DOCUMENT RESUME

ED 070 794 UD 013 104

AUTHOR Green, Donald Ross; Draper, John F.
TITLE Exploratory Studies of Bias in Achievement Tests.
INSTITUTION CTB/McGraw Hill, Monterey, Calif.
PUB DATE Sep 72
NOTE 59p.; Paper presented at the American Psychological Association Annual Convention, Honolulu, Hawaii, September 1972

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Achievement Tests; Bias; Comparative Testing; Educational Testing; Ethnic Groups; *Group Tests; Item Analysis; Minority Groups; *Racial Differences; Research Methodology; *Social Differences; *Test Bias; Testing Problems; Test Validity
IDENTIFIERS California Achievement Tests

ABSTRACT

This paper considers the question of bias in group administered academic achievement tests, bias which is inherent in the instruments themselves. A body of data on the test of performance of three disadvantaged minority groups--northern, urban black; southern, rural black; and, southwestern, Mexican-Americans--as tryout samples in contrast to white, advantaged groups in the same regions, was analyzed using five different general methods for examining tests for bias. In an item tryout, a set of items is administered to a sample of the relevant population and the results are then examined item by item in an effort to pick the more effective items. The first method is an item selection routine using the point biserial correlation for each item as the criterion. The second method, group by score interactions, involves dividing the tryout group into, say, fourths, based on quartiles, and examining the proportion of the cases making each possible response in each of these levels. The third method involves plotting item difficulties so as to locate aberrant items. The fourth method involves estimating and plotting item characteristic curves separately for each group and comparing the plots. The fifth method comprises various intergroup factor analytic approaches. (Author/JM)

ED 070794

EXPLORATORY STUDIES OF BIAS IN ACHIEVEMENT TESTS

Donald Ross Green and John F. Draper

CTB/McGraw-Hill

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

UD 013104

A paper presented at the Annual Meeting
of the American Psychological Association, Honolulu.
September 1972

In this paper we wish to discuss our explorations of methods of assessing test bias. We are hoping that this information can be used to construct less biased tests as well as contribute to an understanding of the nature and sources of bias.

A biased test is popularly understood to be a test which is unfair to identifiable subgroups of the general population in which it is being used. Although many people seem to believe the matter is simple, little is actually known about the nature of bias in tests and even the most widely accepted propositions badly need verification. This is partly because this verification is deceptively difficult to obtain for many kinds and uses of tests. Sometimes quite indirect methods are needed. Williams, for example, is trying to show that the classic IQ tests favor whites and are unfair to blacks by building a similar test favoring blacks and which is unfair to whites. A second source of difficulty lies in the ambiguities in the popular definition of bias given above.

Therefore, before proceeding we wish to make a few preliminary points. First, bias is presumably a potential attribute of all kinds of tests; to keep matters simple we shall limit our discussion to typical achievement and/or ability tests. Second, we will call a test any collection of items intended to measure a single unitary domain, not collections of test batteries and other composite collections. Thus when we say "test" many would say "subtest." This we hope will also simplify matters. Third, we acknowledge that the numbers and kinds of subgroups against which a test may be biased is nearly endless; again to keep matters simple we will limit our discussion to the kinds of subgroups used in the studies being reported here today. In our work we have used groups defined by four kinds of descriptors: a) ethnic identification (black, Mexican-American, white), b) type of housing area (urban, suburban, rural), c) economic status (middle, low),

d) region (Northern, Southern, Southwestern). In our data these categories are confounded.

Finally, we want to establish clearly that the concept of bias as unfairness can be equated directly and more usefully to the proposition that bias occurs when a test measures different things for different sets of individuals. This definition does not conflict with those used by others for test bias in the absence of external validity criteria (e.g., Angoff and Ford 1971, Cleary and Hilton 1968, Potthoff, 1966), and avoids some, but not all, of the pitfalls that arise in discussions of fairness. On the face of it an unfair test must systematically yield scores for some identifiable groups that are improperly high or low; unsystematic error is lack of reliability, not bias, although consistently different amounts of error between groups can be considered a special kind of bias. Bias can occur only when two or more groups obtain scores on a test such that the scores of at least one group are typically less fair than the scores of at least one other group. The question then arises: how can that be? One possible answer is that the test has been applied unfairly or improperly to one group but not the other. Clearly this sort of thing happens and some believe that is the sum total of bias. Certainly it is a serious problem. However it is not our topic here because that source of bias is not inherent in the instrument. The question here concerns bias built into tests.

Are there any ways that bias can occur that are not a consequence of biased administration? The answer appears to be yes if, and only if, the test measures different things for at least two otherwise distinguishable groups such as the ethnic and cultural groups we are most concerned with here. If a test is properly administered under appropriate conditions and yet is biased, the most reasonable explanation is that the test is measuring something different for the different groups; otherwise the results would

be fair. This can occur in several ways. Table 1 indicates briefly a scheme for categorizing these ways.

To determine bias it is not always necessary to consider these different types of bias. This is particularly the case when there are unambiguous external criteria of validity such as in the studies being reported by Caylor here today. But whenever such criteria are lacking (e.g., in scholastic achievement tests) or when the criterion measures may themselves be biased (e.g., the Stanford-Binet used as the criterion of group ability tests) then these categories suggest ways of coming to understand the nature of the bias. Obviously understanding bias is the ultimate goal and is necessary if the bias is to be eliminated. In any case the scheme suggests ways of looking for bias.

Type A is simply unequal reliability; the variance attributable to random error is substantially larger for one group than another. Such a test would yield more inaccurate scores for one group than for the other although the direction of error, unlike the other types of bias, varies randomly within the group. Since one can determine from item tryouts just how much each item contributes to reliability, i.e. to the size of the KR 20 given the remaining items in a set, control of this sort of bias should ordinarily be relatively simple.

Just how common is this sort of bias? We don't know. Our data suggest the amount is very little although it may be a common phenomenon. It does seem unlikely that a test would be biased in only this way. We consider it unlikely partly because no ready explanation for such a phenomenon, if it exists, comes to mind. Note that for simplicity of illustration the other types of bias are illustrated as though the amount of error would be the same for both groups; this also seems unlikely. Unequal reliability would probably accompany the four other types, since in each case, one

Table 1

Types of Bias Illustrated by Hypothetical Proportions of Variance Attributable to Different Sources

Type	Group	Sources of Variance: Error	Factor 1	Factor 2	Factor 3	Description
A	I	10	90	--	--	Unequal reliability
	II	25	75	--	--	
B	I	10	30	10	50	Same factors in different proportion
	II	10	5	40	45	
C	I	10	30	60	--	Additional factor(s) for one group
	II	10	30	30	30	
D	I	10	30	--	60	Some common factors, some factors unique to each group
	II	10	30	60	--	
E	I	10	--	90	--	Nothing in common
	II	10	--	--	90	

really has different tests for the two groups.

The second type would appear on rational grounds to be highly likely while the remaining three can be considered less probable; for a set of test items to actually engage a different set of traits in different groups is more difficult to imagine than engaging the same traits in different proportions. Indeed many of the explanations of bias commonly offered fit this latter situation. For example Williams (1970) has suggested that typical reading comprehension tests measure more vocabulary among blacks than among whites and offers an example of a passage written for blacks which would reverse this. Since it is well known that the paragraph-type reading comprehension tests also measure general background--some people know more about the content of the passages than others--at least three interrelated factors probably enter into scores on such a test. Simple methods of detecting this kind of bias in test construction are not very obvious nor do easily executed corrective measures suggest themselves.

Similar problems arise with each of the remaining types. To be sure, the nature of the bias that would occur given one or another of the types of bias we have listed would be quite different. Nevertheless most of the practical ways of examining tests, especially during construction, do not distinguish among them. However, several of the possible corrective measures such as differential scoring clearly depend upon knowing with which of these types one is dealing. In short our typology serves to highlight some of the problems in assessing bias in addition to describing how it may be that a test is biased.

Let us point out here that bias of types B, C, D, E can easily lead one group to obtain consistently lower scores than the other; this is only

a possibility, not an inevitable consequence.* Consider again the example of a reading comprehension test which might fit Type B. Let it consist of questions about reading passages presented in the test. Let us assume that for middle class suburban white fifth grade children (group I) the set of questions on the passages produce a highly reliable measure with 10% of the variance among scores due to error, 30% due to differential prior knowledge of the content, 10% due to word knowledge--they all know most of the words--and the remaining 50% to reading comprehension skill per se. The same instrument for poor inner city black children (group II) might be 10% error, 5% prior knowledge (perhaps none of them know much about the content), 40% vocabulary with the remaining 45% being reading comprehension. The members of the black group would then uniformly score relatively low because of poor background knowledge and many would have scores relatively low because of poor vocabulary as well. The effect of both variables is a lower average score for the blacks. The first factor, prior knowledge, contributes little variance to black scores because of uniform lack of information while the second factor, vocabulary, contributes little variance to white scores because of uniform knowledge. Clearly an interpretation of the score as an assessment of status in reading comprehension is doubly unfair to blacks given these conditions.

As a matter of fact we do know that most academic tests, both aptitude and achievement, yield consistently higher scores for one set of groups in society in contrast to various other groups such as poor people, blacks, and Chicanos (Coleman, 1966). Some people overgeneralize these results to indicate that the latter groups are inferior to the former. In so doing they are assuming the tests are fair and unbiased (other inappropriate

* Systematic between group differences in the observed scores need not be found when there is bias; in such a case the apparent equality of performance is misleading and would not be found if the bias were eliminated.

assumptions are often made as well); if this assumption is false their conclusions become opinions without any logical basis. Yet it has not been customary practice to examine tests for bias. Obviously one cannot examine all tests for possible bias against all kinds of groups. But given the situation just described it seems painfully clear that a systematic examination for cultural bias of the major published ability and achievement tests now in use in our schools is long overdue.

Simple, readily applied procedures are needed. What we are reporting here are some explorations of ways to proceed toward that end. So far we have tried five sorts of approaches, none of them definitive. They are, (1) the point biserial approach, (2) group by score category interactions for mean item difficulties, (3) the adjusted item difficulty approach, (4) the estimated item characteristic approach, and (5) the intergroup factor approach. The first approach was developed in a study previously reported (Green 1972). Since much of our data came from the source used in that study and since we wish to include a report of an attempt to verify the conclusions reached then we will describe the sample and procedures used there first.

THE POINT BISERIAL APPROACH: INITIAL STUDY

This study compares the results of using three disadvantaged minority groups--northern, urban black; southern, rural black; and southwestern Mexican-American--as tryout samples in contrast to white, advantaged groups in the same regions. In an item tryout a set of items are administered to a sample of the relevant population and the results are then examined item by item in an effort to pick the more effective items.

