Sex Differences in Item Responses on the Graduate Record Examination

Thomas F. Donlon, Marilyn M. Hicks, and Madeline M. Wallmark
Educational Testing Service

This study explores the scope and nature of sex differences in the December 1974 aptitude tests of the Graduate Record Examination by identifying individual test items that differ from the other items in terms of the magnitude of the difference in item difficulty for the sexes. The method is that used in earlier studies by Angoff and Stern (1971) and by Strassberg-Rosenberg and Donlon (1975). In general, limited evidence of sex differences was established.

Sex differences are frequently encountered in testing. For example, Keating and Stanley (1972) found relatively few females gifted in mathematics and science. In earlier years these differences were so common, and so widely anticipated and accepted, that Coffman (1961) observed that comparable scores by samples of men and women were likely to be interpreted as evidence of distorted sampling.

In such a climate, analyses of the factors that were associated with differentials in sex performance were seldom undertaken. In recent years, however, there has been a strong interest in the question of sex differences, particularly as they relate to differentiation in social roles and to the principle of equal rights and privileges. A number of aspects of education have been examined, with a natural focus on testing. Tittle, McCarthy, and Steckler (1974) have provided a rational analysis of the problem, with the empirical results of a content analysis of a number of mathematical achievement tests. Lockheed (1973) has commented from a sociological standpoint, stressing the importance of language and of role stereotyping in the established patterns of sex difference.

Coffman (1961) studied the College Board Scholastic Aptitude Test (SAT) of 1954 and Donlon (1973) studied the SAT of 1964. The mathematical test (SAT-M) was analyzed into four categories of items:

1. Subject Content—problems with a real-world referent, either persons or things;
2. Geometric Figures—problems with an associated geometric diagram;
3. Algebraic Unknowns—problems involving unknowns (e.g., x, y) but *not* in the context of geometry or trigonometry; and
4. Other.

The average sex differences in item p-value for these categories were, respectively, .102, .078, .042, and .065. Evaluating these data, Donlon observed that a test constructor would have the power to vary the scaled score difference between the sexes from 20 to 60 points, depending on the proportions of Subject Content and Algebraic Unknown items.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 1 Winter 1980 pp. 9–20

Strassberg-Rosenberg and Donlon (1975) analyzed the SAT of April 1974, using Angoff's (1972) technique. Essentially, this method seeks to determine outliers, items that lie beyond a critical distance from a reference line established by all the items. Considering all such distances as a distribution, a criterion of 1.5 standard deviations was arbitrarily imposed for both the Verbal (SAT-V) and Mathematics (SAT-M) tests. The outlier analysis tended to confirm the earlier Coffman (1961) and Donlon (1973) results. Items most favorable to males were in the categories World of Practical Affairs and Science. Items most favorable to females were in the categories Human Relations and Aesthetic-Philosophical. Antonyms and Reading Comprehension favored females, and Analogies favored males. Sentence Completion, which was unbiased in Donlon's (1973) study, favored males in Strassberg-Rosenberg and Donlon's (1975) study. Such differences among verbal item types, however, were of very slight magnitude.

The analyses of SAT-M confirmed a general wide advantage for males, but a greater relative success for females on Data Sufficiency items, confirming Donlon's (1973) report. (For reasons other than sex bias, e.g., coachability and form, Data Sufficiency items are no longer used in SAT-M). Further, the results indicated that algebra items are relatively easiest for females; geometry items, relatively hardest. This suggests that spatial visualization may be a factor. Indeed, Maccoby and Jacklin (1974) and Stafford (1961) have advanced hypotheses that certain visualization skills may be linked to sex.

In view of the demonstrations concerning the SAT, it was decided to investigate a form of the Graduate Record Examination Aptitude Test (GREAT). This test is made under somewhat different specifications than the SAT but with many clear parallels, including the use of the same four verbal item types. The population of test takers, of course, is older, college experienced, and more highly selected. Nonetheless, questions of parity of performance between the sexes are clearly germane.

