DOCUMENT RESUME

ED 065 510 TM 001 430

AUTHOR	Echternacht, Gary J.
TITLE	An Examination of Test Bias and Response Characteristics for Six Candidate Groups Taking the ATGSB.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	PR-72-4
PUB DATE	Mar 72
NOTE	33p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Admission Criteria; Business Education; Caucasian Students; Females; *Graduate Students; Males; Negro Students; *Racial Factors; Response Style (Tests); *Sex Differences; *Test Bias; *Test Wiseness
IDENTIFIERS	*Admission Test Graduate Study in Business; ATGSB

ABSTRACT

Differences between black and white, male and female groups who took the Admission Test for Graduate Study in Business (ATGSB) during February 1971 were examined in three studies. The studies were: (1) a study of the biasedness of the test with respect to the groups, (2) a comparison of the mean criterion scores for those candidates who omit items, and (3) a comparison of the response randomness for the sub-groups involved. The ATGSB in this study consisted of seven separate tests: Reading Recall I, Reading Recall II, Antonyms, Analogies, Sentence Completion, Mathematics, and Data Sufficiency. The tests were conducted in two types of settings, a regular test center which charged a fee and at fee-free centers. A total of 2,930 candidates, who were attending institutions in the West, North Central, South, Northeast, and Foreign Institutions, were sampled for each research question. The groups and their sizes were: fee-free males 485, fee-free females 370, regular center black--males 630, females 150, white males 995 and white females 300. The results of Study One: Statistically Defined Test Bias showed that if item-group interaction is accepted as a definition of test bias, then each section of the ATGSB is biased in some way. It appeared that racial bias is contributing more than sex bias in each of the subtests. In Study Two: Omit Behavior, the white group mean criterion score was the lowest among the three groups in the majority of cases. Study Three: Randomness of Response results showed the randomness in chosing distractors was greatest in the fee-free groups and there was more randomness in female groups than male groups. Conclusions, recommendations and graphs are given. (DB)

PR-72-4

AN EXAMINATION OF TEST BIAS AND RESPONSE CHARACTERISTICS FOR SIX CANDIDATE GROUPS TAKING THE ATGSB

Gary J. Echternacht

March 1972

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY

AN EXAMINATION OF TEST BIAS AND RESPONSE CHARACTERISTICS FOR SIX CANDIDATE GROUPS TAKING THE ATGSB

In the recent past there has been a growing concern about the "fairness" of standardized tests with respect to various groups in the testing population. The Research and Development Committee of the Graduate Business Admissions Council, for example, has become concerned about the appropriateness of its admission test for both female and black subgroups. This study examines differences between black and white, male and female groups who took the Admission Test for Graduate Study in Business (ATGSB) during February, 1971. More specifically, this report covers three studies: (1) a study of the biasedness of the test with respect to the above-mentioned groups, (2) a comparison of the mean criterion scores for those candidates who omit items, and (3) a comparison of the response randomness for the sub-groups involved. This study considers six groups of examinees: black - females, black - males, fee-free - females, fee-free - males, white - females, and white - males. The nature of these examinee groups will be discussed later.

Characteristics of the ATGSB

The test used in this study was the regularly scheduled ATGSB administered during February, 1971. The test has five separately timed sections. Two of these sections are identical in terms of content, format, and difficulty. These two sections require the examinee to read three excerpts from the current business literature. After reading these three passages, the examinee is asked a series of

questions about the content of each of the passages. In each case the examinee cannot return to the passages for reference. These sections are termed Reading Recall I and II, respectively. A second verbal section contains discreet verbal items consisting of antonyms, analogies, and sentence completion type items in that order. For the purpose of this study only, this section was considered as though it contained three separate sub-sections, each sub-section characterized by its item type. Another section consists of rather traditional mathematical type items. Questions are asked concerning graphs and charts along with a few questions requiring knowledge of simple algebra. A final section contains items termed "Data Sufficiency". In this section a mathematical problem is given along with some data. The examinee is then required to judge whether this data is sufficient to solve the problem stated. In summary, this study considered the ATGSB to consist of seven separate tests: Reading Recall I, Reading Recall II, Antonyms, Analogies, Sentence Completion, Mathematics, and Data Sufficiency.

The Student Population

The February administration was conducted in two types of settings. One setting was the regular test center where examinees paid a fee to take the test, the other was a free testing given at locations termed fee-free centers. These fee-free centers were predominantly black colleges located primarily in the south. Although the majority of examinees in the fee-free centers were black, other minority and a few white examinees took the test in the fee-free centers. In this study only black fee-free candidates were sampled from the fee-free centers.

In the regular center group, an overwhelming majority of examinees were taking the ATGSB for the first time. The modal age for each candidate group defined by sex and race (black and white) was 22, with the median age for black and white females being 23, for black males 26, and white males 24. Similar descriptive statistics were not available for the fee-free group.

The geographic distribution of students in each candidate group, classified ' the undergraduate college attended, is given in Table 1. The regional classification used is that of the Census Bureau. Chi-square statistics were calculated using the proportions given in the 1970 census for the total population group as the expected figure. Significance at the .05 level was obtained, and it was concluded that relatively more black students in the candidate group attend institutions in the west than might be expected if the total proportion of blacks living in the west is considered to be the norm.

Table 1

Percentage of Candidate Groups Attending
Undergraduate Institutions in
Four Regions and Foreign Areas
(Census Bureau Classification)

Candidate Group	West	North Central	South	Northeast	Foreign
Black - Female	15	17	53*	16	--
White - Female	18	28	21	33	--
Black - Male	15	27	37**	21	1
White - Male	16	31	22	30	2

*71% of the Black - Females attending institutions in the south attended predominantly Black institutions.

**74% of the Black - Males attending institutions in the south attended predominantly Black institutions.

The geographic location of the center where the ATGSB was taken by the candidates was similar to the undergraduate institution attended. These figures are given in terms of percentage of the examinee group taking the test in centers in the usual census regions (Table 2).

Table 2

Percentage of Candidate Groups Taking the ATGSB in Centers in Four Regions and Foreign Areas

Candidate Group	West	North Central	South	Northeast	Foreign
Black - Female	16	22	44	18	--
White - Female	21	24	20	33	2
Black - Male	20	29	26	24	2
White - Male	19	29	22	26	5

It should be noted that these percentages differ only slightly from those reported for the attending colleges. This is due in part to relocation by those out of college and to general student mobility.

In this study, a total of 2930 candidates were sampled for each research question. The individual group sample sizes were: fee-free males 485, fee-free females 370, regular center black - males 630, regular center black - females 150, white - males 995, and white - females 300. Each of these samples were random samples taken from the total examinee population in the regular centers, while the fee-free groups consisted of the entire population.