Would an item tryout using these different groups lead to the selection of different items from the item pool? If so:

(1) Do the different items selected measure different things?

(2) Are the resulting item sets "better" for the minority groups?

(3) Will the relative discrepancy in scores favoring majority groups be reduced by using a minority tryout group?

Method

The data were derived from that obtained during the standardization of the *California Achievement Tests, 1970 Edition* (CAT - 70) published by CTB/McGraw-Hill. The CAT-70 is a general achievement battery with five overlapping levels, four of which were used. The standardization took place early in 1970 and involved over 200,000 students in about 400 schools. The items in the battery came from a variety of sources, but it is fair to say that they were written by and for "middle America." The tryout samples also fit this description. Thus, the tests should favor white middle-class Americans if they favor any group.

All schools participating in the CAT-70 standardization answered questionnaires which provided information on the basic character of the population served. From the data on these questionnaries, seven groups of the schools were drawn for this study. The characteristics and sizes of these groups are shown in Table 2. The samples used in this study are drawn from schools serving pupils highly homogeneous with respect to ethnic background and rather homogeneous with respect to socioeconomic status.

Data Analyses

The basic procedure used for examining the data was an item selection routine using the point biserial correlation for each item as the criterion. Each of the seven groups was treated as a tryout sample with the items in each test functioning as an item pool. For each group on each test at each

Table 2

Characteristics of the Sample Groups

Group Number	Geographic[1] Region	Residential Type	Ethnic Group	Socioeconomic Status	Number of Cases by Grade 1	3	5	8	10
I	North	Residential Suburban	White (97%)[2]	Middle (81%)[2]	299	225	265	328	--
II	North	Central City	Black (99%)	Low (81%)	285	304	278	250	--
III	South	Residential Suburban	White (99%)	Middle (77%)	361	211	293	304	279
IV	South	Rural	Black (100%)	Low (96%)	202	220	171	245	183
V	South	Rural	White (91%)	Low (81%)	323	200	199	296	246
VI	Southwest	Small and Large Cities	Mexican-[3] American (87%)	Low (82%)	146	144	169	399	--
VII	Southwest	City and Suburban	Anglo-American (99%)	Middle (81%)	189	218	249	277	--

[1]The states containing these particular school systems are--North: Illinois, Indiana, Kansas, New Jersey; South: Alabama, Georgia, South Carolina; Southwest: Arizona, Oklahoma, Texas.

[2]Estimated per cent of cases falling in the category.

[3]81% speak mostly Spanish at home.

grade, the "best" half of the items (i.e., those with the highest item-test correlations) were noted. Four kinds of analyses were made.

(1) The number and percent of items chosen for one group in the pair but not for the other was recorded. These items will be called "biased." The number of these biased items in any one comparison suggests the degree to which the two groups interact with the test items in a distinct manner.

(2) Scores for each group in a pair were obtained on both sets of biased items. These two tests may be called the "majority biased test" and the "minority biased test" since they contain the items uniquely best for the respective groups. The correlation between each group's score on the two tests was found and estimates of the variance not common to the two were made to judge how different the sets of items really are in what they measure.

(3) Another analysis consisted of examining and comparing KR 20 reliability estimates.

(4) Finally, mean scores were examined for changes in the relative status of the groups as a result of item selection.

Results

Proportions of biased items. The medians of the proportions of biased items among those selected are shown on Table 3 for all possible pairs of groups. The overall median proportion was approximately .30. The white, middle-class groups appear more like each other (these pairs had lower medians) than they are like the minority groups. The latter also had more in common than they shared with the three majority groups.

Independence of biased item tests. All groups differed from their pairs to some degree by the criterion of proportion of biased items, and some of the differences appear to be substantial. However, it is possible

Table 3

Median Proportions of Biased Items
For Each Pair of Groups

Group	I	II	III	IV	V	VI	VII
I	---	.36	.26	.35	.30	.38	.26
II	.36	---	.33	.26	.25	.25	.41
III	.26	.33	---	.38	.30	.33	.27
IV	.35	.26	.38	---	.30	.30	.41
V	.30	.25	.30	.30	---	.24	.33
VI	.38	.25	.33	.30	.24	---	.43
VII	.26	.41	.27	.41	.33	.43	---

that these sets of biased items still measure much the same thing. To examine this possibility, scores for each individual were obtained on both biased item tests. This was possible since each individual answered all items. The correlations between these two scores were obtained for each group on each test. These correlations varied from -.17 to +.82 with a median of about .5 (.55 for group I and .46 for group II) which leaves a lot of variance unaccounted for. Since the number of biased items was very small in many cases, the reliabilities of the biased tests are typically low; thus the median after correction for attenuation is near .8 (.84 and .77 respectively; range -.30 to +1.00). But even allowing for this, it appears that in many instances the majority and minority tests measure somewhat different things and as a rule do so for both groups involved.

Reliability. As noted earlier one case of bias occurs if the test scores of one group contain substantially more error than they do for another group. The overall median KR 20's on the full-tests for groups I through VII were .91, .91, .91, .92, .93, .90, and .92, respectively. Obviously, there is little evidence of bias by this criterion, although a test-by-test comparison of these reliabilities shows that the figures are higher for the majority group more often than not (97 of 162 comparisons). The data concerning reliabilities after item selection also show a very small amount of bias so defined. These results do not preclude the possibility that other kinds of reliability estimates might show more bias of this sort but they do not make it seem likely.

Changes in test scores. Another way to look at bias is to assert that the scores of some groups are unfairly low because the test does not adequately measure all the relevant abilities or knowledge, and, in particular, does not measure well those relevant attributes on which the group in question happens to score well. If the item pool contains items which measure

these attributes at all, a selection routine using this group might be expected to increase the importance of these attributes in determining the total score, thereby reducing the disadvantage of the group. Therefore, the three minority groups considered here might be expected to do relatively better on the items selected as best for them than they did on the original test. Each group's improvement on each of the nine tests in the battery was compared to the improvement shown by its comparison group. The minority groups showed greater relative improvement consistently in the upper grades, but not in grades 1 and 3. As was the case for proportions of biased items, the southern, rural, white group does not fit the pattern: the item selection procedure helped them as often as it helped the rural blacks, perhaps because their initial scores were more alike to begin with, especially in the lower grades.

The argument of the preceding paragraph would appear to have even greater force for the biased item tests.

The majority biased item tests (note this is the set of items best for the majority) are almost uniformly more difficult for both groups than are the minority biased item tests. The differences between majority group mean scores and minority group mean scores are usually smaller on the minority biased item tests than on the majority biased item tests. Table 4 shows the frequencies of this phenomenon. In most cases the relative advantage of the majority groups was reduced when using items chosen as best for the minority group but was increased when using items chosen as best for themselves.

In short, each analysis indicated bias, apparently small in amount, but clearly suggesting that ordinary item selection procedures may be producing biased tests. One such study hardly proves the point but it does give credibility to the possibility. To examine this possibility more closely we set out to both confirm the results of the initial study and

Table 4

Number of Comparisons in Which Mean Difference on Biased Item Tests Favors Each Group[a]

Grade	Comparison Groups II & I		IV & III		IV & V		VI & VII		Totals Min.	Totals Maj.	χ^2	p
1	5	3[b]	6	3	8	1	8	1	27	8	10.3	.01
3	5	4	5	4	3	6	7	2	20	16	0.4	NS
5	7	1[b]	5	4	7	2	7	2	26	9	8.3	.01
8	8	0[b]	9	0	6	3	5	4	28	7	12.6	.001
10	--	--	6	3	5	4	--	--	11	7	0.9	NS
Totals	25	8	31	14	29	16	27	9	112	47		
χ^2	8.8		6.4		3.8		9.0		26.6			
p	.01		.02		.05		.01		.001			

[a]Let $\bar{Y}_m$ = majority mean on majority test, $\bar{X}_m$ = minority mean on majority test, $\bar{Y}_n$ = majority mean on minority test, and $\bar{X}_n$ = minority mean on minority test. Then, $\bar{Y}_m - \bar{X}_m > (\bar{Y}_n - \bar{X}_n)$ favors minority; $\bar{Y}_m - \bar{X}_m < (\bar{Y}_n - \bar{X}_n)$ favors majority.

[b]Note that analyses were not made for the Total Language of the CAT-70 for this group at this grade. Therefore, comparisons were made for only eight tests.

then to look at other procedures for assessing bias in achievement tests.

THE POINT BISERIAL APPROACH: SUBSEQUENT ANALYSES

First we wish to report on our efforts to verify the outcomes of the original study. To do this the data were examined in several ways. The principle analyses were intended to provide a look at the stability of the data and to yield some cross validations. This was accomplished by redoing portions of the study for random halves of each group and applying that outcome to the other halves.

Thus one-half of the grade 5 northern white group (I) was selected randomly as was one-half of the grade 5 northern black group (II) making four groups. We will call the first half of group I and the first of group II the criterion halves and the remaining half of each group the cross validation halves. Then using the reading comprehension test, the point biserial for the first half of group I and again for the first half of group II were found and the "best" items (those with the highest point biserials for this half of the group) were selected. As before the "biased" items were then determined. This procedure was repeated 100 times for the grade 5 reading comprehension test using the two northern groups, I and II, so that the stability of various statistics could be observed; the other half of each group was used as cross validation data. The same kinds of analyses were done for five other grade test combinations: the grades 3, 5, and 8 reading comprehension test for groups III and IV; the grade 3 computation test and the grade 8 language mechanics test for groups VI and VII.

Proportion of biased items. The first statistic above in need of examination is the proportion of "biased" items summarized in Table 3 which shows the median proportion of biased items for the various pairs of groups. Table 5 adds detail to this picture confirming that groups I, III, and VII,

Table 5

A Comparison of the Median Proportions of Biased Items for Similar and Unlike Pairs of Groups

Test or Grade	Groups Compared						
	Majority[1]	Minority[2]	All Similar Pairs	I vs. Minority[3]	III vs. Minority[4]	VII vs. Minority[5]	All Unlike Pairs
Reading Vocabulary	34	30	32	38	42	48	45
Reading Comprehension	28	28	28	31	32	42	37
Arithmetic Computation	18	27	22	29	28	26	28
Arithmetic Concepts and Problems	27	28	29	41	42	45	42
Language Mechanics	24	22	23	44	38	43	42
Language Usage and Structure	30	33	32	35	44	42	39
Grade 1	25	36	30	40	33	42	40
Grade 3	22	22	22	31	46	36	35
Grade 5	30	24	27	46	40	48	44
Grade 8	28	30	28	36	28	44	36

NOTE: [1]I-III, I-VII, III-VII; [2]II-IV, II-VI, IV-VI; [3]I-II, I-IV, I-VI; [4]III-II, III-IV, III-VI; [5]VII-II, VII-IV, VII-VI.

the majority groups, are similar to each as are the minority groups II, IV, and VI when contrasted to the nine possible minority-majority pairs. The consistency of this result is very strong since it applies without exception to the comparisons of the medians based on all tests within each grade as well as those based on all grades for each kind of subtest. By this criterion it seems clear that less test bias against a group can ordinarily be expected if the items were chosen using data based on a similar group.

