. Thus, a particular ethnic/racial or gender group might appear to score further below average on the multiple-choice questions than on the essays only because the essay scores are less reliable. If the reliability of the essay scores could be increased (perhaps by including more essays on the test), the pattern of relative strengths could be reversed. Because the mean true score for any large subpopulation is equal to the mean observed score for that subpopulation, the subgroup standard score means may be interpreted in terms of the standard deviation of the true scores (i.e., the expected distribution of the test scores if there were no errors of measurement). The standard deviation of the true scores is equivalent to the square root of the reliability (when the observed scores are in standardized form); this follows from the definition of reliability as the ratio of true variance to observed variance.[2] Therefore, means for the various subgroups in true score standard deviation

[1] For ease of data presentation, these groups are referred to as the "high essay" group and the "high multiple-choice" group.

[2] $r_{xx} = s_T^2 / s_x^2$, with standard scores $s_x^2 = 1$, so $r_{xx} = s_T^2$ and $\sqrt{r_{xx}} = s_T$.

TABLE 1

Relationship of GPAs to Performance on Combined AP History Examinations

Score	Group	Weighted Difference	Standard Error	High Essay N	High Essay M	High Essay S.D.	High Multiple-Choice N	High Multiple-Choice M	High Multiple-Choice S.D.	Both High N	Both High M	Both High S.D.	Both Low N	Both Low M	Both Low S.D.
History GPA	Total	–0.01	0.07	117	2.94	0.67	136	2.92	0.65	365	3.28	0.58	286	2.71	0.70
	Males	0.02	0.10	62	2.97	0.67	85	2.90	0.71	202	3.25	0.61	155	2.71	0.62
	Females	0.06	0.09	49	2.86	0.76	55	3 09	0.50	154	3.28	0.64	131	2.66	0.83
Freshman GPA	Total	0.12	0.04	336	2.89	0.57	351	3.00	0.63	896	3.21	0.56	857	2.67	0.62
	Males	0.05	0.06	184	2.88	0.61	202	2.94	0.67	455	3.19	0.54	476	2.63	0.63
	Females	0.14	0.06	148	2.88	0.55	145	3.04	0.64	425	3.26	0.53	391	2.74	0.63
Social sciences/ Humanities GPA	Total	0.17	0.06	279	2.90	0.66	263	3.08	0.71	689	3.29	0.61	684	2.70	0.75
	Males	0.18	0.07	147	2.91	0.70	140	3.08	0.61	340	3.25	0.60	372	2.66	0.79
	Females	0.21	0.07	129	2.87	0.68	116	3.12	0.69	342	3.33	0.60	318	2.75	0.73
English GPA	Total	0.11	0.06	250	3.07	0.60	220	3.16	0.72	596	3.29	0.59	649	2.86	0.66
	Males	0.02	0.09	123	3.05	0.65	124	3.04	0.81	295	3.27	0.58	342	2.87	0.68
	Females	0.15	0.06	120	3.10	0.59	93	3.27	0.56	296	3.34	0.57	309	2.89	0.62

units were estimated by dividing the observed standard score means by the square root of the reliability for each question type

$$z_T = \frac{z_x}{s_T} = \frac{z_x}{\sqrt{r_{xx}}}$$

in the population of all test candidates, r_{xx} = .90 for the multiple-choice questions and .50 for the essays. As noted above, the reliability estimates were conservative, resulting in a slight overadjustment.

Results and Discussion

Results for 38-College Sample

Table 1 compares freshman grade-point averages in selected subject areas for four groups that performed differentially on the combined AP history examinations. Within each college, the mean grade of the high essay/low multiple-choice group was subtracted from the mean grade of the high multiple-choice/low essay group, so positive values of the weighted difference indicate higher grades in the high multiple-choice/low essay group. Note that because extreme groups in this sample were defined separately for men, women, and the total group, the sample size for the total is not equal to the sum of the sample sizes for men and women. Also note that the value in the "weighted difference" column is close, but not identical, to the difference between the "high essay" and "high multiple-choice" columns because the weighted average of differences is not identical to the difference of weighted averages when cell sizes vary (for example, when a college had more students in the high essay group than in the high multiple-choice group).[3] In some cases, the weighted difference may be positive even though the mean is slightly higher in the high essay group.