These samples differed significantly in terms of mean scores. It was decided not to match the samples in terms of total score though, as that would make the group labels misleading, i.e. a low scoring sub-group would be compared with another complete group. The use of a low scoring sub-group would require a redefinition of the groupings that would be contrary to the aims of this study.

Study One: Statistically Defined Test Bias

The problem of defining what is meant by test bias has received considerable attention by Cardall and Coffman (1964), Cleary and Hilton (1968), and Potthoff (1966) among others. Basically, two approaches have been taken: with or without a criterion variable present.

The case of defining test bias with a criterion is most straightforward and logically most appealing. When a criterion variable is present, the definition of test bias simply says that a test is not biased if individuals from different groups who have the same test scores have the same expected criterion scores. Some further difficulties exist if the test is not perfectly valid, but most researchers have continued to persue the problem by considering homogeniety of regressions.

Two other problems do arise, though, when test bias is examined in light of a criterion variable. First, suitable criteria are difficult to define, especially when the tests are admissions tests and the criteria are variables that reflect some notion of successful performance. Secondly, its is a very expensive proposition to collect criterion data, and their collection often renders such research studies to be not feasible due to high project costs.

It naturally follows that most research studies have tried to attack test bias questions without resorting to collecting any criterion variables. This is logically a more difficult task because one gets stalled in the beginning in trying to define test bias. Although there seems to be no generally suitable means to define test bias in the absence of a criterion, several attempts have been made to answer such questions by examining a concept which seems closely related -- that of item-group interaction.

The problem of defining item-group interaction is, in itself, difficult. One can say that there is no bias present in a test if the difference in p - values (the proportion who answer an item correctly) is identical for all items in the test for any two groups. If multivariate statistical tests are made of this hypothesis, difficulties can arise if the variance matrices for each group are not homogeneous or if the p - values are not close to ½. Potthoff gives a number of techniques for handling such situations and a variety of techniques to choose from.

Method for Study One

In this study, a method of estimating bias was needed that was both inexpensive and readily available from the standard item analysis procedures now in use. Thus, p - values were calculated for each item in the test for each group under study. Since p - values can vary from only 0 to 1, and one often concludes the existence of bias when items are taken with p - values close to one of these extremes, a transformation of the p - values, commonly used by ETS, termed delta was used as the unit being studied. Delta is defined as the value Δ satisfying the equation,

$$p = \frac{1}{4\sqrt{2\pi}} \int_{\Delta}^{\infty} e^{-\frac{1}{2}\left(\frac{x - 13}{4}\right)^2} du$$

and is calculated for every item in each item analysis. The delta scale is approximately normal with a mean of 13 and a standard deviation of 4. Thus, a p - value of .5 is associated with a delta of 13.

The definition of test bias used in this study was that of item-group interaction. It was hypothesized that if no bias were present in

a set of items for two groups, the differences in item deltas for these two groups would be distributed as a normal distribution with some unknown mean and some unknown variance. If the differences did not form a normal distribution, bias would be concluded.

The method used to determine whether the differences in deltas were normally distributed was to plot these differences on normal probability paper and estimate whether these plots formed a straight line as would be found had there been no bias; i.e., item delta differences constant subject only to an error term associated with items. Since there were six groups of examinees from which item delta differences were to be calculated, only five group pairs were independent. That is, the item delta differences for any pair of groups could be obtained by knowing the differences for just five independent pairings. The problem then was to select the independent pairings. Since racial bias was considered to be of most importance, it was decided to examine racial bias within sex, and then make an overall comparison between sexes. Therefore, the group pairings were: (1) white-male vs. regular center black-male, (2) white-male vs. fee-free-male, (3) white-female vs. regular center black-female, (4) white-female vs. fee-free-female, and (5) male (pooled over race) vs. female (pooled over race). Each of these comparisons were independent.

The general method of determining whether the points fall on a straight line is a generalization of the Kolomogorov-Smirnov technique for testing for normality. The generalization involves estimating the hypothesized normal distribution parameters with the sample parameters. The hypothetical normal distribution is plotted as a straight line, and confidence bands

are drawn for a given significance level and number of items. If any point falls outside the band, rejection of the normality hypothesis is assumed. The significance level used in this study is the .05 level. Items (points) falling outside the band are noted.

In addition to the plots, repeated measures analysis of variance was run for each subtest with race and sex as factors and items as repeated measures. This analysis was not performed to test for the appropriate item-group interaction effects since group statistics were being used and no appropriate error term could be used in a significance test. The analysis of variance was performed to provide an overall picture of the proportion of the sum of squares that accounted for each line in the analysis of variance table, which would provide a lead as to the magnitude of the item-group interaction with respect to the other factors in the analysis.

Results of Study One

The results of study one could be divided into seven subsections, each subsection dealing with a specific subtest. The repeated measures analysis of variance results are presented in the last part of this section. The presentations following indicate where significant non-normality (test bias) has been concluded, and tries to provide some help in remedying the bias by noting items that fell outside the confidence band in the analysis. In a sense, this is somewhat misleading in that the noted items are items where delta differences differ significantly from the normal distribution specified by the sample estimates. If the differences deviate greatly from a normal distribution, the sample specifications may also form a less than desirable criterion. This should be kept in mind in reading the results. A selection of the plots appears in the appendix.

Table 3 gives a summary of finding points not on the appropriate straight lines for the Reading Recall I subtest. The column labeled frequency denotes the number of times (5 is the maximum) that the particular item was found to lie outside the confidence band--a degree of bias figure. The groups column indicates the respective group pairs where this deviation was found. The final column indicates the nature of the repair that is required in order to make the item not biased, assuming the sample estimates of the mean difference and variance hold for the population. "Difference less" means that the difference between the two groups should be less if test bias is to be eliminated, while "difference more" implies the opposite. In a sense, the last two columns indicate the group favored by the bias. Difference less indications show bias favoring the white group, while difference more indications show bias favoring the regular center black or fee-free group.

Table 3

Deviate Items Found in Reading Recall I

Item No.	Frequency	Groups	Difference Less	Difference More
1	2	2-4, 2-6		2
3	1	2-4	1	
5	1	M-F		1
6	1	2-4	1	
7	1	2-6	1	
8	2	2-6, 2-4	1	1
9	1	2-6	1	
10	1	2-6	1	
11	1	2-6	1	
14	2	2-6, M-F	2	
15	1	1-5	1	
17	1	1-5	1	
19	1	2-6		1
22	1	2-6, M-F		2
23	1	2-6	1	
25	1	2-4, M-F	1	1
27	1	2-6		1
29	2	1-5, 2-6	2	
30	2	1-5, 2-6	2	

Codes are used in the groups column for convenience in presentation. The codes are designated as follows: 1 = white-male; 2 = white-female; 3 = regular center black-male; 4 = regular center black-female; 5 = fee-free-male; 6 = fee-free-female; M = male (pooled data); F = female (pooled data). This notation is used throughout this section.