The stability of the proportion of biased items statistic can also be seen in Table 6 which shows the means and standard deviations over the 100 trials for each of the six tests. The variability seems quite small for most of them.

Another indication of stability is the frequency with which particular items were chosen as biased. Table 7 shows how many times for both the group I and II and the group III and IV comparisons each of the 42 reading comprehension items in the grade 5 test were chosen as biased in favor of whites, in favor of blacks, for both or neither. Thirty-six of the 42 items were categorized in the initial study exactly the same way as they were most of the time for the random halves of group I and II while 40 of 42 were the same for the group III and IV comparison. Clearly most items are consistently assigned; contradictory assignment such as being chosen for both groups and also being rejected for both are rare. Furthermore, particular items tend to be categorized the same in both the group I vs. II and group III vs. IV comparisons. In fact the point biserials for groups I and III correlate .61; the figure for II and IV is .85. The corresponding item difficulty correlations are .98 and .94. In short both the number of biased items and which items they are tend to be stable because the items have similar characteristics in similar groups.

Table 6

Number of Biased Items Found in Follow-up Studies

Groups	Grade	Test	Initial Study		100 Repeated Trials		
			No. of Items	(%)	Mean	(%)	S.D.
I & II	5	Reading Comprehension	10	(48)	8.9	(43)	1.4
III & IV	3	Reading Comprehension	6	(26)	8.3	(36)	1.3
III & IV	5	Reading Comprehension	9	(43)	9.3	(44)	1.3
III & IV	8	Reading Comprehension	9	(39)	7.1	(32)	1.5
VI & VII	3	Arithmetic Computation	9	(25)	12.2	(34)	1.8
VI & VII	8	Language Mechanics	14	(39)	10.7	(30)	1.8

Table 7

Frequency of Selection of Each Item in the
Grade 5 Reading Comprehension Test for the Criterion Halves*

Groups	Items Chosen For	Item Numbers																				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
I & II	Both Groups	-	-	2	-	65	27	-	33	92	97	3	-	-	29	1	26	45	90	8	1	-
	Group I Only	-	-	-	-	28	33	3	7	-	3	96	67	-	64	62	2	1	10	23	98	10
	Group II Only	52	78	59	20	5	24	29	56	8	-	-	-	-	2	4	65	51	-	20	-	-
	Neither Group	48	22	39	80	2	16	68	4	-	-	1	33	100	5	33	7	3	-	49	1	90
III & IV	Both Groups	10	-	31	-	37	8	38	7	91	37	3	14	-	88	2	41	14	90	5	5	-
	Group III Only	-	-	1	-	26	75	-	1	-	23	69	-	-	6	15	32	2	5	4	93	11
	Group IV Only	90	100	68	100	22	1	54	84	9	33	-	86	-	6	5	14	61	5	49	-	-
	Neither Group	-	-	-	-	15	16	8	8	-	7	28	-	100	-	78	13	23	-	42	2	89

Groups	Items Chosen For	Item Numbers																				
		22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
I & II	Both Groups	-	-	48	57	2	45	97	2	31	-	95	-	1	17	68	36	83	66	3	-	-
	Group I Only	1	27	-	2	-	-	1	89	47	-	3	74	95	-	12	1	6	6	41	10	9
	Group II Only	-	-	52	37	98	54	2	-	7	-	2	-	-	83	18	60	11	27	7	-	-
	Neither Group	99	73	-	4	-	1	-	9	15	100	-	26	4	-	3	3	-	-	49	90	91
III & IV	Both Groups	-	-	97	30	93	99	100	1	-	-	10	-	-	10	94	8	91	20	6	-	-
	Group III Only	1	31	-	43	-	1	-	96	29	-	25	11	34	-	6	81	-	79	83	-	34
	Group IV Only	-	-	3	17	7	-	-	-	-	-	17	-	-	90	-	-	9	-	2	-	-
	Neither Group	99	69	-	10	-	-	-	3	71	100	52	89	66	-	-	11	-	1	9	100	66

*The underlining indicates the selections made in the initial study.

Independence of biased item tests. The data from the initial study indicated that the biased item set favoring each group usually measured different things. The new analysis permits cross validation of this result since the halves of the groups not used in choosing the items also obtained scores on these sets of items. This cross validation was done for the grade 5 Reading Comprehension Test in groups I and II. In the initial study the correlations between the two sets of the biased items were .55 and .36 for groups I and II, respectively. The median correlations for the 100 criterion halves were .54 and .35, respectively. For the blacks the two tests measure substantially different things. The medians for the cross validation halves were .57 and .40. Since the size of these coefficients are about what was obtained initially, the correction for attenuation should also yield about the same results. The cross validation correlations do tend to be slightly higher for both groups indicating a some what lesser tendency for the two tests to measure different things but the differences are not sufficient to alter the interpretation that the two sets of items tend to measure rather different things in both groups. Again the results of the initial study are confirmed.

Changes in test scores. The final matter to be verified is the phenomenon considered in Table 4, the advantage to a group in mean score relative to the other group of having the test consist of items chosen as best for them. In Table 8 the relevant data from the initial study is compared to the corresponding means for the criterion halves and the 100 cross validation halves. The outcome of the initial study is fully supported for the first and last of the six tests considered but either unsupported or contradicted by the data for the other four. From these results we conclude that biased item tests as we have defined them do not necessarily yield relatively higher or lower scores than do other item sets for any group. On the other hand, in some cases, as illustrated by the Language Mechanics

Table 8

Mean Difference Between Mean of Mean Item Difficulties for Majority and Minority Groups on Majority and Minority Biased Tests for the Criterion and Cross Validation Groups on Six Tests

GRADE	GROUPS	TEST	STUDY*	MAJORITY TEST		MINORITY TEST		DIFFERENCE		MINORITY GAIN**		FREQUENCY OF POSITIVE MINORITY GAIN
				Majority Group	Minority Group	Majority Group	Minority Group	Majority Group	Minority Group	Mean	S.D.	
5	I & II	Reading Comprehension	1)	.620	.350	.870	.610	.250	.260	.010	--	--
			2)	.648	.371	.845	.584	.197	.213	.016	.038	63
			3)	.656	.366	.844	.585	.188	.219	.031	.038	75
3	III & IV	Reading Comprehension	1)	.817	.467	.817	.450	.000	-.017	-.017	--	--
			2)	.820	.493	.845	.506	.025	.013	-.012	.033	34
			3)	.821	.489	.841	.507	.020	.018	-.002	.033	44
5	III & IV	Reading Comprehension	1)	.722	.311	.856	.500	.134	.189	.055	--	--
			2)	.681	.320	.874	.496	.193	.176	-.017	.035	32
			3)	.681	.318	.870	.493	.189	.175	-.014	.035	34
8	III & IV	Reading Comprehension	1)	.556	.333	.778	.611	.222	.278	.056	--	--
			2)	.552	.350	.692	.480	.140	.130	-.010	.049	40
			3)	.554	.346	.693	.482	.139	.136	-.003	.052	43
3	VI & VII	Arithmetic Computation	1)	.822	.400	.922	.656	.100	.256	.156	--	--
			2)	.877	.568	.900	.588	.023	.020	-.003	.110	55
			3)	.878	.571	.900	.588	.022	.017	-.005	.115	52
8	VI & VII	Language Mechanics	1)	.643	.371	.828	.721	.185	.350	.165	--	--
			2)	.673	.437	.786	.633	.113	.196	.083	.051	97
			3)	.675	.441	.785	.633	.110	.192	.082	.052	92

*1) Initial Study; 2) Criterion Halves; 3) Cross Validation Halves.
**Minority group mean difference less majority group mean difference.

test a pronounced advantage does occur. Note that a .08 mean difference in item difficulty is the equivalent of a six point change in score, which is about sixteen percentile points around the median.

This rehash of the procedures and data of the initial study still leaves the interpretations somewhat ambiguous. The point biserial approach appears to show some bias in some CAT tests against minority groups but in very small amounts in all but a couple of instances. Since the items examined were all preselected on the basis of data from a single "standard" (i.e., heterogeneous) tryout sample, it is quite possible that these data produce an underestimation of the amount of bias. Furthermore, it is plain that the interpretation of differences in point biserials as bias is not in itself unambiguous unless one can adequately account for the role of difficulty in these differences. In some instances items have low point biserials because of floor or ceiling effects. However, examination of the distributions of item difficulties obtained before and after selection based on point biserials does not show much change although extremely easy and extremely difficult items tend to be eliminated. The distributions vary from excellent to terrible for the tests and groups considered, but these distributions do not seem to be directly related to any conclusions drawn about bias so far. We will consider the many questions that arise about difficulty again later in the paper since some of what follows throws light on the matter.

In any case it is obvious that using differential item-test correlations as the criterion of bias is not the only reasonable approach to assessing bias in an achievement test and so the next step is to consider other approaches.

GROUP BY SCORE LEVEL INTERACTIONS

A customary way of looking at item analysis data is to divide the tryout group into fourths, or fifths based on quartiles or quintiles, respectively, and examine the proportion of the cases making each possible

response in each of these levels. Given different tryout groups such data can be examined for interaction by a chi square test. This can be done for each possible response. Black and standard groups were used in item tryouts for a number of tests now under construction at CTB. Chi square tests for interactions between black and white tryout group response patterns were undertaken for these data and for the group I and group II grade 5 data from the initial study. Table 9 gives the information obtained on two versions of an item meant for a first grade oral usage test. The first version of the item is as follows: Look at the first picture. This girl can draw. Now look at the second picture. Listen to this sentence. *This is the picture she drawed.* I will say it again. *This is the picture she drawed.* Mark your answer. The second version is identical except that the sentence reads "This is the picture she has drawn." The first version produces a significant interaction while the second does not largely because it does not function well in either group. Figures 1 and 2 illustrate the distribution by fifths.

Table 10 indicates the frequency with which interactions were found for various tests. The reason for the high frequency of "biased" items in the CAT grade 5 tests and the Science Skills tests is not apparent although one can easily imagine why phonological discrimination and oral usage items appear to be discriminatory.

