DOCUMENT RESUME

ED 400 330 TM 025 735

AUTHOR Bridgeman, Brent; McHale, Frederick
TITLE Gender and Ethnic Group Differences on the GMAT Analytical Writing Assessment.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-96-2
PUB DATE Feb 96
NOTE 35p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Admission (School); Asian Americans; Black Students; College Entrance Examinations; *Cultural Differences; *Ethnic Groups; Higher Education; Hispanic Americans; Limited English Speaking; *Minority Groups; Racial Differences; *Sex Differences; White Students; *Writing Tests
IDENTIFIERS Analytical Tests; *Graduate Management Admission Test; Latinos

ABSTRACT

Gender and ethnic group differences on the Analytical Writing Assessment that is part of the Graduate Management Admissions Test were evaluated. Data from the first operational administration for 36,583 examinees in October 1994 were used. Standardized differences from the White male reference group were computed separately for men and women in four ethnic groups: (1) White; (2) Asian American; (3) African American; and (4) Hispanic/Latino. Within the White, African American, and Hispanic/Latino groups, women received higher scores than men on the Analytical Writing Assessment; in the Asian American group, men received higher writing scores, but the difference was not as great as on the Verbal score. Examinees whose best language was not English scored relatively higher on the Analytical Writing Assessment than on the Verbal measure. Simulations of eligibility for an admissions pool suggested that the addition of the Analytical Writing score would noticeably increase the number of women in the pool, but would have virtually no impact on ethnic minorities. Rater and score reliability were reasonably consistent across ethnic and gender groups. (Contains 4 tables, 9 figures, and 11 references.) (Author/SLD)

RESEARCH REPORT



# GENDER AND ETHNIC GROUP DIFFERENCES ON THE GMAT ANALYTICAL WRITING ASSESSMENT

**Brent Bridgeman**
**Frederick McHale**



**Educational Testing Service**
**Princeton, New Jersey**
**February 1996**

# Gender and Ethnic Group Differences on the GMAT Analytical Writing Assessment

Brent Bridgeman and Frederick McHale

January, 1996

## Abstract

Gender and ethnic group differences on the Analytical Writing Assessment that is part of the Graduate Management Admissions Test were evaluated. Data from the first operational administration in October of 1994 were used. Standardized differences from the White male reference group were computed separately for men and women in four ethnic groups: White, Asian American, African American, and Hispanic/Latino. Within the White, African American, and Hispanic/Latino groups, women received higher scores than men on the Analytical Writing Assessment; in the Asian American group, men received higher writing scores, but the difference was not as great as on the Verbal score. Examinees whose best language was not English scored relatively higher on the Analytical Writing Assessment than on the Verbal measure. Simulations of eligibility for an admissions pool suggested that the addition of the Analytical Writing score would noticeably increase the number of the women in the pool, but would have virtually no impact on ethnic minorities. Rater and score reliability were reasonably consistent across ethnic and gender groups.

## Gender and Ethnic Group Differences on the GMAT Analytical Writing Assessment

Beginning in October 1994, the Graduate Management Admissions Test (GMAT) included an Analytical Writing assessment. This assessment was designed to assess productive communication and analytical skills that cannot be assessed directly with multiple-choice questions. The writing assessment consists of two separately timed writing tasks. The performance of women and minority groups on this new assessment is of particular interest. An agreement between the College Board et al. and the State of New York (College Entrance Examination Board, et al. against Mario Cuomo, et al. in the U. S. District Court [Northern District of New York]) stipulated that "GMAC shall conduct a study and prepare a report concerning the performance of women and members of minority groups on essay questions measuring developed and other skills." This report is the first step in establishing the psychometric characteristics of the new writing assessment for women and minorities, and it is but one component in a broader on-going effort to better understand the role of test scores and other information in admissions to graduate management programs.

During the first year of implementation of the new writing assessment, criterion performance indicators (e.g., grades in graduate management courses or drop-out rates) on the students who took the test in October of 1994 will not yet be available. Although analysis of such data will ultimately be necessary (and will be conducted), a considerable amount of useful information concerning the functioning of the new test for women and minority groups can be collected during the first year. This report focuses on data from the October 1994 administration.

One set of analyses addresses mean differences among various subgroups on the new writing scale. Merely noting the score of each gender/ethnic group is of little interest without a frame of reference showing how these groups perform on other measures; the multiple-choice scores provide a useful comparative framework. Previous research suggests that essay assessments produce smaller gender differences than do multiple-choice tests. Evidence from Great Britain (Murphy, 1982), Australia (Bell & Hay, 1987), and Ireland (Bolger & Kellaghan, 1990) consistently indicates that males have a relative advantage, on average, on multiple-choice tests. Mazzeo, Schmitt, and Bleistein (1993) found a similar male advantage on several of the Advanced Placement (AP) examinations that are taken by high school students who are seeking college credit or placement into advanced college courses based on college-level courses that they have completed. In several different subject areas, average scores of males and females were nearly equal on the essay portion of the examination while males had significantly higher average scores on the multiple-choice portion. This difference remained even after correcting for the differential reliability of the two question types, and removing items from the multiple-choice test on which males did particularly well had very little impact on the observed gender differences. Differences were especially striking on the United States History examination, with estimated true score means for males and females essentially equivalent on the essays (difference of less than .02 in standard deviation units) but with the mean for males more than .3 standard deviation units higher than the mean for females on the multiple-choice portion of the test.

Less evidence is available regarding ethnic group differences on essay tests in comparison with multiple-choice tests. A recent study of the AP examination in U. S. History (Bridgeman & Morgan, 1994) suggests that ethnic group differences and gender differences should be

considered together; within each major ethnic group studied (White, African American, Asian American, and Latino) men received higher standardized scores on the multiple-choice questions than on the essays, with women in each ethnic group showing the opposite pattern. The same pattern held among students who reported that English was not their best language. The difference was quite small for men, but women whose best language was not English scored substantially higher on the essay than on the multiple-choice portion of the examination. Students who are not native speakers of English might be expected to have more difficulty with an essay examination. However, on an essay, examinees can avoid use of unfamiliar vocabulary and grammatical structures; on a multiple-choice test, failure to understand nuances of structure and vocabulary can lead to incorrect answers.

Although the multiple-choice scores provide a useful reference point, it is important to remember that, by design, the essays are measuring a somewhat different construct. If subgroups show different patterns of relative strengths and weaknesses on multiple-choice and essay tests, it does not necessarily mean that either question format is inherently unfair or biased. It may simply mean that groups differ in the type of test activities on which they perform best. Nevertheless, knowledge of such variations in relative strengths of particular groups provides a useful first step in helping admissions officers make better informed decisions.

When criterion data become available, it will be possible to evaluate the extent to which the essay tests over- or under-predict criterion performance for various subgroups, that is, whether women and minorities perform better or worse than would be predicted from their test scores. Although such studies are useful, they must be interpreted very cautiously (Linn, 1984; Humphreys, 1987). Given that test and criterion scores are imperfectly correlated, regression

effects make it virtually inevitable that a reasonable test will overpredict for groups that score low on the criterion and underpredict for groups that are high on the criterion. If the correlation of test and criterion were .5 (which is realistic), two groups would have to differ twice as much (in standard score units) on the predictor as on the criterion in order for the regression model to show no "bias."[1] As Humphreys (1987) has suggested, elimination of this type of "bias" is impossible as a general rule and should not be a goal of test producers. Thus, the straightforward analysis of mean differences between groups on the new measure (in the context of how much these groups typically differ on other measures) may be more indicative of an everyday understanding of test bias than a regression study. Of course, a regression study is still needed to determine whether the new measure predicts as accurately for minority groups and women as for White men.