Method

Description of the Instrument

The GREAT of December 1974 was administered in three separately timed sections. The first was a 25-minute section containing 55 Vocabulary, or verbal omnibus, items: Antonyms, Analogies, and Sentence Completions. The second was a 50-minute section containing 40 Reading Comprehension items based on 6 passages. The third was a 75-minute section containing 55 Mathematics questions.

There are content classification systems for each item type. For the Vocabulary, or verbal omnibus, items these are Aesthetic-Philosophical, World of Practical Affairs, Science, and Human Relationships. For the Reading Comprehension items, there are content codes for the passage content and for the questions. Passage content differentiates biological science, physical science, argumentative, humanities, social studies, synthesis, and narrative; a minority-centered passage is also prescribed. Individual questions are further classified as (1) understanding of central ideas; (2) understanding of supporting idea; (3) appreciation of intended inference; (4) application of idea beyond the passage; (5) judgment of style or tone; and (6) judgment of author error or illogic.

The Mathematics section is divided between Regular, Mathematics and Data Interpretation. The former are a diverse collection of questions drawing on fundamental algebra and geometry. The latter are, in each case, based upon tables, charts, or graphs and thus constitute a related set of three to five questions.

Procedure

The study contrasted the performance of 1,720 white males and 1,735 white females who took the GREAT in December 1974. These samples were in turn randomly selected from the total population that took the test. A separate item analysis was performed for each of the two groups—male and female. This item analysis is the standard Educational Testing Service proce-

dure; it yields an index of item difficulty, Δ, defined as $\Delta = 4z + 13$, where z is a normal deviate above which lies a proportion of the area under the curve equal to the proportion of examinees answering the item correctly. This proportion is based upon the group reaching the item, so that those not completing the test are not included in the determination of p values for later items. However, the estimates of Δ are adjusted to constitute estimates for the total group. Note that unlike p values, Δ values *increase* with increasing difficulty.

In the Angoff approach, Δ values for the contrasting male and female groups are plotted and typically result in an elliptical plot. The major axis of this ellipse is determined as the line that minimizes the squares of the perpendicular distance of the points to the line. Formulas for this calculation are presented in Angoff and Stern (1971).

The procedure provides a measure of the perpendicular distance (D) of each item point from the line. The standard deviation of the distribution of such differences is a function of item × group interactions; items which depart most extremely from the clusters of points are the major contributors to the interaction. These outliers are the items that fail to conform to the general relationship for item difficulty between the sexes. The definition of outliers is developed in terms of the standard deviation of the distances (D), in terms of the standard scores of this distribution. The decision to use a specific value of these z scores in defining a criterion for an outlier is arbitrary. In this study a level of 1.5 standard deviation units was selected. Such a level avoids undue capitalization on chance factors but should identify differences of practical significance.

It should be noted that the analysis focuses on *specific* interactions. The items may differ in difficulty more or less systematically for the two groups. This kind of difference will be indicated by the parameters of the trend line, the major axis of the ellipse. It is possible to compare the trend lines for different kinds of items and to characterize the different levels of the parameters for different item types. In the present study such comparisons were made for major subgroups of items.

Results and Discussion

The results are presented in terms of the three major subdivisions of the test—Vocabulary, Reading Comprehension, and Mathematics—with certain subanalyses developed by considering subdivisions, groups defined by common characteristics of content or format.

The most intuitive approach compares the sexes with regard to the average item difficulty. Table 1 presents these data within Vocabulary and Mathematics sections for each test section and for the item type distinctions. In general, Vocabulary was very slightly more difficult for females, Reading Comprehension was about at parity, and Mathematics was more difficult for females by about one Δ unit. Within the Vocabulary section Antonyms showed the greatest difference, a bias against females of approximately .15 Δ unit. Within the Mathematics section the Data Interpretation items as a subset showed somewhat greater sex differences than the Regular Mathematics item type. The mathematical differences, as might be expected, favored males.

The standard deviations of the within-sex distributions of item difficulty were closely consistent. Only the values for Reading Comprehension and for Data Interpretation suggested much difference in dispersion. The item types varied widely in their average difficulty for both sexes, with a range of about 3 Δ units for the verbal material (from 10.97 for Reading Comprehension to 14.17 for Analogies) and of about 1.2 Δ units for the two kinds of mathematical material (from 11.31 for Regular Mathematics to 12.54 for Data Interpretation).