The one characteristic the science tests and the CAT test share that the others do not is that they have a uniformly large white-black difference in item difficulty. Very few of the CAT items have group by fifth plots that cross and few of them look much like either of the figures given. Instead they tend to look like that for item 16 (see Figure 15) of the grade 5 reading comprehension test. The point biserials are almost the same, .39 and .36, but the chi square is 11.7. Clearly difficulty is functioning differently

Table 9

Item Analysis Data on Two Versions of an Oral Usage Item for Standard and Black Tryout Groups

STANDARD
Form X Oral Usage
Item 17 Grade 1.2

	A	B*
Percent	.53	.45
High 5th	.12	.88
Mid 5th	.23	.77
Mid 5th	.54	.44
Mid 5th	.88	.09
Low 5th	.89	.05
Choice N	126	106
Test Mean	27.1	38.2

Item Statistics
Difficulty = .447
Point Biserial = .657
Biserial = .827
Select = .008

Summary Data
N = 237 Mean = 31.96
KR 20 = .87 S.D. = 8.09

BLACK
Form X Oral Usage
Item 17 Grade 1.2

	A	B*
Percent	.69	.26
High 5th	.39	.61
Mid 5th	.78	.22
Mid 5th	.80	.17
Mid 5th	.81	.14
Low 5th	.70	.13
Choice N	124	46
Test Mean	26.8	32.1

Item Statistics
Difficulty = .256
Point Biserial = .325
Biserial = .443
Select = .005

Summary Data
N = 180 Mean = 27.58
KR 20 = .80 S.D. = 7.01

$\chi^2 = 41.8$

STANDARD
Form Y Oral Usage
Item 17 Grade 1.2

	A*	B
Percent	.79	.18
High 5th	.77	.23
Mid 5th	.67	.33
Mid 5th	.81	.16
Mid 5th	.90	.10
Low 5th	.80	.11
Choice N	201	47
Test Mean	30.6	33.4

Item Statistics
Difficulty = .788
Point Biserial = .080
Biserial = .112
Select = .004

Summary Data
N = 255 Mean = 30.71
KR 20 = .85 S.D. = 7.77

BLACK
Form Y Oral Usage
Item 17 Grade 1.2

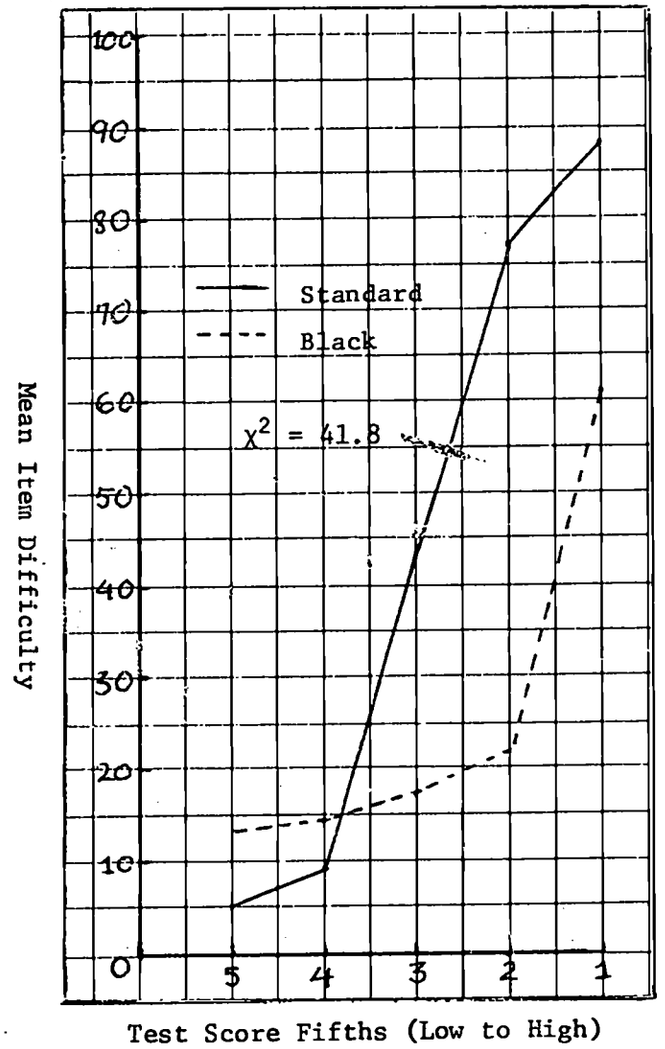
	A*	B
Percent	.79	.16
High 5th	.85	.15
Mid 5th	.87	.07
Mid 5th	.80	.17
Mid 5th	.90	.10
Low 5th	.61	.24
Choice N	120	24
Test Mean	28.8	26.8

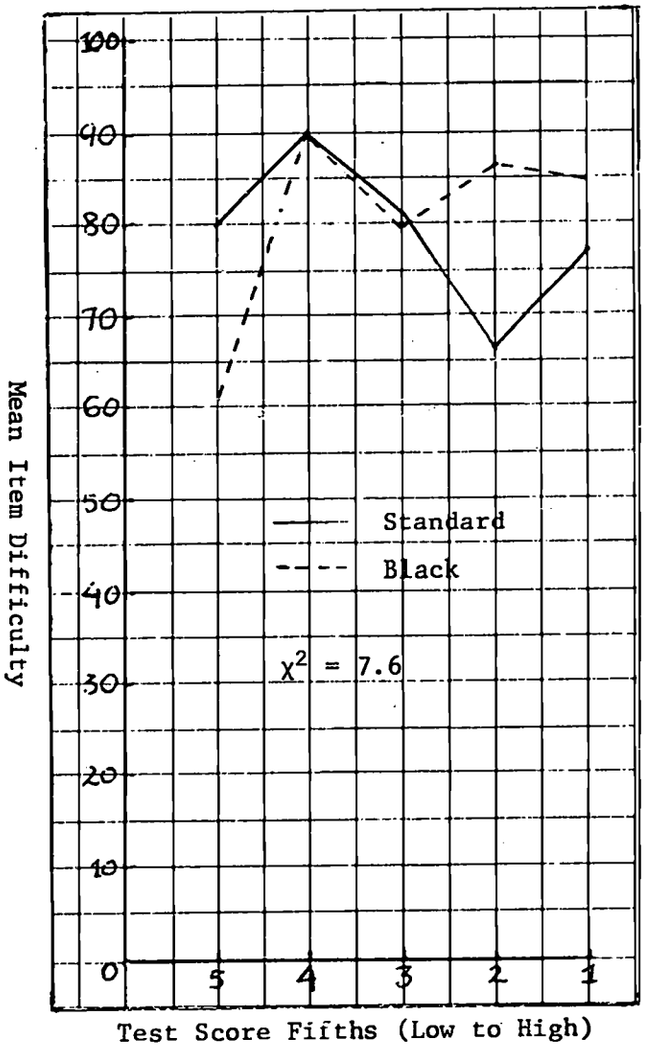
Item Statistics
Difficulty = .795
Point Biserial = .171
Biserial = .242
Select = .001

Summary Data
N = 151 Mean = 27.95
KR 20 = .80 S.D. = 6.97

$\chi^2 = 7.6$

* correct response





Figures 1 and 2. Mean Item Difficulties in Score Categories for Grade 1 Standard and Black Tryout Groups on Two Versions of an Oral Usage Item.

If the product $A_c' A_c$ is temporarily defined as a diagonal matrix Λ_c by requiring that the columns of A_c and F_{ci} are orthogonal, we have

$$Y_i' Y_j = F_{ci} \Lambda_c F_{cj}'.$$

Then the expression for the symmetric matrix

$$Y_i' Y_j Y_j' Y_i = F_{ci} \Lambda_c F_{cj}' F_{cj} \Lambda_c F_{ci}' = F_{ci} \Lambda_c^2 F_{ci}',$$

is in the form of an easily soluable eigenvalue-eigenvector problem.

Then we may solve for F_{cj} from the expression

$$F_{cj}' = \Lambda_c^{-1} F_{ci}' Y_i' Y_j \qquad \text{since}$$

$$Y_i' Y_j = F_{ci} \Lambda_c F_{cj}',$$

$$F_{ci}' Y_i' Y_j = \Lambda_c F_{cj}' \qquad \text{and}$$

$$\Lambda_c' F_{ci}' Y_i' Y_j = F_{cj}'.$$

Given Y_i, Y_j, F_{ci}, and Λ_c we may solve for A_c by the expression in adjoined matrices below.

$$\left[Y_i \vdots Y_j \right] \quad \begin{bmatrix} F_{ci} \\ \cdots \\ F_{cj} \end{bmatrix} \quad \Lambda_c^{-\frac{1}{2}} = A_c.$$

Now the expressions

$$(Y_i - A_c F_{ci}')' (Y_i - A_c F_c') = F_i \Lambda_i F_i', \text{ where } \Lambda_i = A_i' A_i,$$

$$(Y_i - A_c F_{ci}') = A_i F_i', \text{ and } (Y_i - A_c F_{ci}') F_i = A_i = Y_i F_i$$

allow us to solve for A_i and F_i which completes the estimation of parameters.