The probable lower reliability of the essay scores compared to the multiple-choice scores will tend to attenuate group differences more on the essays. Thus, a particular ethnic/gender group might appear to be further below average on the multiple-choice questions than on the essays only because the essay scores are less reliable; if the essay score were made more reliable (perhaps by including more essays on the test) the pattern of relative strengths could reverse. Scores adjusted to take account of such reliability differences are sometimes called true scores.

---

[1]The regression model of test bias was proposed by Humphreys (1952) but is usually known as the Cleary (1968) definition. This model proposes that for an unbiased test the groups being compared should have equal variances of errors of prediction, equal slopes of regression lines, and equal intercepts of the regression lines. Although this model may work well when two groups have identical means on the criterion, when groups differ on the criterion they will be regressing to different means. Under these conditions (different means on the criterion), tests or measuring instruments that would seem to be unbiased in the everyday use of the word would still be biased by the Cleary definition. For example, a tape measure could be shown to be a "biased" measure of height. As long as men are taller than women on average, a tape measure will underpredict height for men and overpredict it for women. If the tape measure were not very accurate (say it measured to only the nearest inch), the under- or overprediction could be quite noticeable.

The estimated true score differences are useful for estimating how much groups may differ in the underlying skills represented by the different question formats. However, from a practical point of view, the unadjusted scores may be more relevant because decisions about individuals must be based on observed scores and not on unobserveable true scores.

## Method

Subjects

Subjects were the 36,583 examinees who took the GMAT in October of 1994 and who provided information on their gender and population subgroup (only U.S. citizens were asked to provide subgroup information). Specifically, in the directions for the registration form, they were asked to "fill in the space for the group to which you belong." The choices were: 1) American Indian/Alaskan Native/Other native American group, 2) Asian/Asian American, 3) Black/African American, 4) White (non-Hispanic), 5) Mexican American/Chicano, 6) Puerto Rican, 7) Other Hispanic, Latin American, or Latino, 8), Other. Because of relatively small sample sizes, Groups 5 and 7 were combined to form a single Hispanic/Latino group. Group 6 (Puerto Rican) was not included in the Hispanic/Latino group because the mean test score of this small group (352 examinees) was significantly below that of the other Hispanic/Latino groups, and it was the only group with a substantial majority of examinees whose best language was not English. Group 1 was excluded because it contained only 230 examinees. Thus, sample sizes for the 4 subgroups used in the analyses were substantial (see Table 1). Although these subgroup classifications are widely used, within-group homogeneity should not be implied. Substantial variation in both

socioeconomic and test score variables exists within each of these subgroups, and they presumably also differ on a number of other variables that were not measured.

Measures

The Verbal (V) and Quantitative (Q) scores on the GMAT were based on five-choice multiple-choice questions that were scored with a formula that corrects for random guessing (one point for a correct answer, minus one-fourth of a point for a wrong answer, and no points for an unanswered question). The V and Q scores were reported on scales that range from 0 to 60, with a score of 30 defined as average based on a baseline group that took the test in 1954. A Total score (based on the sum of the 61 Verbal and 52 Quantitative questions) was reported on a scale ranging from 200 to 800 with 500 representing the average score in the 1954 baseline group.

The Analytical Writing score was based on two 30-minute writing tasks. One task called for the analysis of an issue and the other task called for the analysis of an argument. Each task was rated on a scale from 1 (fundamentally deficient) to 6 (outstanding), with a score of 0 assigned to papers that were totally illegible or obviously not written on the assigned topic. According to the scoring guide, an outstanding paper on the analysis of an issue topic "presents a cogent, well-articulated analysis of the complexities of the issue and demonstrates mastery of the elements of effective writing; an outstanding paper on the analysis of an argument topic "presents a cogent, well-articulated critique of the argument and demonstrates mastery of the elements of effective writing." Sample topics and a copy of the scoring guide are provided in the Bulletin of Information that is given to all registrants for the examination.

Two readers independently rated the first essay task, and a different two readers independently rated the second task. For both tasks, if rater 1 and rater 2 disagreed by more than

one point, a third rater was used. The third rater's score was averaged with the other score that was closest to it, and the remaining score was dropped. If the third rater exactly split the difference between the scores assigned by raters 1 and 2, the third rater's score was used. The 0-6 scores from the two raters were averaged and multiplied by 10 to form the score for each essay task. The scores on the two tasks were then averaged and rounded to the nearest multiple of 5 to form the Analytical Writing score that was reported to examinees and institutions.

Procedures

In order to facilitate comparisons among test scores that may have different means and standard deviations, gender and ethnic differences are presented as standardized differences. That is, the difference is expressed in standard deviation units. A standardized difference (d) of .5, for example, would mean that the score of the specified group was half of a standard deviation above the mean of the comparison group. The equation used was:

$$d = \frac{M_{sg} - M_{wm}}{\sqrt{\frac{S_{sg}^2 + S_{wm}^2}{2}}}$$

The test score mean ($M$) for the White male ($wm$) group was subtracted from the mean for the particular subgroup of interest ($sg$), so that positive numbers indicate higher scores for the subgroup and negative numbers indicate higher scores for White males. The denominator for $d$ may be computed in a number of different ways (see Stanley, 1992); the advantage of the unweighted average variance ($S^2$) approach used here is that it is independent of subgroup sample size.

A supplementary analysis compared examinees whose best language was not English (according to their self-reports) with examinees whose best language was English. For this analysis, subjects from all domestic ethnic groups (including American Indian, Puerto Rican, and "other") were used. Instead of using a White male comparison group, standardized differences were computed relative to a male English-best group.

Differences corrected for unreliability (or true score differences) were estimated by dividing the standardized difference (*d*) by the standard deviation of the true scores, as estimated by the square root of the reliability, separately for each of the four scores (Analytical Writing [AW], Verbal, Quantitative, and Total). Reliability estimates were based on KR-20 (adapted for use with formula scores) for the Verbal , Quantitative, and Total scores, and on the correlation between the two tasks for Analytical Writing. The weighted average of the reliabilites in separate subgroups is sometimes used to make the true score adjustments (e.g., Mazzeo, Schmitt, & Bleistein, 1993). However, because the White group is by far the largest in the GMAT population, it dominates weighted averages. Furthermore, reliabilities are fairly consistent across subgroups, although within-group reliabilities tend to be lower than reliabilities in the population as a whole. Therefore, we used reliability estimates from the White subgroup which were .53, .82, .80, and .88 for AW, Verbal, Quantitative, and Total respectively. These are all based on lower-bound reliability estimates; actual reliabilities may be higher, in which case the adjustments would be smaller.

Although adjusted and unadjusted standard score differences may be of some theoretical interest, their direct impact on admissions decisions is not immediately apparent. In order to demonstrate how the inclusion or exclusion of the Analytical Writing score could impact on the

ethnic/gender composition of a class, we established a hypothetical admissions situation in which only the top 20% (or top 50%) of the applicants would constitute the eligible pool. The top 20% (or 50%) was established first by the Verbal (V) and Quantitative (Q) scores alone, then by adding the college undergraduate grade point average (UGPA) or the new Analytical Writing score (AW) to the V and Q scores, and finally by a combination of V, Q, UGPA, and the AW score. UGPA was multiplied by 20 to give it a standard deviation that was comparable to the other scores[2], and the composites were then formed by a simple addition of the scores. When criterion data become available, other weighting schemes may be used, but for the current illustrative purposes, this approximately equal weighting scheme is sufficient.

Analyses of reliability differences

In addition to considering the mean differences among groups, possible reliability differences across groups were also investigated. Two types of reliability were of primary concern, rater reliability and score reliability. Rater reliability refers to the consistency with which two independent readers assign scores to the same task; score reliability refers to the consistency of scores from one question (or prompt) to another.

