ED 385 543 TM 023 959

| | |
|---|---|
| AUTHOR | Mazzeo, John; And Others |
| TITLE | Sex-Related Performance Differences on Constructed-Response and Multiple-Choice Sections of Advanced Placement Examinations. College Board Report No. 92-7. |
| INSTITUTION | College Entrance Examination Board, New York, N.Y.; Educational Testing Service, Princeton, N.J. |
| REPORT NO | ETS-RR-93-5 |
| PUB DATE | 93 |
| NOTE | 37p. |
| AVAILABLE FROM | College Board Publications, Box 886, New York, NY 10101-0886 ($12). |
| PUB TYPE | Reports - Research/Technical (143) |
| EDRS PRICE | MF01/PC02 Plus Postage. |
| DESCRIPTORS | Advanced Placement; College Entrance Examinations; *Constructed Response; Ethnic Groups; High Schools; *High School Students; *Multiple Choice Tests; Racial Differences; *Scores; *Sex Differences; Test Items; Test Reliability; Test Results |
| IDENTIFIERS | *Advanced Placement Examinations (CEEB) |

ABSTRACT

This report describes three exploratory studies of the performance of males and females on the multiple-choice and constructed-response sections of four Advanced Placement Examinations: United States History, Biology, Chemistry, and English Language and Composition. Analyses were ca:'ied out for each racial or ethnic group with a sample size of at least 200. Gender differences associated with differences in the score reliabilities of the two types of assessment were studied. Analyses were also conducted to assess the extent to which sex-related differences in multiple-choice scores could be attributed to differentially functioning items favoring males. Exploratory analyses were also undertaken to determine whether patterns of sex-related items could be ob erved for constructed- response questions. There was little support for the "different-reliabilities" hypothesis, and fairly small numbers of multiple-choice items exhibited sex-related differential functioning. The third study suggested that topic variability may have a greater effect on sex-related differences than the variability associated with particular question types or content areas. Fourteen figures and 22 tables present analysis results. An appendix presents four tables of summary statistics. (Contains 28 references.) (Author/SLD)

College Board Report

No. 92-7

# Sex-Related Performance Differences on Constructed-Response and Multiple-Choice Sectiors of Advanced Placement Examinations

John Mazzeo
Alicia P. Schmitt
Carole A. Bleistein

# Sex-Related Performance Differences on Constructed-Response and Multiple-Choice Sections of Advanced Placement Examinations

John Mazzeo
Alicia P. Schmitt
Carole A. Bleistein

College Board Report No. 92-7
ETS RR No. 93-5

College Entrance Examination Board, New York, 1993

John Mazzeo is Analysis Director, NAEP Trial State Assessment, Division of Statistics and Psychometrics, at ETS.
Alicia P. Schmitt is a Senior Measurement Statistician in the College Board division at ETS.
Carole A. Bleistein is a Measurement Statistician in the College Board division at ETS.

## Acknowledgments

Funding for this research was provided by the Joint Staff Research and Development Committee of the College Board and Educational Testing Service (ETS) and ETS's Program Research Planning Council. The interpretations of results and conclusions are those of the authors and should not be viewed as statements of ETS or College Board positions.

The authors would like to thank Samuel Livingston and Neil Dorans for their technical consultation and Brent Bridgeman of ETS and Joseph J. Stevens of the University of New Mexico for their reviews of an earlier draft. Any remaining errors in analysis or interpretation exist despite their efforts. The authors would also like to thank Judith Alfort, Ted Blew, Karen Damiano, Behroz Manackshana, Bernadette McIntosh, and Evelyn Weiss of ETS for their help with the production of the various plots and tables included in the report.

Researchers are encouraged to freely express their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board is a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,800 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the programs of the College Board and participate in the determination of its policies and activities.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101-0886. The price is $12.

## CONTENTS

Abstract ........ 1

Introduction ........ 1

Review of Relevant Research ........ 2
AP Examinations ........ 2
Other Tests ........ 4

Selection and Format of AP Examinations Studied ........ 5

Study 1: Mode-of-Assessment Analyses across Racial or Ethnic Groups ........ 6
Comparison of Standardized Differences for Each Mode of Assessment ........ 6
Analyses of Multiple-Choice Scores Conditioned on Constructed-Response Scores ........ 8

Study 2: Differential Item Functioning (DIF) Analyses ......,........ 15
Identifying DIF Items ........ 16
Impact of DIF on Sex-Related Differences in Multiple-Choice Scores ........ 20

Study 3: Variability in the Size of Sex-Related Performance Differences by Type of Constructed-Response Question ........ 22
Procedure ........ 23
Results and Discussion ........ 23

Summary and Conclusions ........ 26

References ........ 27

Appendix. Summary Statistics by Sex and Racial or Ethnic Group ........ 29

## Figures

1. Standardized sex-related performance differences on multiple-choice and constructed-response sections for large volume 1986 AP Examinations. ........ 3

2. Standardized sex-related performance differences on multiple-choice and constructed-response sections for 1987 AP Examinations. ........ 3

3. A. U.S. History multiple-choice means conditioned on constructed-response scores for white examinees. ........ 9

B. Average differences (male-female) in History multiple-choice means conditioned on constructed-response scores for white examinees. ........ 9

4. A. U.S. History multiple-choice means conditioned on constructed-response scores for Asian American examinees. ........ 10

B. Average differences (male-female) in History multiple-choice means conditioned on constructed-response scores for Asian American examinees. ........ 10

5. A. U S. History multiple-choice means conditioned on constructed-response scores for black examinees. .......... 10

B. Average differences (male-female) in History multiple-choice means conditioned on constructed-response scores for black examinees. .......... 10

6. A. U.S. History multiple-choice means conditioned on constructed-response scores for Me ican American examinees. .......... 11

B. Average differences (male-female) in History multiple-choice means conditioned on constructed-response scores for Mexican American examinees. .......... 11

7. A. U.S. History multiple-choice means conditioned on constructed-response scores for other Hispanic examinees. .......... 11

B. Average differences (male-female) in History multiple-choice means conditioned on constructed-response scores for other Hispanic examinees. .......... 11

8. A. Biology multiple-choice means conditioned on constructed-response scores for white examinees. .......... 12

B. Average differences (male-female) in Biology multiple-choice means conditioned on constructed-response scores for white examinees. .......... 12

9. A. Biology multiple-choice means conditioned on constructed-response scores for Asian American examinees. .......... 13

B. Average differences (male-female) in Biology multiple-choice means conditioned on constructed-response scores for Asian American examinees. .......... 13

10. A. Biology multiple-choice means conditioned on constructed-response scores for black examinees. .......... 13

B. Average differences (male-female) in Biology multiple-choice means conditioned on constructed-response scores for black examinees. .......... 13

11. A. Chemistry multiple-choice means conditioned on constructed-response scores for white examinees. .......... 14

B. Average differences (male-female) in Chemistry multiple-choice means conditioned on constructed-response scores for white examinees. .......... 14

12. A. Chemistry multiple-choice means conditioned on constructed-response scores for Asian American examinees. .......... 15

B. Average differences (male-female) in Chemistry multiple-choice means conditioned on constructed-response scores for Asian American examinees. .......... 15

13. A. English multiple-choice means conditioned on constructed-response scores for white examinees. .......... 16

B. Average differences (male-female) in English multiple-choice means conditioned on constructed-response scores for white examinees .......... 16

14. A. English multiple-choice means conditioned on constructed-response scores for Asian American examinees. .......... 16

B. Average differences (male-female) in English multiple-choice means conditioned on constructed-response scores for Asian American examinees. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

## Tables

1. Standardized Differences in Males' and Females' Scores on the English Placement Text . . . . . . . . 4

2. Standardized Differences in Males' and Females' Mean Scores on the ECT-E . . . . . . . . . . . . . . . . . 4

3. Observed- and True-Score Differences in Males' and Females' Means in Standard Deviation Units for the California Bar Exam . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

4. Standardized Differences by Racial or Ethnic Group for the U.S. History Exam . . . . . . . . . . . . . . . 7

5. Standardized Differences by Racial or Ethnic Group for the Biology Exam . . . . . . . . . . . . . . . . . . . . 7

6. Standardized Differences by Racial or Ethnic Group for the Chemistry Exam . . . . . . . . . . . . . . . . . 8

7. Standardized Differences by Racial or Ethnic Group for the English Language and Composition Exam . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

8. Sex-Related Differences in Predicted History Multiple-Choice Means Conditioned on Constructed-Response Scores for Mexican American Examinees . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

9. Sex-Related Differences in Predicted Biology Multiple-Choice Means Conditioned on Constructed-Response Scores for Asian American Examinees . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

10. Sex-Related Differences in Predicted Chemistry Multiple-Choice Means Conditioned on Constructed-Response Scores for White Examinees . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

11. DIF Classification Criteria . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17

12. Number of Items in Each DIF Category for the U.S. History Exam . . . . . . . . . . . . . . . . . . . . . . . . . 19

13. Number of Items in Each DIF Category for the Biology Exam . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

14. Number of Items in Each DIF Category for the Chemistry Exam . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

15. Number of Items in Each DIF Category for the English Language and Composition Exam . . . . . 19

16. Observed- and True-Score Standard Differences between Males and Females on U.S. History Constructed-Response and Multiple-Choice Test Components . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21

17. Observed- and True-Score Standard Differences between Males and Females on Biology Constructed-Response and Multiple-Choice Test Components . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21

18. Observed- and True-Score Standard Differences betwe Males and Females on Chemistry Constructed-Response and Multiple-Choice Test Components . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21

19. Standardized Sex-Related Differences by Constructed-Response Question Type for Four U.S. History Exam Forms . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23

20. Standardized Sex-Related Differences by Constructed-Response Question Type for Three Chemistry Exam Forms . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24

21. Standardized Sex-Related Differences by Constructed-Response Question Type for Three Biology Exam Forms ........ 25

22. Standardized Sex-Related Differences by Constructed-Response Question Type for Three English Language and Composition Exam Forms ........ 25

## ABSTRACT

A number of studies in which scores on multiple-choice and constructed-response tests have been analyzed in terms of the sex of the test takers have indicated that the test performance of females relative to that of males was better on constructed-response tests than on multiple-choice tests. This report describes three exploratory studies of the performance of males and females on the multiple-choice and constructed-response sections of four Advanced Placement (AP) Examinations: United States History, Biology, Chemistry, and English Language and Composition. The studies were intended to evaluate some possible reasons for the apparent relationship between test format and the magnitude of sex-related differences in performance.

For the first study, analyses were carried out to evaluate the extent to which such differences could be attributed to differences in the score reliabilities associated with these two modes of assessment. For the second study, analyses of the multiple-choice sections and follow-up descriptive analyses were conducted to assess the extent to which sex-related differences in multiple-choice scores could be attributed to the presence of differentially functioning items favoring males. For the third study, a set of exploratory analyses was undertaken to determine whether patterns of sex-related differences could be observed for different types of constructed-response questions.

The results of the first study provided little support for the "different-reliabilities" hypothesis. Across all exams and all ethnic groups, there were substantial differences between the scores of males and females even after taking into account differences in the reliabilities of the two sections. The results of the second study indicated that fairly small numbers of items exhibited substantial amounts of sex-related differential item functioning (DIF), and removing these items resulted in almost no reduction in the magnitude of sex-related differences on the multiple-choice sections. The results of the third study identified some consistent patterns across ethnic and racial groups regarding which questions females will perform best on, relative to males. However, taken as a whole, the results of the third study suggest that topic variability may have a greater effect than the variability associated with particular question types or broadly defined content areas.

## INTRODUCTION

Over the last two decades there has been a continuing interest in differences in cognitive abilities between the sexes (e.g., Maccoby and Jacklin 1974; Benbow 1988). In their review of the literature, Wilder and Powell (1989) report that although such differences on tests of verbal ability and achievement have diminished over the last two decades, there are still a number of fairly consistent and substantial differences favoring males on tests of quantitative ability and on tests of achievement in mathematics and science.

In a recent review article comparing multiple-choice and constructed-response tests, Traub and MacRury (1990) discuss three studies (Bolger 1984; Murphy 1980, 1982) that indicate the presence of a different kind of sex-related difference having to do with the mode of assessment. As will be illustrated below, results similar to those reported by Traub and MacRury have been found in other studies in which scores on multiple-choice and constructed-response tests have been analyzed in terms of the sex of the test takers. Despite differences across studies in test content, the test performance of females relative to that of males was better on constructed-response tests than on multiple-choice tests. The results in these studies followed one of three patterns:

1. The average performance of males was higher than that of females on both types of tests, but the difference was smaller on the constructed-response test.
2. The average performance of females was higher than that of males on both types of exams, but the difference was greater on the constructed-response test.
3. The average performance of males was higher than that of females on the multiple-choice section, but the reverse was the case on the constructed-response section.

The consistency of the results across these studies is surprising when one considers that the term "constructed response," as typically used, encompasses a wide range of formats from simple fill-in items to complex performance assessments (Bennett in press).

Any explanation for the occurrence of such results will most likely be complex and multifaceted. There are a number of hypotheses that, alone or in combination, might explain why the performance of females, relative to that of males, is better on constructed-response tests. Perhaps the most obvious one might be termed the "different-skills" hypothesis. A number of educational researchers (e.g., Fredriksen and Collins 1989; Fredriksen 1984) have argued that the two modes of assessment demand different sets of academic and subject-matter competencies. The keen interest that currently exists in alternative measurement formats is based in large part on the belief that multiple-choice tests are ill-suited to the measurement of the higher level outcomes that are of interest to most educators (Stiggins 1991). Results from the College Board's AP Examinations are consistent with this view in that, for many subject areas, only moderate correlations between multiple-choice and constructed-response sections of the test are found (College Board 1988, 53). Sex-related differences in performance profiles across the two modes of assessment most likely reflect real disparities in the average level of achievement obtained by males and females with respect to these different competencies.

However, the two modes of assessment also have strengths and weaknesses as measurement instruments, and there are alternative explanations that attribute at least a portion of these sex-related differences in performance profiles to factors other than the academic and subject-matter competencies the exams are intended to measure. For example, one such alternative explanation for the first pattern of results described above might be termed the "different-reliabilities" hypothesis. Multiple-choice tests often produce higher reliability coefficients than constructed-response tests of approximately equivalent temporal length (see, for example, College Board 1988, 47). Lower constructed-response section reliabilities result at least partly from the fact that, in a given amount of time, one can ask a greater number of multiple-choice questions than constructed-response questions. In addition, constructed-response questions, particularly essays, must be scored by human readers, which introduces a degree of subjectivity (and thus unreliability) into the scoring process. The different-reliabilities hypothesis asserts that even if males and females differ to roughly the same degree with respect to the academic and subject-matter competencies assessed by multiple-choice and constructed-response tests, such differences will be reflected to a lesser degree in scores on constructed-response tests because of the attenuating influence of measurement error.

A second class of alternative explanations might be termed "method-bias" hypotheses. Multiple-choice and constructed-response test scores may be influenced by sources of variation that are related to sex but unrelated to the academic and subject-matter competencies that are the intended focus of the test. The reported patterns of sex-related differences on multiple-choice and constructed-response tests may be due to the operation of these "construct-irrelevant" factors. Some construct-irrelevant factors might be pervasive in nature, affecting all items in a multiple-choice format or all questions in a constructed-response format; others might affect only particular types of multiple-choice items or constructed-response questions.

As an example of the second type of factor, suppose that a relatively small number of multiple-choice items are affected by sources of variation unrelated to the academic and subject-matter competencies that are the intended focus of the test (e.g., familiarity with certain concepts or vocabulary associated with stereotypical male activities) and constructed-response test items are not affected (or are affected to a lesser degree) by these unrelated sources. If females are, on average, worse than males with respect to these construct-irrelevant factors, they will perform less well, relative to males, on multiple-choice measures of educational outcomes.

As another example, suppose that scores on essay questions in the social sciences (such as history or political science) are determined not only by the accuracy of historical (or political) facts and the sophistication of a student's argument (i.e., what might be termed "construct-relevant" factors), but also by the length of a student's response and the neatness of his or her handwriting (i.e., "construct-irrelevant" factors). If females are better, on average, than their male counterparts with respect to the construct-irrelevant factors, their performance relative to males may appear better on constructed-response tests than on multiple-choice tests.

This report describes a series of exploratory studies of the performance of males and females on the multiple-choice and constructed-response sections of several AP Examinations. Four examinations were selected for study: United States History, Biology, Chemistry, and English Language and Composition. Three separate studies were carried out. For the first study, a set of descriptive analyses was performed to evaluate the consistency of mode of assessment-related differences between the sexes across several ethnic and racial groups; two additional sets of analyses were carried out to evaluate the extent to which such differences could be attributed to differences in the score reliabilities associated with these two modes of assessment. For the second study, differential item functioning analyses of the multiple-choice sections and follow-up descriptive analyses were conducted to assess the extent to which sex-related differences in multiple-choice and constructed-response scores could be attributed to construct-irrelevant factors operating at the multiple-choice item level. For the third study, a set of exploratory analyses was undertaken to determine whether patterns of sex-related differences could be observed for different types of constructed-response questions.

The report is organized as follows. First, some background material is given on the nature and pervasiveness of male-female differences related to assessment mode on the 1986 and 1987 AP Examinations. This is followed by a brief review of selected research studies in which similar results were obtained. Second, the selection and format of the four AP Examinations used in the studies are described. Then the methods and results of each of the three studies are discussed. The final section of the report is a summary and discussion of the results.