A major test of the equivalence of the test material for the two groups is the basic correlation between item difficulties for males and females. Table 2 presents these correlations together with the slope and intercept parameters for the major axis. In general, a very high order

Table 1
Comparisons of Average Item Difficulty

Test	N	Mean	S.D.	Mean	S.D.	Mean Difference
Vocabulary	55	12.94	3.16	13.00	3.15	-0.06
Antonyms	20	12.89	3.29	13.04	3.25	-0.15
Sentences	17	11.71	1.94	11.73	2.10	-0.02
Analogies	18	14.17	3.45	14.16	3.39	+0.01
Reading	40	10.97	1.92	10.96	1.78	+0.01
Mathematics	55	11.63	2.99	12.64	3.05	-1.01
Regular Math	41	11.31	3.02	12.23	2.96	-0.92
Data Interpretation	14	12.54	2.71	13.82	3.00	-1.28

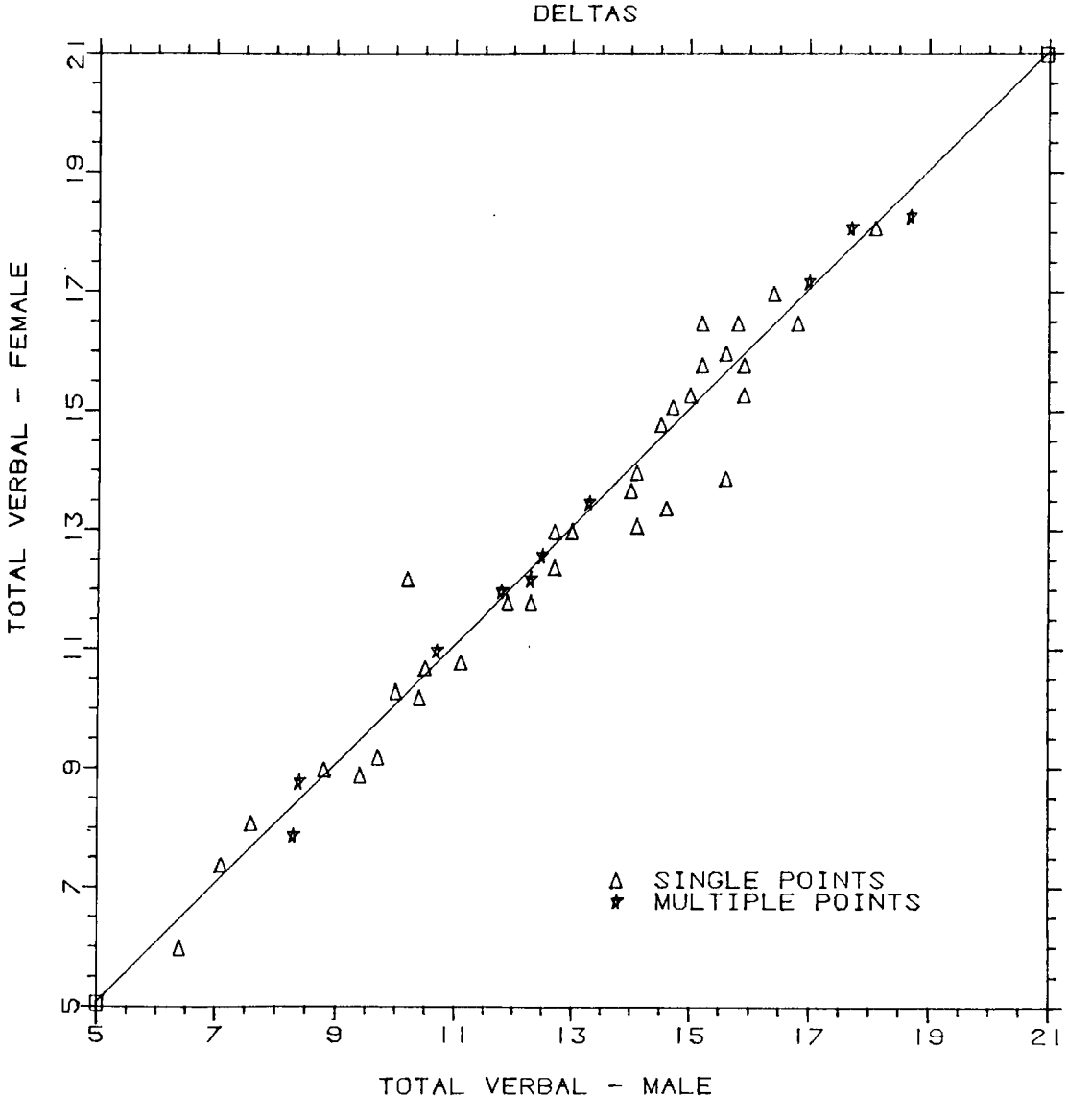
of consistency in item difficulty was observed. The correlations were all above .960, the lowest correlation being in Reading Comprehension with .966. This high order of correlation reflects a consistent relationship such that items were about as difficult or easy relative to each other within the male groups as they were within the female groups. This is best indicated by the slope parameter, which approximated 1.00 in each case. This slope parameter tended to deviate from unity most markedly for Sentence Completion (1.08), for Data Interpretation (1.11), and for Reading Comprehension (.93). Such deviations indicate that group superiority is related to the level of item difficulty, with males doing increasingly better on difficult Sentence Completion and Data Interpretation items and females doing better on more difficult Reading items.