History grades were nearly identical for students in the high essay/low multiple-choice and high multiple-choice/low essay groups. Thus, students who scored high on the essay questions (and low on the multiple-choice questions) could expect to be as successful in their college history courses as students with the opposite pattern. Differences between groups were generally somewhat greater with respect to the other grade-point averages. The greatest differences (favoring students in the high multiple-choice group) appeared in social sciences/humanities grades, perhaps because multiple-choice tests frequently play a more important role in determining final grades in these courses. Ekstrom and Villegas (1994), in a sample of introductory courses at 14 colleges, found that multiple-choice tests were used in 57 percent of the psychology courses but in only 26 percent of the history courses and 16 percent of the English courses. Small differences, or differences favoring the high essay group, might then be expected in English courses where essay tests are relatively more im-

[3]Suppose average grades were much higher at College A than at College B. Further suppose that, within each college, grades in the high essay and high multiple-choice groups were identical, but College A had more students in the high essay group while College B had more students in the high multiple-choice group. Computing a weighted average across both colleges for the essay groups and the multiple-choice groups separately (i.e., the column average) shows a higher average for the high essay groups, but the weighted average of the difference column is zero:

	High Essay N	High Essay M	High Multiple-Choice N	High Multiple-Choice M	Difference N	Difference M
College A	10	3.0	5	3.0	15	0.0
College B	5	2.0	10	2.0	15	0.0
Weighted M	15	2.7	15	2.3	30	0.0

TABLE 2

Relationship of HSGPA and Test Scores to Performance on Combined AP History Examinations

Grade or Score	Group	Weighted Difference	Standard Error	High Essay			High Multiple-Choice			Both High			Both Low		
				N	M	S.D.	N	M	S.D.	N	M	S.D.	N	M	S.D.
HSGPA	Total	0.04	0.03	293	3.57	0.37	312	3.61	0.34	810	3.69	0.35	752	3.48	0.39
	Males	0.03	0.04	164	3.56	0.36	172	3.58	0.34	408	3.66	0.34	406	3.44	0.40
	Females	0.07	0.04	127	3.59	0.37	135	3.70	0.32	392	3.75	0.32	348	3.51	0.38
SAT-V	Total	60	5	336	559	63	351	618	70	898	636	63	859	532	72
	Males	52	6	184	575	63	202	631	70	456	638	63	477	538	70
	Females	50	7	148	548	62	145	608	62	425	630	69	394	528	74
SAT-M	Total	37	6	336	607	78	351	644	73	898	646	70	859	593	82
	Males	16	7	184	638	74	202	657	70	456	660	70	477	617	76
	Females	26	8	148	574	85	145	610	70	425	620	73	394	570	77
TSWE	Total	0.9	0.4	336	54	6	351	55	6	898	56	5	859	52	7
	Males	0.8	0.5	184	54	5	202	55	5	456	56	5	477	51	7
	Females	1.2	0.5	148	53	5	145	55	6	425	56	4	394	52	6

portant. Indeed, the difference in the English GPA for males was very small, although the difference for females was unexpectedly large. Nevertheless, differences for all the grade-point averages were quite small in absolute terms.

As shown in Table 2, differences between groups in HSGPA were also quite small, although this finding must be interpreted cautiously because HSGPA was uniformly high for this sample of students who had taken the AP examinations in history. Note that students who scored in the lowest third on both the essay and multiple-choice sections (both low) still had HSGPAs of 3.48, and the FGPA of this group (see Table 1) was 2.67. In marked contrast, the 60-point weighted difference on the SAT-V was more than 10 times the standard error, and almost one within-group standard deviation, compared to less than one-tenth of a standard deviation for history grades. Differences between groups on TSWE, a multiple-choice test of writing-related skills, were small, although they may have been affected by the ceiling on the test (the maximum possible score is 60). When the groups were broken down by gender, the findings essentially paralleled those for the total sample. Differences for groups as defined by scores on the AP Biology Examination are summarized in Table 3. Note that the N's were substantially smaller not only because fewer students took the AP Biology Examination than the combined history examinations, but also because the correlation between the essay and multiple-choice sections of the AP Biology Examination was considerably higher (.73 versus .48 to .53), resulting in substantially fewer students who scored high in one format and low in the other. Despite these differences, Table 3 presents the same message as Tables 1 and 2. Students in both the high essay and the high multiple-choice groups did equally well in college, although students in the high multiple-choice group received much higher SAT scores.