## REVIEW OF RELEVANT RESEARCH

### AP™ Examinations

Each spring the College Board offers AP Examinations in 28 subject areas to high school students enrolled in corresponding AP courses. With two exceptions (the General and Drawing portfolio evaluations in Stu 'o Art), each examination consists of both a multiple-choice and a constructed-response section. The item types contained in the constructed-response sections include extended essay questions (in the English and History exams), word problems (in the science and mathematics exams), and performance tasks (in the foreign language exams).

Data from the 1986 administration of the AP Examinations were used to obtain separate summary statistics for males and females on the multiple-choice and constructed-response sections of 11 of the larger volume AP Examinations. These data are presented in Figure 1. Each exam is represented as a single point. The x-coordinate of each point (on the horizontal axis) corresponds to the standardized mean difference in scores between males and females on the multiple-choice portion of that exam. The y-coordinate of each point (on the vertical axis) corresponds to the standardized mean difference in scores on the constructed-response section. The standardized differences were calculated as follows:

(1) $$StD_m = \frac{\bar{X}_m - \bar{X}_f}{S_t}$$

where,

(2) $$S_t = \sqrt{\frac{(n_m - 1)S_m^2 + (n_f - 1)S_f^2}{n_m + n_f - 2}}$$

and where $\bar{X}$, $S^2$, and $n$ refer to the mean, variance, and sample size, respectively, of the group indicated by the subscript ($m$ = males, $f$ = females). Also shown in Figure 1 is the 45-degree line. Points that appear below this line indicate exams in which the average performance of females, relative to that of males, is better on the constructed-response portion of the exam (i.e., one of the three patterns described earlier holds when performance is expressed in terms of within-section standard-deviation units).

On average, males scored higher than females on the multiple-choice sections of all 11 exams displayed in Figure 1. The exams were in subject areas spanning the humanities (English Literature and Composition, English Language and Composition), the social sciences (United States History and European History), the natural sciences (Biology, Chemistry, Physics B, Physics C: Mechanics, and Physics C: Electricity and Magnetism), and mathematics (Calculus AB and Calculus BC). The mean standardized differences varied considerably from rather small differences (less than .1 for the English exams) to substantial ones (greater than .5 for one of the physics exams).

The average score of male examinees was also higher than that of females on the constructed-response sections for eight of these tests, the exceptions being English Language and Composition, English Literature and Composition, and European History However, for 10 of the 11 exams (Calculus AB being the one exception), the points fall to the right of the 45-degree line, indicating that females performed better, relative to males, on the constructed-response section than they did on the multiple-choice section. The differences in section performance are particularly noteworthy for two exams. The difference between males and females on the constructed-response portions of the United States History and European History exams was approximately zero. In contrast, on the multiple-choice portion of the same exam, the difference exceeded .35 for both exams.

Figure 2 contains a similar plot based on data from the 1987 administration for all AP Examinations offered at that time which included a constructed-response section. Again,

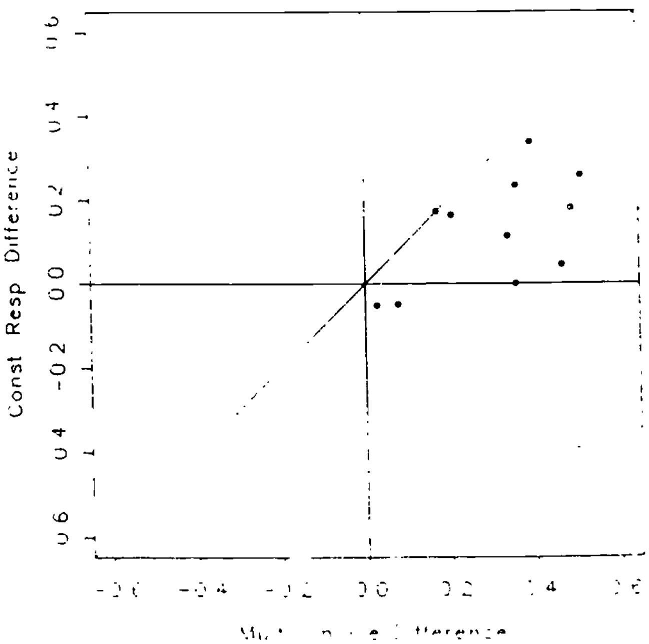

**Figure 1. Standardized sex-related performance differences on multiple-choice and constructed-response sections for large volume 1986 AP Examinations.**

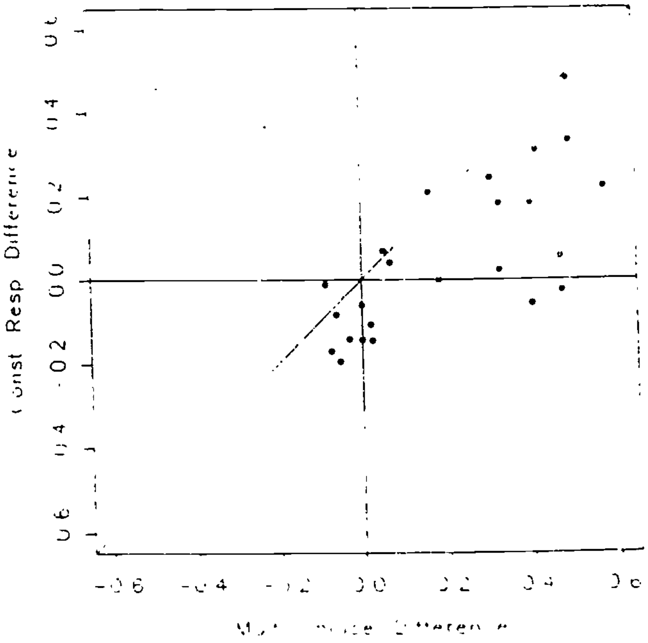

**Figure 2. Standardized sex-related performance differences on multiple-choice and constructed-response sections for 1987 AP Examinations.**

BEST COPY AVAILABLE

almost all the points (18 out of 24) in Figure 2 clearly fall to the right of the 45-degree line, indicating that females performed better relative to males on the constructed-response sections. Also, as in 1986, the sex-related differences on multiple-choice and constructed-response sections were particularly noteworthy for the United States History and European History exams. Discrepancies in sex-related differences on multiple-choice and constructed-response sections are also quite large for the exams in Government and Politics, Physics B, and English Language and Composition.

Thus, to a large extent, results from AP Examinations are consistent with those of the studies discussed in Traub and MacRury (1990), which indicate that the average test performance of females relative to that of males is often better on constructed-response tests than on multiple-choice tests. This phenomenon was particularly apparent on the AP Examinations in History and Government and Politics, though it was not restricted to these subject areas.

## Other Tests

In addition to the studies described in Traub and MacRury (1990), a small number of other studies have looked at differences in male and female performance on multiple-choice and constructed-response tests. Breland and Griswold (1981) compared the scores of males and females on several basic skills measures, including an English placement test that consisted of three multiple-choice sections (reading, sentence construction, and logic and organization) and one essay measure. Means and standard deviations on all four sections of the test were reported for the cohort of students who entered California state universities and colleges in 1977. Standardized differences based on these data were calculated according to equation (1) and are given in Table 1.

The standardized differences on the three multiple-choice sections were all fairly small (.05 or less) with males performing slightly better on the logic and organization and reading sections and females performing slightly better on the sentence construction section. In contrast, the standardized difference on the essay section was considerably larger. Though females did better than males, as was the case for the multiple-choice sentence construction section, the standardized difference on the essay was about eight times larger. It is important to note that the essay section is probably a less reliable measure than any of the multiple-choice sections. Other things being equal, one would expect the less reliable measure to produce smaller differences between the groups, not the larger differences observed in the data presented by Breland and Griswold.

**Table 1. Standardized Differences in Males' and Females' Scores on the English Placement Test**

| *Test Section* | *Difference* |
|---|---|
| Reading | .01 |
| Sentence construction | .05 |
| Logic and organization | .04 |
| Essay | -.39 |

Calculations are based on data from the Breland and Griswold (1981) study

Positive differences indicate test performance of males was higher than that of females

Petersen and Livingston (1982) presented the results of a study of the relationship between the essay and multiple-choice sections of the Admissions Testing Program's English Composition Test with Essay (ECT-E). The test consists of a 70-item multiple-choice section and a single 20-minute essay that is independently scored by two graders using a six-point scale. Petersen and Livingston reported summary statistics by sex for the ECT-E sections for all candidates taking the ECT-E in December 1977. Statistics were reported separately for white (24,574 males and 24,346 females), black (817 males and 1,320 females), Mexican American (350 males and 299 females), and Asian American (1,268 males and 1,058 females) examinees. Standardized differences for the multiple-choice and essay sections based on these summary statistics are given in Table 2.

With one exception, females outperformed males on both the multiple-choice and constructed-response portions of the test, and the size of the difference was similar for both sections. However, in all ethnic groups, the performance of females relative to the performance of males was better on the constructed-response section than on the multiple-choice section. As pointed out earlier, other things being equal, one would expect the less reliable essay measure to produce smaller differences between the groups rather than larger differences.

More recently, Klein (1989) has reported test score differences between males and females on the California Bar Exam. The exam consisted of three parts: (1) a 200-item multiple-choice test (the Multistate Bar Examination), (2) six one-hour essay questions, and (3) two three-hour written performance test problems. Klein determined standardized mean observed-score differences (differences between male and female means divided by the total-group standard deviation) and standardized mean true-score differences (differences between means divided by the square root of the

**Table 2. Standardized Differences in Males' and Females' Mean Scores on the ECT-E**

| *Test Section* | *Whites* | *Blacks* | *Mexican Americans* | *Asian Americans* |
|---|---|---|---|---|
| Multiple choice | -.12 | .07 | .07 | -.14 |
| Essay | .18 | -.18 | -.19 | .21 |

Calculations are based on data from the Petersen and Livingston (1982) study

Positive differences indicate test performance of males was higher than that of females

Table 3. Observed- and True-Score Differences in Males' and Females' Means in Standard Deviation Units for the California Bar Exam

|  | *Standardized Difference* |  |
|---|---|---|
| *Section* | *Observed Score* | *True Score* |
| Multiple choice | .19 | .20 |
| Essay | -.12 | .14 |
| Performance | -.21 | -.28 |

Calculations are from the Klein (1989) study, with signs reversed to be consistent with differences obtained in other studies.
Positive differences indicate test performance of males was higher than that of females.

total group reliability) for each of the three sections of the exam. These standardized differences (reversed in sign to be consistent with those reported in Tables 1 and 2 and Figures 1 and 2) are presented in Table 3. Males performed better than females on the multiple-choice section of the California Bar Exam. However, females performed better on both the essay test and the performance test. In fact, the largest difference between the sexes is in favor of females on the performance test.

In summary, studies discussed in Traub and MacRury (1990), data from the AP Examinations in 1986 and 1987, and studies by Breland and Griswold (1981), Petersen and Livingston (1982), and Klein (1989) all suggest that, for a variety of standardized tests, females perform better relative to males on constructed-response sections of these tests than they do on multiple-choice sections. In some cases, in particular on a large number of the AP Examinations, males scored better than females on both the multiple-choice and constructed-response sections of the exams, but the standardized differences were smaller on the constructed-response sections than on the multiple-choice sections. In other cases, males scored better than females on the multiple-choice sections and females scored better on the constructed-response sections. In the cases in which females scored higher on both the multiple-choice and constructed-response sections of the same exam, the standardized differences were larger on the constructed-response section than on the multiple-choice section.

## SELECTION AND FORMAT OF AP EXAMINATIONS STUDIED

Four AP Examinations were selected for detailed study: United States History, Biology, Chemistry, and English Language and Composition. These particular examinations were selected for a variety of reasons. First, in order to avoid confounding sex-related differences with possible racial or ethnic differences, it was our intention to study such differences within particular racial or ethnic subgroups. In addition, we wanted to base our analyses on fairly large sample sizes so that our results could be expected to be relatively stable and questions concerning their statistical significance would be relatively unimportant. Thus, the examinations selected had large enough candidate volumes in 1987 to allow analyses based on fairly large numbers of cases. For each exam, the groups studied were restricted to those for which there were at least 300 male and 300 female examinees.

Second, since the effect of the multiple-choice factors under study might be sufficient to account for small discrepancies in section performance (but not large discrepancies—or perhaps not all the large discrepancies), we wanted to study examinations that exhibited a range of discrepancies in male-female performance differences on multiple-choice and constructed-response sections. Thus, the examinations selected varied with respect to the degree to which sex-related differences were discrepant on the multiple choice and constructed-response sections.

Third, in order for our results to have some degree of generality, we wanted to include a variety of subject areas spanning the humanities, social sciences, and natural sciences. In addition, the particular examinations selected contain a range of constructed-response question types. Both the History and English exams contain holistically graded essay questions. The Biology exam contains four essay questions that are analytically scored. The Chemistry exam contains word problems (parts 1 and 2), short answers (part 3), and essays (part 4), all of which are analytically scored.

The United States History exam consists of a 100-item five-option multiple-choice section and a two-part constructed-response section. The multiple-choice section is formula scored (rights minus one-fourth wrong), but negative formula scores are converted to 0; therefore multiple-choice scores can range from 0 to 100. The first part of the constructed-response section consists of a single mandatory document-based question. Examinees are provided with sample historical documents and are asked to construct an argument based on these documents. The second part allows the examinees to choose one out of six possible standard thematic history questions. Each constructed-response part is independently scored by a single reader using a 0 to 15 scale; therefore total constructed-response section scores can range from 0 to 30. Estimates of the total-group reliabilities are .89 for the multiple-choice section and .53 for the constructed-response section (Morgan and Flesher 1987).

The Biology exam consists of a 120-item five-option multiple-choice section and a four-question constructed-response section. The multiple-choice section is scored in the same manner as the United States History exam, with scores ranging from 0 to 120. Each constructed-response question is independently scored by a single reader using a 0 to 10 scale; therefore total constructed-response section scores range from 0 to 40. Constructed-response questions cover topics in molecules and cells (question 1), genetics and evolution (question 2), and organisms and populations (questions 3 and 4). Estimated total-group reliabilities for the multiple-choice and constructed-response sections are .93 and .78, respectively (Mazzeo et al. 1987).

The English Language and Composition exam consists of a 60-item five-option multiple-choice section and a three-question constructed-response section. The multiple-choice section is formula scored as described above and results in scores ranging from 0 to 60. Each constructed-response question is independently graded by a single reader using a 0 to 9 scale; therefore total constructed-response section scores range from 0 to 27. The estimated total-group reliability of the multiple-choice section of the English exam is .87, and that for the constructed-response section is .55 (Morgan et al. 1987).

The Chemistry exam has an 80-item five-option multiple-choice section, which is formula scored in the manner described above, resulting in scores ranging from 0 to 80. There is also a four-part constructed-response section: Part one consists of a single mandatory word problem, part two requires examinees to answer one of two word problems, part three requires examinees to indicate the reactants and products for five out of eight chemical reactions, and part four requires examinees to answer three out of five essay questions. Each constructed-response part is independently scored by a single reader. Parts one and two are scored on a 0 to 9 scale, part three is scored on a 15-point scale, and each of the questions in part four is scored on a 0 to 8 scale. Total constructed-response section scores range from 0 to 57. Estimated reliabilities are .93 and .80 for the respective multiple-choice and constructed-response sections (Eignor and Bleistein 1987).

## STUDY 1: MODE-OF-ASSESSMENT ANALYSES ACROSS RACIAL OR ETHNIC GROUPS

Study 1 was directed toward two questions: (1) Is the pattern of sex-related differences on multiple-choice and constructed-response sections consistent across the larger ethnic or racial subgroups that take the AP Examinations? (2) Is the performance of females relative to males still better on the constructed-response sections after taking into account the differences in the reliabilities of the two modes of assessment?

The first question was answered by determining standardized differences on multiple-choice and constructed-response sections for each of the ethnic or racial groups for which large samples of examinees were available. The second question was addressed by two separate sets of analyses. The first analysis compared the obtained standardized differences for each mode of assessment to estimates of the standardized "true-score" differences. The magnitude of the obtained differences (i.e., observed-score differences) is in part a function of the reliabilities of the multiple-choice and constructed-response sections. Estimated true-score differences are observed-score differences that have been corrected for the effects of differing reliabilities. The second set of analyses examined the multiple-choice test performance of males and females conditioned on constructed-response scores. If the average multiple-choice performance of males and females differed for individuals of *equal* constructed-response section proficiency, then reliability differences would appear to be precluded as a plausible explanation for such differences.

Unfortunately, comparing the multiple-choice performances for individuals conditioned on their *observed* constructed-response scores is not the same as comparing individuals with the same degree of proficiency. The observed constructed-response score contains measurement error. As a result, males and females with the same *observed* constructed-response score do not generally have the same average *true score* and, hence, are not generally matched with respect to proficiency. For example, if the *marginal* constructed-response observed-score mean of males is higher than that of females, then constructed-response true-score means for males *conditional* on constructed-response scores will also be higher for males than for females. In other words, in this example, for a group of examinees with any given constructed-response score, the males are on average more proficient. Since multiple-choice and constructed-response scores on AP Examinations are positively correlated, one would expect males to exhibit higher multiple-choice scores.

In an attempt to identify potential artifacts in our results due to the above considerations, additional analyses were carried out that compared estimates of the linear regression of multiple-choice scores on constructed-response *true scores*. Although the regression of multiple-choice scores on constructed-response true scores is most likely nonlinear (due to the bounded nature of both variables), the results of these analyses were used in a heuristic fashion to provide an estimate of the extent to which matching on a "fallible" constructed-response criterion could explain differences in the multiple-choice performance of males and females conditioned on constructed-response observed scores.