As can be seen by examining the table, 19 of the 30 items showed differences that fell outside the confidence bands at least once. Half of these items involved the white-female vs. fee-free-female comparision (13 of 26). The items noted were found only once or twice. Four of the 26 noted item differences were attributable to the white-female vs. regular center black-female comparison. The same was true for both the white-male vs. fee-free-male comparisons and the male vs. female comparisons. No evidence of racial bias was found in the white-male vs. regular center black-male comparisons.

In summary, this section of the ATGSB appears to be biased. Most of that bias occurs when female groups are considered, or when fee-free groups are considered. The reference to particular items does not indicate that there is definitely bias present in those items, but rather they indicate items showing bias when the distribution of item delta differences is specified by the sample estimates of the parameters.

In the Reading Recall II section, bias similar to that found previously in Reading Recall I was found. The results are summarized in Table 4. As before most of the bias involves the female comparisons (18 of 20 racial comparisons), and a substantial number involve fee-free candidates (11 of 21). Six of the items show bias in 2 of the 5. Item 23 tends to distinctly favor the white group. Sex bias was also found to be significant in

the pooled male vs. pooled female comparison. There was no bias concluded from the white-male vs. fee-free-male comparison.

Table 4

Deviate Items Found in Reading Recall II

Item No.	Frequency	Groups	Difference Less	Difference More
1	2	1-3, 2-6	1	1
2	1	2-6		1
4	1	2-6	1	
5	2	2-6, M-F	2	
8	1	2-4	1	
10	1	2-4		1
11	1	2-6	1	
16	1	2-6		1
18	2	2-6, 2-4	2	
20	2	2-6, M-F		2
23	2	2-4, 2-6	2	
24	2	1-3, 2-6	1	1
26	1	2-4	1	
27	1	2-4	1	
28	2	2-4, 2-6	2	
30	1	M-F		1

The Antonym section (Table 5) shows proportionately more biased items than do either of the Reading Recall sections. Only 2 of 14 items did not contain any bias. For this section, racial bias within the male group was much more frequent than in the past two sections. Only two items showed any racial bias within the female group, and these indicators both involved the fee-free group. Also, indicators of sex bias were found to a larger extent than in previous subtests. Actually, the major source of sex bias in this section involved item 2 and that item heavily favored the female group.

Table 5

Deviate Items Found in Antonyms

Item No.	Frequency	Groups	Difference Less	Difference More
1	3	1-3, 1-5, M-F	2	1
2	2	1-5, M-F	1	1
3	3	1-3, 1-5, M-F	1	2
4	1	M-F		1
6	2	1-5, M-F	2	
7	2	1-3, 1-5	2	
8	1	M-F		1
9	2	1-5, M-F	1	1
10	1	1-5	1	
11	1	1-5		1
12	1	2-6		1
13	1	2-6	1	

In the Analogies section, Table 6 shows that relatively few items were noted to be biased. Bias must still be concluded, as the overall significance test of normality of delta differences was rejected. Both items involved in the sex bias indicate a favoring of the male group, while the three items involved in the race bias favor the regular center black and fee-free candidates. No evidence of bias was found in the white-male vs. fee-free-male comparison.

Table 6

Deviate Items Found in Analogies

Item No.	Frequency	Groups	Difference Less	Differency More
1	2	M-F, 2-4	1	1
2	1	M-F	1	
10	1	1-3		1
13	1	2-6		1

Only one comparison showed significant bias for the Sentence Completion section. That comparison was the white-female vs. fee-free-female, as can be seen by examining Table 7. Items 3 and 8 tended to favor white-females, while items 4 and 12 tended to favor fee-free-females. All other comparisons showed no significant deviation from the normality hypothesis.

Table 7

Deviate Items Found in Sentence Completion

Item No.	Frequency	Groups	Difference Less	Difference More
3	1	2-6	1	
4	1	2-6		1
8	1	2-6	1	
12	1	2-6		1

The results for the Data Sufficiency section appear in Table 8. As with the previous two sections, the extent of the bias present in the section is less than the first three sections. Only four items displayed bias. Each of these items showed bias favoring the white group within sex. There were no significant results for either the white-male vs. regular center black-male and the male vs. female comparisons.

Table 8

Deviate Items Found in Data Sufficiency

Item No.	Frequency	Groups	Difference Less	Difference More
7	1	2-4	1	
9	1	2-6	1	
10	1	2-4	1	
12	1	1-5	1	

By far the most noted extent of test bias occurred in the Mathematics section (Table 9). Of the 54 items in the section, 33 have some indicator

of bias. A large proportion of the bias favors the lower scoring group, as indicated by the relatively frequent occurrence of deviation in the difference less column. In each case where the male vs. female difference was significant, the bias favored the female group. In the within sex comparisons for racial bias, 30 out of 40 noted items involved the fee-free group, 9 in the male group, 21 in the female group. Item 45 was the only item noted in each group comparison as deviating from the hypothesized distribution.

Table 9

Deviate Items found in Mathematics

Item No.	Frequency	Groups	Difference Less	Difference More
1	1	2-4		1
2	1	2-6		1
3	1	1-5		1
4	1	2-4	1	
6	1	2-6	1	
8	1	2-6	1	
9	2	2-6, M-F		2
12	1	2-6		1
15	1	2-6	1	
19	3	1-3, 1-5, 2-6	2	1
20	1	1-5	1	
22	1	M-F		1
24	1	1-5	1	
25	1	2-6	1	
26	1	1-5	1	
28	1	2-6	1	
32	1	M-F		1
33	2	1-5, 2-6	1	1
37	1	2-6		1
38	1	2-6	1	
40	3	1-5, 2-6, M-F	1	2
41	1	1-3		1
45	5	1-3, 2-4, 1-5, 2-6, M-F		5
46	1	2-6	1	
47	2	1-3, 2-6		2
48	2	2-6, M-F	1	1
49	2	1-3, M-F		2
50	1	2-6	1	
51	1	2-6		1
52	2	2-6, M-F		2
53	3	2-4, 2-6, M-F		3
54	3	1-3, 1-5, 2-6		3

In summarizing the results of Study One, one clear fact stands out. If the notion of item-group interaction is accepted as a definition of test bias, then each section of the ATGSB is biased in some way. Of the 35 comparisons made (7 subsections x 5 comparisons per subsection), 25 were found to be significant, indicating test bias.