Because each candidate writes two essays and each essay is read by two readers, reliability estimates should be straightforward. However, to the extent that the two tasks are not designed to be strictly parallel in content, the score reliability estimate will be a lower bound (i.e., the actual reliability will be no lower than the estimate). Previous research suggests that score reliability is likely to be a greater concern than rater reliability (Linn, Baker, & Dunbar, 1991).

[2]Standard deviations were 8.4, 9.1, 9.7, and 8.8 for V, Q, AW, and UGPA respectively.

## Results

Table 1 shows the sample sizes (*n*), means (*M*) and standard deviations (*SD*) for both AW tasks and for the total Analytical Writing score (AW Total) for males and females in each ethnic group. For comparison, means and *SD*s for the V, Q and Total (V and Q) scores are also included.

Figure 1 shows the standardized differences (d) for White females compared to the White male comparison group. (Because results were similar for the two AW tasks, and because only the AW Total is reported to individuals or institutions, only the AW Total [labeled simply AW] is presented in the Figures.) Consistent with the findings for other tests cited above, women had a slight advantage on the writing task relative to their multiple-choice Verbal scores. Indeed, it was an absolute advantage, not just a relative one; White women scored slightly higher than White men on the writing tasks.

In the Asian American sample in Figure 2, scores were relatively higher for AW than for Verbal, though the difference was quite small. (For scores below the White male mean, shorter bars correspond to relatively higher scores.) Asian American women did not score higher than Asian American men on the AW score, but the gender difference was smaller on the AW score than on the Verbal score.

The relative superiority of women on the AW score also is evident in the African American sample in Figure 3. On average, both men and women in this subgroup scored relatively higher on the AW essays than on the Verbal measure, and African American women scored higher than African American men on AW, but lower on Verbal.

As shown in Figure 4, the pattern in the Hispanic/Latino group was essentially the same as in the White and African American samples; women scored higher than men on AW but lower on Verbal. Overall, AW scores were closer to the White male mean than the other scores.

The differences corrected for reliability differences between the scores are presented in Figures 5-8. The relative advantage of the AW score over the Verbal score for women is still evident for all subgroups. However, for men, the size of the ethnic group differences in the constructs assessed by writing tasks appear to be virtually identical to the differences found in the multiple-choice Verbal score.

Figure 9 shows the analysis for the examinees for whom English is not their best language. Consistent with previous findings (Bridgeman & Morgan, 1994), the essay format does not seem to disadvantage students whose best language is not English, relative to their performance on the Verbal section.

Table 2 shows the ethnic/gender composition for a pool of candidates for admission for a hypothetical program that considers only the top 20% of the applicants. If the top 20% were determined only on V and Q GMAT scores, the pool would be only 28% female. If GMAT Analytical Writing (AW) scores and UGPA were also considered (with each of the four scores given roughly equal weight), the pool would be 36% female. Despite this rather dramatic impact on the gender composition of the eligible pool, note that the ethnic composition would be virtually identical for any of the four sets of predictors.

As indicated in Table 3, selecting a pool from the top half, rather than from the top 20%, had a surprisingly small impact on the percentage from each ethnic group in the pool, although the percentage of women increased somewhat Indeed, the 39% female representation in this pool

selected using all four scores approached the female representation in the total unselected initial pool of applicants, which was 41% female.

Rater reliability and score reliability estimates are in Table 4. Rater reliability is the average of the rater 1 and rater 2 correlation for tasks 1 and 2 (adjusted by Spearman-Brown because two raters are used[3]). Score reliability is the correlation of task 1 and task 2, adjusted by Spearman-Brown. The score reliability estimate includes both raters and topics as potential sources of error. Both rater reliability and score reliability were slightly higher in the other three ethnic subgroups than in the White subgroup. To some extent, these higher reliabilities could reflect the greater variability of scores in the non-White ethnic groups. The standard error of measurement, which takes the within-subgroup score variability into account, suggested that measurement errors are fairly comparable across all subgroups. Gender differences in reliability (or measurement error) within each subgroup were trivial.

## Discussion

Consistent with previous findings in other testing programs (e.g., Mazzeo, Schmitt, & Bleistein, 1993), women appeared to perform relatively well with a format that requires written responses. Indeed, in three out of the four ethnic groups studied (White, African American, and Hispanic/Latino), women had higher Analytical Writing scores than men; in the fourth subgroup (Asian American), men had slightly higher scores, though their advantage, in standard score units, was less on the Analytical Writing score than on either the Verbal or Quantitative scores.

[3]Because of the use of a third reader to mediate score discrepancies of more than one point, overall rater reliability is slightly higher than that estimated from the simple correlation of rater 1 and rater 2. For 91% of the essays, the scores of raters 1 and 2 were within one point, so a third reader was needed for only 9% of the essays.

Including the Analytical Writing score as part of an admissions screening battery would substantially increase the number of women in the eligible pool, but would have only a minimal impact on the number of ethnic minorities in the pool. Rater and score reliability for the Analytical Writing measure were at least as high for the other ethnic/gender subgroups as for White males.

## References

Bell, R.C., & Hay, J.A. (1987). Differences and biases in English language examination formats. British Journal of Educational Psychology, 57, 212-220.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. Journal of Educational Measurement, 27, 165-174.

Bridgeman, B., & Morgan, R. (1994). Consequences of discrepancies between performance on multiple-choice and essay questions (College Board Report No. 94-5; ETS RR-94-41). New York: College Entrance Examination Board.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115-124.

Humphreys, L. G. (1952). Individual differences. Annual Review of Psychology, 3, 131-150.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. Journal of Applied Psychology, 71, 327-333.

Linn, R. L. (1984). Selection bias: Multiple meanings. Journal of Educational Measurement, 21, 33-48.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20, 15-21.

Mazzeo, J., Schmitt, A.P., & Bleistein, C.A. (1993). Sex-related differences on constructed response and multiple-choice sections of Advanced Placement Examinations (College Board Report No. 92-7; ETS RR-93-5). New York: College Entrance Examination Board.

Murphy, R.J.L. (1982). Sex differences in objective test performance. British Journal of Educational Psychology, 52, 213-219.

Stanley, J. C. (1992). Differences on the College Board Achievement Tests and the Advanced Placement examinations: Effect sizes versus some upper-tail ratios. In N. Colangelo, S. G. Assouline, & D. L. Ambroson (Eds.), Talent development: Proceedings from the 1991 Henry B. and Jocelyn Wallace National Research Symposium on Talent Development (pp. 52-59). Unionville, NY: Trillium Press.