## Comparison of Standardized Differences for Each Mode of Assessment

### Procedures

Means and standard deviations for the multiple-choice and constructed-response sections of each exam were obtained separately for males and females within each racial or ethnic group. These summary statistics were then used to calculate standardized differences between males and females for the multiple-choice and constructed-response portions of each exam. Standardized differences were obtained in both observed-score and true-score metrics. Standardized observed-score differences were calculated according to equation (1) above. Estimates of the standardized true-score difference for each section were calculated by

$$(3) \qquad StD_t = \frac{(\bar{X}_m - \bar{X}_f)}{\sqrt{\dfrac{(n_m - 1)\hat{\rho}_m S_m^2 + (n_f - 1)\hat{\rho}_f S_f^2}{n_m + n_f - 2}}}$$

where $\hat{\rho}_m$ and $\hat{\rho}_f$ are the estimated reliabilities for males and females, respectively.

Estimates of reliabilities for the multiple-choice and constructed-response sections of all four examinations were already available from the 1987 Test Analysis Reports (Morgan and Flesher [1987] for United States History; Mazzeo et al. [1987] for Biology; Eignor and Bleistein [1987] for Chemistry; Morgan et al. [1987] for English Language and Composition)[1]. However, the reliability coefficients given in these reports were based on representative samples of the *entire* group of candidates that took each of the exams in 1987. The formula for estimating standardized true-score differences requires separate reliability estimates for males and females for each of the ethnic or racial groups being studied. Approximations of these subgroup reliability coefficients were obtained by making adjustments to the reported reliability coefficient based on an assumption of equal average standard errors of measurement across subgroups. Specifically, estimates of subgroup-specific reliability coefficients were obtained from the reported coefficients by

$$(4) \qquad \hat{\rho}_s = 1 - \frac{(1 - \hat{\rho}_t)S_t^2}{S_s^2}$$

where $\hat{\rho}_t$ and $S_t$ refer to the total-group reliability and standard deviation estimates and $\hat{\rho}_s$ and $S_s$ refer to the corresponding estimates for the subgroup of interest.

### Results

The Appendix contains multiple-choice and constructed-response summary statistics (sample sizes, means, standard deviations, and estimated reliability coefficients) for the AP United States History, Biology, Chemistry, and English Language and Composition Examinations. Results are reported separately for males and females within each ethnic or racial group. Tables 4–7 present observed-score and true-score standardized differences for each applicable group.

**Table 4. Standardized Differences by Racial or Ethnic Group for the U. S. History Exam**

| | *Standardized Differences* | | | |
|---|---|---|---|---|
| | *Observed Score* | | *True Score* | |
| | *Multiple Choice* | *Constructed Response* | *Multiple Choice* | *Constructed Response* |
| White | .34 | .01 | .36 | .02 |
| Asian American | .26 | − .01 | 27 | − .01 |
| Black | .28 | .03 | .30 | .05 |
| Mexican American | .50 | .15 | .53 | .22 |
| Other Hispanic | .33 | .07 | .35 | .10 |

Positive differences indicate test performance of males was higher than that of females.

For the United States History exam (Table 4), results are reported for white, Asian American, black, Mexican American, and other Hispanic examinees. The average multiple-choice test score for male examinees was higher than the female average for all five groups, and the difference exceeded a quarter of a standard deviation. For four of the five ethnic or racial groups, males also scored higher on the constructed-response sections of the test, but the differences were considerably smaller in magnitude and, in some cases, negligible. For Asian Americans, females scored slightly higher on the constructed-response section (-.01), but males scored considerably higher on average on the multiple-choice section than did females (.26). In other words, for all five groups, the average performance of females relative to that of males was always better on the constructed-response section.

Also reported in Table 4 are the estimated standardized true-score differences for the U.S. History exam. For all five ethnic or racial groups, the differences in the reliability of the multiple-choice and constructed-response sections

**Table 5. Standardized Differences by Racial or Ethnic Group for the Biology Exam**

| | *Standardized Differences* | | | |
|---|---|---|---|---|
| | *Observed Score* | | *True Score* | |
| | *Multiple Choice* | *Constructed Response* | *Multiple Choice* | *Constructed Response* |
| White | .33 | .16 | .34 | .18 |
| Asian American | .32 | 21 | .33 | 24 |
| Black | .36 | .20 | 37 | 23 |

Positive differences indicate test performance of males was higher than that of females.

[1] For each exam, multiple-choice section reliabilities were obtained using coefficient alpha on a sample of data from the operational administration of the exam. The multiple-choice section reliabilities reflect sources of measurement error associated with item sampling. The constructed-response section reliabilities were obtained using coefficient alpha with each constructed-response part treated as a separate item. Since each separate constructed-response part is scored by a different reader, the constructed-response section reliabilities reflect sources of measurement error due to both item and reader sampling.

Coefficient alpha assumes that all test parts are essentially tau equivalent (Lord and Novick 1968). This condition is unlikely to hold for either the items in the multiple-choice sections or the parts of the constructed-response sections of the AP Examinations studied here. When such conditions do not hold, the resulting reliability coefficients are lower-bound estimates and may underestimate the actual alternat -forms reliabilities. As a result, the reported standardized true-score differences are probably overestimates.

**Table 6. Standardized Differences by Racial or Ethnic Group for the Chemistry Exam**

| | *Standardized Differences* | | | |
|---|---|---|---|---|
| | *Observed Score* | | *True Score* | |
| | *Multiple Choice* | *Constructed Response* | *Multiple Choice* | *Constructed Response* |
| White | 42 | 31 | .44 | .35 |
| Asian American | .32 | .22 | .33 | .25 |

Positive differences indicate test performance of males was higher than that of females.

account for very little of the discrepancy in the size of the sex-related differences obtained from the two modes of assessment. Even for Mexican Americans, the group with the largest difference on the constructed-response section, the multiple-choice true-score difference (.53) is still over twice as large as the corresponding constructed-response difference (.22).

For the Biology exam (Table 5), results are reported for whites, Asian Americans, and blacks. In all three groups, males scored higher on both the multiple-choice and the constructed-response sections of the test, but the typical pattern is again observed. The largest discrepancy in the magnitude of the sex-related differences across assessment modes occurred for whites, where the difference on the multiple-choice section (.33) was over twice as large as the difference on the constructed-response section (.16). The smallest discrepancy occurred among Asian Americans, where the difference on the multiple-choice section (.32) was about 1.5 times as great as the difference on the constructed-response section (.21).

As was the case for the History exam, the differences in the reliability of the Biology multiple-choice and constructed-response sections do not account for much of the discrepancy in sex-related differences on these two sections. For example, the estimated true-score standardized difference for whites on the multiple-choice section is still about two times greater than the corresponding difference on the constructed-response section. In the case of Asian Americans, the multiple-choice true-score difference is still about 1.4 times larger than the corresponding constructed-response difference.

**Table 7. Standardized Differences by Racial or Ethnic Group for the English Language and Composition Exam**

| | *Standardized Differences* | | | |
|---|---|---|---|---|
| | *Observed Score* | | *True Score* | |
| | *Multiple Choice* | *Constructed Response* | *Multiple Choice* | *Constructed Response* |
| White | 17 | 02 | 19 | 02 |
| Asian American | 18 | .14 | .20 | -.19 |

Positive differences indicate test performance of males was higher than that of females.

For the Chemistry exam (Table 6), results are reported for whites and Asian Americans. In both groups, males scored higher on both the multiple-choice and constructed-response sections of the test. For whites, the difference on the multiple-choice section (.42) is only about 40 percent larger than the difference on the constructed-response section (.31). For Asian Americans, the difference on the multiple-choice section was about 50 percent larger than the difference on the constructed-response section.

The magnitude of the discrepancies in sex-related differences on the Chemistry multiple-choice and constructed-response sections is considerably smaller than that observed for either U.S. History or Biology Despite the smaller magnitude, differences in section reiiabilities still do not appear to explain these discrepancies. For whites, the estimated true-score standardized difference is still 25 percent larger than the corresponding difference on the constructed-response section. For Asian Americans, the multiple-choice true-score difference is still about 30 percent larger.

For the English Language and Composition exam (Table 7), results are reported for whites and Asian Americans. White males scored higher than white females on both the multiple-choice and constructed-response sections of the test. However, the standardized difference on the constructed-response section is close to 0 (.02) compared to a difference on the multiple-choice section of about a fifth of a standard deviation. Asian American females scored somewhat higher than Asian American males on the constructed-response section (-.14), but Asian American males scored higher to almost the same degree on the multiple-choice section (.18).

As was the case for the three other exams, differences in reliability among multiple-choice and constructed-response sections explain very little of the discrepancy in sex-related differences on the English Language and Composition exam. For whites, the estimated true-score standardized difference on the multiple-choice section is still about nine times larger than the corresponding difference on the constructed-response section. For Asian Americans, where the direction of the difference changes across assessment modes, the size of true-score differences is more discrepant than that of the observed-score analogues.

## Analyses of Multiple-Choice Scores Conditioned on Constructed-Response Scores

### Procedure

Estimates of the regression of multiple-choice scores on constructed-response scores were obtained in two steps. In the first step, separate bivariate frequency distributions of constructed-response and multiple-choice scores for males and females were obtained for each ethnic or racial group studied. Each of these bivariate frequency distributions was then smoothed using generalized log-linear models (Rosenbaum and Thayer 1987). The models provide the "smoothest" bivariate distribution that preserves certain prespecified features of the observed data. A single smoothing model was used for all exams and all ethnic or racial groups. The smoothing model used preserved (1) the marginal means, marginal variances, and marginal skewnesses of the constructed-response and multiple-choice score distributions; (2) the marginal frequencies in three groups of multiple-choice and constructed-response scores; and (3) the correlation between multiple-choice and constructed-response scores. In the second step, estimates of expected multiple-choice means conditioned on constructed-response scores were obtained directly by taking the mean of the smoothed conditional multiple-choice frequencies.

As discussed earlier, comparing the conditional multiple-choice score means of males and females can be misleading when there is a substantial difference between the groups in mean scores on the conditioning variable (in this case, the constructed-response test score). Therefore, for the ethnic or racial groups with the largest sex-related differences in constructed-response scores, we also obtained estimates of the linear regression of multiple-choice scores on constructed-response true scores according to

$$(5) \qquad \mathcal{E}(X_{mc}|X_{cr}) = \bar{X}_{mc} + \frac{rS^2_{mc}(X_{cr} - \bar{X}_{cr})}{\hat{\rho}_{cr}S^2_{cr}}$$

where $r$ is the within-subgroup correlation between multiple-choice and constructed-response scores and $\hat{\rho}_{cr}$ is the within-subgroup reliability of constructed-response scores. The predicted multiple-choice values for selected constructed-response true-score values were then compared for males and females. The magnitudes of these differences were also compared to sex-related differences in the conditional observed-score means.

### Results

Separate plots of smoothed multiple-choice means conditioned on constructed-response scores are given in Figures 3A-7A for the U.S. History exam. Figures 3B-7B show corresponding plots of conditional mean differences between the sexes. For both kinds of plots, results are shown for only that portion of the constructed-response score range for which sufficient data existed. For all five groups, multiple-choice means were higher for males than for females throughout the constructed-response score range. For whites (Figure 3B), Asian Americans (Figure 4B), blacks (Figure 5B), and other Hispanics (Figure 7B), sex-related differences in conditional multiple-choice means ranged between roughly two and five points, with differences decreasing at higher constructed-response scores for whites and Asian Americans but increasing at higher constructed-response scores for blacks and other Hispanics.

The largest sex-related differences in conditional means were obtained for Mexican American examinees (Figure 6). The difference at a constructed-response score of 11 (the male constructed-response mean is 11.1 and the female constructed-response mean is 10.5) is about 7 points. However, the Mexican American group also evi-

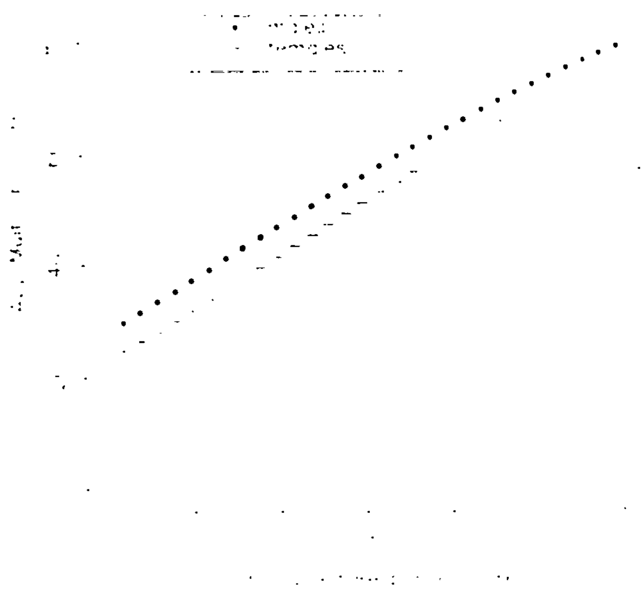

**Figure 3A. U.S. History multiple-choice means conditioned on constructed-response scores for white examinees.**

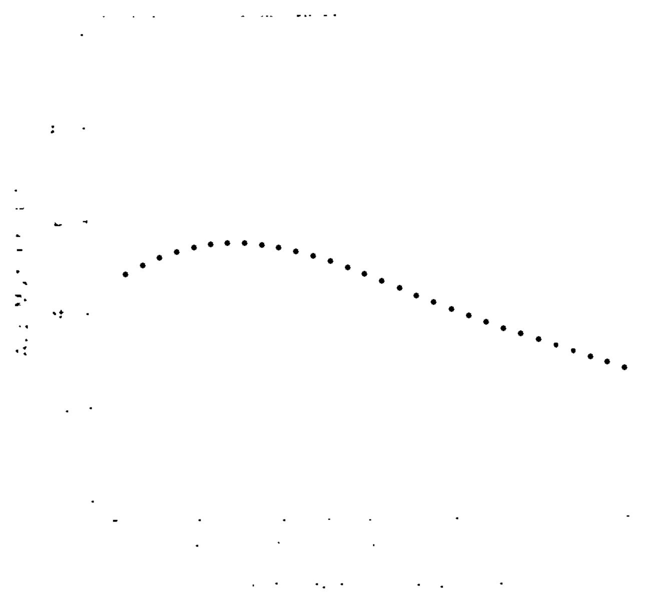

**Figure 3B. Average differences (male-female) in History multiple-choice means conditioned on constructed-response scores for white examinees.**

BEST COPY AVAILABLE

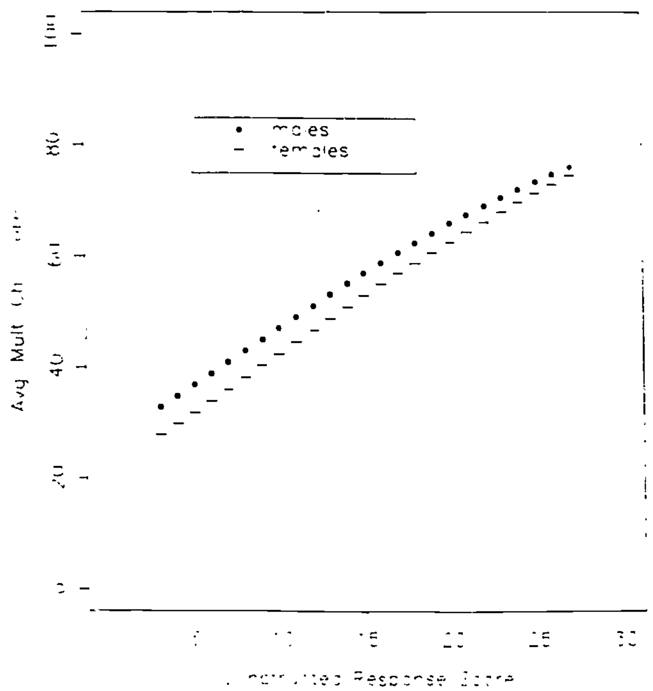

Figure 4A. U.S. History multiple-choice means conditioned on constructed-response scores for Asian American examinees.

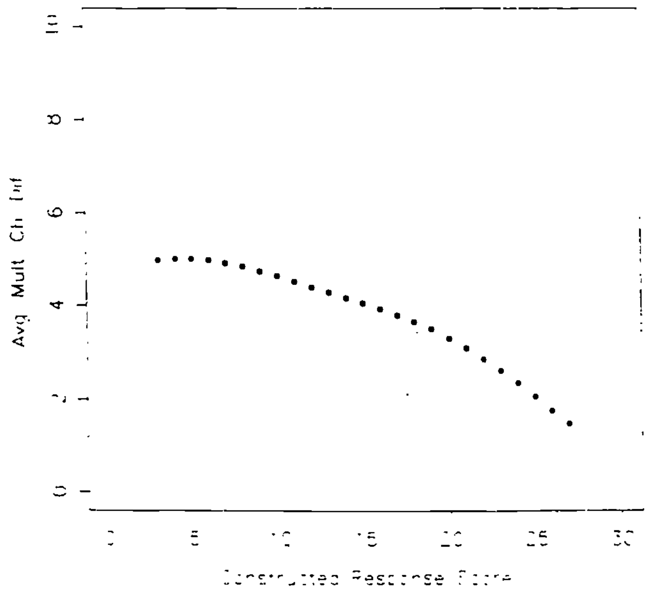

Figure 4B. Average differences (male-female) in History multiple-choice means conditioned on constructed-response scores for Asian American examinees.