Items were noted whose item delta difference fell outside the confidence bands for the set of item delta differences. These were items that differed from the hypothetical normal distribution where the parameters for that distribution were taken to be the sample estimates. These items were noted merely in order to provide a clue as to the nature of the bias present, rather than to assert that the addition of these items to the test caused the test to become biased.

Most of the indications of racial bias within sex seemed to occur in the female groupings (12 of the 20 significant within sex results) and in the fee-free comparisons (11 of 20). Generally, the bias present favored no one particular race. Four of the 7 male vs. female comparisons were significant. No one sex seemed to be favored over the other.

The item deltas for the six groups were also analyzed by analysis of variance. The structure of the analysis was conceived to be a 2 x 3 factorial (sex and race as factors) with repeated measures (items). Since there were no error terms available for a significance test, only the percentage of the total sum of squares attributable to the various factors in the analysis were given. These were given in order to display the importance of each factor in relation to the others; e.g., determining whether a sex bias (sex x item interaction) or a racial bias (race x item interaction) seems more immanent. The percentages appear in Table 10.

The differences in means among the racial groupings (white, black regular center, and fee-free) and the variability in item deltas is more apparant from the high percentages of sums of squares attributable to race and items respectively. It appears that racial bias is contributing more than sex bias in each of the subtests. The degree of sex x race interaction and sex x item interaction appears minimal in comparison to the remaining factors. The three factor interaction also appears to be minimal.

Table 10

Percentage of Total Sum of Squares Attributable to Factors in ANOVA

	Factor						
	Sex	Race	Sex x Race	Items	Sex x Items	Race x Items	Sex x Race x Items
Reading Recall I	1.59	25.79	.62	64.62	.90	5.75	.73
Reading Recall II	.53	32.45	.71	56.52	.80	7.87	1.12
Antonyms	.02	16.62	.54	75.24	1.35	5.41	.82
Analogies	.02	17.96	.92	75.95	.66	4.05	.45
Sentence Completion	.12	23.29	.68	72.88	.33	1.78	.91
Mathematics	2.07	30.14	.01	61.33	.91	4.47	1.07
Data Sufficiency	.47	21.98	.15	71.54	.52	4.43	.92

Study Two: Omit Behavior

In a study of culturally deprived youth, Flaugher and Pike (1970) determined that, because of the inappropriate difficulty level of the test being studied, higher scoring students in a low scoring group omitted large numbers of items, which was opposite of the pattern in a middle-scoring group. This study attempted to determine whether the mean criterion scores (section scores) for those who omit differed among the three groups: white, regular black, and fee-free black. In each case, only within-sex differences were considered.

Method

The standard item analysis program calculates an index of the average ability level, mean criterion score, for the group of examinees choosing each option, including omitting. The mean criterion score is on a scale with a mean of 13.0 and a standard deviation of 4.0, corresponding to the delta scale for item difficulty value.

For example, if the criterion used was the score on the total test, the mean score of the total sample would be assigned a value of 13.0. If the average score for the group choosing a particular option was above the sample mean, the group's mean criterion score would be greater than 13.0; if their average was below the sample average, it would be less than 13.0.

If we consider all possible ranking patterns constructed in such a way that the first digit indicates the ranking, in terms of mean criterion score, of the white group, the second digit indicating the ranking of the regular center black group, and the third digit indicating the fee-free group ranking, there are six possible patterns.

Results

For example in the first pattern, denoted 1, 2, 3, the criterion score for the white group was lowest and for the fee-free group, it was the highest. Under null conditions of a random pattern in omit mean criterion scores, approximately 1/6 of the items of any given section should fall in each category defined by these six orderings. If evidence can be provided to show that this is not the case, we can conclude that there is some systematic difference in mean criterion scores for the three groups under study.

In this study the criterion scores were ranked from low to high; i.e., the lowest mean criterion score received a rank of one. The number of items in each ranking pattern for both males and females appear in Table 11. In cases where ties were found, the ties were broken by using random digits from a table.

In examining the frequency of occurrence for the various patterns, in Table 11 it appears to be quite obvious that these patterns are occurring in a non-random fashion. For example, the fee-free group almost never has the lowest mean criterion score, as indicated by a one in the third digit of the patterns. On the other hand, the white group mean criterion score was the lowest among the three groups in the majority of cases. There appears to be some doubt over whether the regular black or white group has the lowest mean criterion score for the analogy and sentence completion type items. The frequencies of pattern occurrence is the same for both males and females, with women tending to show a slightly wider distribution of patterns.

Table 11

Frequency of Ranking Pattern of

Omit Mean Criterion Scores

Test	Pattern 123	132	213	231	312	321
Reading Recall I						
Male	16	6	5	0	1	1
Female	14	6	5	1	3	0
Reading Recall II						
Male	18	4	6	1	0	0
Female	10	13	2	1	3	0
Antonyms						
Male	12	0	0	0	0	0
Female	9	3	0	0	0	0
Analogies						
Male	6	1	4	0	2	0
Female	8	3	1	0	1	0
Sentence Completion						
Male	4	0	6	0	1	1
Female	4	1	3	1	1	2
Mathematics						
Male	45	2	5	0	1	0
Female	29	13	5	4	2	0
Data Sufficiency						
Male	10	1	2	0	1	0
Female	4	7	2	0	0	1

In addition to obtaining counts of the ranking pattern of mean criterion scores for those who omit, three-way analyses of variance were performed for each type of test with sex, race, and items as factors and the respective criterion scores as observations. Although the items were correlated to an extent, it was believed that if any

race effects were found, some strength could be added to the above argument. Using this type of analysis, the group averages were tested for equality over the other two factors.

In each case a significant race effect was found, and a sex x race interaction was found for analogies and antonyms. The third order interaction was used as the error term and the tabled results appear in Appendix II. These results should be taken with extreme caution and are presented only as supplementary evidence (weak as it is) of rejecting the null hypothesis of random differences in mean criterion score for those who omit.

Study Three: Randomness of Response

When exploring the differences in performance on a multiple-choice test for two groups of students, a third possible source of difference is varying test-taking strategies. Flaugher and Pike (1970) reported one such study, which investigated the randomness of response that was evident in a group of low-scoring high school students. Through the use of an index of randomness in responding (Pike & Flaugher, 1970) they found that less randomness was characteristic of a particular item type within the test, that of the verbal analogy item. The present study was conducted to replicate and extend these particular findings on a new population.