Table 1

Sample Sizes, Means and Standard Deviations of GMAT Scores by Subgroup

| Ethnic Group | Gender | *n* | AW Task 1 | | AW Task 2 | | AW Total | | Verbal | | Quantitative | | Total(V & Q) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *M* | *(SD)* | *M* | *(SD)* | M | (SD) | M | (SD) | M | (SD) | M | (SD) |
| White | M | 14,041 | 39 | (10) | 40 | (9) | 41 | (8) | 31 | (7) | 33 | (8) | 538 | (100) |
| White | F | 9,210 | 39 | (10) | 42 | (8) | 42 | (8) | 30 | (7) | 29 | (8) | 505 | (97) |
| Asian American | M | 1,777 | 33 | (12) | 34 | (12) | 35 | (11) | 26 | (9) | 35 | (9) | 511 | (113) |
| Asian American | F | 1,421 | 32 | (12) | 33 | (12) | 34 | (11) | 23 | (9) | 32 | (8) | 473 | (104) |
| African American | M | 1,088 | 32 | (11) | 35 | (10) | 35 | (9) | 24 | (8) | 24 | (8) | 428 | (104) |
| African american | F | 1,408 | 33 | (11) | 36 | (9) | 36 | (8) | 24 | (7) | 22 | (7) | 405 | (96) |
| Hispanic/Latino | M | 700 | 34 | (11) | 37 | (9) | 37 | (9) | 27 | (8) | 29 | (9) | 485 | (108) |
| Hispanic/Latino | F | 562 | 35 | (11) | 39 | (9) | 38 | (9) | 26 | (7) | 26 | (8) | 451 | (98) |

Table 2

For Top 20% as Defined by Various Score Composites, Percent of Total from Each Ethnic/Gender Group

| Scores in Composite | White M | White F | Asian American M | Asian American F | African American M | African American F | Hispanic/ Latino M | Hispanic/ Latino F | Combined Ethnic M | Combined Ethnic F |
|---|---|---|---|---|---|---|---|---|---|---|
| V+ Q | 61 | 24 | 7 | 3 | 1 | 1 | 2 | 1 | 72 | 28 |
| V + Q+ AW | 58 | 27 | 6 | 3 | 1 | 1 | 2 | 1 | 68 | 32 |
| V + Q + UGPA | 56 | 29 | 7 | 3 | 2 | 1 | 2 | 1 | 66 | 34 |
| V + Q + UGPA + AW | 55 | 31 | 6 | 3 | 1 | 1 | 1 | 1 | 64 | 36 |

Note. - - Percents for ethnic/gender groups do not add to 100% because of examinees not classifed into one of the four ethnic groups studied. These examinees are included in the combined ethnic group.

Table 3

For Top 50% as Defined by Various Score Composites, Percent of Total from Each Ethnic/Gender Group

| Scores in Composite | White | | Asian American | | African American | | Hispanic/ Latino | | Combined Ethnic | |
|---|---|---|---|---|---|---|---|---|---|---|
| | M | F | M | F | M | F | M | F | M | F |
| V + Q | 55 | 28 | 6 | 3 | 2 | 1 | 2 | 1 | 66 | 34 |
| V + Q+ AW | 54 | 31 | 5 | 3 | 2 | 2 | 2 | 1 | 63 | 37 |
| V + Q + UGPA | 51 | 31 | 6 | 4 | 1 | 2 | 2 | 1 | 62 | 38 |
| V + Q + UGPA + AW | 51 | 32 | 5 | 3 | 1 | 2 | 2 | 1 | 61 | 39 |

Note. - - Percents for ethnic/gender groups do not add to 100% because of examinees not classifed into one of the four ethnic groups studied. These examinees are included in the combined ethnic group.

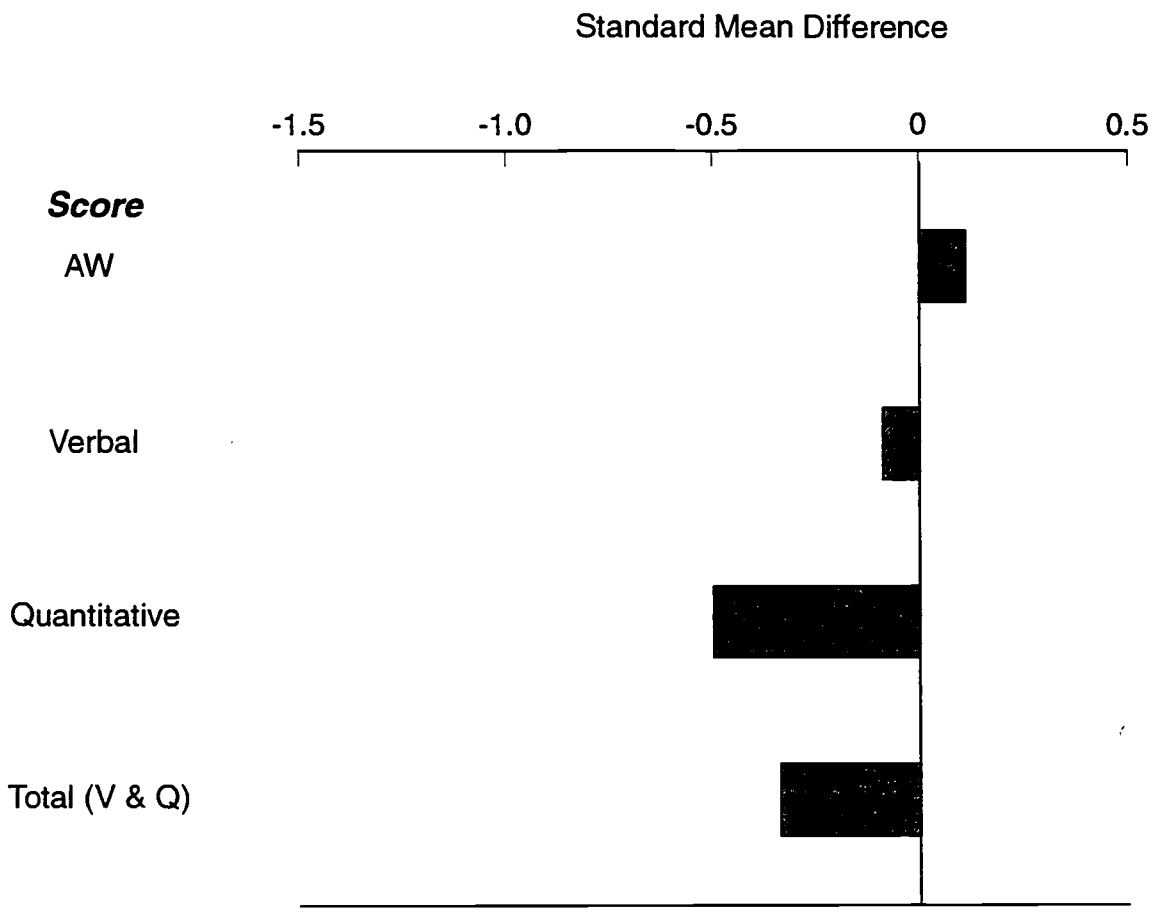
Table 4

Rater and Score Reliability for Each Subgroup

| Subgroup | Rater Reliability | Score Reliability | Standard Error of Measurement |
|---|---|---|---|
| White Male | .75 | .54 | 5.4 |
| White Female | .73 | .51 | 5.4 |
| Asian Am. Male | .85 | .77 | 5.2 |
| Asian Am. Female | .85 | .77 | 5.1 |
| African Am. Male | .80 | .62 | 5.5 |
| African Am. Female | .77 | .56 | 5.5 |
| Hispanic/Latino Male | .77 | .64 | 5.2 |
| Hispanic/Latino Female | .78 | .64 | 5.3 |