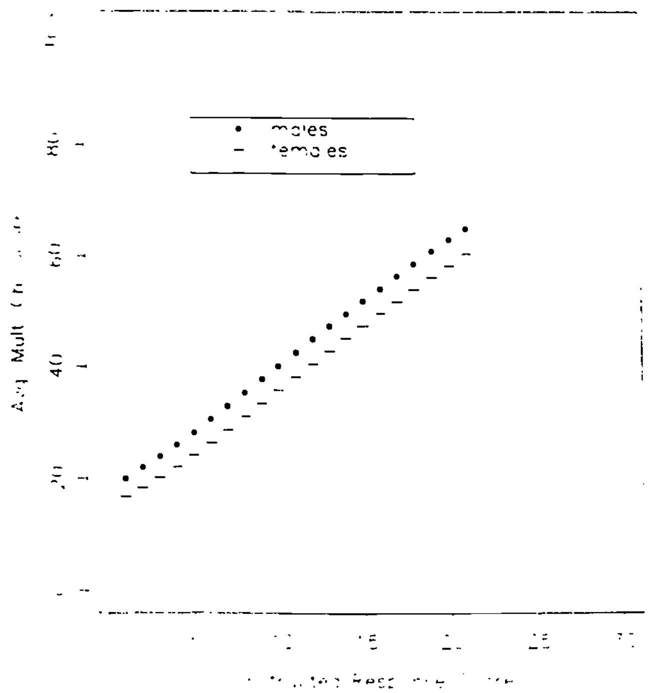

Figure 5A. U.S. History multiple-choice means conditioned on constructed-response scores for black examinees.

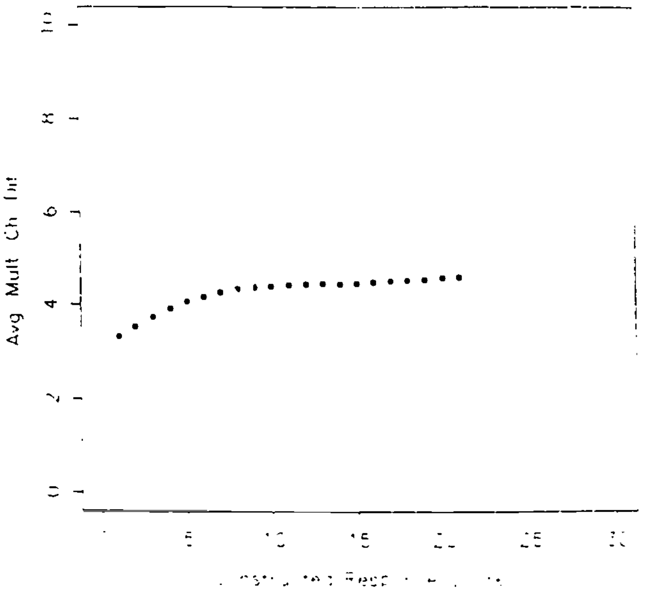

Figure 5B. Average differences (male-female) in History multiple-choice means conditioned on constructed-response scores for black examinees.

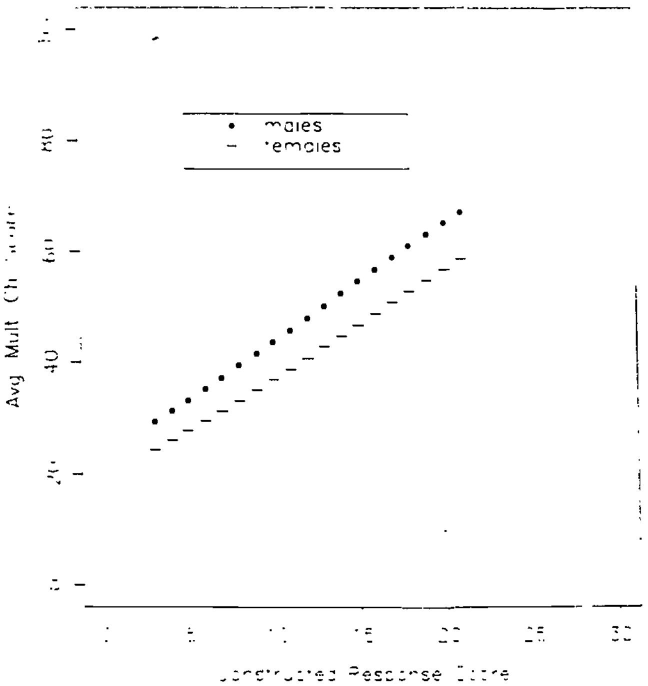

Figure 6A. U.S. History multiple-choice means conditioned on constructed-response scores for Mexican American examinees.

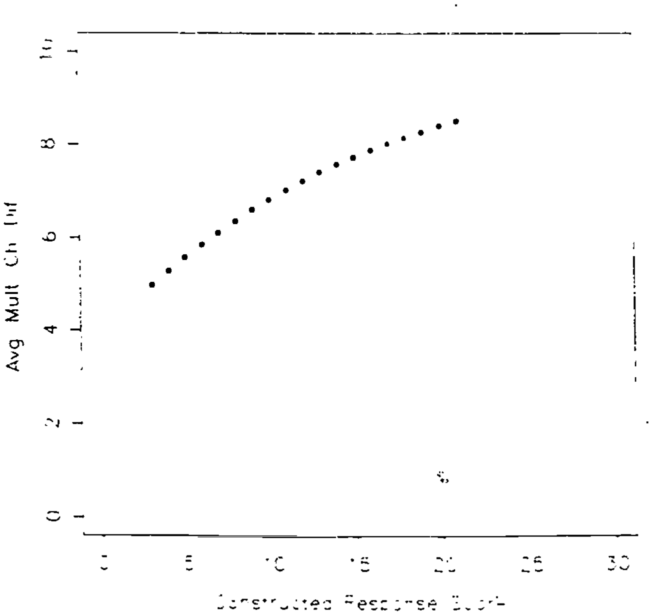

Figure 6B. Average differences (male-female) in History multiple-choice means conditioned on constructed-response scores for Mexican American examinees.

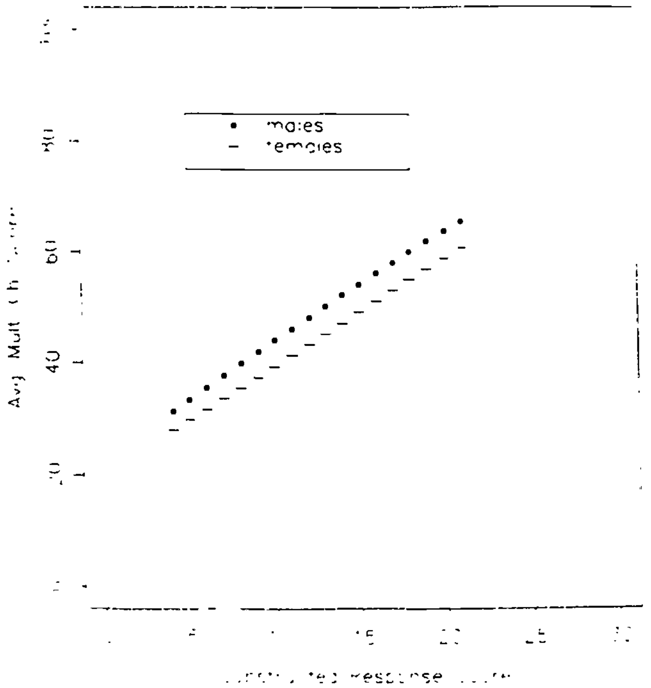

Figure 7A. U.S. History multiple-choice means conditioned on constructed-response scores for other Hispanic examinees.

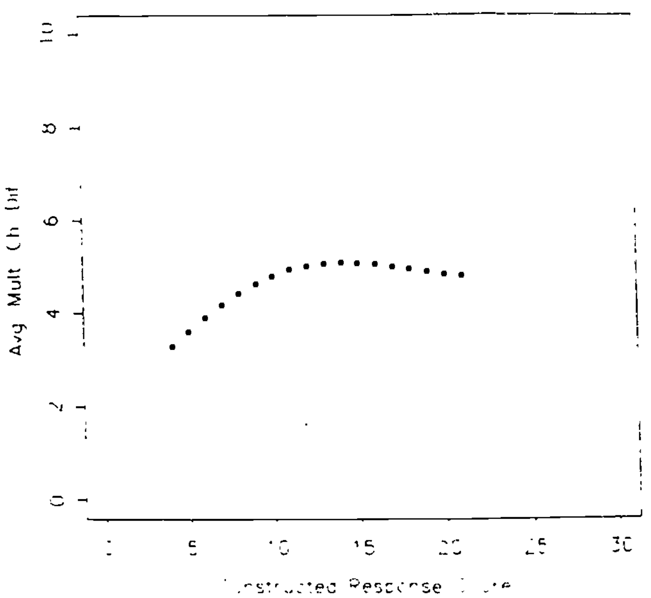

Figure 7B. Average differences (male-female) in History multiple-choice means conditioned on constructed-response scores for other Hispanic examinees.

enced the largest sex-related difference in *marginal* constructed-response means, with males outperforming females by approximately .12 observed-score standard deviation units. Based on the discussion above, one would expect some sex-related differences in multiple-choice means conditional on constructed-response observed scores, since the males at a given constructed-response observed score are likely to have a higher average constructed-response true score than females at the same constructed-response observed score.

In order to approximate how much of the difference in conditional multiple-choice means might be due to conditioning on a fallible measure, separate estimates of multiple-choice means conditioned on constructed-response true scores were obtained for Mexican American males and females. Table 8 gives estimates of the mean sex-related difference in multiple-choice scores at four different constructed-response *true scores*. Also given for comparison purposes are the differences between male and female smoothed multiple-choice means at the corresponding constructed-response *observed scores*.

As expected, sex-related differences in average multiple-choice scores conditioned on constructed-response true scores are less than the corresponding differences conditioned on observed scores. For example, the average multiple-choice score of males and females with constructed-response true scores of 7 (about one standard deviation below the mean) is 3.6 points. Although this is about 40 percent smaller than the corresponding sex-related difference conditioned on observed scores, it is still substantial. At a constructed-response true score of 11, the sex-related difference conditioned on true scores is still almost 6 points. The results of this analysis suggest that substantial sex-related differences in conditional multiple-choice averages are likely to remain even after taking into account the fallible nature of the conditioning variable. Since these differences in marginal constructed-response means for the remaining groups are considerably smaller, the effects of conditioning on a fallible measure would be expected to explain less of the difference in conditional means for these groups. For white and Asian American examinees in particular, such effects would be negligible, since the male and female constructed-response means are nearly identical.

**Table 8. Sex-Related Differences in Predicted History Multiple-Choice Means Conditioned on Constructed-Response Scores for Mexican American Examinees**

| *Constructed-Response True Score* | *Difference Conditioned on Constructed-Response Observed Score* | *Difference Conditioned on Constructed-Response True Score* |
|---|---|---|
| 4 | 5.3 | 2.0 |
| 7 | 6 1 | 3 6 |
| 11 | 7.0 | 5 7 |
| 14 | 7.6 | 7.2 |

Positive differences indicate test performance of males was higher than that of females.

For the Biology exam, plots of smoothed multiple-choice means conditioned on constructed-response scores are given in Figures 8A-10A and the corresponding difference plots are given in Figures 8B-10B. Conditional means

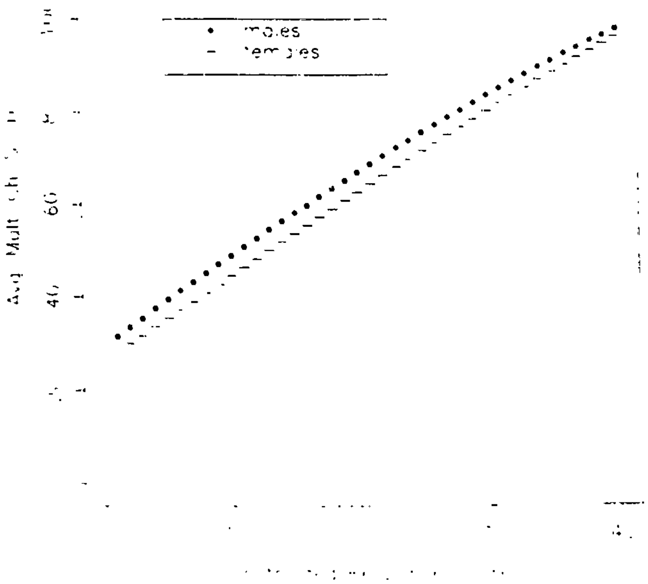

Figure 8A. Biology multiple-choice means conditioned on constructed-response scores for white examinees.

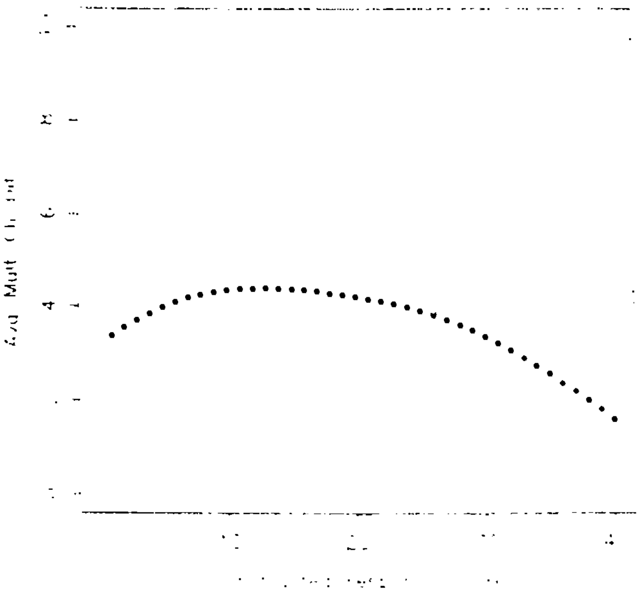

Figure 8B. Average differences (male-female) in Biology multiple-choice means conditioned on constructed-response scores for white examinees.

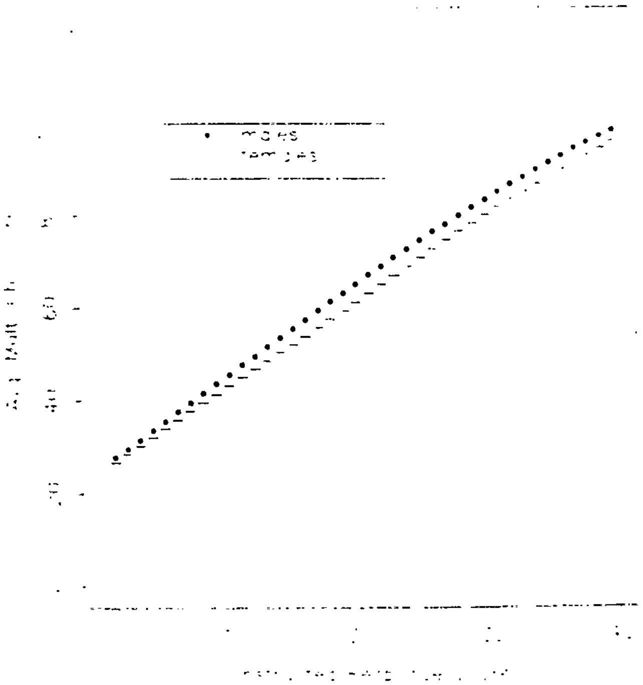

Figure 9A. Biology multiple-choice means conditioned on constructed-response scores for Asian American examinees.

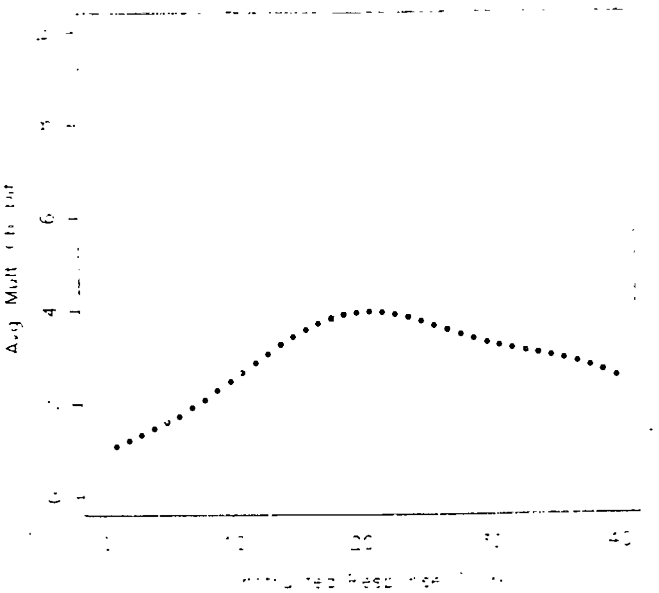

Figure 9B. Average differences (male-female) in Biology multiple-choice means conditioned on constructed-response scores for Asian American examinees.