Table 10

Proportion of Items in Various Tests
Showing Group by Score Interactions

Test Name	Grade	Number of Items Examined	Interactions	
			Number	Proportion
PRIMARY READING				
Letter Names	1	5	0	0
Letter Sounds	1	37	9	.24
Letter Sounds	2	12	1	.08
Visual Discrimination	1	10	1	.10
Visual Discrimination	2	5	0	0
Listening Comprehension	1	20	6	.30
Totals		89	17	.19
READING				
Word Reading	2	34	8	.24
Phonic Errors	2	30	4	.13
Reading Comprehension	2	30	6	.20
Reading Comprehension	3	65	17	.26
Reading Vocabulary	3	39	15	.38
Totals		198	50	.25
LANGUAGE				
Phonological Discrimination	1	52	28	.54
Oral Usage	1	92	35	.38
Oral Language	2	24	4	.17
Punctuation & Capitalization	3	34	16	.47
Language Usage	3	24	4	.17
Spelling	3	34	3	.09
Totals		260	90	.35
SCIENCE				
Science	3	40	31	.78
Science	6	43	34	.79
Science	8	18	13	.72
Science	10	75	45	.60
Science	12	52	36	.69
Totals		228	159	.70
SOCIAL SCIENCE				
Social Science	3	49	18	.37
Social Science	6	76	22	.29
Social Science	8	28	6	.21
Social Science	10	77	16	.21
Totals		230	62	.27
MATHEMATICS				
Mathematics	1	42	8	.19
Math Computation	2	58	7	.12
Math Computation	3	58	17	.29
Math Concepts & Applications	2	40	7	.18
Math Concepts & Applications	3	62	10	.16
Totals		260	49	.19
C.A.T.				
Reading Vocabulary	5	33	27	.82
Reading Comprehension	5	39	32	.82
Math Computation	5	63	52	.82
Math Concepts & Problems	5	36	27	.75
Totals		171	138	.81

This method may be generalized to the case where there are m sub-groups of interest. The overall model is then

$$\left[Y_1 \vdots Y_2 \vdots \cdots \vdots Y_m \right] =$$

$$\left[A_c \vdots A_1 \vdots A_2 \vdots \cdots \vdots A_m \right] \begin{bmatrix} \Lambda_c^{\frac{1}{2}} & 0 & 0 & 0 & \cdots & 0 \\ 0 & \Lambda_1^{\frac{1}{2}} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \Lambda_2^{\frac{1}{2}} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & \cdot & & & \cdot \\ 0 & 0 & 0 & & \cdots & \Lambda_m^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} F_{ci}' & F_{c_2}' & \cdots & F_{cm}' \\ F_1 & 0 & 0 \cdots & 0 \\ 0 & F_2' & 0 \cdots & 0 \\ 0 & 0 & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & F_m' \end{bmatrix}$$

The F_{ci} may be estimated from the recursive set of equations

$$Y_1' Y_2 \; Y_2' Y_3 \; Y_3' \cdots Y_m \; Y_m' Y_1 = F_{ci} \; \Lambda^m \; F_{ci}'$$

$$\Lambda^{-(m-1)} \; F_{ci}' \; Y_1' \; Y_2 \; Y_2' \cdots Y_{m-1} \; Y_{m-1}' \; Y_m = F_{cm}'$$

$$\Lambda^{-(m-2)} \; F_{ci}' \; Y_1' \; Y_2 \; Y_2' \cdots Y_{m-2} \; Y_{m-2}' \; Y_{m-1} = F_{cm-1}'$$

$$\vdots$$

$$\Lambda^{-1} \; F_{ci}' \; Y_i' \; Y_2 = F_{c_2}'$$

then

$$\left[Y_1 \vdots Y_2 \vdots \cdots \vdots Y_m \right] \begin{bmatrix} F_{c^1} \\ \cdots \\ F_{c^2} \\ \cdots \\ \cdot \\ \cdot \\ \cdot \\ \cdots \\ F_{cm} \end{bmatrix} \quad \Lambda_c^{-\frac{1}{2}} = A_c \text{ and for each } Y_i - A_c \; F_{ci}',$$

A_i and F_i may be estimated as before.

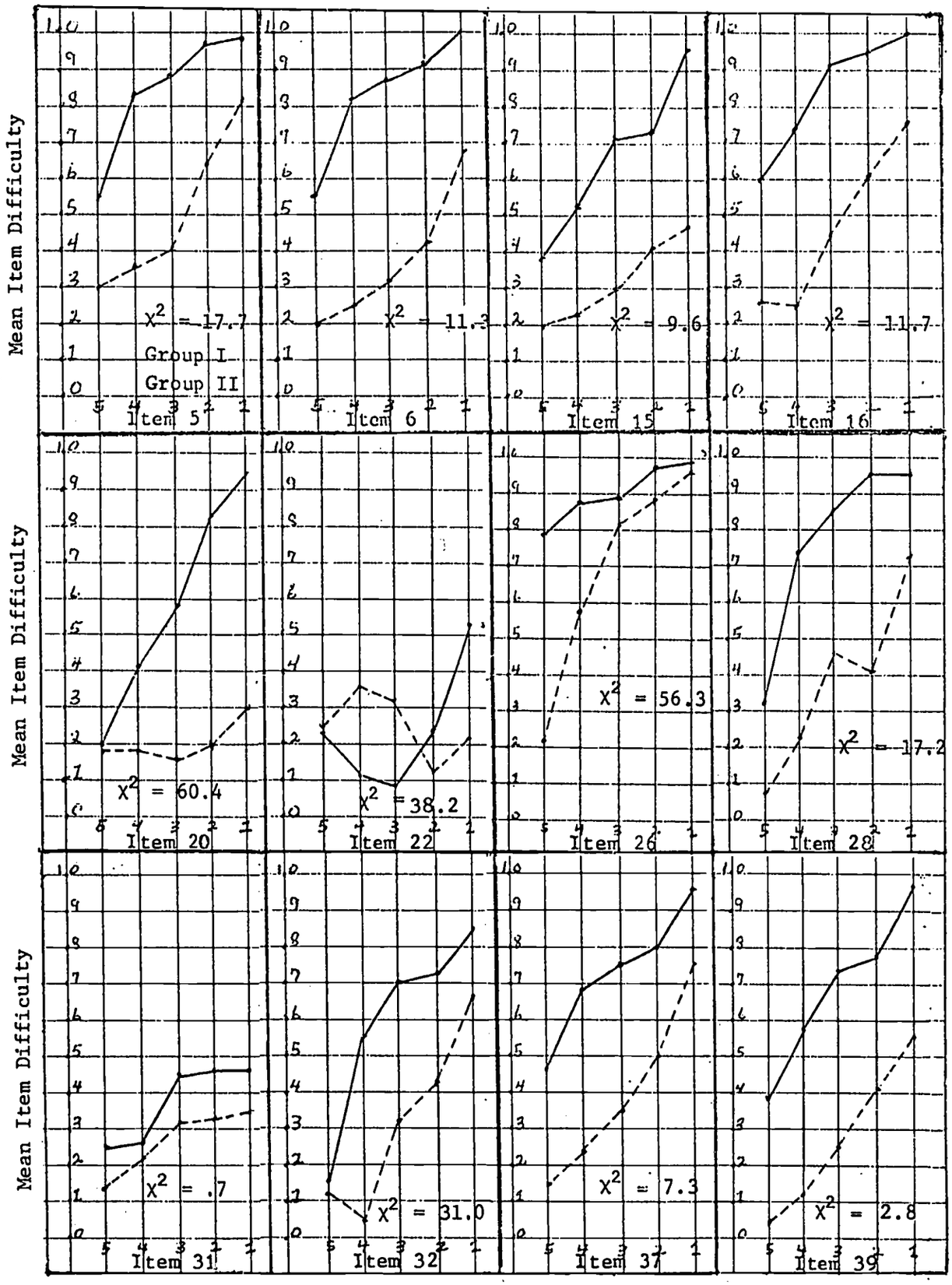


Figure 15. Mean Item Difficulties in Total Score Fifths (Low to High) for Groups I and II on Several Items of the Grade 5 Reading Comprehension Test.

Results of the Application of the Inter-Group Analysis

Computer programming problems precluded having an inter-group analysis for three or more groups available at this time. We do have however, the results of two different two group analyses. One analysis was done for previously defined groups I and II as a black-white inter-group analysis, the other analysis involved the two white groups I and III so as to establish a benchmark for the interpretation of the white-black inter-group results.

The first of these two analyses were based on the data from 270 black and 360 white fifth grade students on the 42 items in the CAT-70 reading comprehension test. For the other analysis the group I data were put together with data on the same test obtained from 396 5th grade students in group III. For both analyses the first stage of parameter estimation was to estimate the eigenvalues of the common inter-group space. In both cases plots of the eigenvalues indicate that the common space should be considered to consist of three dimensions; in both cases three dimensions accounted for almost 80% of the common space variance.

In the next step of parameter estimation three columns of each of the matrices which were with respect to sources of variation common to groups were estimated. This was done for each of the two analyses and then the eigenvalues of the group specific spaces were estimated. In general it appeared there were three dimensions as well to each group specific space.

The third step was to estimate three columns of each group specific matrix and then to determine the various proportions of variance of the total space which could be explained by each source. Tables 11 and 12 contain the proportions of test variance and Tables 13 and 14 contain frequencies of items which fall into classes according to the proportion of item variance attributable to the group specific source.

in the two approaches. Let us, therefore, consider difficulty more directly.

ADJUSTED DIFFERENCES IN ITEM DIFFICULTIES

Angoff has examined a way of looking at plots of item difficulties so as to locate aberrant items in a subtest and thus to examine them for unfairness and exclusion. As a modification of this procedure, it is suggested that the item-test biserial correlations be incorporated in the procedure so as to estimate linear test score-item score regression whereby adjusted item difficulties may be formed in a manner analogous to the way in which adjusted means are formed in an Analysis of Covariance. Such a procedure would allow the effect of differential item-test correlations as well as differential item difficulty to influence the location of aberrant items. For example, if such a procedure were employed, an item which would be deviant in terms of item difficulties and which would have a low item-test correlation for one or both groups would show up as relatively more deviant than a similar item with a high item-test correlation. It may be argued that an item showing more aberrance in adjusted item difficulty is indeed more deviant than one could have inferred from the unadjusted plot.

Such an argument rests upon the contention that given two items with equal differences in difficulty for two groups, the item which is more strongly related to the test score is more likely to be reflecting group differences in the behaviors the test accesses than is likely for the item which is less strongly related to the test score.

Both adjusted and ordinary item difficulties were calculated from a set of primary level item tryout data obtained from black and standard tryout samples. Figures 3 and 4 show the relationship between the two plotting procedures. Note that the two do not produce the same ordering of the items as to aberrance.