Method

In many kinds of paper and pencil tests an examinee is presented a list of items each followed by the possible answers, or alternatives. The examinee may use a number of strategies in answering an item. He may know the answer (or think he knows the answer) and mark that alternative, he might be able to eliminate some of the alternatives and guess among the remaining alternatives, he may guess completely, or he may omit the item. The ability to eliminate some of the alternatives as not being plausible and then responding may be referred to as the amount of non-randomness present in the responses, the more alternatives eliminated, the less the randomness in that response.

In order to measure the amount of randomness present, the proportion of examinees responding to each of the distractors was studied. If the distribution of responses to the various distractors was flat, that is,

evenly distributed, then there was evidence of complete guessing on the part of the group of examinees who did not choose the correct response. To the extent that the distribution differed from this, a non-random guessing pattern was assumed. In order to measure the flatness of a group of probabilities for a finite set of categories, Shannon (1949) developed a quantity denoted H, termed entropy, where:

$$H = \sum_{i=1}^{r} p_1 \log p_i \quad .$$

The concept of entropy was first developed in thermo-dynamics but has become the subject of information theory and has been applied in the mathematical theory of communication by Shannon and Weaver (1949).

For the purpose of this study the definition of entropy used was

$$H = (P - \sum_{i=1}^{r} p_i \log p_i / P) / \log r \quad ,$$

where P is the proportion of examinees answering the item incorrectly, p_i is the proportion of examinees who chose distractor i, and r is the number of distractors. This form was used so that when all distractors were chosen with equal frequency, the entropy value was one, and when all examinees who missed the item responded with the same distractor, the entropy value was zero.

Results

The item entropy was calculated for each item on each sub-test for each of the six groups under study. The mean of these item entropies was then calculated for each of the six groups. These means appear in Table 12. The racial and sex differences are significant at the .05 level for each section in question, the analyses appearing in Appendix III.

The patterns seem similar for each section and both races within those sections. That is, the entropy for the regular center black students tends to be slightly lower than that for the white group, while the entropy for the fee-free group tends to be relatively high for each section. Basically, this says that the randomness in choosing distractors is greatest in the fee-free group. Apparently all distractors appear equally likely to candidates who do not know the correct answer. The randomness for the remaining two groups appears to be less, with a slight nod being given to the regular center black group. In these groups, students tend to discriminate more among the distractors, possibly being able to eliminate some alternatives as being implausible or being especially attracted to a "strong" distractor.

In examining sex differences within race, there appears to be more randomness in the female groups as compared to the male group. This conclusion holds for 19 of the 21 within race comparisons made over the 7 sections.

Table 12

Mean Group Values of Entropy

Test	White	Regular Center Black	Fee-Free
Reading Recall I			
Male	.795	.798	.869
Female	.843	.850	.890
Reading Recall II			
Male	.753	.727	.820
Female	.813	.790	.834
Antonyms			
Male	.863	.843	.920
Female	.886	.883	.930
Analogies			
Male	.851	.825	.900
Female	.879	.859	.893
Sentence Completion			
Male	.876	.855	.915
Female	.923	.883	.903
Mathematics			
Male	.753	.727	.820
Female	.813	.790	.834
Data Sufficiency			
Male	.675	.651	.803
Female	.754	.724	.800

Conclusions

1. If the definition of test bias is taken to be that of item-group interaction, every subsection of the ATGSB appears to be biased in some fashion.
2. Most of the indications of test bias with respect to race occurs in the female group. That is, more significant results were obtained from the race within female comparisons than from the race within male comparisons.
3. In comparing male and female samples, pooled over the different racial groupings, no one sex was favored over the other in a uniform fashion.
4. In the race within sex comparisons, no one race was favored uniformly over the other. Stated another way, even though test bias was concluded to exist, the bias favored neither Negroes nor Caucasians.
5. In general, the white group exhibited the lowest mean criterion scores for those who omit particular items, and the fee-free group exhibited the highest mean criterion scores for those who omit particular items. This pattern appears less frequently in the case of analogies and sentence completion, with the results being similar for both male and female groups.
6. There appears to be more randomness of response in the fee-free group. Regular center black candidates tend to show only a reduced tendency to respond at random to the distractors. In general, each group tends to respond with equal frequency to the distractors.

Recommendations

1. Since the ATGSB appears to be a biased test, steps need to be taken that will at least reduce the degree of bias present in the test. This could be accomplished, in part, during the pretesting of the test. Within sex racial differences in item deltas should be calculated for each item in the pretest. Steps should then be taken so that the final forms developed have normally distributed item delta differences.
2. Further research needs to be conducted into the area of test bias in the ATGSB. This research should be conducted with a set of criterion variables and possibly be longitudinal in nature. Further research in the area of test bias where no criterion is present is not recommended at this time.
3. Future analyses should ignore any fee-free candidates, as this group tends to display characteristics different from the regular center black candidates.

References

Echternacht, G. J. A quick method for determining test bias. Research Bulletin RB 72-xx. Princeton, N. J.: Educational Testing Service, 1972.

Flaugher, R. L., & Pike, L. W. Reactions to a very difficult test by an inner-city high school population: a test and item analysis. Unpublished manuscript. Princeton, N. J.: Educational Testing Service, 1970.