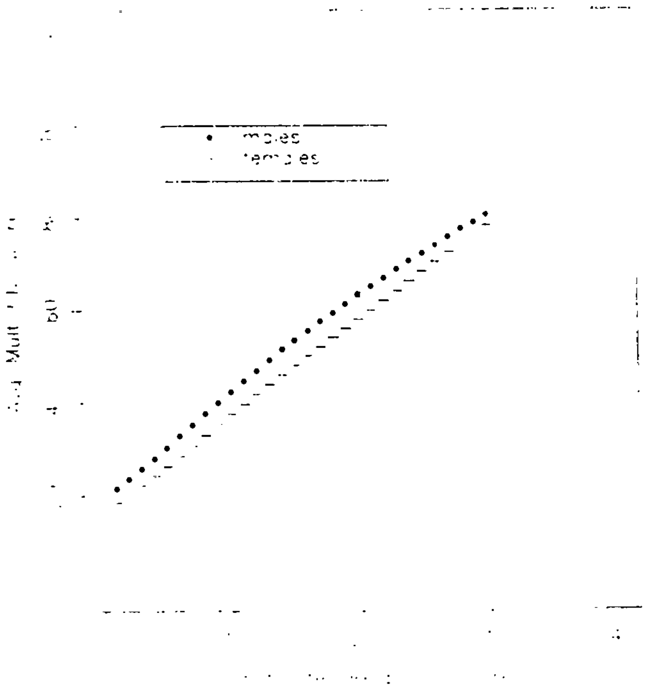

Figure 10A. Biology multiple-choice means conditioned on constructed-response scores for black examinees.

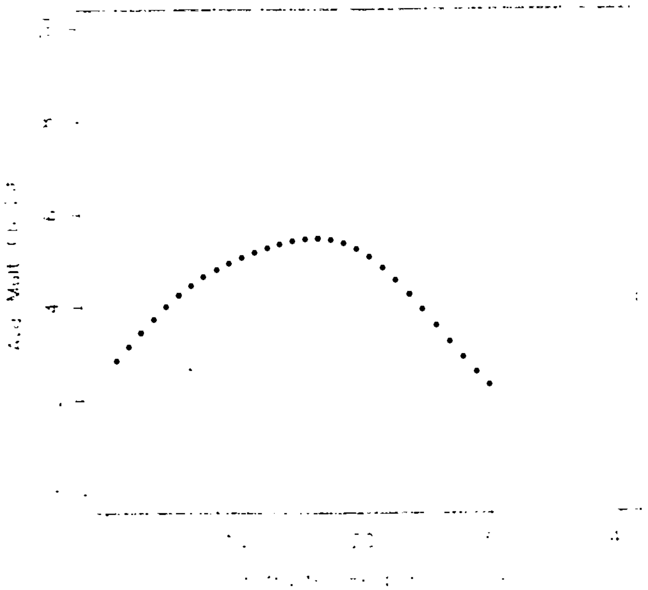

Figure 10B. Average differences (male-female) in Biology multiple-choice means conditioned on constructed-response scores for black examinees.

are again higher for males than females at all constructed-response score points for all three racial or ethnic groups. For white and Asian American examinees, the differences in conditional multiple-choice means are, for the most part, between 2 and 4 points, with slightly smaller differences at high scores for whites and low scores for Asian Americans. Differences for blacks are slightly larger, particularly between constructed-response score points 10 to 20.

The largest sex-related difference in marginal constructed-response means was observed for the Asian American group, with males outperforming females by approximately .2 standard deviation units. Table 9 gives sex-related differences in conditional multiple-choice score means for four different levels of constructed-response true scores for Asian American examinces. Again, it is evident that substantial male-female differences in multiple-choice performance conditioned on constructed-response true scores still exist. At all constructed-response true-score levels shown, the differences in estimated multiple-choice performance are at least 2 points.

Figures 11A and 12A show Chemistry multiple-choice means cunditioned on constructed-response scores of white and Asian American examinees, and Figures 11B and 12B contain the corresponding difference plots. Conditional means for males are again higher than those for females throughout the constructed-response score range for both groups. For the most part, the differences fall between 1.5 and 3 points for white and Asian American examinees.

**Table 9. Sex-Related Differences in Predicted Biology Multiple-Choice Means Conditioned on Constructed-Response Scores for Asian American Examinees**

| *Constructed-Response True Score* | *Difference Conditioned on Constructed-Response Observed Score* | *Difference Conditioned on Constructed-Response True Score* |
|---|---|---|
| 12 | 2.8 | 2.2 |
| 19 | 3.9 | 2.2 |
| 21 | 4.0 | 2.3 |
| 28 | 3.4 | 2.2 |

Positive differences indicate test performance of males was higher than that of females.

The largest sex-related difference in Chemistry marginal constructed-response means was observed for the white group, with males outperforming fema'cs by approximately .3 observed-score standard deviation units. Table 10 gives sex-related differences in Chemistry multiple-choice score means for whi.e males and females conditioned on four different levels of constr:cted-response true scores. As was the case with the previous subjects, differences still exist after making adjustments for the unreliability of the matching criterion.

Figures 13A and 14A contain plots of conditional multiple-choice means for the English exam for whites and Asian Americans, and Figures 13B and 14B contain the corresponding difference plots. Sex-related differences for white examinees (Figure 13) appear to be smaller than those observed for Asian American examinees. For the former group the differences range between 0 and 2 points; for the latter group the differences are between 1 and 3.5 points. For white examinees the difference in conditional means at

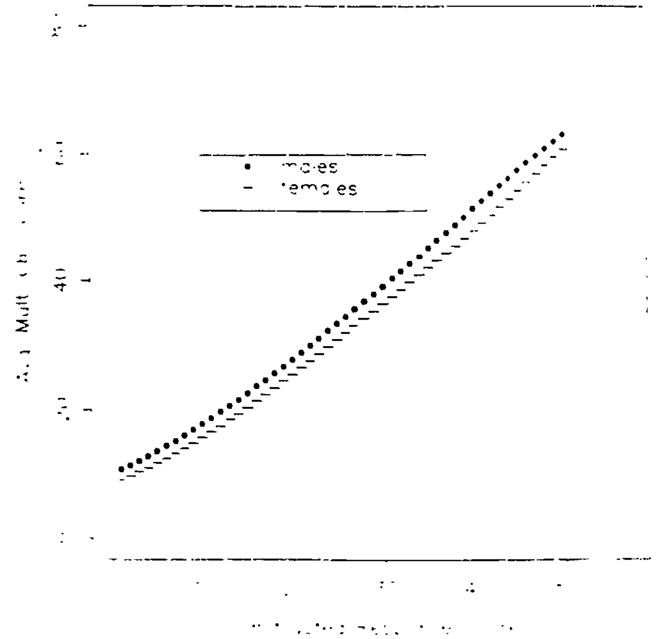

**Figure 11A. Chemistry multiple-choice means conditioned on constructed-response scores for white examinees.**

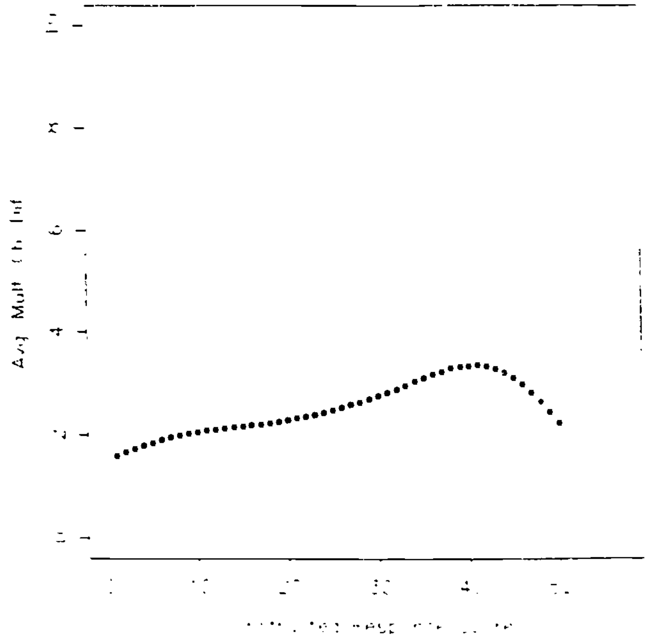

**Figure 11B. Average differences (male-female) in Chemistry multiple-choice means conditioned on constructed-response scores for white examinees.**

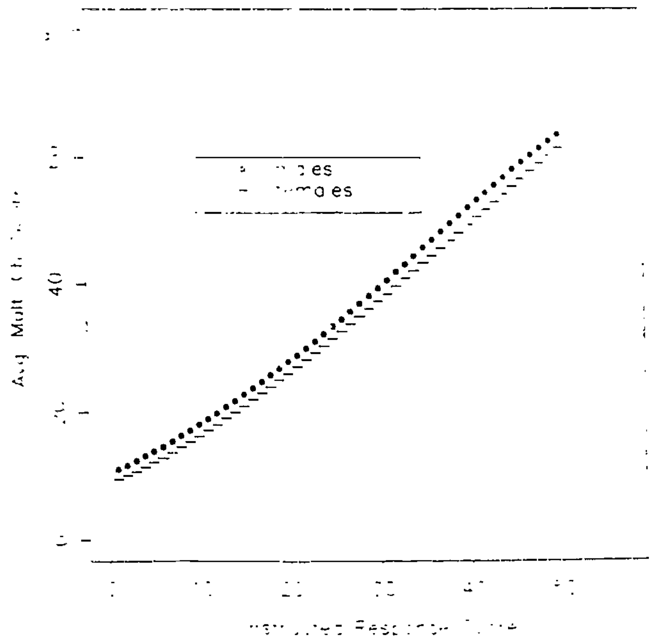

**Figure 12A. Chemistry multiple-choice means conditioned on constructed-response scores for Asian American examinees.**

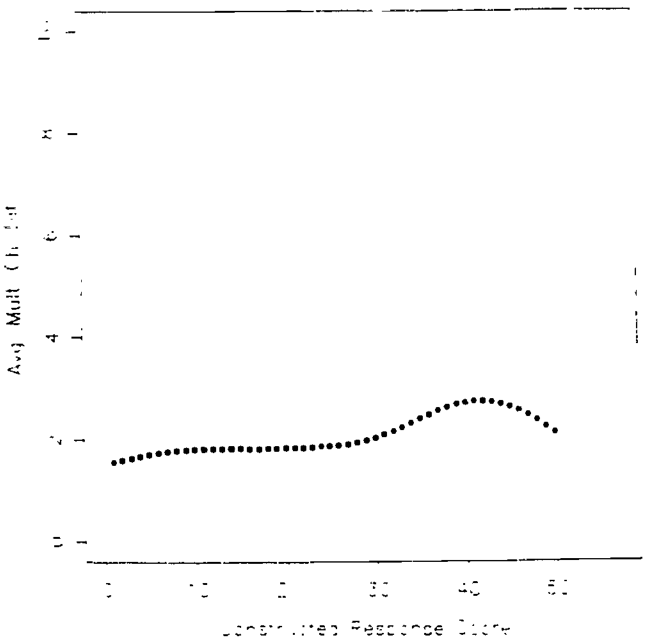

**Figure 12B. Average differences (male-female) in Chemistry multiple-choice means conditioned on constructed-response scores for Asian American examinees.**

a constructed-response score of 14 (the male and female constructed-response means are about 14.2) is about 2 points. For Asian American examinees (Figure 14), the difference in conditional means at a constructed-response score of 14 (the male constructed-response mean is about 13.9 and the female constructed-response mean is about 14.3) is approximately 3 points. Analyses comparing the regression of multiple-choice scores on constructed-response true scores were omitted for this exam, since the size of sex-related differences in marginal construc'ed-response scores was small relative to those examined for previous exams.

To summarize, sex-related differences in average multiple-choice score performance conditioned on constructed-response observed scores were evident in all four exams and in all groups studied. Analyses in selected ethnic or racial groups in which nontrivial sex-related differences in marginal constructed-response means are present suggest that some portion of these conditional differences might be attributable to imperfect matching on the conditioning variable (i.e., constructed-response observed score); however, even after making some adjustments for this artifact, it appears that nontrivial sex-related differences in multiple-choice test scores conditioned on constructed-response scores remain. These results, taken together, further discredit the notion that differences in reliabilities between the sections account for the disparities in the magnitude of sex-related differences on multiple-choice and constructed-response sections.

**Table 10. Sex-Related Differences in Predicted Chemistry Multiple-Choice Means Conditioned on Constructed-Response Scores for White Examinees**

| *Constructed-Response-True Score* | *Difference Conditioned on Constructed-Response Observed Score* | *Difference Conditioned on Constructed-Response True Score* |
|---|---|---|
| 14 | 2.1 | 0.9 |
| 23 | 2.4 | 1.4 |
| 27 | 2.6 | 1.5 |
| 35 | 3.1 | 1.9 |

Positive differences indicate test performance of males was higher than that of females

## STUDY 2: DIFFERENTIAL ITEM FUNCTIONING (DIF) ANALYSES

One of the aims of the research described in this report was to explore the degree to which the better performance of males relative to females on the multiple-choice sections of AP Examinations might be due to the presence of differentially functioning items that favor males. Put in a slightly different way, would the test performance of females relative to that of males look more similar on the multiple-choice and constructed-response tests if the multiple-choice test were purged of any differentially functioning items? Before such a question can be answered, it is first necessary to determine the extent and identity of differentially functioning multiple-choice items.

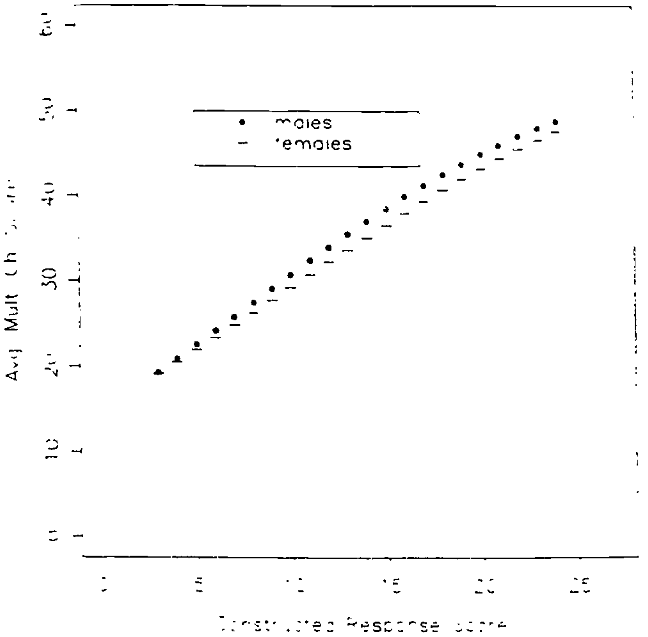

Figure 13A. English multiple-choice means conditioned on constructed-response scores for white examinees.

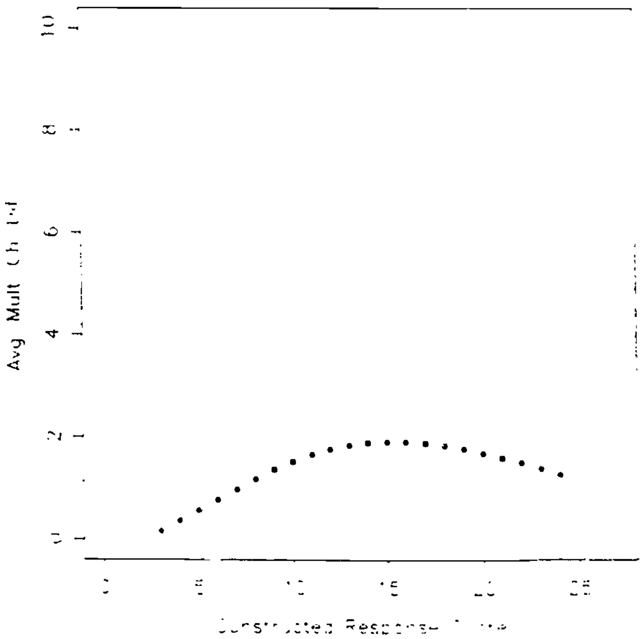

Figure 13B. Average differences (male-female) in English multiple-choice means conditioned on constructed-response scores for white examinees.

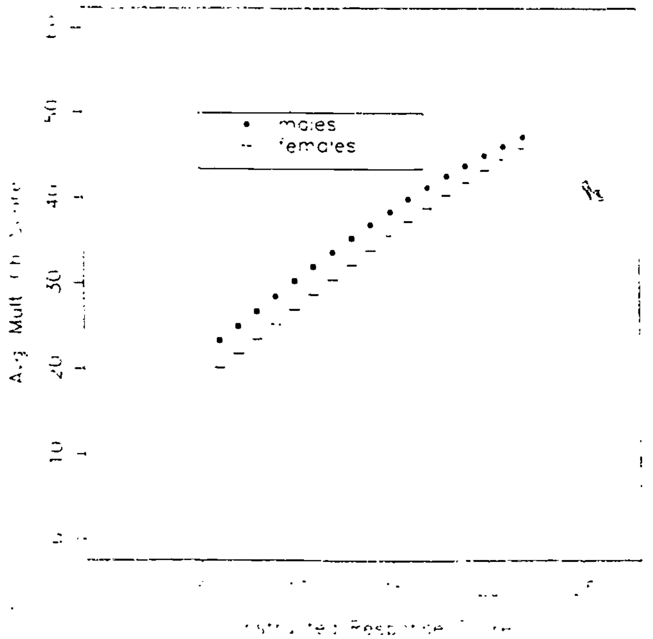

Figure 14A. English multiple-choice means conditioned on constructed-response scores for Asian American examinees.