Table 11

Proportions of Variance for the Group I-II Comparison

Source	Group	
	I	II
Common Intergroup	67	60
Group Specific	11	14
Residual and Error	22	26

Table 12

Proportions of Variance for the Group I-III Comparison

Source	Group	
	I	III
Common Intergroup	70	70
Group Specific	9	8
Residual and Error	21	22

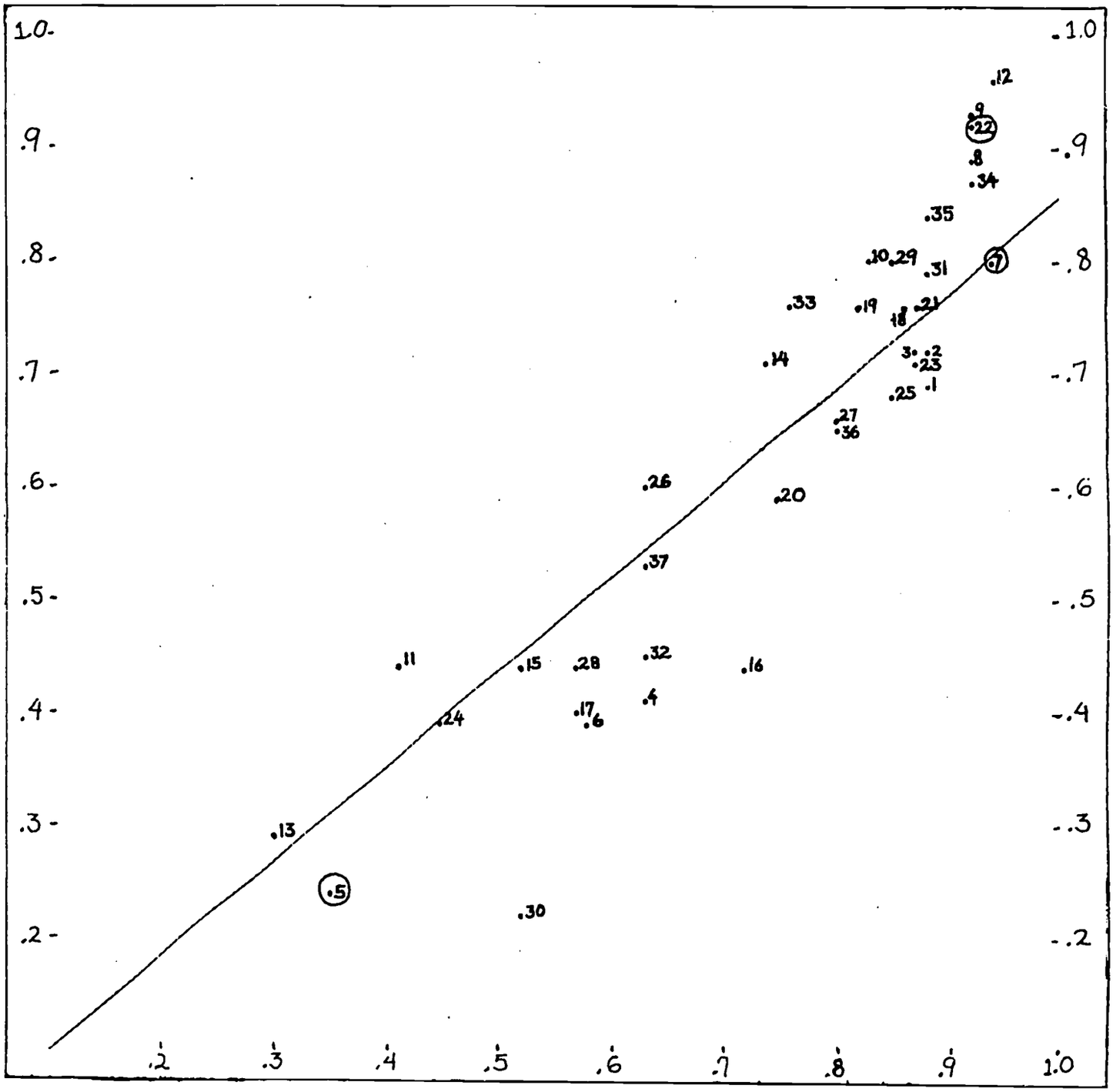


Figure 3. Cross Plots of Item Difficulties on a Phonological Discrimination Test for Grade 1 Standard and Black Tryout Groups.

Table 13

Frequencies of Items within Categories of
Group Specific Variance Accounted for by Groups I and II

Proportion of Group Specific Variance	Group I	Group II
$0 < X \leq .05$	17	16
$.05 < X \leq .10$	10	10
$.10 < X \leq .15$	7	2
$.15 < X \leq .20$	2	7
$.20 < X \leq .25$	3	2
$.25 < X \leq .50$	3	4
$.50 < X \leq 1.00$	0	1

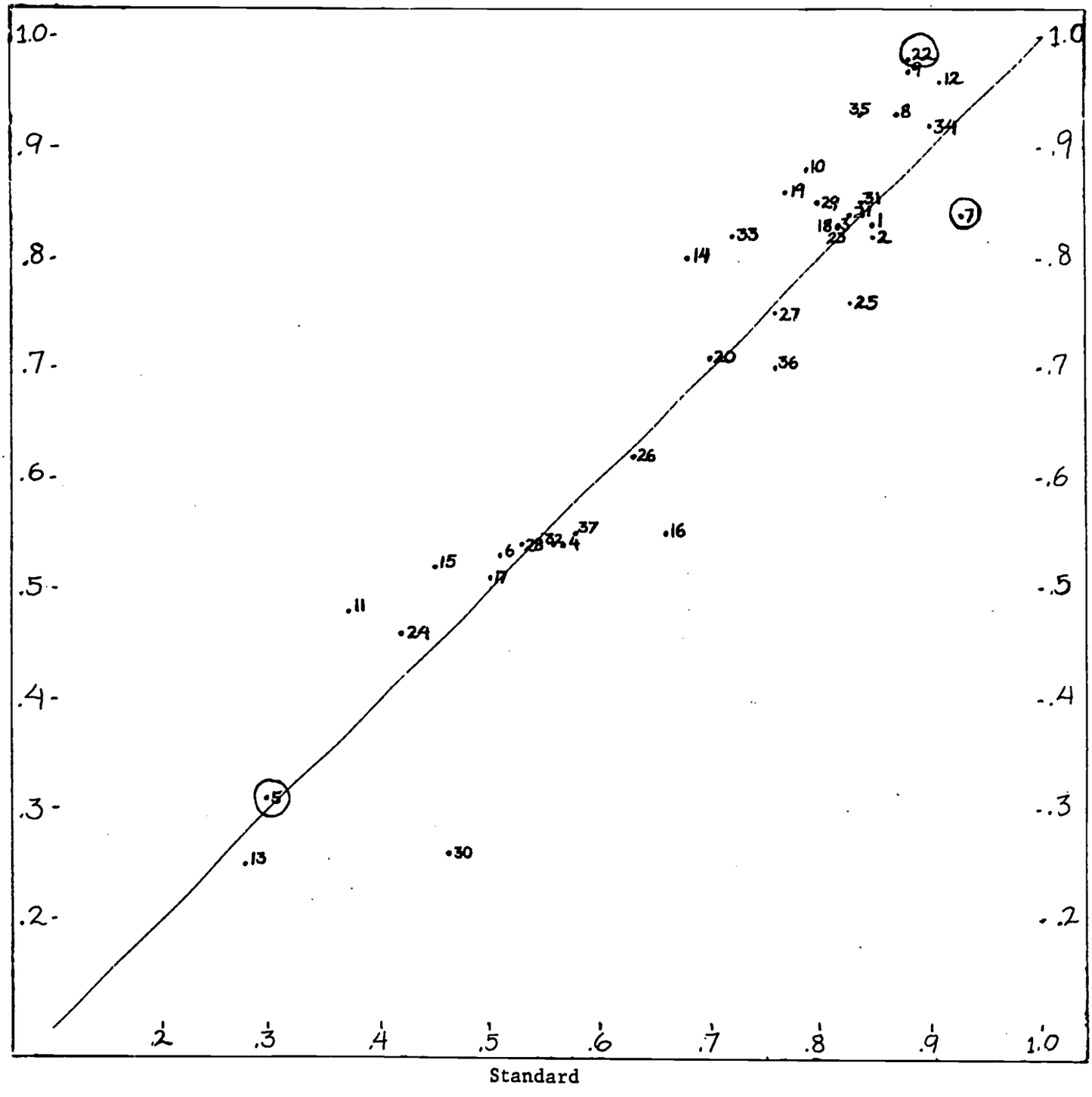


Figure 4. Cross Plots of Adjusted Item Difficulties on a Phonological Discrimination Test for the Grade 1 Standard and Black Tryout Groups Shown in Figure 3.

Table 14

Frequencies of Items within Categories of
Group Specific Variance Accounted for by Groups I and III

Proportion of Group Specific Variance	Group I	Group III
$0 < X \leq .05$	17	16
$.05 < X \leq .10$	12	13
$.10 < X \leq .15$	4	5
$.15 < X \leq .20$	4	2
$.20 < X \leq .25$	2	4
$.25 < X \leq .50$	3	2
$.50 < X \leq 1.00$	0	0

ESTIMATION OF ITEM CHARACTERISTIC CURVES

Another way of looking for items which may be in some sense unfair is to estimate and plot item characteristic curves separately for each group and to compare the plots. Item characteristic curves are essentially representations of nonlinear regressions of the probability of correct response on the latent trait which a test attempts to measure. If the test score is taken as a reasonable estimate of the trait, we may estimate the regression of the probability of correct response to an item on the test score by means of a higher order polynomial and plot the polynomial function as our estimate of the characteristic curve. The plots of characteristic curves obtained separately for groups of interest may then be superimposed and inspected for possible group by item interaction.

One weakness of this approach is that it requires a large number of subjects in each group in order to achieve estimates of any quality. If sufficient data can be obtained, however, the procedure provides a graphic representation which is easily inspected and which provides detail beyond the distribution by fifths. Our own experience has been that a few items which appeared to be acceptable when their distributions by fifths were examined, had estimated item characteristic curves which indicated that they were less than desirable. Figure 5 contains the most egregious example of all the estimated item characteristic curves which we plotted. Note that there is a group by item interaction and that the curves are not monotonic or constantly increasing as is desirable of such curves. We have a hypothesis for the behavior of the curves in Figure 5 with respect to a reading comprehension item. The information required to answer the item is in the second sentence of a paragraph. The "topic sentence" read without the rest of the paragraph would lead one to select one of the incorrect foils. It is hypothesized that those students who scored in the lower half of their

An examination of Tables 11 and 12 lead to an estimate of approximately 5% of group specific variance in the Group II model, beyond the benchmark 9 or 10% that could be expected from very similar groups. For the test overall this would seem to be neither an absolution of nor an indictment for unfairness. Tables 13 and 14 indicate much the same on the item level. If one were to arbitararily establish 25% as an undue amount of group specific item variance then there are 9 unfair items out of 42, since some of the same items are unfair for more than one group. Of the nine, two had greater than 50% of their item variance attributable to residual or errors and are thus just plain bad items. Excluding those items there are still more items, 4, indicated as unfair for Group II than any other group. One of those items indicated as unfair by Group II is item 22 for which you have seen the plot of its estimated item characteristic curve (Figure 5).

Figures 8 through 11 are the Group I and II plots of estimated item characteristic curves for the items which were indicated unfair by the Group II analyses and Figure 12 is the plot of one of the items indicated unfair by the Group I analysis.