Figure 1

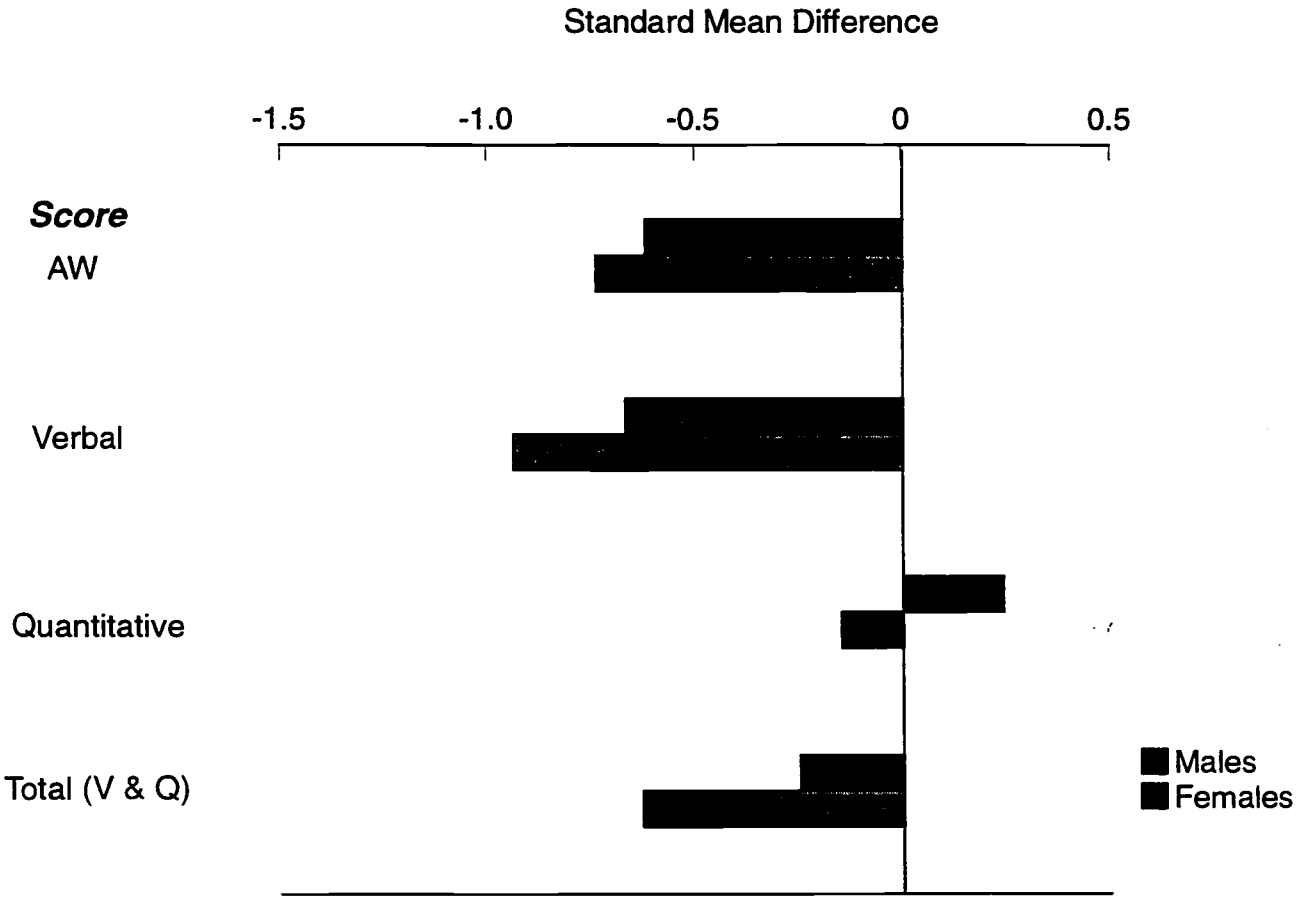
Standardized Difference from White Male Mean for White Female Examinees




n = 9,210 White females;14,041 in White male comparison group; maximum standard error = .01

Figure 2

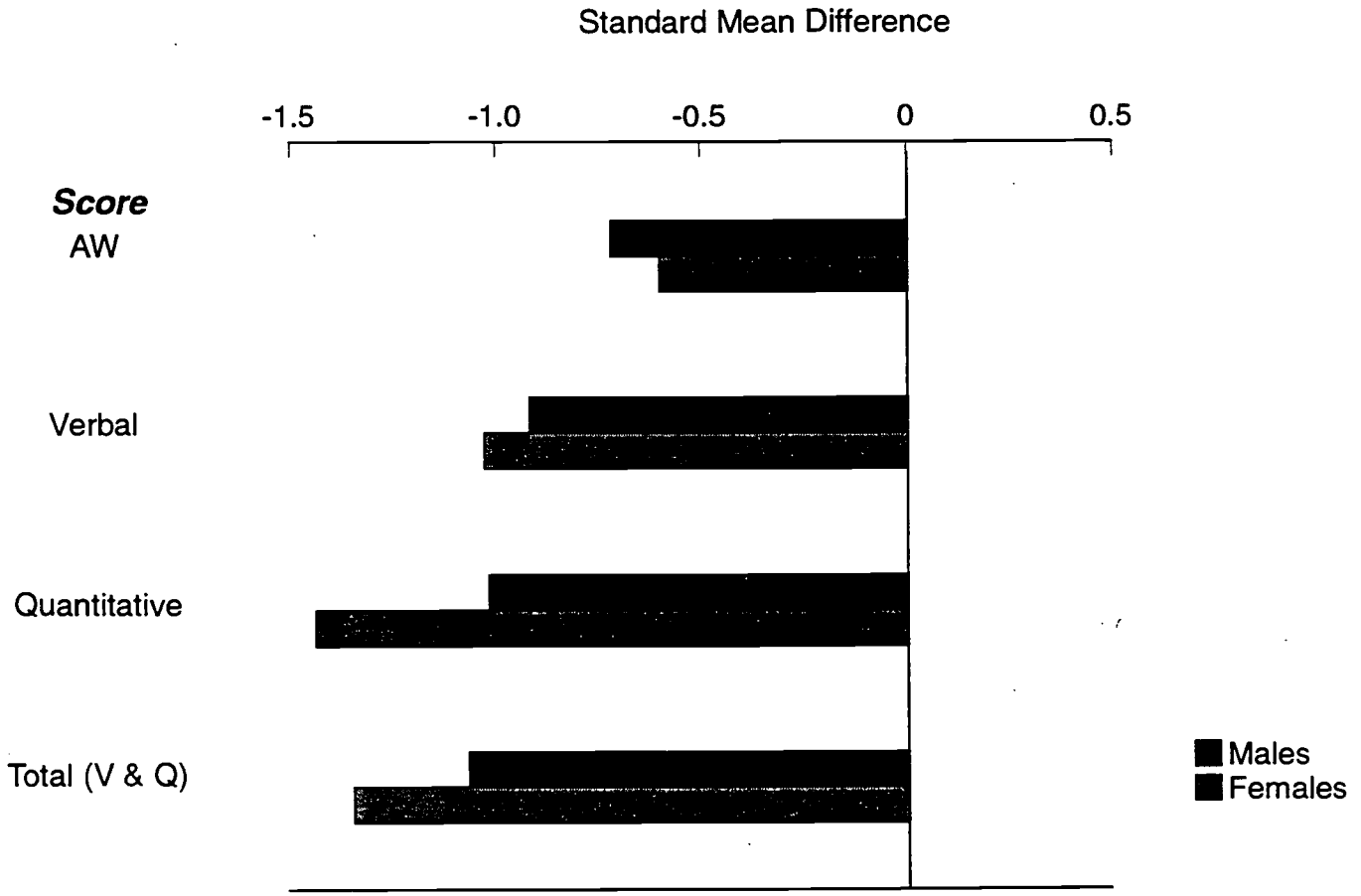
Standardized Difference from White Male Mean for Asian American Examinees




n = 1,777 Male and 1,421 Female; 14,041 in White male comparison group; maximum standard error = .03.