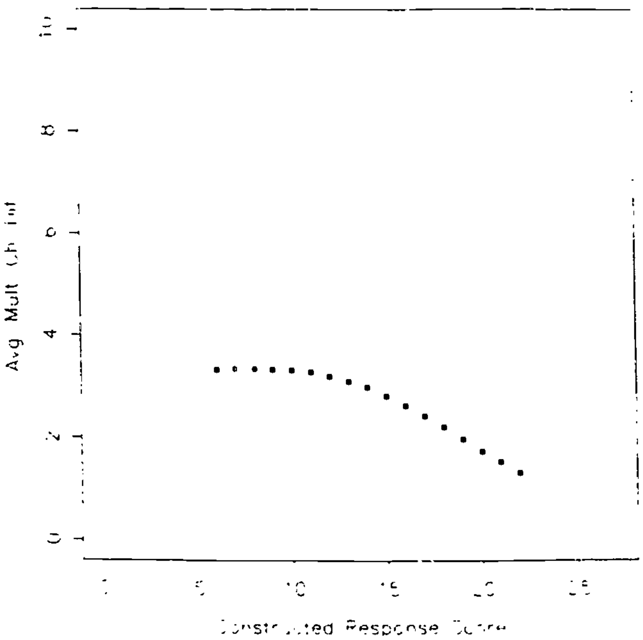

Figure 14B. Average differences (male-female) in English multiple-choice means conditioned on constructed-response scores for Asian American examinees.

## Identifying DIF Items

For the present study, DIF was operationally defined as differences in item performance between groups that were matched with respect to total multiple-choice score. The two groups being compared are sometimes referred to as the focal group (here, females) and the reference group (here, males). In Study 2, DIF analyses were carried out for the multiple-choice sections for each of the forms included in Study 1. For each exam, separate DIF analyses were conducted for each racial or ethnic group.

### DIF Methods

Two different indices were used to identify items that were performing differentially for males and females. The main index used was the Mantel-Haenszel (MH) statistic (Holland and Thayer 1988). The statistic is used to identify items for which the odds of a correct response differ for focal group and reference group examinees that have been matched on some measure of proficiency. A second index of DIF, based on the standardization approach of Dorans and Kulick (1986), was also used to supplement the results of the MH analyses. This index, referred to as the standardized formula score difference (SFD), identifies items for which the item formula score means differ for matched groups of focal and reference group examinees.

The MH procedure provides a test of the null hypothesis that the ratio of reference-group-to-focal-group odds of correctly answering an item is equal to one for all levels of the matching variable versus the alternative hypothesis that the constant odds ratio differs from one. The MH test statistic is based on a weighted average of the odds ratios at each level of a matching variable (in the case of DIF analyses, each score level of the criterion variable). For a particular score level, the conditional odds ratio is defined as

$$(6) \qquad \hat{\alpha}_s = \frac{R_{rs}W_{fs}}{R_{fs}W_{rs}}$$

where $R_{rs}$ and $W_{rs}$ are the proportion correct and proportion incorrect for the reference group and $R_{fs}$ and $W_{fs}$ are similarly defined for the focal group.

It should be noted that in classifying examinee item responses, a distinction was made between "intentional omissions" and items that were "not reached." Omissions following the last answered item in the test were treated as not reached. Following Schmitt and Bleistein (1987), examinees classified as not reaching an item were omitted from the calculation of the $\hat{\alpha}_s$ values in an attempt to remove certain artifacts in these DIF statistics caused by differences in speededness for the focal and reference groups. Omissions prior to the last answered item were treated as intentional. Intentional omissions were treated as incorrect responses.

The Mantel-Haenszel estimate of the constant odds ratio is defined as

$$(7) \qquad \hat{\alpha}_{MH} = \frac{\Sigma a_s\hat{\alpha}_s}{\Sigma a_s}$$

where $a_s$ is a statistically optimal weight associated with score level $s$ and $\Sigma$ is the summation operator. At Educational Testing Service (ETS), the MH estimator is typically transformed to the "Δ-metric" used in the ETS test development process. The Δ-metric has a mean of 13 and a standard deviation of 4. Holland and Thayer (1988) converted $\hat{\alpha}_{MH}$ into a difference in deltas via

$$(8) \qquad \hat{\Delta}_{MH} = -2.35 \ln[\hat{\alpha}_{MH}]$$

This estimate provides a measure of DIF effect size on the Δ scale. A value of zero indicates no DIF, positive values favor the focal group, and negative values favor the reference group.

**Table 11. DIF Classification Criteria**

| | $\lvert\Delta_{MH}\rvert < 1$ | $1 \le \lvert\Delta_{MH}\rvert \le 1.5$ | $1.5 \le \lvert\Delta_{MH}\rvert$ |
|---|---|---|---|
| $\lvert\Delta_{MH}\rvert$ not significantly different from 0 (.05 level) | A | A | A |
| $\lvert\Delta_{MH}\rvert$ significantly different from 0 but not significantly greater than 1 (.05 level) | A | B | B |
| $\lvert\Delta_{MH}\rvert$ significantly greater than 1 (.05 level) | (empty) | B | C |

In the present investigation, the $\Delta_{MH}$ statistic was used to classify items into one of three categories on the basis of the standard criteria used by operational testing programs at ETS. The classification system and associated criteria are given in Table 11. At ETS, category A items are considered to have negligible amounts of DIF, and category C items are often removed from operational tests. Category B items are often examined for research purposes but rarely removed from tests.

As mentioned above, a second DIF index based on the standardization approach (Dorans and Kulick 1986) was also examined. In the traditional standardization analysis, an item is said to exhibit DIF when the probability of correctly answering the item is lower or higher for examinees from the focal group than for a matched group of examinees from the reference group. The traditional standardization approach provides numerical indices for quantifying DIF in the proportion-correct metric. The index used in the current study (SFD) provided a measure of DIF effect size in the formula-score metric (since AP multiple-choice sections are formula scored with a correction for guessing).

The SFD index is a weighted average of the difference between focal group and reference group item formula score means. For a given score level, the difference between focal group and reference item means is given by

$$(9) \qquad D_s = [R_{fs} - \frac{(W_{fs} - O_{fs})}{k-1}] - [R_{rs} - \frac{W_{rs} - O_{rs})}{k-1}]$$

where $k$ is the number of response options for the multiple-choice item under study and $O_{fs}$ and $O_{rs}$ are the proportions of students omitting the item in the focal and reference groups, respectively. As with the MH estimator, examinees who did not reach the item were excluded from the calculation of the $D_s$ values.

The SFD statistic is defined as

$$(10) \qquad SFD = \frac{\sum b_s D_s}{\sum b_s}$$

where $b_s$ is the weighting factor at score level $s$ used to weight differences in the item formula score means between the focal group and the reference group. Although a variety of weighting schemes are possible, $b_s = N_{fs}$ (the number of focal group examinees at score level $s$) has been used in practice because it gives the greatest weight to the $D_s$ values at those score levels most often attained by the focal group under study. Use of $N_{fs}$ means that SFD equals the difference between the observed item formula score of the focal group on the item and the imputed performance of selected reference group members who are matched in proficiency to the focal group members.

The SFD index can range from $-k/(k-1)$ to $k/(k-1)$. All exams studied here consisted of five-option multiple-choice items. Therefore, SFD can range from $-1.2$ to $1.2$. Negative values indicate that the item disadvantages the focal group and positive values indicate that the item favors the focal group. Based on previous experience with the standardization approach, SFD values outside the interval $(-.1, .1)$ range were considered sizable.

In order to remove some artifacts caused by potential differences in the speededness of the test for focal and reference groups, examinees not reaching an item were excluded from the calculation of both the $\alpha_s$ and $D_s$ values. However, additional artifacts due to differential speededness can still arise because of the effect of the number of not-reached items on the matching criterion. For the principal DIF analyses that were carried out, not-reached items were treated in a manner identical to omitted responses in the calculation of the total test formula score used to match examinees (i.e., examinees receive 0 points for an item that is either omitted or not reached). In the presence of differential speededness, the total test matching variable may itself be biased as a measure of proficiency. For example, if the test is more speeded for the focal group than the reference group, focal group members may be expected to get lower total test scores than reference group examinees with equal subject-matter knowledge because of their slower pace.

As a check on potential differential speededness in the matching score, a second standardization-based index was used to identify items that exhibited differential not-reached rates for the focal and reference groups. This index was used to identify items near the end of the test that exhibited differential speededness. If a sizable number of such items was identified, a new "unspeeded" matching variable was defined that excluded the differentially speeded items.

The index used is referred to as the standardized not-reached difference (SND). The index is similar in form to SFD. First, a difference in the proportion of examinees not reaching an item is determined for each of the score levels of the criterion variable, i.e.,

$$(11) \qquad DNR_s = PNR_{fs} - PNR_{rs}$$

where $PNR_{fs}$ and $PNR_{rs}$ are the proportions of individuals at score level $s$ who did not reach the item for the focal and reference groups, respectively. These individual score level differences are summarized across score levels by applying the same standardized weighting function as was applied to the SFD,

$$(12) \qquad SND = \frac{\sum b_s DNR_s}{\sum b_s}$$

For items at the end of a separately timed section of a test, these standardized differences provide a measurement of the differential speededness of a test. Values of SND can range from $-1$ to $+1$, with negative numbers favoring the focal group (i.e., indicating lower rates of not reaching the item). Experience with this index suggests that values outside the interval $(-.05, .05)$ should be considered sizable.

### Samples

DIF analyses were carried out using data from slightly different samples than those used for Study 1. Original analysis plans called for an evaluation of other variables such as years of study of natural sciences, social sciences, and history. In order to facilitate these analyses, the AP Examination file for 1987 was matched to the Scholastic Aptitude Test (SAT) extract file to obtain student descriptive information. This reduced sample was used for the DIF analyses reported here.[2]

Separate DIF analyses, with females as the focal group and males as the reference group, were carried out for each racial or ethnic group that contained a focal or reference group with a sample size of at least 200. Using this criterion, DIF analyses for the United States History examination were carried out for whites, Asian Americans, blacks, Mexican Americans, and other Hispanics. For Biology, DIF analyses were carried out for whites, Asian Americans, and blacks. For the Chemistry and English examinations, DIF analyses were carried out for whites and Asian Americans.

### Results

Results from the DIF analyses are presented in Tables 12–15 for the History, Biology, Chemistry, and English examinations, respectively. Each of these tables presents a summary of the number of items categorized as A, B, or C by the MH procedure. Furthermore, within each of these categories, items were identified as either positive or negative. Negative B or C categories identify those items on which females performed differentially worse than matched males; positive B and C categories identify items on which females performed differentially better. The number of items categorized as A by the MH procedure but with SFD values greater than .10 or less than $-.10$ are also shown. Last, the number of items with SND greater than .05 or less than -.05 is given.

For the History exam, there was little or no evidence of differential speededness. Only one item had an SND value that exceeded .05 in absolute value. This occurred for the Mexican American DIF analyses. As a result of the lack of differential speededness evidence, DIF analyses for all

[2] Subsequent evaluation of student descriptive information from the SAT extract files indicated no differences between males and females. Therefore, plans for the additional analyses were dropped. Further examination of the score differences and patterns of results for the full samples and the reduced samples suggested that redoing DIF analyses using the full sample of data was unwarranted.

**Table 12. Number of Items in Each DIF Category for the U.S. History Exam**

| *DIF Category* | *White* | *Asian American* | *Black* | *Mexican American* | *Other Hispanics* |
|---|---|---|---|---|---|
| C + | 0 | 0 | 0 | 0 | 0 |
| B + | 1 | 2 | 4 | 3 | 6 |
| A | 94 | 93 | 92 | 90 | 89 |
| B - | 3 | 4 | 4 | 6 | 4 |
| C - | 2 | 1 | 0 | 1 | 1 |
| SFD < - .10* | 0 | 0 | 1 | 0 | 2 |
| SFD > .10* | 0 | 0 | 0 | 2 | 1 |

*Number of items exceeding indicated criterion that were *not* already flagged by MH D-DIF criteria.

**Table 13. Number of Items in Each DIF Category for the Biology Exam**

| *DIF Category* | *White** | *Asian American** | *Black*† |
|---|---|---|---|
| C + | 0 | 0 | 1 |
| B + | 0 | 3 | 3 |
| A | 119 | 113 | 109 |
| B – | 1 | 4 | 6 |
| C – | 0 | 0 | 1 |
| SFD < – .10‡ | 0 | 0 | 2 |
| SFD> .10‡ | 0 | 0 | 3 |

*Total test score used as the matching criterion.
†Last nine items removed from the matching criterion.
‡Number of items exceeding indicated criterion that were *not* already flagged by MH D-DIF criteria.

five groups used total test formula score as a matching variable.

Results at the top of Table 12 indicate that little DIF favoring males exists for any of the five racial or ethnic groups. For the white group, only 2 out of the 100 items in the analyses were categorized as negative C (i.e., large DIF favoring males). For Asian Americans, Mexican Americans, and other Hispanics, only 1 out of the 100 items was so classified. Further, only slightly less than half the remaining items classified as category B items favored males.

Table 13 contains results for the Biology exam. Based on the SND index, there was no evidence of differential speededness for Asian Americans and whites. However, for blacks, the last nine items in the test exhibited differential not-reached rates for males and females. All nine items had SND values less than – .05, indicating that a greater proportion of black females reached these items than did a matched group of black males. Because of this differential speededness effect, the matching criterion for the DIF analyses of this group excluded these nine items.

Again, little evidence of DIF was found for the Biology exam. For whites only 1 of 120 items was identified as being anything other than a category A item. For Asian Americans, seven items were identified as category B, but only about half of these exhibited DIF favoring males. Although a somewhat larger number of items were outside category A for blacks (11), only two of these were categorized as C and only one favored males.

**Table 14. Number of Items in Each DIF Category for the Chemistry Exam**

| *DIF Category* | *White* | *Asian American* |
|---|---|---|
| C + | 0 | 0 |
| B + | 0 | 1 |
| A | 76 | 76 |
| B - | 2 | 2 |
| C - | 2 | 1 |
| SFD< - .10* | 0 | 0 |
| SFD > .10* | 0 | 0 |

*Number of items exceeding indicated criterion that were *not* already flagged by MH D-DIF criteria.

**Table 15. Number of Items in Each DIF Category for the English Language and Composition Exam**

| *DIF Category* | *White* | *Asian American* |
|---|---|---|
| C + | 0 | 0 |
| B + | 2 | 1 |
| A | 58 | 59 |
| B - | 0 | 0 |
| C – | 0 | 0 |
| SFD< – .10* | 1 | 1 |
| SFD > .10* | 0 | 0 |

*Number of items exceeding indicated criterion that were *not* already flagged by MH D-DIF criteria.

Results for the Chemistry exam are given in Table 14. No differential speededness effect was observed for either group; therefore total test score was used as the matching variable for both Asian American and white group analyses. As with the previous exams, the amount of DIF appears negligible. For both groups, 76 of the 80 items were in category A. Of the four remaining items, one item for Asian Americans and two items for whites were categorized as C. All three of these items favored males.

Results for the English exam are given in Table 15. Again, no evidence of differential speededness was observed, and total test score was used as the matching criterion for both Asian American and white group analyses. For Asian Americans, 59 of the 60 items were classified as category A. The corresponding number for whites was 58 out of 60. None of the items was classified in category C, and only one item for each group (the item identified by the SFD index) favored males. In short, no evidence of DIF favoring males was obtained for the English Language and Composition exam.

## Impact of DIF on Sex-Related Differences in Multiple-Choice Scores

The analyses reported in the previous section indicate that the multiple-choice sections of the examinations under study contained relatively few differentially functioning items. Even fewer items were detected that exhibited DIF that disadvantaged females. Consequently, one would expect that removing these differentially functioning items would have only a negligible impact on the size of the sex-related differences on the multiple-choice test. Therefore, the presence of differentially functioning items in the multiple-choice section would appear to have little to do with the fact that female performance, relative to that of males, is better on the constructed-response portions of these exams than on the multiple-choice portions. The analyses described in this section were carried out to illustrate this conclusion.

### Procedure

A set of "modified" multiple-choice scores was defined by deleting multiple-choice items identified as functioning differentially for three of the four exams (History, Biology, and Chemistry). The English exam was omitted because of the apparent absence of any differentially functioning items. Three different modified scores were defined for each focal group by deleting items from the test using the following three criteria:

1. Deleting only items classified as C.
2. Deleting *all* items classified as B + , B − , C + , or C − on the basis of the MH statistic value or exceeding the criteria used for the SFD index.
3. Deleting only items in the B − and C − categories or items exceeding the SFD criteria with values less than or equal to − .10.

Deletion criterion (1) was intended to provide an example of the effect of removing DIF items using standard ETS operational criteria for item deletion. Deletion criterion (2) was used to provide a more liberal estimate of the overall impact that differentially functioning items might have on the magnitude of the observed sex-related differences. Criterion (2) is more liberal since both B and C items were deleted and category A items with SFD values outside the interval ( − .10, .10) were also deleted. Deletion criterion (3) was used to provide a kind of "upper-bound" estimate of the amount of differences that might be attributable to DIF, since only those C items favoring males were deleted, and B items favoring males and A items with SFD values less than − .10 were also removed.

Standardized observed-score differences between males and females were computed for each modified score according to equation (3). In addition, standardized true-score differences were also estimated for each modified score according to equation (4). The required reliability estimates for males and females in each ethnic or racial group were obtained using the Spearman-Brown formula in conjunction with the appropriate male or female racial or ethnic group reliability coefficients given in Tables 4–6.

### Results

Results are presented in Tables 16–18 for the History, Biology, and Chemistry exams, respectively. Each table contains both observed-score and estimated true-score standardized differences between males and females (for each group) on the total constructed-response and multiple-choice sections (repeated from Tables 4–6) as well as for each of the modified multiple-choice scores defined for each ethnic group.