The intercept parameters are more difficult to interpret in terms of sex differences, since they are determined in part by the slope characteristics and cannot be straightforwardly viewed as adjustment constants between the groups. When

Table 2
Parameters of the Bivariate Distribution of Item Difficulty

Test	N	Correlation	Slope	Intercept
Vocabulary	55	.98	1.00	+.08
Antonyms	20	.98	.99	+.30
Sentences	17	.99	1.08	.98
Analogies	18	.98	.98	+.24
Reading	40	.97	.93	+.79
Mathematics	55	.98	1.02	+.78
Regular Math	41	.98	.98	1.14
Data Interpretation	14	.97	1.11	-.10

Figure 1
Δ Plot for Vocabulary Section



the slope is very near unity, of course, the intercept has this interpretation; and for the total set of 55 Vocabulary items with a slope of very nearly 1.00, this indicates that in aggregate they tended to be more difficult for females, confirming the interpretation of the difference in average Δ.

The differences in the characteristics of the lines cannot be linked in any fundamental way to the overt characteristics of the item types. Thus, the parameters for Sentence Completion, 1.0849 and −.9764, were somewhat different from those for Reading Comprehension, .9274 and +.7864, but the item characteristics which determine this result cannot be readily explained. They strongly underscore, however, the need to examine components of tests for sex differences, since the analysis of total item sets tends to obscure more specific differences.

Vocabulary Analysis

Figure 1 presents a plot of the Δ values for the 55 items of the Vocabulary section. Applying the criterion of 1.5 to the distribution of normalized *D*-values, five outliers were identified: Items 5, 6, 16, 34, and 40. Table 3 presents descriptive data on these five items. In general, support for previously reported content factors is obtained. The content of the male-biased material is scientific or practical; the content of the female-biased material is person-oriented. There are, of course, very few cases. The number of outliers is not markedly different from those found in the earlier study of the SAT.

Table 4 presents the average *D*-value for each of the three item types that are combined in the Vocabulary section. These values show that the Antonym items favored the females, while the Analogy and Sentence Completion items favored males. These results are directly contrary to the patterns observed in the SAT analysis by Strassberg-Rosenberg and Donlon (1975). This suggests that item types per se are not linked to relative difficulty for the sexes but that the content of items is the determining factor. Since the item types tend to be controlled for content with an effort to establish content balance, the finding of pattern "swings" from one test to another may indicate only a variation due to item sampling.

Figure 2 presents a plot of the Δ values for the 40 items in the Reading Comprehension section. Applying the criterion of ± 1.5 units to the distribution of normalized *D*-values, six outliers were identified: Items 4, 10, 11, 12, 15, and 17. Table 5 presents descriptive data on these six items. In general, the determiners seem to be a combination of passage characteristics and item characteristics. Thus, five of the six outliers were accounted for by two of the six passages; but within these passages not all items were outliers.