Consider what the exclusion of the items plotted in Figures 8 through 12 would do to total scores. The items in Figures 9 and 10 show a clear separation between the Group I and the Group II curves over the major portion of their score range with Group I above the curve for Group II. Thus it is clear that the exclusion the two items plotted in Figures 9 and 10 would result in an increase in the probability of a higher relative score (relative to an overall mean score) for almost all of the individuals in Group II. Figure 11 on the other hand shows a separation only in the lower score range and thus the exclusion of this item would increase the probability of a higher relative score for only those members of Group II who score in that lower range. The exclusion of the item plotted in Figure

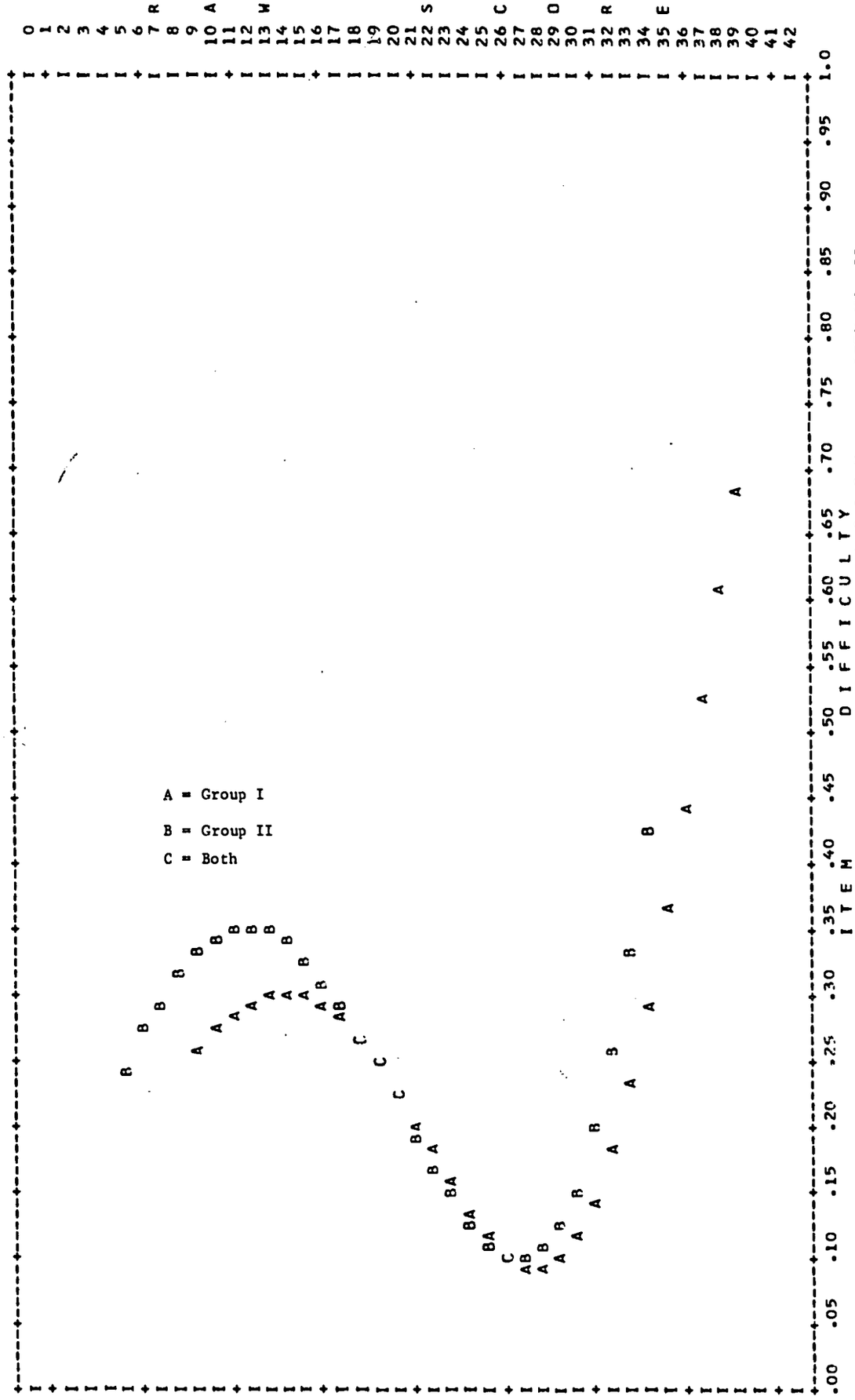


Figure 5. The Estimated Item Characteristic Curve for Item Number 22.

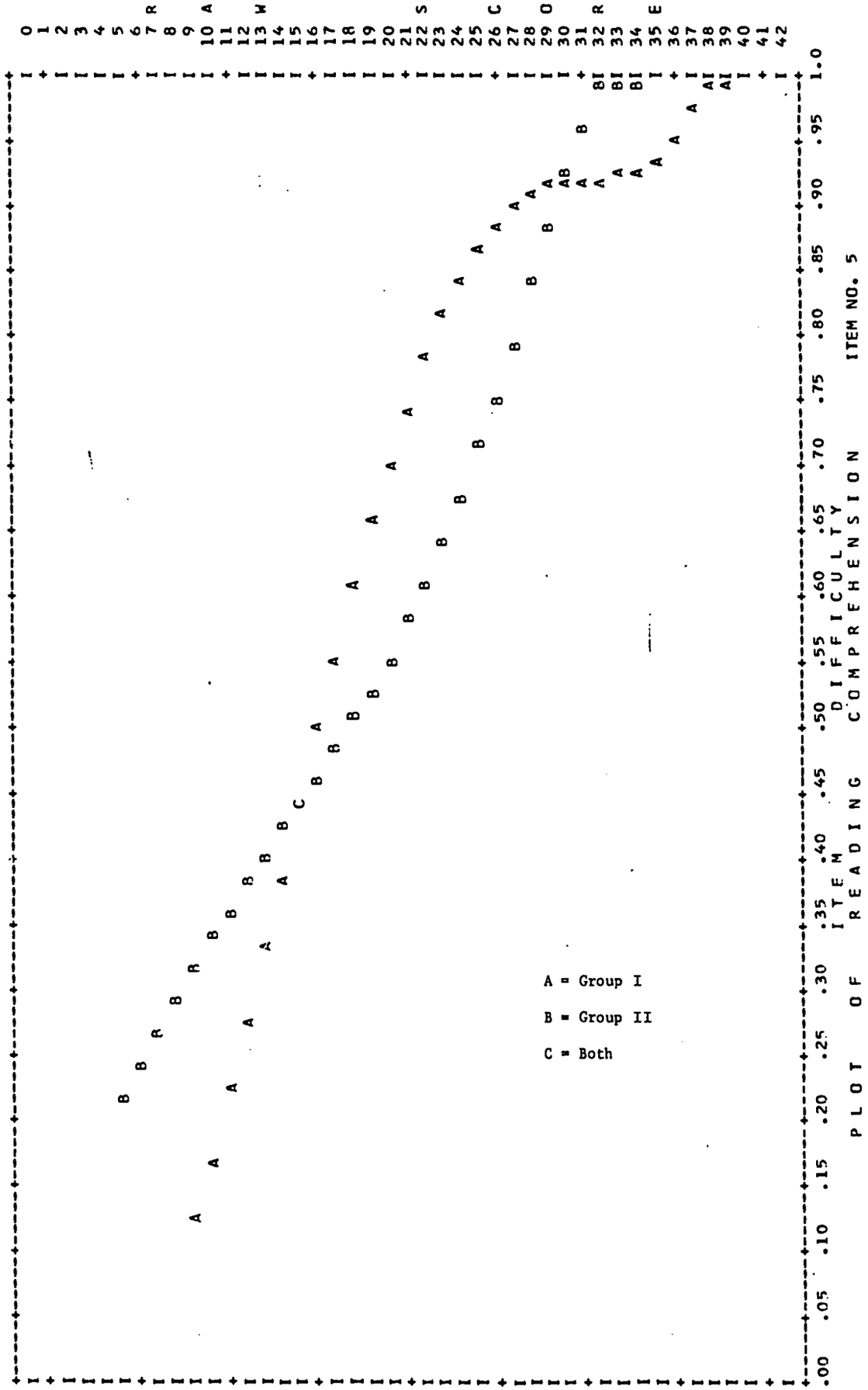


Figure 8. The Estimated Item Characteristic Curve for Item Number 5.

groups tended to read the passage slowly but completely and therefore were not lead astray, while those scoring a bit higher overall tended to scan the passage too rapidly and were misled. However, the highest scoring students read quickly yet assimilated the entire passage and thus were not fooled. Lest you get the wrong impression, Figure 6 contains a more typical pair of plots and Figure 7 contains the best of 42 pairs of plots where "best" means the plots which appear most like the standard conception of a good ogive.

INTERGROUP FACTOR ANALYTIC APPROACHES

In this section, a type of factor or component analysis will be outlined which it is suggested may be useful in the examination of achievement tests for bias. In order to see why any type of factor analytic method would be appropriate, consider the nature of an achievement test. From an achievement test score one infers the location of a student in the domain of the test by means of a conglomerate statistic based on the evaluations of a series of responses to items, items which may be considered stimulus aggregates. Each separate response evaluation is itself an inference based on assumptions about 1) the conceived behavioral domain of the test, 2) the relevance of the item to that domain for the subject who responds, and 3) the mutual understanding of respondent and response evaluator about the general rules of response evaluation.