Results for University of California Campus Sample

Table 4 presents data on the 694 students who took the AP English Literature and Composition Examination and were enrolled in the regular freshman English course at a campus of the University of California. The results parallel those in the other samples with near equivalence in grades but substantial differences in SAT-V scores.

As shown in Table 5, not only the means but also the distribution of grades were equivalent in the high essay/low multiple-choice and high multiple-choice/low essay groups. Not surprisingly, there were over twice as many A-/A students in the both high group as in the both low group. Table 5 also shows data for a remainder group consisting of students who were not included in the four main groups. English grades for this group were indistinguishable from grades for the high essay/low multiple-choice and high multiple-choice/low essay groups. Thus students who were mid-level performers on both the essay and multiple-choice sections received about the same grades in regular freshman English as students who received mid-level AP scores by doing well in one format and poorly in the other.

Results for Sample with SAT and AP Scores Only

The relationship of test scores and grades to performance on the AP U.S. History Examination is presented in Table 6. Out of a total of 58,596 students in the file, 3,602 scored in the top third on the essays and the bottom third on the multiple-choice questions; 2,281

TABLE 3

Relationship of GPAs and Test Scores to Performance on AP Biology Examination

GPA or Score	Weighted Difference	Standard Error	High Essay N	High Essay M	High Essay S.D.	High Multiple-Choice N	High Multiple-Choice M	High Multiple-Choice S.D.	Both High N	Both High M	Both High S.D.	Both Low N	Both Low M	Both Low S.D.
Science/math GPA	0.09	0.14	40	2.59	0.90	39	2.68	0.81	249	3.03	0.76	242	2.29	0.81
Freshman GPA	0.08	0.07	48	2.80	0.43	43	2.84	0.48	274	3.17	0.53	275	2.62	0.56
Social sciences/ Humanities GPA	0.05	0.14	36	2.82	0.64	28	3.01	.069	206	3.26	0.56	230	2.57	0.72
English GPA	0.02	0.10	40	2.74	0.66	24	3.05	0.28	170	3.29	0.59	210	2.91	0.54
HSGPA	0.06	0.04	39	3.59	0.31	38	3.63	0.29	237	3.66	0.35	247	3.50	0.35
SAT-V	56	11	48	558	73	43	618	63	275	621	69	276	527	74
SAT-M	77	13	48	594	90	43	661	58	275	657	64	276	576	70
TSWE	3.0	0.8	48	53	6	43	56	5	275	55	6	276	51	7

TABLE 4

Relationship of GPAs and SAT-V Score to Performance on AP English Literature and Composition Examination

GPA or Score	Group	High Essay N	High Essay M	High Essay S.D.	High Multiple-Choice N	High Multiple-Choice M	High Multiple-Choice S.D.	Both High N	Both High M	Both High S.D.	Both Low N	Both Low M	Both Low S.D.
English GPA	Total	74	3.19	0.53	71	3.13	0.62	76	3.26	0.50	73	3.00	0.52
	Males	31	3.17	0.56	43	3.07	0.68	35	3.22	0.54	30	3.03	0.60
	Females	43	3.21	0.48	28	3.23	0.39	41	3.29	0.47	43	2.97	0.46
Freshman GPA	Total	74	3.06	0.48	71	3.12	0.51	76	3.08	0.50	73	2.78	0.51
	Males	31	3.10	0.46	43	3.07	0.57	35	3.06	0.53	30	2.87	0.59
	Females	43	3.03	0.48	28	3.20	0.39	41	3.10	0.48	43	2.71	0.45
SAT-V	Total	74	511	52	71	583	59	76	569	51	73	481	72
	Males	31	520	48	43	580	64	35	567	53	30	490	65
	Females	43	503	54	25	586	51	41	571	51	43	475	77

scored in the top third on the multiple-choice questions and the bottom third on the essays. Grades in college courses were not available for this sample; the grades in Table 6 are high school grades as reported by students on the SDQ. Because high school grades tend to be high for nearly all students who take AP Examinations, the differences in grades must be interpreted cautiously. Consistent with the findings in the other samples, very large differences were found for the SAT-V and substantial differences for other multiple-choice tests. High school grades, which are typically determined by a combination of multiple-choice tests, constructed-response tests, and other non-test indicators, were somewhat higher in the high multiple-choice group, although the difference for English grades was only .15 in pooled standard deviation units (*d*) as compared with 1.08 for the SAT-V. The only test score based exclusively on essay performance was the student's essay score on the AP English Literature and Composition Examination; this was also the only score for which performance was higher for the high essay group on the AP U.S. History Examination.