For the History exam, the standardized observed- and true-score differences on the modified scores were only slightly smaller than those obtained for the total multiple-choice section. For each racial or ethnic group, the largest percentage reduction was obtained for the subscores defined by deletion criterion (3), as expected. All multiple-choice differences, including those associated with the criterion (2) score, remained considerably larger than the constructed-response differences. For example, deleting all negative DIF items for the other Hispanic subgroup reduced the standardized observed-score difference by about 16 percent from .33 to .29. This was the largest percentage reduction observed for the History exam. Despite this reduction, the difference of .29 is still about four times larger than the standardized difference on the constructed-response section.

Results for the Biology exam were quite similar. The deletion of DIF items did little to reduce sex-related differences on the Biology multiple-choice section. The largest reduction was observed for black examinees using deletion

**Table 16. Observed- and True-Score Standard Differences between Males and Females on U.S. History Constructed-Response and Multiple-Choice Test Components**

| *Standardized Differences** | | | | |
|---|---|---|---|---|
| *Actual* | | *Corrected Multiple Choice†* | | |
| *Constructed Response* | *Total Multiple Choice* | *—All DIF Items* | *—All Negative DIF Items* | *—C Items* |
| | | **White** | | |
| .01 | 34 | .31 | 30 | .32 |
| (.02) | (.36) | (.33) | (.32) | (.34) |
| | | **Asian American** | | |
| – .01 | .26 | .24 | .23 | .25 |
| ( – .01) | (.27) | (.25) | (.24) | (.26) |
| | | **Black** | | |
| 03 | .28 | .27 | .24 | |
| (.05) | (.30) | (.28) | (.26) | — |
| | | **Mexican American** | | |
| .15 | .50 | .46 | .44 | .49 |
| (.22) | (.53) | (.50) | (.47) | (.52) |
| | | **Other Hispanic** | | |
| .07 | .33 | .33 | .29 | .32 |
| (.10) | (.35) | (.35) | (.30) | (.34) |

Positive differences indicate test performance of males was higher than that of females.

*True-score differences given in parentheses.

†Corrected standardized score multiple-choice differences are based on subscores that were refined by deleting DIF items as indicated in the column subheadings.

**Table 17. Observed- and True-Score St:ndard Differences between Males and Females on Biology Constructed-Response and Multiple-Choice Test Components**

| *Standardized Differences** | | | | |
|---|---|---|---|---|
| *Actual* | | *Corrected Multiple Choice†* | | |
| *Constructed Response* | *Total Multiple Choice* | *—All DIF Items* | *—All Negative DIF Items* | *—C Items* |
| | | **White** | | |
| 16 | .33 | .32 | | |
| (.18) | (.34) | (.34) | — | — |
| | | **Asian American** | | |
| .21 | 32 | .32 | .31 | |
| (.24) | (.33) | (.33) | (.32) | — |
| | | **Black** | | |
| 20 | .36 | .34 | .31 | .36 |
| (.23) | (.37) | (.36) | (.33) | (.37) |

Positive differences indicate test performance of ma' s was higher than that of females.

*True-score differences given in parentheses.

†Corrected standardized score multiple-choice differences are based on subscores that were refined by deleting DIF items as indicated in the column subheadings.

criterion (3). The standardized difference was reduced by approximately 12 percent from .36 to .31 for observed scores, and from .37 to .33 for true scores. Despite these reductions, differences on the criterion (3) subscores were still 1.5 times greater than constructed-response differences in terms of observed scores and 1.4 times greater in terms of true scores. Operational DIF flagging criteria resulted in the definition of a modified score only for blacks, and virtually no reduction in standardized differences resulted.

For the Chemistry exam, the largest reduction (7 percent) for A' ian Americans using deletion criterion (3) produced a score for which standardized differences, in terms

**Table 18. Observed- and True-Score Standard Differences between Males and Females on Chemistry Constructed-Response and Multiple-Choice Test Components**

| *Standardized Differences** | | | | |
|---|---|---|---|---|
| *Actual* | | *Corrected Multiple Choice*† | | |
| *Constructed Response* | *Total Multiple Choice* | *--All DIF Items* | *-- All Negative DIF Items* | *--C Items* |
| | | **White** | | |
| 31 | .42 | .39 | | .40 |
| (.35) | (.44) | (.41) | — | .42 |
| | | **Asian American** | | |
| 22 | .32 | .30 | 30 | .31 |
| (.25) | (.33) | (.32) | (.31) | .32 |

Positive differences indicate test performance of males was higher than that of females.

*True-score differences given in parentheses.

†Corrected standardized score multiple-choice differences are based on subscores that were refined by deleting DIF items as indicated in the column subheadings.

of true score, were 25 percent larger than those observed for the constructed-response section. Again, modified subscores defined according to operational DIF flagging criteria resulted in only small reductions in standardized differences.

In conclusion, although some small reductions in multiple-choice standardized differences were obtained by defining modified multiple-choice scores using liberal definitions of differential item functioning, these reductions did not eliminate the discrepancies in magnitude of sex-related differences on the multiple-choice test as compared to constructed-response section differences. Modified scores based on ETS operational definitions of differentially functioning items effected almost no reduction in standardized multiple-choice differences.

## STUDY 3: VARIABILITY IN THE SIZE OF SEX-RELATED PERFORMANCE DIFFERENCES BY TYPE OF CONSTRUCTED-RESPONSE QUESTION

The results from the first two studies indicate that differences in reliability and the presence of differentially functioning items in multiple-choice sections explain little of the pattern of sex-based differences found on the multiple-choice and constructed-response sections of AP Examinations. In Study 3, a final set of analyses was carried out to examine whether a pattern in the size of sex-related differences on individual constructed-response questions or question types could be found across racial or ethnic groups and across test forms. The analyses were carried out with the hope of generating hypotheses for future studies.

The constructed-response tests included in the History and Chemistry exams contain more than one *type* of question. As described in Study 1, each History exam contains both a mandatory document-based question (DBQ) and a set of thematic essays, of which each examinee is required to choose and answer one. The Chemistry exam contains three types of constructed-response questions: word problems, reactions, and analytically scored short essays. The topics for questions differ from form to form for both the History and Chemistry exams. Therefore, in examining patterns of sex-related differences for these two exams, it was of particular interest to determine whether, despite form-to-form changes in specific topics as well as possible cohort-to-cohort changes in relevant proficiencies, certain question types consistently produced larger (or smaller) differences than the remaining question types. If such consistencies were found, future studies could examine the task demands of the different question types more closely and provide a better understanding of the nature and meaning of the sex-related differences obtained.

The constructed-response tests included in the Biology and English exams contain a single question type. Each form of the Biology constructed-response test contains a different sample of topics; however, each form must include topics from each of several broadly defined content areas. The English exam is somewhat less structured, with topics varying more freely from form to form of the exam. For the Biology exam, the question of interest was whether, despite form-to-form changes in the particular topics, sex-related differences were consistently larger or smaller for particular broad content areas. For the English exam, evidence was sought regarding whether particular topics produced consistently large (or small) sex-related differences across test forms and racial or ethnic groups. By examining the nature of the content and topics exhibiting consistent sex-related differences, future studies might provide further insight into the meaning and nature of these differences.

## Procedure

Summary statistics (by sex within each racial or ethnic group) were obtained for each part of the constructed-response tests and for the multiple-choice test for each of the exams studied. As with earlier studies, results were examined only for those racial or ethnic groups that exceeded the sample size requirements delineated in Study 1. Based on these statistics, standardized observed-score differences between males and females on each part of the constructed-response tests were obtained according to equation (6). For reference purposes, standardized differences on the multiple-choice section of each form were also determined. For each exam, standardized differences were determined for the test form used in Studies 1 and 2 (Form J) and for several additional test forms. The additional test forms were analyzed to allow for an evaluation of the stability across years of any interesting pattern of results. Two additional test forms were studied for the Chemistry, English, and Biology exams (Forms H and I). For the History exam, data from a fourth test form (Form K) were also analyzed.

The constructed-response tests for the History, Chemistry, and English exams each contained a different and unique set of questions, but the tests were identical to Form J (described in Study 1) with respect to the number of questions asked, the format of the questions, and grading scales used. However, the constructed-response tests for Forms H and I of the Biology exam differed from that of Form J. The Form J constructed-response test contained four mandatory questions, each of which was scored on a 10-point scale. The constructed-response tests for Forms H and I consisted of three parts each containing two essay questions. Examinees were required to answer only one of the two questions in each part, and all questions were scored on a 15-point scale.

## Results and Discussion

### Analyses by Question Type: U.S. History and Chemistry

Table 19 contains standardized differences for all five racial or ethnic groups for each type of constructed-response question for Forms J, I, H, and K. For Form J, there was little evidence of a pattern in constructed-response sex-related differences. With one exception (Mexican Americans), differences were similar in magnitude for the DBQ and thematic essays, and there was little consistency across groups with respect to which type of constructed-response question females did best on. Somewhat different results were obtained for the remaining three forms. Within each test form there was a good deal of consistency across racial or ethnic groups regarding which part of the constructed-response

**Table 19. Standardized Sex-Related Differences by Constructed-Response Question Type for Four U.S. History Exam Forms**

*Form J*

| *Question Type* | *White* | *Asian American*[5] | *Black* | *Mexican American* | *Other Hispanic* |
|---|---|---|---|---|---|
| Document-based | 01 | .00 | .02 | 06 | .07 |
| Thematic | .03 | - .01 | .04 | .18 | .05 |

*Form I*

| *Question Type* | *White* | *Asian American* | *Black* | *Mexican American* | *Other Hispanic* |
|---|---|---|---|---|---|
| Document-based | - .06 | - 09 | - 04 | - 16 | - .08 |
| Thematic | .06 | .05 | .07 | 12 | 18 |

*Form H*

| *Question Type* | *White* | *Asian American* | *Black* | *Mexican American* | *Other Hispanic* |
|---|---|---|---|---|---|
| Document-based | 05 | .10 | .14 | .17 | .04 |
| Thematic | 03 | .07 | .00 | 05 | .10 |

*Form K*

| *Question Type* | *White* | *Asian American* | *Black* | *Mexican American* | *Other Hispanic* |
|---|---|---|---|---|---|
| Document-based | 11 | 12 | 09 | 15 | 22 |
| Thematic | 09 | 06 | 06 | 05 | 13 |

Positive differences indicate test performance of males was higher than that of females.

test females performed best on relative to males. However, for one of these forms it was the thematic essays, and for the other two it was the DBQ.

For Form I, females did better than males on the DBQ, but the reverse was true on the thematic question for all racial or ethnic groups. For Form H, differences favoring males were slightly greater in magnitude on the DBQ than on the thematic essays, with one exception (the other Hispanic group). This is opposite the result obtained for Form I. The results for Form K also show a consistent pattern. Sex-related differences on the two question types differ substantially in magnitude for only three of the five groups. However, as with Form H, differences favoring males were larger on the DBQ than on the thematic essay for all five groups.

To summarize for the History exam, within a test form there is some evidence that consistent patterns of sex-related differences by question type occur. However, across test forms no clear relationship emerged. Since the questions assigned to the constructed-response tests are unique to each form, one interpretation of these results is that the knowledge and skills required by particular topics may have more to do with the magnitude of sex-related differences than do the task demands peculiar to the question types. An alternative explanation is that the relative preparation of males and females with respect to the proficiencies relevant for success on the two question types may differ from cohort to cohort or that both cohort and topic differences contribute to the lack of consistent results. Further research is required to better determine the nature and meaning of the results reported here.

The constructed-response tests for the Chemistry exam contain three types of constructed-response questions: word problems (parts 1 and 2), reactions (part 3), and short essays (part 4). The results for the Chemistry exam are presented in Table 20 for whites and Asian Americans. Males scored higher than females across all question types and test forms.

For Form J, the most striking result was that for both groups, sex-related differences were a good deal larger on the short essay questions of part 4 than on the remaining portions of the test. In fact, the standardized differences on the constructed-response questions of part 4 were almost as large as the differences evident on the multiple-choice section For Forms I and H, the magnitude of the sex-related differences was more similar in size on the four parts of the constructed-response test. However, differences were largest on the short essay portion of the test. In fact, part 4 standardized differences were almost as large as those evident on the multiple-choice sections.

Despite a consistent pattern of results regarding which question type produced the largest sex-related differences, as with the History exam, there is some evidence that the form-to-form changes in the topics associated with the particular question type in any given form may be equally relevant to understanding the specific meaning of the differences. The degree of variability by question type in the magnitude of the differences changed substantially across forms. This suggests that question type interacts in some way with the particular topic to influence the size of the sex-related differences that were obtained. Further analysis of the nature of these short essay tasks and the implications for understanding the nature of sex-related differences in Chemistry exam performance may be warranted.

**Table 20. Standardized Sex-Related Differences by Constructed-Response Question Type for Three Chemistry Exam Forms**

| | *Form J* | |
|---|---|---|
| *Question Type* | *White* | *Asian American* |
| Word problem 1 | .19 | 13 |
| Word problem 2 | .22 | .14 |
| Equations | .27 | .12 |
| Essays | .36 | .28 |
| | *Form I* | |
| *Question Type* | *White* | *Asian American* |
| Word problem 1 | 06 | .05 |
| Word problem 2 | .06 | .04 |
| Equations | .04 | .02 |
| Essays | .10 | .07 |
| | *Form H* | |
| *Question Type* | *White* | *Asian American* |
| Word problem 1 | .05 | .03 |
| Word problem 2 | .05 | .02 |
| Equations | .04 | .02 |
| Essays | .07 | .04 |

Positive differences indicate test performance of males was higher than that of females.

### Analyses by Content and Topic: Biology and English Language and Composition

The constructed-response tests of the Biology exam consist of analytically scored essays. The essays cover topics from different, somewhat broadly defined content areas. Thus, for the Biology exam it was possible to examine the performance of females relative to males across these different content areas. A question of particular interest was whether females did consistently better relative to males in any single content area across test forms.

The topics for the Biology exam Forms H and I were classified in a slightly different manner than those for Form J. In the current (Form J) system, question topics are classified into one of three broad content areas: (1) cells and molecules, (2) genetics and evolution, and (3) organisms and populations. The old classification system also had three categories: (1) cellular and molecular biology, (2) organismal biology, and (3) populational biology. The categories of the older system do not map directly into those of the new system. For example, topics related to basic biolog-

**Table 21. Standardized Sex-Related Differences by Constructed-Response Question Type for Three Biology Exam Forms**

| | *Form J* | | |
|---|---|---|---|
| *Question Type* | *White* | *Asian American* | *Black* |
| Cell/Mol. #1 | .03 | .08 | .02 |
| Organismal #1 | .27 | .27 | 32 |
| Organismal #2 | .07 | .09 | 14 |
| Cell Mol. #2 | .16 | .22 | .19 |
| Cell Mol. Avg. | .10 | .15 | 11 |
| Organismal Avg. | 17 | 18 | .23 |
| | *Form I* | | |
| *Question Type* | *White* | *Asian American* | *Black* |
| Cell/Mol. | .07 | .12 | - .02 |
| Organismal | .13 | .11 | .18 |
| Population | .08 | .09 | 06 |
| | *Form H* | | |
| *Question Type* | *White* | *Asian American* | *Black* |
| Cell/Mol. | .12 | .09 | .02 |
| Organismal | .15 | .13 | .06 |
| Population | - .02 | .01 | - .05 |

Positive differences indicate test performance of males was higher than that of females.

**Table 22. Standardized Sex-Related Differences by Constructed-Response Question Type for Three English Language and Composition Exam Forms**

| | *Form J* | |
|---|---|---|
| *Question Type* | *White* | *Asian American* |
| Topic 1 | .11 | .01 |
| Topic 2 | - .09 | - .18 |
| Topic 3 | .02 | - .12 |
| | *Form I* | |
| *Question Type* | *White* | *Asian American* |
| Topic 1 | .13 | - .06 |
| Topic 2 | - .01 | - .04 |
| Topic 3 | - .18 | - .04 |
| | *Form H* | |
| *Question Type* | *White* | *Asian American* |
| Topic 1 | .04 | .01 |
| Topic 2 | - .09 | - .04 |
| Topic 3 | - .01 | - .01 |

Positive differences indicate test performance of males was higher than that of females.

ical chemistry are classified in category (1) for both systems. However, topics related to genetics and heredity were classified as cellular and molecular biology under the old system but as genetics and evolution under the new system. In order to avoid confusion in describing the results, questions on all forms were classified with respect to the old category system.

The results for the Biology exam are given in Table 21 for whites, Asian Americans, and blacks. For Form J, there were two topics from each of two of the content areas used with the old classification scheme, cellular and molecular biology and organismal biology. Standardized differences are shown separately for each of the Form J questions. In addition, the average standardized difference across questions for the two content areas is also given. Averaged over topics, Form J sex-related differences were larger for the organismal biology area than for the cellular and molecular biology area for all racial or ethnic groups. Further, the rank ordering of Form J questions, in terms of the size of differences, was identical for all three groups. However, the particular ordering obtained and the amount of across-question variability in the size of the sex-related differences suggest that their magnitude may have more to do with particular topic areas than with the more broadly defined content area. For example, for all three groups, differences were larger on question 4 (cellular and molecular biology) than on question 3 (organismal biology).