There are no formal distinctions among item types in the Reading Comprehension section in the way that the Vocabulary section and the Mathematics section are differentiated. How-

Table 3
D-Values and Content Code for the Five Vocabulary Outliers

Item	Group Favored	D-Value	Content	Item Type
5	M	−0.88	Practical Affairs	Analogy
6	F	+1.24	Human Relationships	Analogy
16	F	+0.88	Human Relationships	Antonym
34	F	+0.75	Practical Affairs	Analogy
40	M	−1.37	Science	Antonym

Figure 2
Δ Plot for Reading Comprehension Section

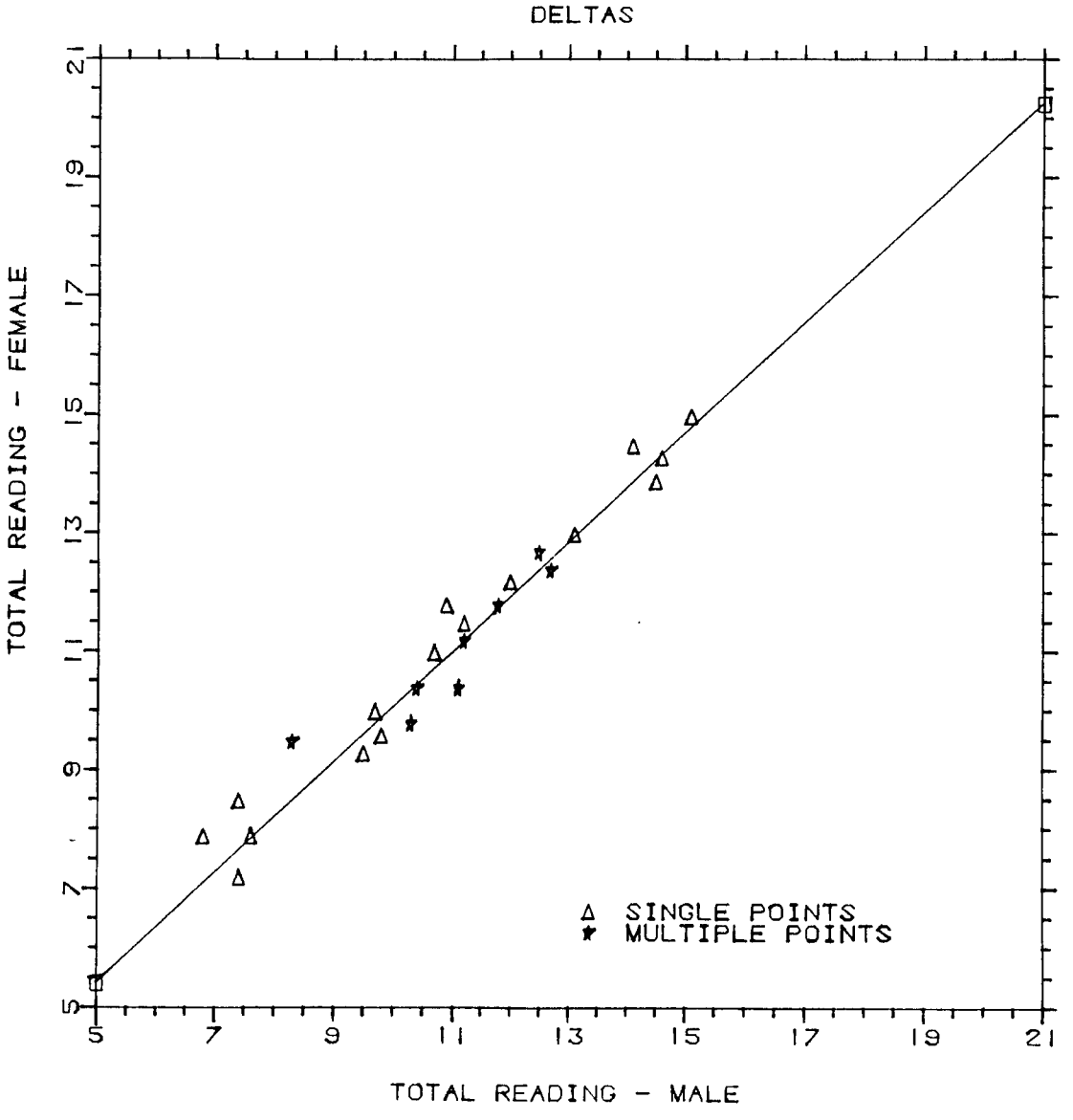


Table 4
Average D-Value by Vocabulary Test

Item Type	N	Average D-Value	Group Favored
Antonyms	20	-.07	M
Analogies	17	+.05	F
Sentences	18	+.03	F

ever, the items are associated with six different passages varying in content characteristics in ways that are logically linked to sex stereotyping. Thus, science passages would not be predicted to be easier for females, while humanities passages would be more difficult for males. Table 6 presents the average *D*-values for the six passages.

Mathematics Test

Figure 3 presents a plot of the Δ values for the 55 items in the Mathematics section. Applying the criterion of the + 1.5 units to the distribution of normalized *D*-values, nine outliers were identified: Items 2, 32, 33, 34, 36, 44, 51, 52, and 55. Three of these—Items 32, 33, and 34—are Data Interpretation items associated with a given graph. The remainder are Regular mathematics: Table 7 presents descriptive data on these nine Mathematics outliers. Five outliers were relatively favorable to females, four to males. Those favoring males were linked to one graph that described changes in personal income in the United States over several years. Of the five female-favoring items, four had the requirement for dealing with algebraic unknowns that was considered by Donlon (1973) to be relatively more favorable to females. However, Item 44, which favored males, also had this requirement.

The two formal item types within the Mathematics section are distinguished in Table 8, and average *D*-values are presented. These values confirm that the Data Interpretation items were relatively more difficult for females than other Mathematics items. Precisely why this is so is unclear. While a hypothesis concerning spatial abilities has been advanced in connection with geometric material, the spatial demands for graphic interpretation seem modest. A more plausible hypothesis concerns the general familiarity with graphic presentations and states that males receive more practice in this type of activ-

Table 5
D-Values and Content Code for the Six Reading Test Outliers

Item	D-Value	Passage Content	Item Content	Group Favored
4	+0.61	Humanities	Intended Inference	F
10	-0.75	Physical Science	Main Idea	M
11	-0.66	Physical Science	Main Idea	M
12	-0.59	Physical Science	Factual Statement	M
15	+0.57	Argumentative	Main Idea	F
17	+0.54	Argumentative	Main Idea	F