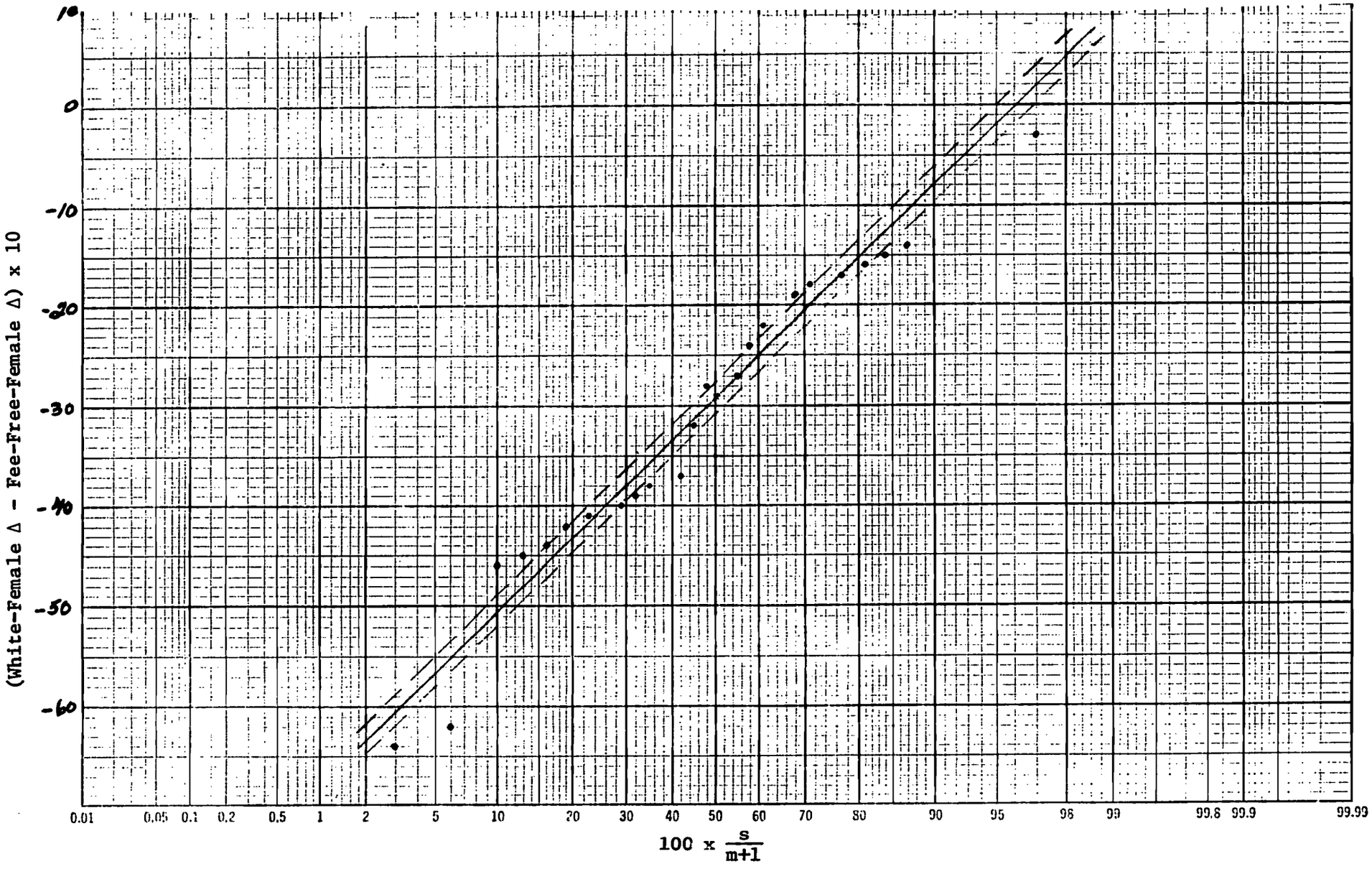
Pike, L. W., & Flaugher, R. L. Assessing the meaningfulness of group responses to multiple choice test items. Paper delivered at 78th APA Annual Convention, Miami, 1970.

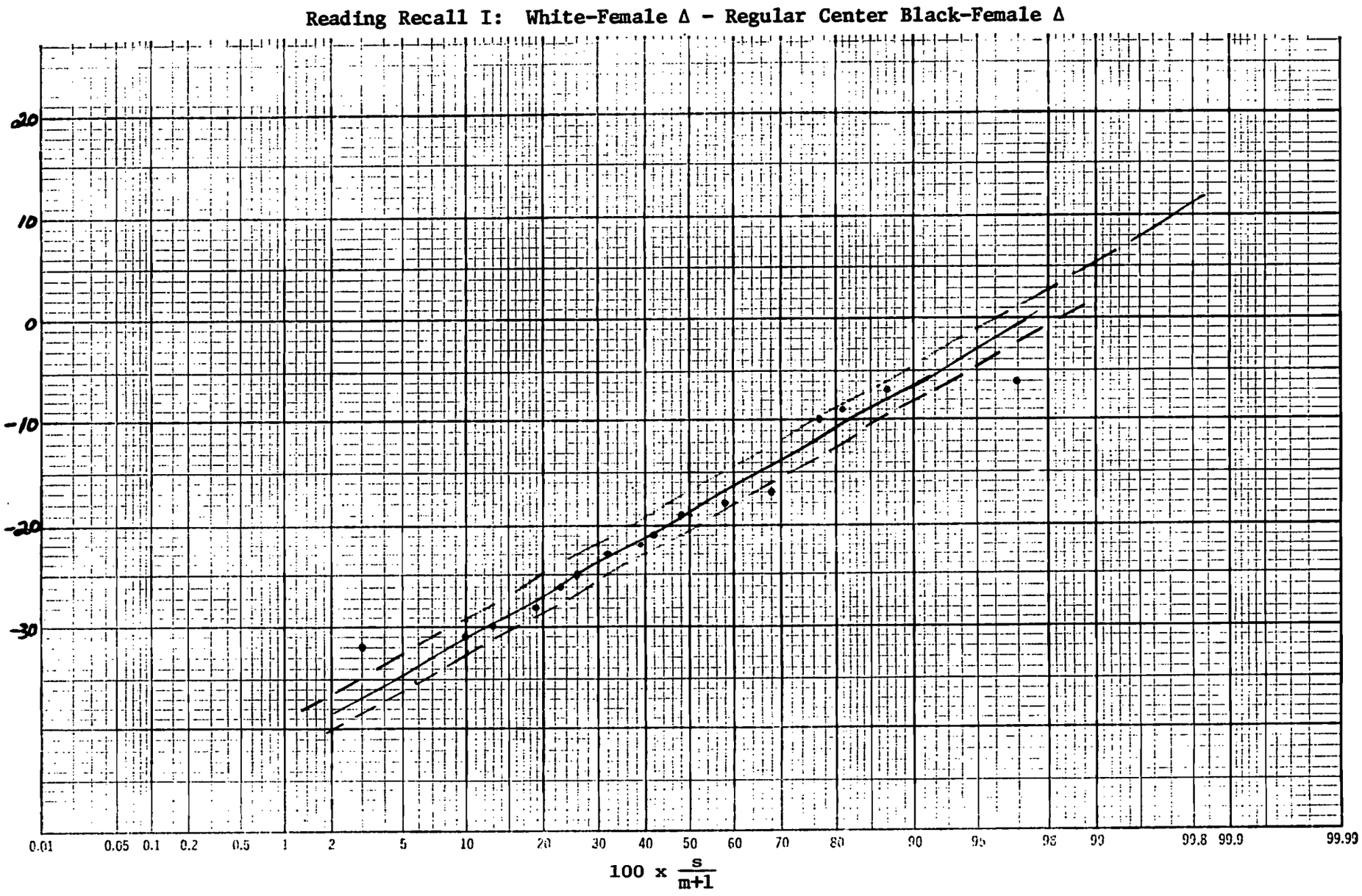
Potthoff, R. F. Statistical aspects of the problem of biases in psychological tests. Institute of Statistics Mimeo Series, No. 479. Chapel Hill: University of North Carolina, 1966.