Table 7 is comparable to Table 6, except that the groups were drawn from the 73,270 students who took the AP English Literature and Composition Examination. The differences between test scores were even larger than those shown in Table 6, although differences in high school grades were smaller. The only score favoring the high essay group was the essay score on the AP U.S. History Examination.

Table 8 shows the percentages of students, by sex, ethnic/racial background, and best language in the high essay and high multiple-choice groups on the AP U.S. History Examination. A higher percentage of men was in the high multiple-choice group than in the high essay group; for women the opposite was true. The percentages of each ethnic/racial group in the high essay category were quite consistent, ranging from a low of 5.2

TABLE 5

Relationship of English Grades to Performance on AP English Language and Composition Examination

Group	English Grade B– or lower	English Grade B, B+	English Grade A–, A
High essay	17 (23)*	35 (47)	22 (30)
High multiple-choice	18 (25)	32 (45)	21 (30)
Both high	18 (24)	30 (39)	28 (37)
Both Low	29 (40)	32 (44)	12 (16)
Remainder	94 (24)	188 (47)	148 (30)

*Number in parentheses is percent of total group (row).

TABLE 6

Relationship of Test Scores and Grades to Performance on AP U.S. History Examination

Grades or Score	High Essay			High Multiple-Choice			
	N	M	S.D.	N	M	S.D.	d
SAT-V	3,062	510	74	2,281	591	77	1.08
SAT-M	3,062	574	92	2,281	628	92	0.59
TSWE	3,062	51	6.8	2,281	54	6.0	0.46
ECT	1,680	543	83	1,293	588	80	0.55
HSGPA	2,931	3.61	0.47	2,241	3.71	0.50	0.21
English grade	2,904	3.52	0.54	2,219	3.61	0.53	0.15
AP English Literature and Composition: Multiple-Choice	169	24.1	8.4	199	31.5	8.0	0.90
AP English Literature and Composition: Essay	169	15.5	3.1	199	14.7	3.3	-0.25

TABLE 7

Relationship of Test Scores and Grades to Performance on AP English Literature and Composition Examination

Grades or Score	High Essay			High Multiple-Choice			
	N	M	S.D.	N	M	S.D.	d
SAT-V	4,175	510	62	2,540	617	62	1.73
SAT-M	4,175	566	92	2,540	633	86	0.75
TSWE	4,175	51	6.0	2,540	56	4.2	0.93
ECT	2,285	541	68	1,411	614	69	1.07
HSGPA	4,013	3.70	0.45	2,448	3.79	0.48	0.09
English grade	3,952	3.67	0.48	2,460	3.68	0.50	0.02
History/social sciences grade	3,949	3.68	0.49	2,452	3.70	0.51	0.04
AP U.S. History: Multiple-Choice	201	47.7	14	186	60.8	14	0.93
AP U.S.: History Essay	201	13.6	3.8	186	13.1	3.4	-0.14

percent for white students to a high of 6.1 percent for American Indian and Latino American students. The percentages in the high multiple-choice group were somewhat more variable, ranging from 2.2 percent for African American students to 4.1 percent for white students. Students whose best language was not English were much more strongly represented in the high essay group than in the high multiple-choice group (7.3 percent versus 2.9 percent). Although students who are not native speakers of English might be expected to have difficulty expressing their thoughts in English on an essay examination, their strong representation in the high essay group may reflect the greater examinee control inherent in essay tests. Students can express themselves using familiar vocabulary and grammatical structures in an essay examination, whereas failing to understand the nuances of vocabulary and structure in a multiple-choice question may lead to an incorrect response.