Figure 3
Δ Plot for Mathematics Section

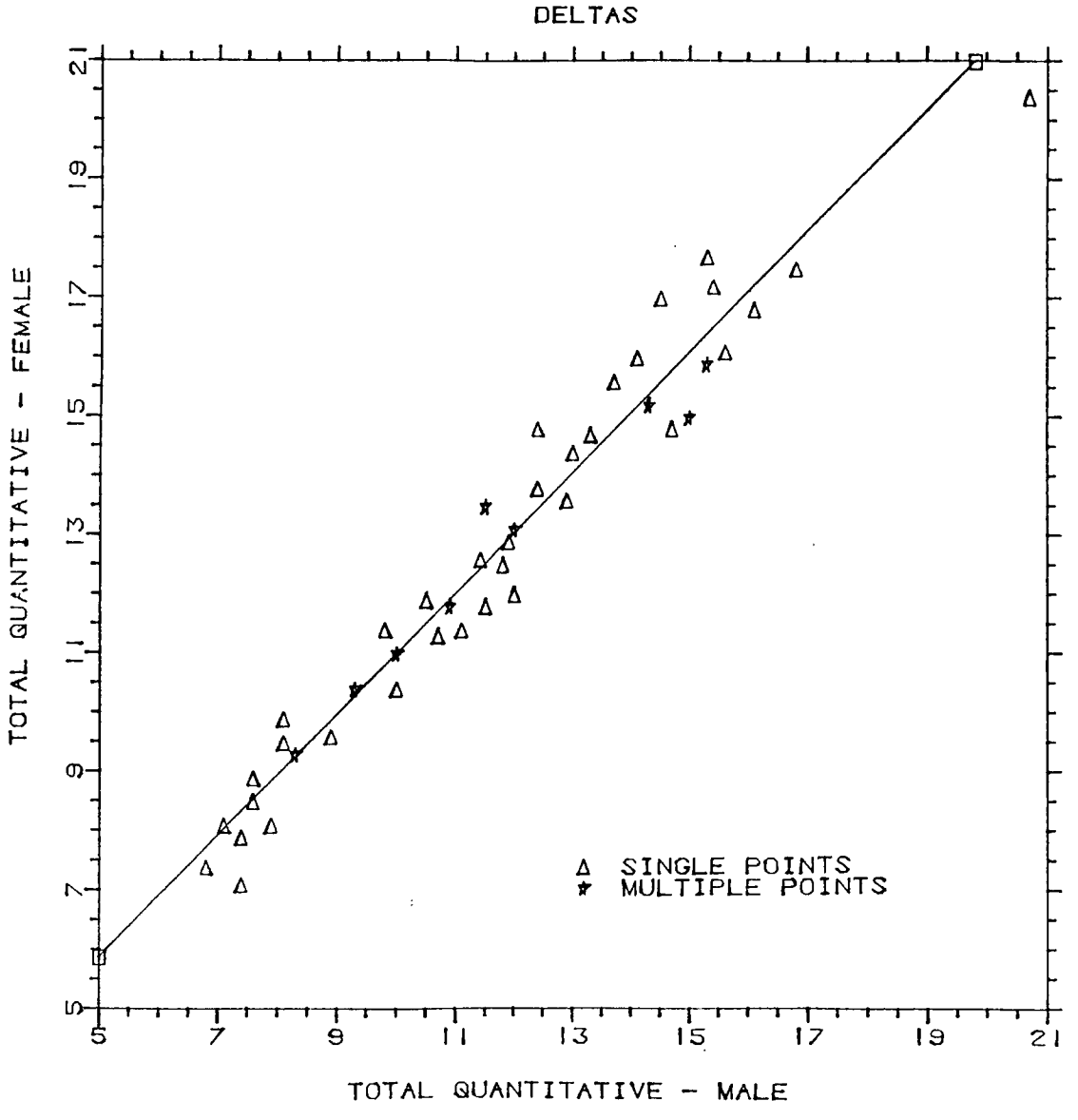


Table 6
Average D-Values for the Six Reading Passages

Passage Content	Items	Average D-Value	Group Favored
Humanities	7	+.26	F
Physical Science	6	-.55	M
Argumentative	6	+.27	F
Social Science	7	-.09	M
Biological Science	7	+.04	F
Synthesis	7	+.04	F

ity than females and have, in effect, more current skills.

Table 9 presents the limited data that might bear on the question of item difficulty and references to human agents. The use of male examples and references is predominant in the materials of educational measurement, as established by Tittle et al. (1974); there have been expressed concerns that such content imbalance would influence item difficulty. The GREAT Mathematics section has very few items with any kind of reference to persons, and these references are trivial. Table 9 presents the *D*-values for five such items. No pattern seems discernable. There were no outliers, but there were modest *D*-distances. Item 16, which refers to a woman college president and her salary, is biased somewhat against females. Although no conclusion can be reached, the evidence tends to run counter to a hypothesis that content influences sex bias, at least at the superficial level that content is involved in these items.