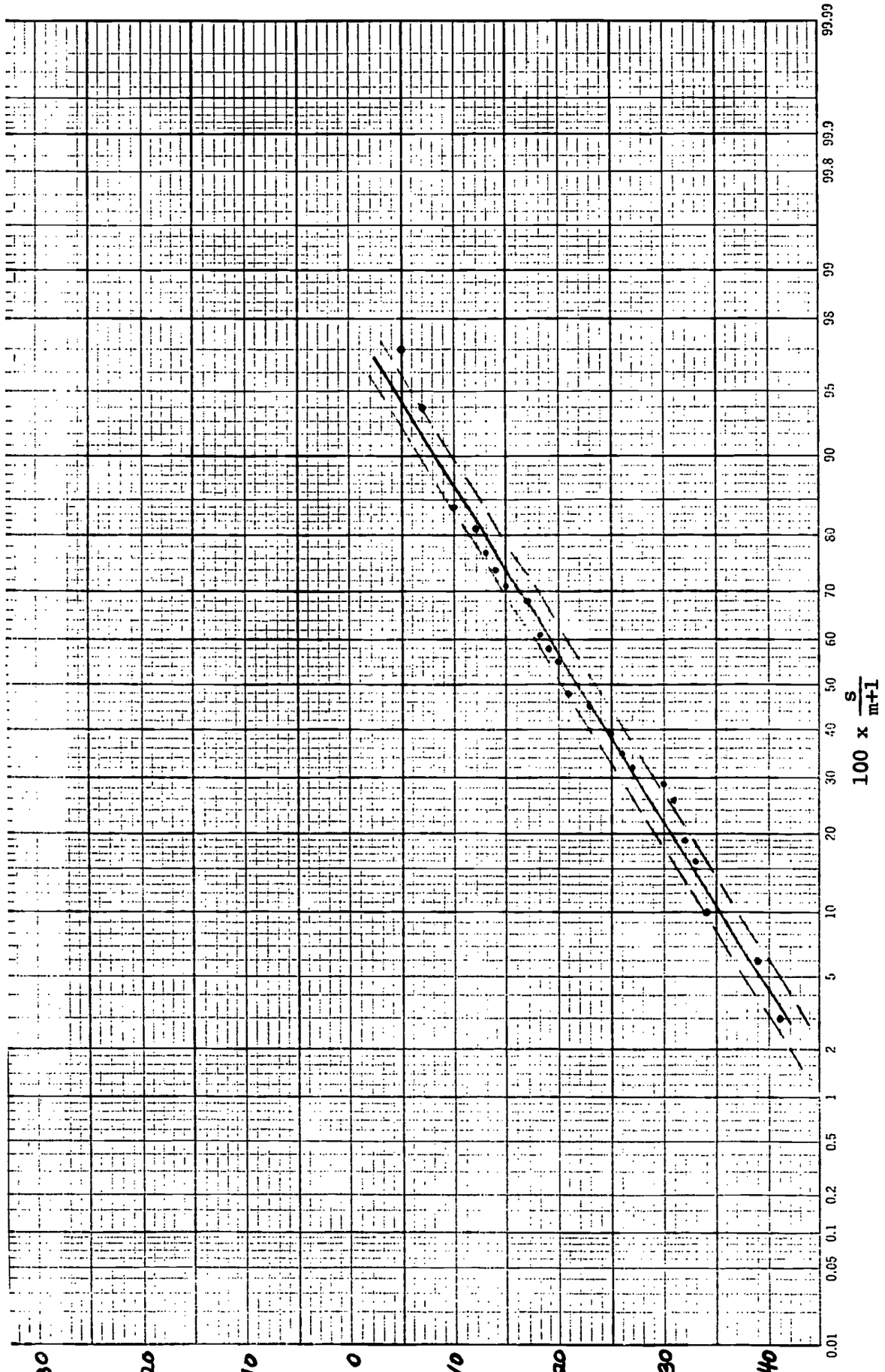
Shannon, C. E., & Weaver, W. The mathematical theory of communication. Urbana, Ill.: University of Illinois Press, 1949.

Appendix

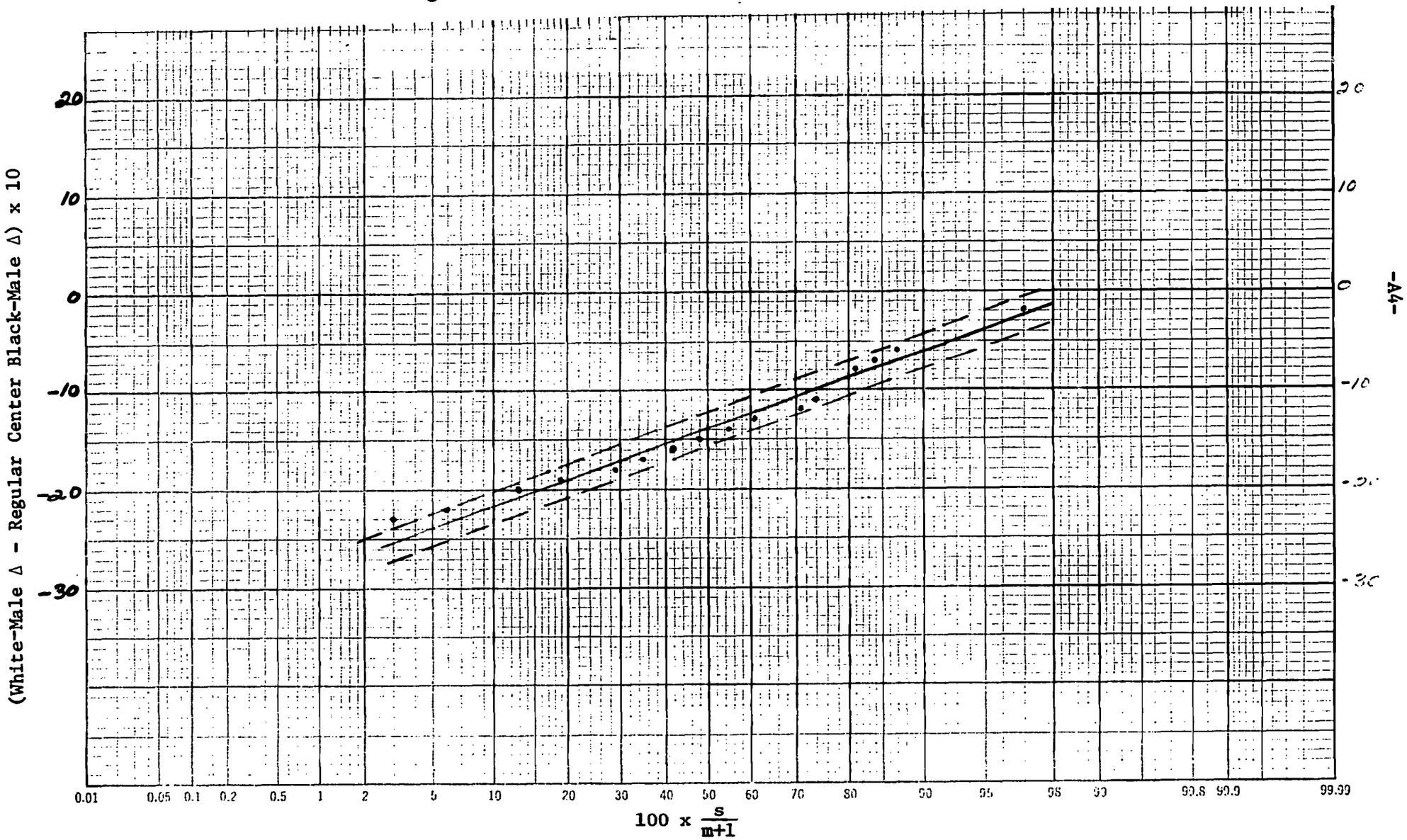
Reading Recall I: White Female Δ - Fee-Free-Female Δ