Figure 3

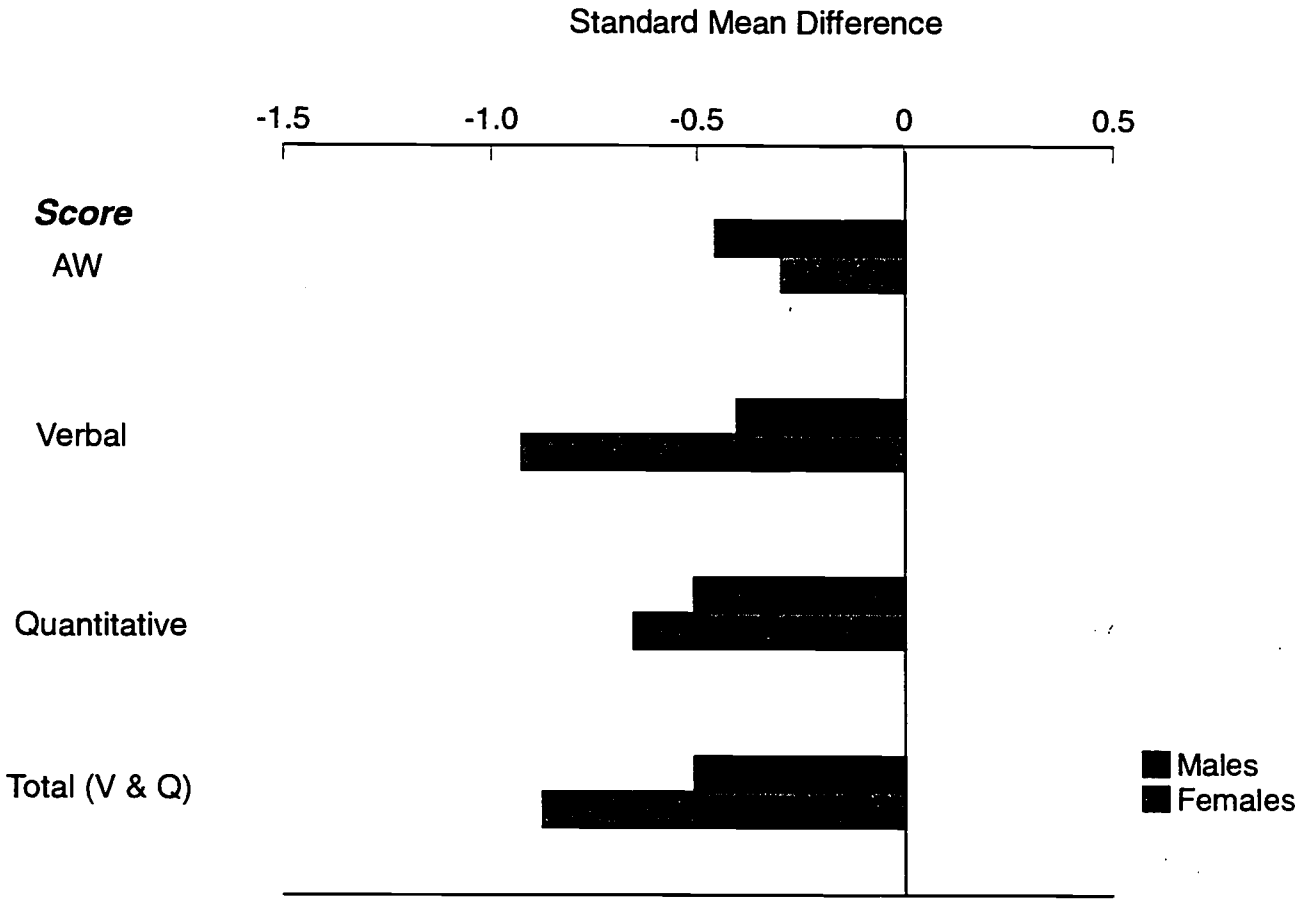
Standardized Difference from White Male Mean for African American Examinees




n = 1,088 Male and 1,408 Female; 14,041 in White male comparison group; maximum standard error = .03.

Figure 4

Standardized Difference from White Male Mean for Hispanic/Latino Examinees




n = 700 Male and 562 Female; maximum standard error = .05.