Results for Forms I and H showed some similarity with those for Form J. For example, for both Forms I and H, sex-related differences were again largest on the part of the constructed-response test pertaining to organismal biology for all three racial or ethnic groups. However, each part of the test reflects scores on one of two student-selected topics. An analysis of the relative degree of topic variability within each of the content areas was beyond the scope of this project. Thus, although a consistent pattern was evident, with organismal biology topics on average yielding larger differences favoring males, the Form J results suggest that topic variability within each content area may be larger than the variability across topic areas.

The results for the English Language and Composition exam are presented in Table 22 for whites and Asian Americans. For Form J, substantial variation across questions in the magnitude of sex-related differences was evident for both groups. Some favor males and others show an advantage for females. For both racial or ethnic groups, female performance relative to that of males was best for question 2, which required an analysis of how a passage written by a female author enriches our sense of childhood, and worst for question 1, which required the student to take a stand on the issues of personal relations versus causes or patriotism.

For Form I, sex-related differences were about the same size on all constructed-response questions for Asian Americans. For whites, however, the magnitude of the differences changed substantially across questions. Females performed worse than males on question 1, which asked students to compare the styles of two passages written by native Americans about the harshness of the American prairie, but better than males on question 3, which required an evaluation of the assertion that human nature wants patterns, standards, and structure in behavior. In fact, sex-related differences in performance on questions 1 and 3, though different in sign, are greater in magnitude than is the difference on the multiple-choice test.

For Form H, variation in the size and direction of sex-related differences across topics was found for both groups, although the degree of variation was again less for Asian Americans than for whites. A difference favoring males was found for question 1, which asked for an analysis of the style and rhetoric of two passages about the launching of the first Soviet space satellite. A difference favoring females was found on question 2, which required a comparison of two drafts of a passage concerning the effect of the experience of war on the author's attitude toward language. The passage speaks of the hollowness of abstract concepts such as glory and honor when compared to the concrete reality of the human suffering and death that result from wars.

The results across forms suggest considerable variability by topic in the size of sex-related differences for whites, a lesser degree of variability for Asian Americans, and some agreement across racial or ethnic groups regarding which topics produce the largest differences for two of the three forms. These results, coupled with those of the other three exams, suggest that specific topics play a substantial role with respect to the magnitude of sex-related differences. In addition, examination of the content of the topics reveals a potentially interesting pattern. Questions based on passages related to topics such as patriotism, space satellites, and the ruggedness of the American prairie produced the largest differences favoring males. One might conjecture that these topics are stereotypically more male oriented. However, such an explanation is clearly speculative, and research much more carefully done than the exploratory analyses reported here would be necessary to evaluate such conjectures.

## SUMMARY AND CONCLUSIONS

This report described three exploratory studies of the performance of males and females on the multiple-choice and constructed-response sections of four AP Examinations: United States History, Biology, Chemistry, and English Language and Composition. The studies were intended to evaluate some possible reasons for the apparent relationship between test format and the magnitude of male-female differences in performance.

The first study focused on the extent to which such differences could be attributed to differences in the score reliabilities associated with these two modes of assessment. One way that we evaluated the plausibility of this hypothesis was by examining discrepancies in the size of sex-related differences on multiple-choice and constructed-response sections after correcting for differences in the reliabilty of the two sections. The results of this first line of analysis provided little support for the "different-reliabilities" hypothesis. Across all exams and all racial or ethnic groups, substantial sex-related differences remained even after taking into account differences in the reliabilities of the two sections. A second way in which the "different-reliabilities" hypothesis was evaluated was by examining multiple-choice scores conditioned on constructed-response scores. Again, across all four exams and across racial or ethnic groups, conditional multiple-choice scores of females were lower than those of males throughout the constructed-response score range. Both sets of results suggest that, at least for AP Examinations, little of the relationship between format and the magnitude of sex-related differences in performance is due to reliability differences associated with the two different item types.

A second kind of hypothesis concerning the genesis of discrepancies in sex-related differences across multiple-choice and constructed-response sections is what we have termed a "method-bias" hypothesis. In particular, we examined whether the larger differences observed on multiple-choice tests might be partly the result of the presence of substantial numbers of items exhibiting DIF. Since males perform better than females on the multiple-choice sections, one might expect that most items exhibiting large degrees of DIF would be items that favor males.

The results of Study 2 suggest that the presence of differentially functioning items also has little to do with the relationship between test format and the magnitude of sex-related differences on multiple-choice and constructed-response sections for any of the exams studied. Using the standard ETS procedures for identifying differentially functioning items and standard ETS criteria for evaluating the magnitude of such effects, fairly small numbers of items exhibited substantial amounts of sex-related DIF. Among these groups of items, both items favoring males and items favoring females were found.

As an illustrative follow-up, new "refined" multiple-choice scores were calculated by deleting the DIF items that were identified, and differences between males and females on these modified subscores were compared. Three criteria for defining DIF items, and hence three types of modified subscores, were employed. The first criterion was close to the standard ETS criterion. The second criterion represented a somewhat more liberal definition of DIF. The third criterion represented an extreme definition in which only items exhibiting DIF favoring males were removed. Even using the most extreme DIF criterion for creating modified subscores, sex-related differences on the multiple-choice sections were reduced only slightly. Almost no reduction in the magnitude of sex-related differences on the multiple-choice sections was observed using the standard ETS DIF criteria or any of the more liberally defined criteria. These results held both across exams and across racial or ethnic groups.

In order to generate some possible hypotheses for future studies, a final set of exploratory analyses was conducted to investigate the degree to which the discrepancies in sex-related differences on multiple-choice and constructed-response sections might differ for individual constructed-response questions or question types. For a given exam, if a particular question or question type consistently (i.e., for all or most racial or ethnic groups) produced larger discrepancies than other questions or question types, future examinations of the nature of these constructed-response questions might provide some clues as to the source of these discrepancies.

Analyses of the History and Chemistry exams focused on whether particular types of constructed-response questions resulted in consistently larger sex-related differences. For the History exam, no clear relationship emerged. For the Chemistry exam, the short essay question did result in consistently larger sex-related differences than the remaining question types for all three test forms studied. However, for two of the three forms, little variability across question types was found. Analyses for the Biology and History exams focused on whether particular content areas or topics result in consistently larger sex-related differences. Results for both exams suggest that specific topics may influence the magnitude of sex-related differences.

The results of the current study suggest that the major factor accounting for the relatively better performance of females on constructed-response tests may be a construct-relevant one. Constructed-response tests likely demand different sets of competencies than their multiple-choice counterparts, and sex-related differences in performance profiles across the two modes of assessment most likely reflect real disparities in the average level of achievement obtained by males and females with respect to these different competencies. Although the studies described in this report are somewhat limited in scope, investigating only two factors (reliability and DIF), other related studies point to similar conclusions.

For example, Breland (1991) examined the degree to which holistic essay scores on AP American and European History Examinations could be predicted from three classes of variables: directly construct relevant, indirectly construct relevant, and construct irrelevant. Significant prediction was obtained from both kinds of construct-relevant variables but not from the construct-irrelevant variables. Bridgeman and Lewis (1991) examined the degree to which multiple-choice and constructed-response sections of AP Examinations in History, English, and Biology predict grades in sequent courses. They found that, despite their lower levels of reliability, the constructed-response section scores predicted sequent course grades about as well as their more reliable multiple-choice counterparts. Both these studies involve AP Examinations, and clearly more research on a broader class of instruments and differing populations of examinees is needed.

The better relative performance of females on constructed-response tests has important implications for high-stakes standardized testing. Currently, a large amount of standardized testing occurs in a multiple-choice format, and important educational decisions are often made based at least partly on the test results. If both types of tests measure important education outcomes, equity concerns would dictate a mix of the two types of assessment instruments. In addition to cost concerns and efficiency, the extent of reliance on one or the other format should be at least partly a function of the relative importance of the types of education outcomes assessed by each.

As the field of educational measurement moves toward a greater use of constructed-response formats, it will be important to continue to obtain evidence that will provide a more thorough understanding of the nature of the competencies measured by each of these assessment methods and the degree to which these competencies correspond to relevant education outcomes. The authors believe that such efforts need to go beyond "face validity" and reflect more empirically oriented construct and criterion-related validation efforts.

## REFERENCES

Benbow, C. P. 1988. "Sex Differences in Mathematical Reasoning Ability in Intellectually Talented Preadolescents: Their Nature, Effects, and Possible Causes." *Behavioral and Brain Sciences* 11: 169–232.

Bennett, R. E. In press. "On the Meanings of Constructed-Response." In *Construction versus Choice in Cognitive Measurement*, edited by R. E. Bennett and W. C. Ward. Hillsdale, N.J.: Erlbaum.

Bolger, N. 1984. "*Gender Differences in Academic Achievement According to Method of Measurement*." Paper presented at the annual meeting of the American Psychological Association, Toronto.

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

Breland, H. M., and P. A. Griswold. 1981. *Group Comparisons for Basic Skills Measures*. College Board Report No. 81–6. New York: College Entrance Examination Board.

Bridgeman, B., and C. Lewis. 1991. *Sex Differences in the Relationship of Advanced Placement Essay and Multiple-Choice Scores to Grades in College Courses*. Research Report No. RR 91–48. Princeton, N.J.: Educational Testing Service.

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

Dorans, N. J., and E. Kulick. 1986. "Demonstrating the Utility of the Standardization Approach to Assessing Unexpected Differential Item Performance on the Scholastic Aptitude Test." *Journal of Educational Measurement* 23: 355–68.

Eignor, D. E., and C. A. Bleistein. 1987. *Test Analysis: College Board Advanced Placement Examination, Chemistry 3JBP*. ETS Statistical Report No. SR-87-151. Princeton, N.J.: Educational Testing Service

Frederiksen, J. R., and A. Collins. 1989. "A Systems Approach to Educational Testing." *Educational Researcher* 18 (9): 27–32.

Frederiksen, N. 1984. "The Real Test Bias: Influences of Testing on Teaching and Learning." *American Psychologist* 39: 193–202.

Holland, P. W. 1981. "Proposed Modification of IANA80." Unpublished memorandum issued January 14, 1987.

Holland, P. W., and D. T. Thayer. 1988. "Differential Item Functioning and the Mantel-Haenszel Procedure." In *Test Validity*, edited by H. Wainer and H. I. Braun. Hillside, N.J.: Erlbaum.

Klein, S. P. 1989. "Does Performance Testing on the Bar Examination Reduce Differences in Scores among Sex and Racial Groups?" Unpublished manuscript.

Lord, F. L., and M. R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison-Wesley.

Maccoby, E. E., and C. N. Jacklin. 1974. *The Psychology of Sex Differences*. Stanford, Calif.: Stanford University Press.

Mazzeo, J., D. McLean, R. Flesher, and P. Bigham. 1987. *Test Analysis: College Board Advanced Placement Examination. Biology. 3JBP*. ETS Statistical Report No. SR–87–47. Princeton, N.J.: Educational Testing Service.

Morgan, R. L., and R. Flesher. 1987. *Test Analysis: College Board Advanced Placement Examination. American History. 3JBP*. ETS Statistical Report No. SR–87–143. Princeton, N.J.: Educational Testing Service.

Morgan, R. L., R. Flesher, A. Nellikunnel, and D. McLean. 1987. *Test Analysis: College Board Advanced Placement Examination. English Language and Composition, 3JBP*. ETS Statistical Report No. SR–87–169. Princeton, N.J.: Educational Testing Service.

Murphy, R. J. L. 1980. "Sex Differences in GCE Examination Entry Statistics and Success Rates." *Education Studies* 6: 169–78.

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

Petersen, N. 1988. "DIF procedures for use in statistical analysis." Unpublished memorandum issued September 14, 1988.

Petersen, N., and S. L. Livingston. 1982. *English Composition Test with Essay: A Descriptive Study of the Relationship between Essay and Objective Scores by Ethnic Group and Sex*. ETS Statistical Report No. SR–82–96. Princeton, N.J.: Educational Testing Service.

Rosenbaum, P. R., and D. T. Thayer. 1987. "Smoothing the Joint and Marginal Distributions of Scored Two-Way Contingency Tables in Test Equating." *British Journal of Mathematical and Statistical Psychology* 40: 43–49.

Schmitt, A. P., and C. A. Bleistein. 1987. *Factors Affecting Differential Item Functioning for Black Examinees on Scholastic Aptitude Test Analogy Items*. Research Report No. 87–23. Princeton, N.J.: Educational Testing Service.

Stiggins, R. J. 1991. "Facing the Challenges of a New Era of Educational Assessment." *Applied Measurement in Education* 4: 263–73.

Traub, R. E., and K. MacRury. 1990. Antwort-Auswahl vs. Freie-Antwort-Aufgaben Bei Lernerfolgstestes. In *Test und Trends 8: Jarbuch der Paedagogischen Diagnostik*, edited by K. Ingekamp and R. S. Jager. Weinheim, Germany: Beltz-Verlag Publishing Co. (English-language version, entitled *Multiple-Choice vs. Free-Response in the Testing of Scholastic Achievement*, is available from the authors.)

Wilder, G. Z., and K. P. Powell. 1989. *Sex Differences in Test Performance: A Survey of the Literature*. College Board Report No. 89–3. New York: College Entrance Examination Board.

## APPENDIX. SUMMARY STATISTICS BY SEX AND RACIAL OR ETHNIC GROUP

### Table A-1. Multiple-Choice and Constructed-Response Summary Statistics by Sex and Racial or Ethnic Group for AP United States History

| *Self-Reported Racial/Ethnic Group* | *Sex* | *Sample Size* | *Multiple Choice* | | | *Constructed Response* | | |
|---|---|---|---|---|---|---|---|---|
| | | | *Mean* | *S.D.* | *Rel.* | *Mean* | *S.D.* | *Rel.* |
| White | Male | 27,613 | 51.08 | 15.24 | .878 | 12.31 | 4.16 | .522 |
| | Female | 25,202 | 45.91 | 15.31 | .879 | 12.27 | 4.04 | .493 |
| Asian American | Male | 3,132 | 52.08 | 16.29 | .893 | 12.63 | 4.25 | .544 |
| | Female | 2,805 | 47.89 | 16.07 | .890 | 12.67 | 4.21 | .535 |
| Black | Male | 970 | 41.23 | 16.85 | .900 | 10.42 | 4.14 | .519 |
| | Female | 1,440 | 36.53 | 16.62 | .897 | 10.27 | 4.29 | .552 |
| Mexican American | Male | 420 | 46.28 | 16.59 | .897 | 11.13 | 4.00 | .485 |
| | Female | 509 | 38.15 | 15.97 | .889 | 10.52 | 4.06 | .498 |
| Other Hispanic | Male | 484 | 48.21 | 16.36 | .894 | 12.05 | 4.15 | .520 |
| | Female | 462 | 42.89 | 15.83 | .887 | 11.75 | 3.99 | .482 |

### Table A-2. Multiple-Choice and Constructed-Response Summary Statistics by Sex and Racial or Ethnic Group for AP Biology

| *Self-Reported Racial Ethnic Group* | *Sex* | *Sample Size* | *Multiple Choice* | | | *Constructed Response* | | |
|---|---|---|---|---|---|---|---|---|
| | | | *Mean* | *S.D.* | *Rel.* | *Mean* | *S.D.* | *Rel.* |
| White | Male | 8,985 | 62.08 | 19.77 | .918 | 17.55 | 8.58 | .761 |
| | Female | 2,805 | 55.62 | 19.83 | .919 | 16.19 | 8.72 | .768 |
| Asian American | Male | 1,585 | 66.24 | 21.36 | .930 | 20.91 | 9.06 | .785 |
| | Female | 1,406 | 59.57 | 20.39 | .923 | 19.02 | 8.83 | .773 |
| Black | Male | 330 | 48.29 | 22.96 | .940 | 13.27 | 9.06 | .785 |
| | Female | 616 | 40.59 | 20.77 | .926 | 11.60 | 8.03 | .726 |

### Table A-3. Multiple-Choice and Constructed-Response Summary Statistics by Sex and Racial or Ethnic Group for AP Chemistry

| *Self-Reported Racial Ethnic Group* | *Sex* | *Sample Size* | *Multiple Choice* | | | *Constructed Response* | | |
|---|---|---|---|---|---|---|---|---|
| | | | *Mean* | *S.D* | *Rel.* | *Mean* | *S.D.* | *Rel.* |
| White | Male | 7,014 | 36.40 | 16.78 | .925 | 26.82 | 12.99 | .794 |
| | Female | 3,075 | 29.50 | 15.44 | .911 | 22.84 | 12.50 | .772 |
| Asian American | Male | 1,579 | 39.59 | 16.92 | .926 | 29.48 | 12.76 | .788 |
| | Female | 770 | 34.40 | 15.73 | .914 | 26.72 | 12.10 | .763 |

### Table A-4. Multiple-Choice and Constructed-Response Summary Statistics by Sex and Racial or Ethnic Group for AP English Language and Composition

| *Self-Reported Racial Ethnic Group* | *Sex* | *Sample Size* | *Multiple Choice* | | | *Constructed Response* | | |
|---|---|---|---|---|---|---|---|---|
| | | | *Mean* | *S.D.* | *Rel* | *Mean* | *S.D.* | *Rel.* |
| White | Male | 5,675 | 37.13 | 10.78 | .867 | 14.24 | 3.44 | .551 |
| | Female | 7,878 | 35.32 | 10.30 | .854 | 14.19 | 3.20 | .482 |
| Asian American | Male | 641 | 36.43 | 11.10 | .874 | 13.89 | 3.53 | .573 |
| | Female | 757 | 34.40 | 11.16 | .876 | 14.36 | 3.35 | .523 |