Table 9 shows the standard score means and estimated true standard score means (multiplied by 100 to eliminate the need for decimal points) on the AP U.S. History Examination for males and females in six ethnic/racial groups and for students who reported that English was not their best language. The numbers in the table indicate how far a particular group is above or below the average for the entire sample. For example, essay scores for white males were .04 standard deviation units above average, and their multiple-choice scores were .21 standard deviation units above average. In terms of true score standard deviation units, white females scored .03 points below average on the essay questions and .17 points below average on the multiple-choice questions. A positive number in the far right column indicates that the group performed relatively better on the multiple-choice section than on the essay section, with corrections for differences in reliability.

For every group, females' essay scores were higher than their multiple-choice scores, and for every group except the small group of students of Puerto Rican background, males' multiple-choice scores were higher than their essay scores. Males whose best language was not English did only slightly better on the multiple-choice questions than on the essay questions; females in this group received much higher scores on the essay than on the multiple-choice questions. The results would be virtually the same for the unadjusted standard score means as for the true score means, except that African American males received almost the same un-

TABLE 8

Percentages of Students with Selected Background Characteristics in High Essay and High Multiple-Choice Groups on AP U.S. History Examination

Group	N	Percentage in High Essay/ Low Multiple-Choice Group	Percentage in High Multiple-Choice/Low Essay Group
Male	30,432	4.2	5.2
Female	28,164	6.4	2.7
White	43,658	5.2	4.1
African American	2,243	5.7	2.2
American Indian	197	6.1	3.0
Asian American	6,500	5.8	3.9
Latino American	2,172	6.1	3.1
English not best language	756	7.3	2.9
Total	58,596	5.3	4.0

TABLE 9

Standard Score Differences between Essay and Multiple-Choice Scores on AP U.S. History Examination

Group	N	Standard Scores		True Standard Scores		Difference Between Essay and Multiple-Choice True Scores
		Essay	Multiple-Choice	Essay	Multiple-Choice	
White						
Male	22,888	4	21	6	22	16
Female	20,770	−2	−16	−3	−17	−14
African American						
Male	830	−40	−43	−57	−45	12
Female	1,413	−48	−80	−68	−84	−16
Asian American						
Male	3,394	11	20	15	21	6
Female	3,106	5	−14	7	−14	−21
Mexican American						
Male	479	−30	−22	−42	−23	19
Female	385	−38	−58	−53	−61	−8
Puerto Rican						
Male	124	−6	−25	−8	−26	−18
Female	121	−46	−68	−65	−72	−7
Other Latinos						
Male	550	−5	2	−7	2	9
Female	504	−25	−55	−36	−58	−22
English not best language						
Male	436	−5	−2	−8	−2	6
Female	320	−19	−51	−26	−53	−27

Note: Standard scores are *z*-scores multiplied by 100 to eliminate decimals. True standard scores were estimated by dividing the standard score representing each group mean by the square root of the reliability. (The reliability of the essay score was .5 and the reliability of the multiple-choice score was .9.)

adjusted standard score on the essay as on the multiple-choice questions. Clearly, generalizations about the relative performance of different ethnic/racial groups on essay and multiple-choice examinations could be distorted unless gender within ethnic group is considered, especially if one gender is overrepresented in a particular ethnic group (as African American females were on the AP U.S. History Examination). Ignoring gender, one might conclude that African American students score relatively higher on essay examinations, but the within-gender analyses make it clear that this is true only for females.

Conclusions

Success in college requires a number of distinct skills, some of which may be best assessed with essay tests while others may be best assessed with multiple-choice tests. This study found that students whose scores on selected AP Examinations were relatively high on essay and relatively low on multiple-choice questions were about as successful in their college courses as students with the opposite pattern, especially in those courses where grades were not determined by multiple-choice tests. Students who performed relatively weakly on the multiple-choice portion of an AP Examination were also relatively weak on the other multiple-choice tests considered. Thus the findings here are consistent with the correlation-based conclusions of Bridgeman and Lewis (1994), indicating the roughly equal effectiveness of essay and multiple-choice tests in predicting course grades, and the superiority of multiple-choice scores for predicting success on other multiple-choice tests.

For the AP Examinations studied, students with mid-level scores resulting from excellent performance on essay questions and poor performance on multiple-choice questions can be expected to perform about as well in college courses as students whose mid-level performance resulted from the opposite pattern of strength and weakness or from average performance on both parts of the examination. Because these conclusions are based on averages over courses with differing writing demands, they do not preclude the possibility that within certain writing-intensive courses students in the high essay group may be at a slight advantage, while in courses that are assessed primarily with multiple-choice tests, students in the high multiple-choice group might have an advantage.