Figure 5

Estimated True Difference from White Male
Mean for White Female Examinees

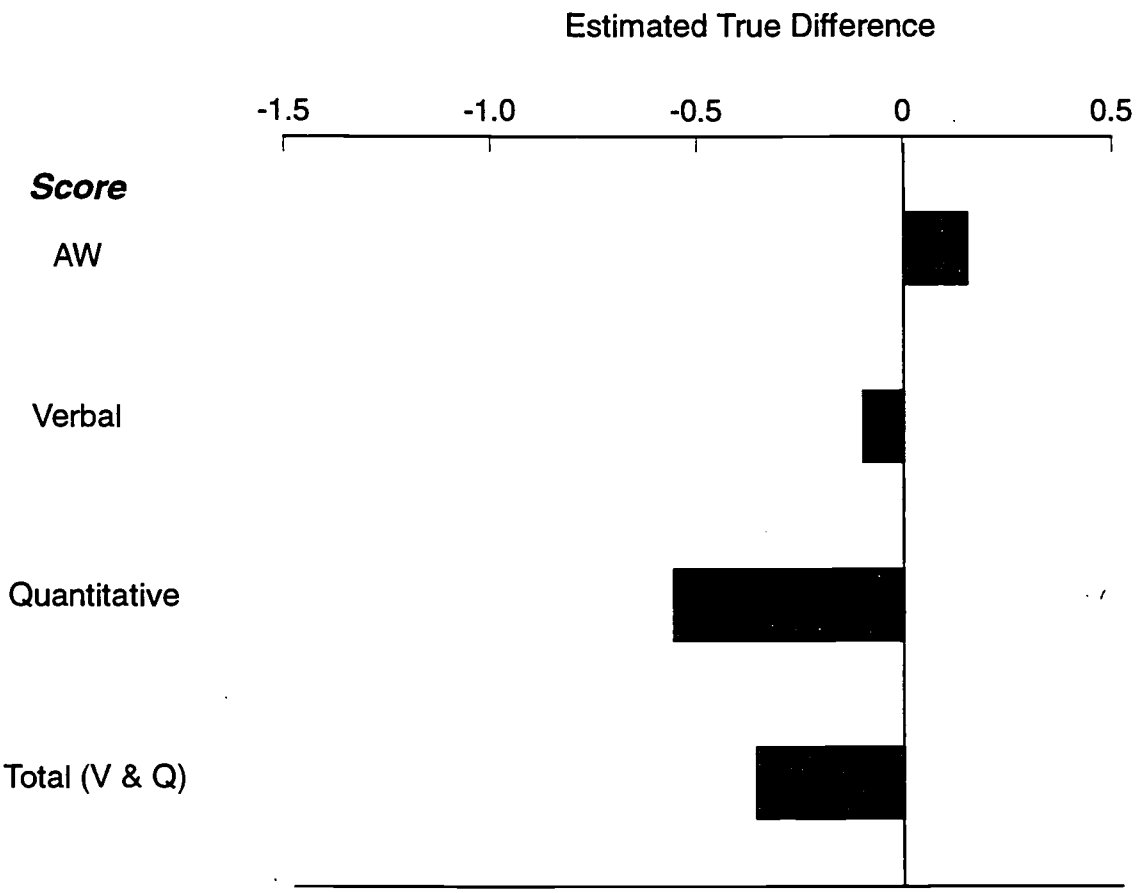


Figure 6

Estimated True Difference from White Male Mean for Asian American Examinees

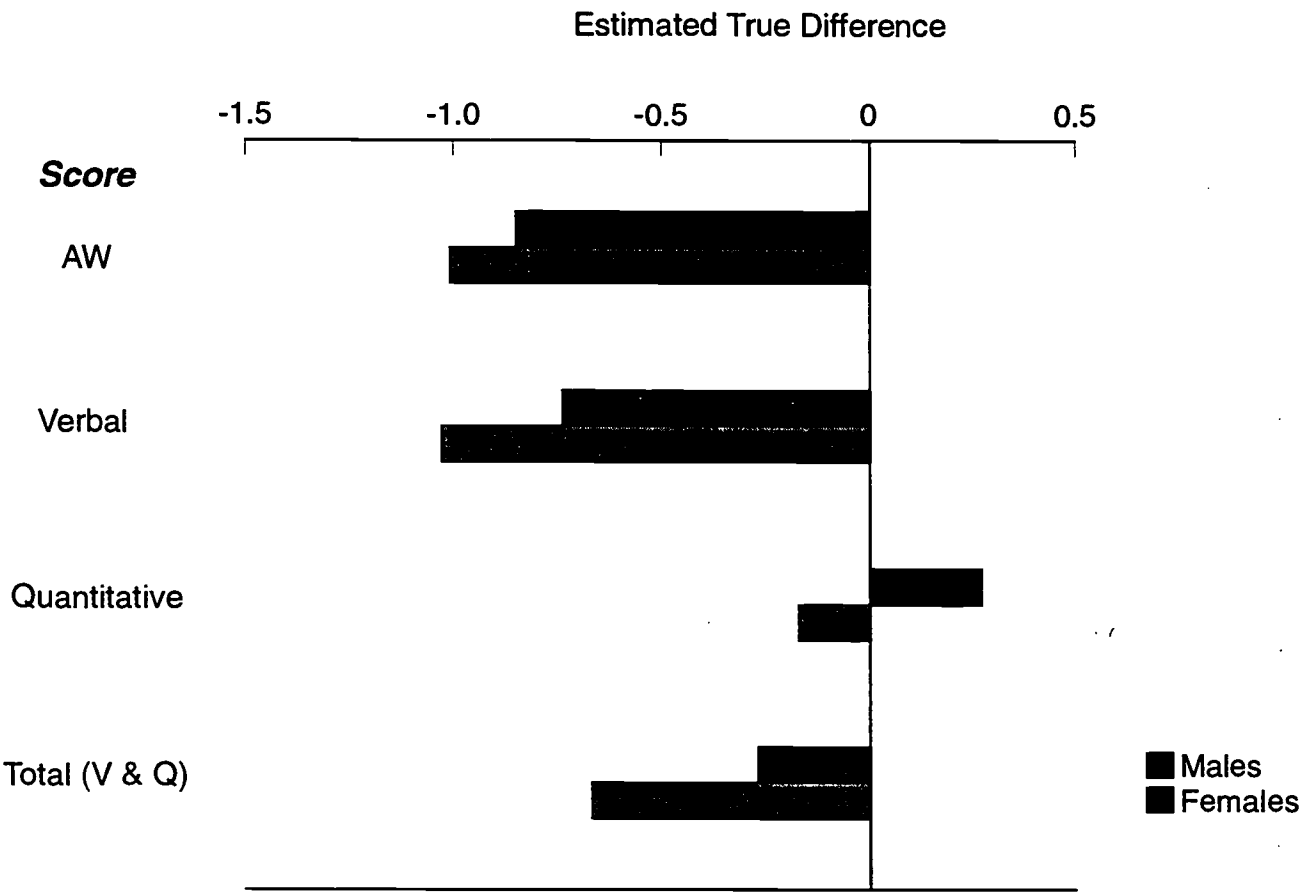


Figure 7

Estimated True Difference from White Male Mean for African American Examinees

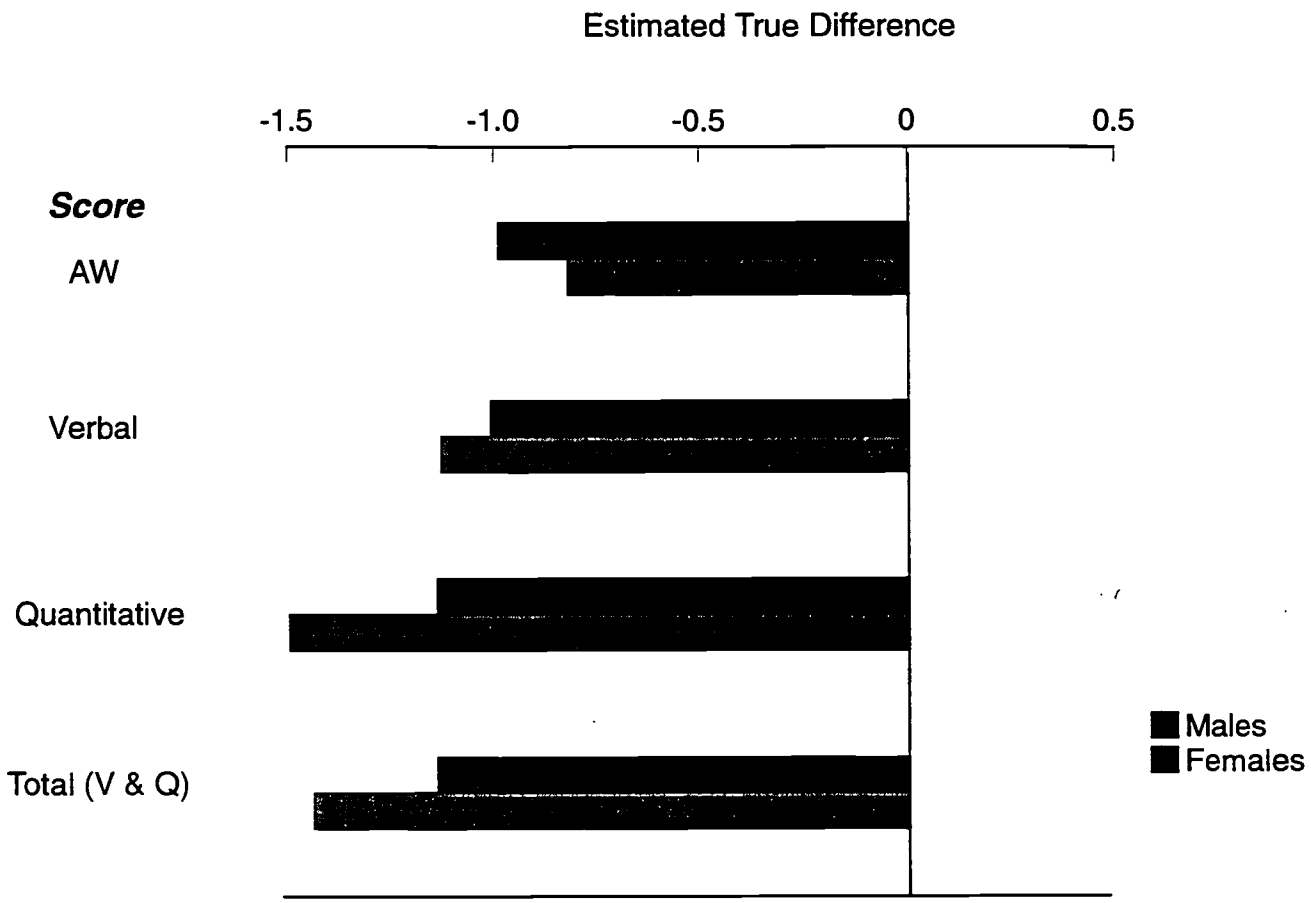


Figure 8

Estimated True Difference from White Male Mean for Hispanic/Latino Examinees

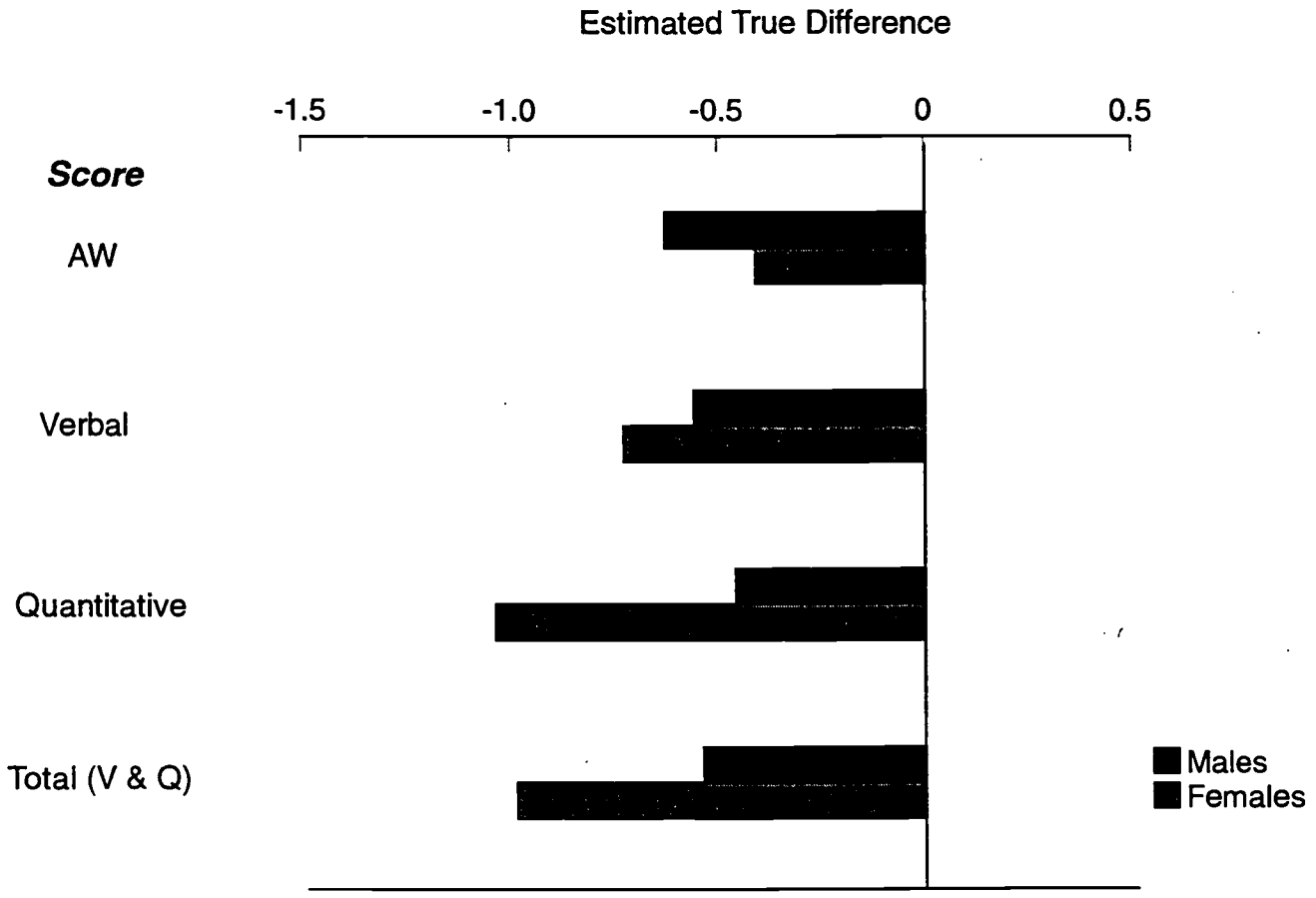