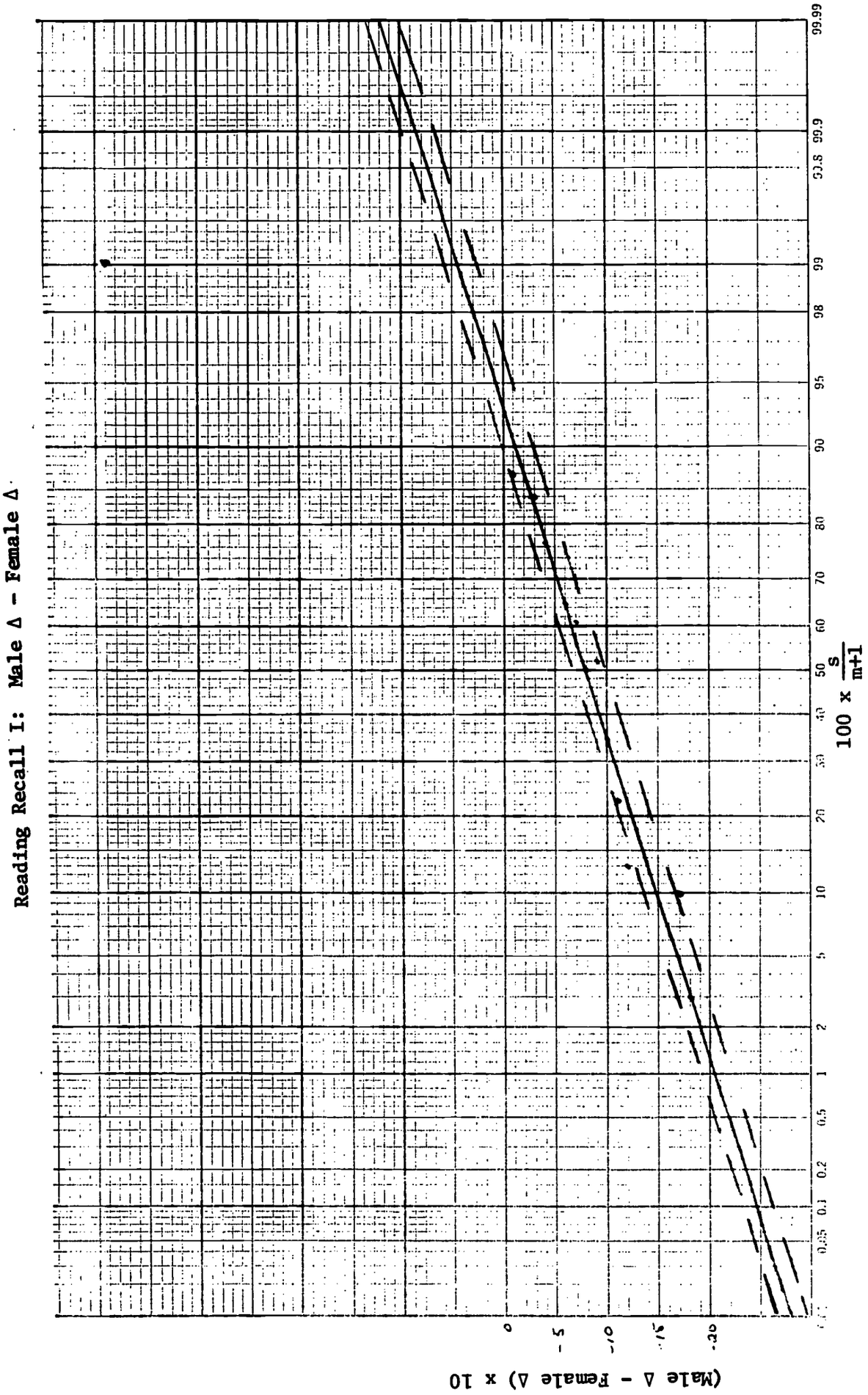



Reading Recall I: White-Female Δ - Regular Center Black-Female Δ
(White-Female Δ - Regular Center Black Female Δ) x 10
20
10
0
-10
-20
-30
0.01
0.05
0.1
0.2
0.5
1
2
5
10
20
30
40
50
60
70
80
90
95
98
99
99.8
99.9
99.99
100 x s/m+1


Reading Recall I: White Male Δ - Fee-Free-Male Δ
(White-Male Δ - Fee-Free-Male Δ) x 10
30
20
10
0
-10
-20
-30
-40
0.01
0.05
0.1
0.2
0.5
1
2
5
10
20
30
40
50
60
70
80
90
95
98
99
99.8
99.9
99.99
100 x s/m+1

Reading Recall I: White Male Δ - Regular Center Black-Male Δ




Reading Recall I: Male Δ - Female Δ
(Male Δ - Female Δ) x 10
0
-5
-10
-15
-20
99.99
99.9
99.8
99
98
95
90
80
70
60
50
40
30
20
10
5
2
1
0.5
0.2
0.1
0.05
100 x s/m+1