The finding of smaller gender differences for the essay section than for the multiple-choice section of the AP U.S. History Examination is consistent with previous results (Mazzeo, Schmitt, and Bleistein 1993; Bridgeman and Lewis 1994). In addition, this analysis makes explicit the relationship of gender within

ethnic/racial group to performance on both types of question. This relationship was demonstrated by expressing the mean of each group as a deviation from the overall mean in both the observed score and true score metrics. Within each ethnic/racial group, and even in the group whose best language was not English, females scored relatively higher on the essay questions than on the multiple-choice questions. And the true scores for males in each group, except the Puerto Rican group, were higher on multiple-choice questions.

Although the results were quite consistent across the AP Examinations studied, generalizations to other examinations and populations can be made only after further research is conducted. In particular, the current results may be limited by the relatively high competence of AP students compared to college freshmen in general. An AP student in the low essay group in this study probably has writing skills that are well above average. The academic performance of students with poor writing skills may be considerably lower than the performance of AP students whose writing skills are low relative only to other AP students. Similarly, the academic backgrounds of students in various ethnic/racial groups (and in the group whose best language was not English) who choose to take particular AP courses may differ significantly from the backgrounds of students in these groups in the population as a whole.

References

Bell, R.C., and J.A. Hay. 1987. "Differences and Biases in English Language Examination Formats." *British Journal of Educational Psychology* 57:212–20.

Bleistein, C.G., M.D. Damiano, and R.B. Flesher. 1989. *Test Analysis: College Board Advanced Placement Examination, United States History.* Princeton, N.J.: Educational Testing Service. Unpublished Statistical Report No. SR-89-129.

Bolger, N., and T. Kellaghan. 1990. "Method of Measurement and Gender Differences in Scholastic Achievement." *Journal of Educational Measurement* 27:165–74.

Breland, H.M. 1991. *A Study of Gender and Performance on Advanced Placement History Examinations.* College Board Report No. 91-4; ETS Research Report 91-61. New York: College Entrance Examination Board.

Bridgeman, B., and C. Lewis. 1994. "The Relationship of Essay and Multiple-Choice Scores with Grades in College Courses." *Journal of Educational Measurement* 31:37–50.

Chiu, K., B. Maneckshana, and R.B. Flesher. 1989. *Test Analysis: College Board Advanced Placement Examination, English Literature and Composition.* Princeton, N.J.: Educational Testing Service. Unpublished Statistical Report No. SR-89-118.

College Board. 1988. *Technical Manual for the Advanced Placement Program.* New York: College Entrance Examination Board.

Eignor, D.R., R.B. Flesher, and D. McClean. 1984. *Test Analysis: College Board Advanced Placement Examination, American History.* Princeton, N.J.: Educational Testing Service. Unpublished Statistical Report No. SR-84-104.

Ekstrom, R.B., and A.M. Villegas. 1994. *College Grades: An Exploratory Study of Policies and Practices.* Princeton, N.J.: Educational Testing Service.

Livingston, S., D. McClean, and R.B. Flesher. 1985. *Test Analysis: College Board Advanced Placement Examination, American History.* Princeton, N.J.: Educational Testing Service. Unpublished Statistical Report No. SR-85-165.

Mazzeo, J., and R.B. Flesher. 1985a. *Test Analysis: College Board Advanced Placement Examination, European History.* Princeton, N.J.: Educational Testing Service. Unpublished Statistical Report No. SR-85-182.

Mazzeo, J., and R.B. Flesher. 1985b. *Test Analysis: College Board Advanced Placement Examination, Biology.* Princeton, N.J.: Educational Testing Service. Unpublished Statistical Report No. SR-85-143.

Mazzeo, J., A.P. Schmitt, and C.A. Bleistein. 1993. *Sex-Related Performance Differences on Constructed-Response and Multiple-Choice Sections of Advanced Placement Examinations.* College Board Report No. 92-7. New York: College Entrance Examination Board.

Murphy, R.J.L. 1982. "Sex Differences in Objective Test Performance." *British Journal of Educational Psychology* 52:213–19.