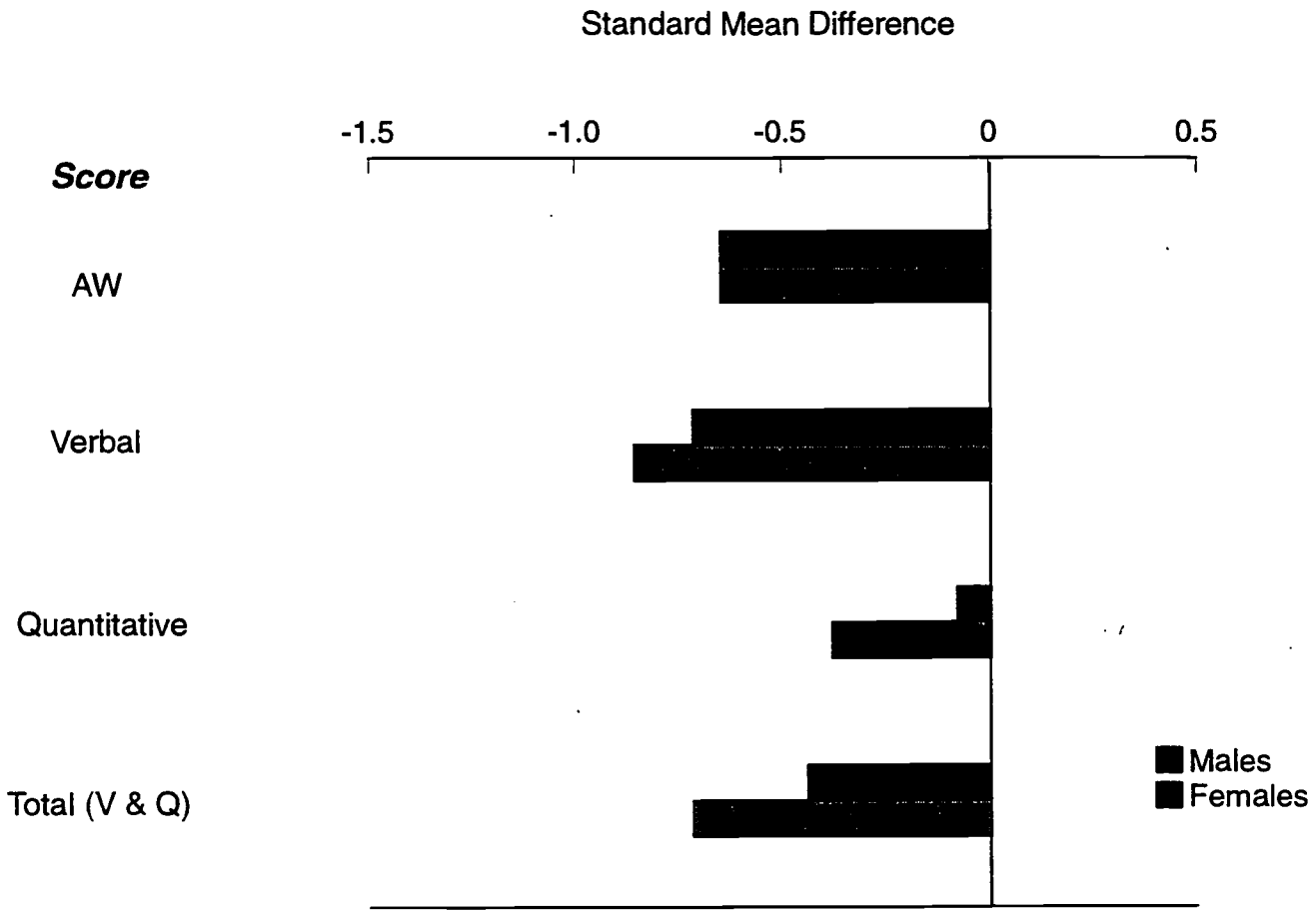
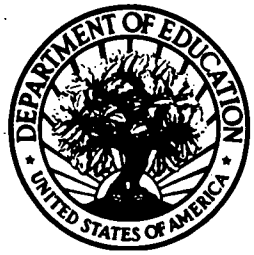


Figure 9

Standardized Differences from English Best Male Mean for English Not-Best Examinees




n = 3,087 Male and 2,193 Female; 16,308 in English Best Male Comparison group; maximum standard error = .02




**U.S. DEPARTMENT OF EDUCATION**
*Office of Educational Research and Improvement (OERI)*
*Educational Resources Information Center (ERIC)*

ERIC®

# NOTICE

## REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").