In evaluating these results, it should be remembered that the use of random selection does

Table 7
D-Values and Content Code for the Nine Mathematics Outliers

Item Type	Item Number	Group Favored	D-Value	Assemblers Content Code
Regular Math	2	F	+1.92	Geometry
	36	F	+1.60	Arithmetic
	44	M	-2.14	Geometry
	51	F	+1.53	Miscellaneous
	52	F	+1.70	Geometry
	55	F	+2.35	Algebra
Data Interpretation	32	M	-2.23	(Economics Graph)
	33	M	-2.05	(Economics Graph)
	34	M	-1.55	(Economics Graph)

Table 8
Average D-Values for the Mathematical Item Type

Item Type	N	Average D-Value	Group Favored
Regular Math	41	+.06	F
Data Interpretation	14	-.18	M

not create samples with equivalent total score distributions, so that the samples may be differentially biased by sex. Further, the possibility of quantitative score differences deriving from differences in spatial ability remains. It is simply judged that the quantitative material used in the test is not predominantly reflecting this ability.

Discussion and Conclusions

The analysis presented here reveals the presence of a few items within the GREAT that are unusually more difficult for males or females than most similar items. The overall effect of their counterbalancing, through content specifications, is to produce tests with no strong sources of bias through content imbalance. The analysis did not reveal any significant ways in which the observed score differences could be altered, except possibly by the replacement of Data Interpretation material by more Regular Mathematics items. The logic for such replacement, however, would not derive from any compelling inequity but from a reexamination of educational values and egalitarian values. As in the SAT-M, it *is* possible to influence the magnitude of the observed difference between the sexes by controlling the selection of item material. Fewer Data Interpretation items, for example, could decrease the size of the difference. But skill in handling such material as complicated graphs in Data Interpretation is a bona fide outcome of education and should not be excluded from the sample of skills without an adequate reason.

Further, the sex difference may not be sex based. Not all males succeed on Data Interpretation; not all females fail. The current approach in analyses such as the one conducted in this study, while reassuringly presenting evidence that the GREAT is not overly unfair to one of the sexes, tends to focus attention on sex differences in behavior without an analysis of the educational and experiential concomitants

Table 9
D-Values for Items With Human Referents

Item	Reference	Group Favored	D-Value
16	The president of a college earned if she doubled her earnings	M	-.04
39	A man's take home pay	M	-.25
45	Fifty students taking a test	M	-.59
49	A dealer . . . his gross profit	F	+.12
53	A class of twenty students . . . given a quiz	M	-.50

that govern such behavior. A superior study would examine the observed differences, particularly in the most stereotypical areas, such as mathematics or science, from the standpoint of educational background. With the continuing interest in appraising potential without the biased expectations of stereotyping, such studies will probably be done in the future.

References

Angoff, W. H. *A technique for the investigation of cultural differences*. Paper presented at the annual meeting of the American Psychological Association, Honolulu, September 1972.

Angoff, W. H., & Stern, J. *The equating of the scales for the Canadian and American Scholastic Aptitude Tests* (ETS PR 71-24). Princeton, NJ: Educational Testing Service, 1971. (CEEB RDR 71-72, No. 4)

Coffman, W. E. Sex differences in responses to items in an aptitude test. *National Council on Measurement in Education Yearbook*, 1961, *18*, 117-124.

Donlon, T. F. *Content factors in sex differences on test questions* (ETS RM-73-28). Princeton, NJ: Educational Testing Service, 1973.

Keating, D. P., & Stanley, J. C. Extreme measures for the exceptionally gifted in mathematics and science. *Educational Researcher*, 1972, *1*, 113-7.

Lockheed, M. *Sex bias in educational testing: A sociologist's perspective*. Paper presented at the International Symposium on Educational Testing, The Hague, Netherlands, July 1973.

Maccoby, E. E., & Jacklin, C. N. *The psychology of sex differences*. Stanford, CA: Stanford University Press, 1974.

Stafford, R. E. Sex differences in spatial visualization as evidence of sex-linked inheritance. *Perceptual and Motor Skills*, 1961, *13*, 428.

Strassberg-Rosenberg, B., & Donlon, T.F. *Content influences on sex differences in performance on aptitude tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC, 1975.

Tittle, C.K., McCarthy, K., & Steckler, J.F. *Women and educational testing*. Princeton, NJ: Educational Testing Service, 1974.

Acknowledgment

The support of the Graduate Record Examinations Board in the conduct of this research is gratefully acknowledged.

Author's Address

Send requests for reprints or further information to Thomas F. Donlon, Educational Testing Service, Princeton, NJ 08541.