If the domain of a test is such that only one type of achievement behavior is accessed by subjects in responding to the items, it is likely that the items will each relate in differing amount to the domain for a subject. The evaluations of item-domain relationships can be conceived of as forming a pattern. If two groups have different patterns of item-domain relations, it is possible that the domains of behavior accessed by

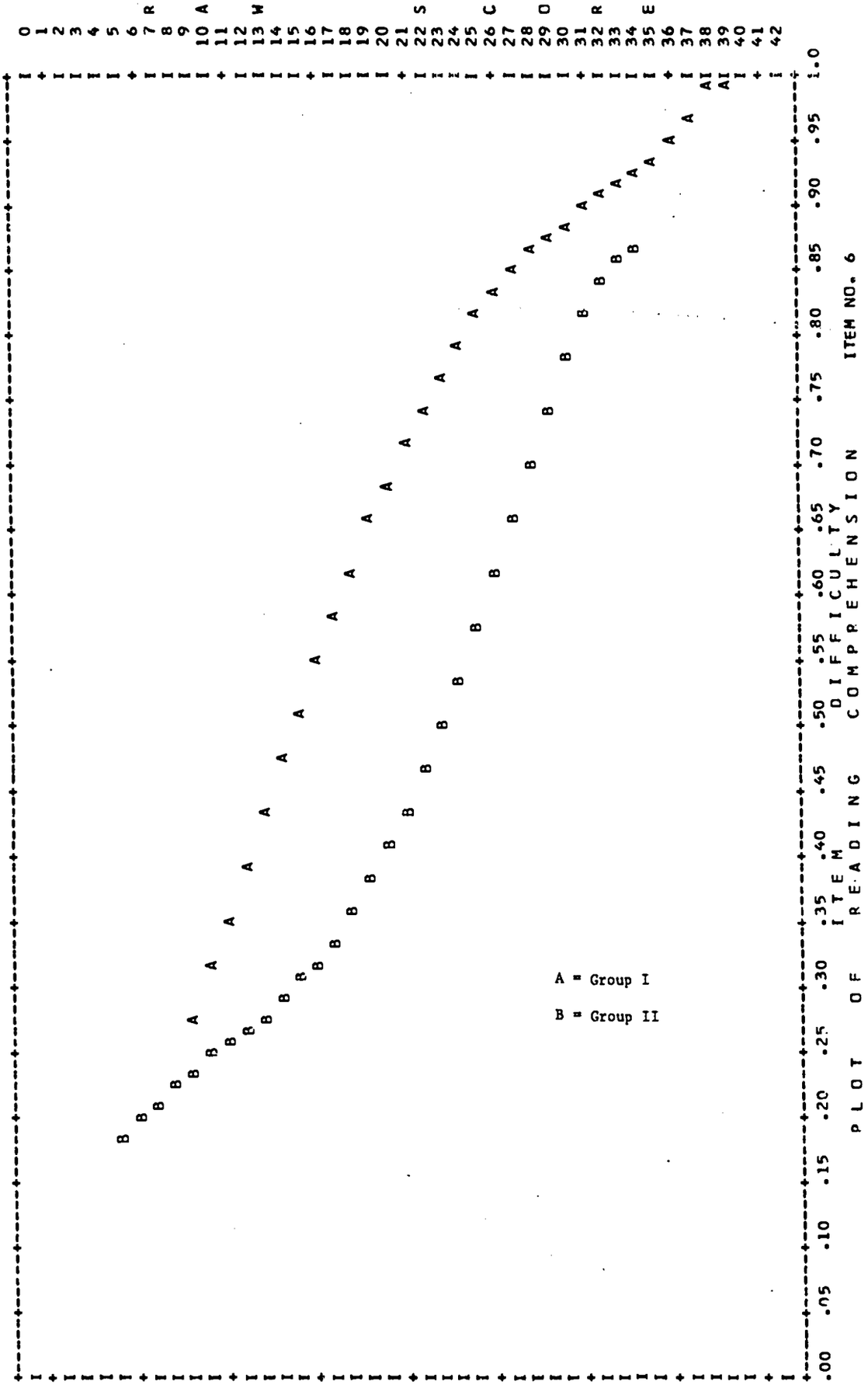


Figure 9. The Estimated Item Characteristic Curve for Item Number 6.

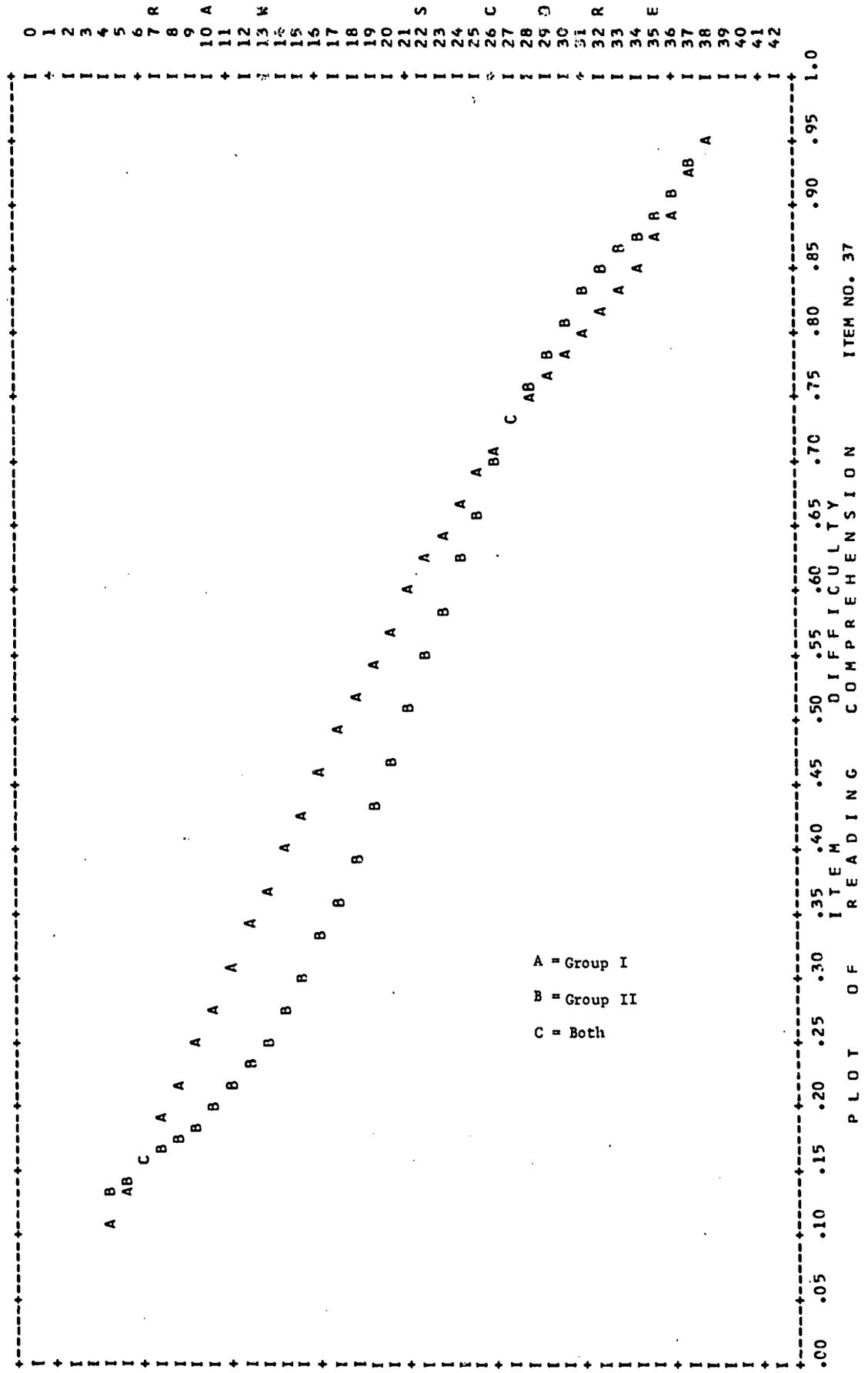


Figure 6. The Estimated Item Characteristic Curve for Item Number 37.

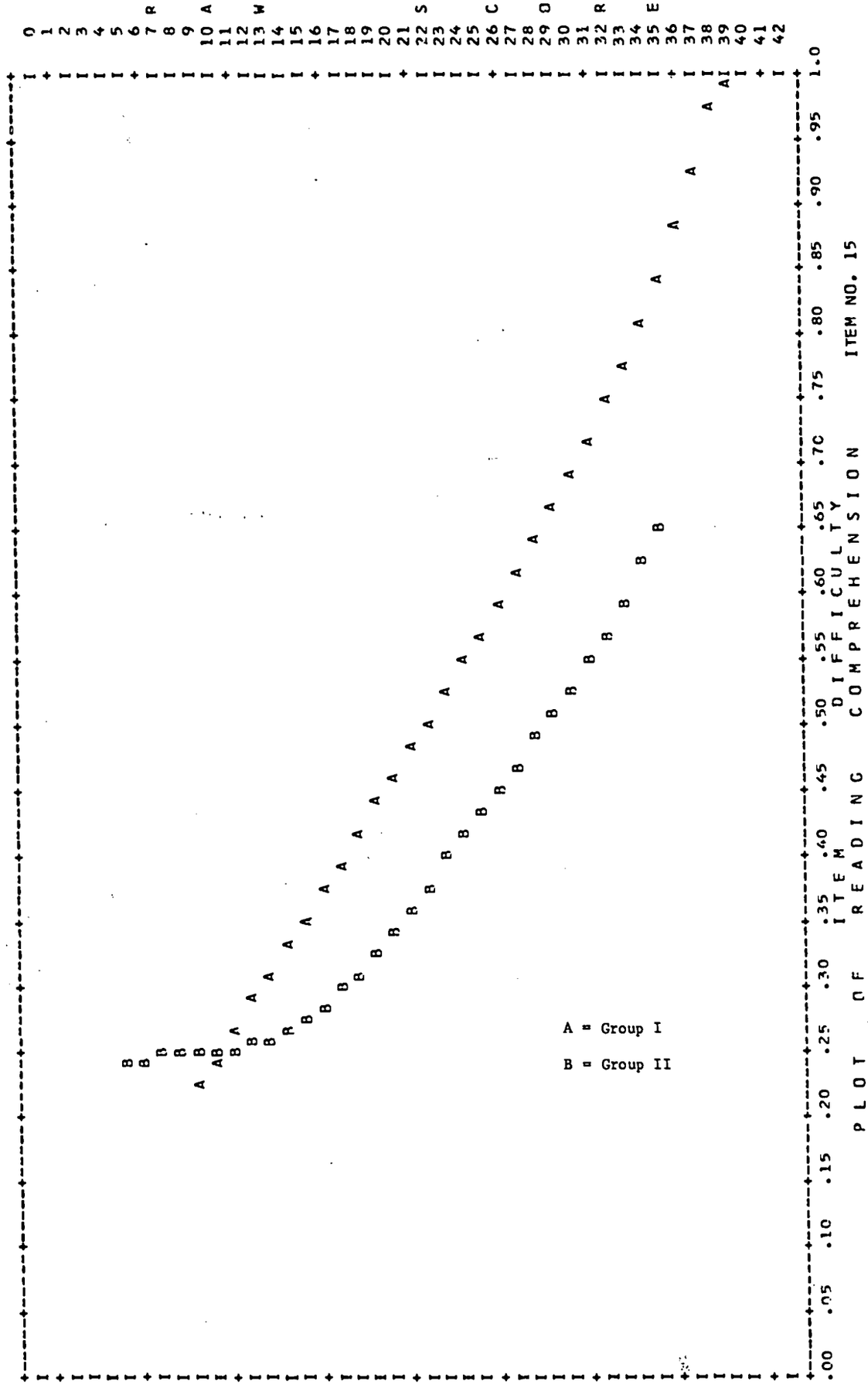


Figure 10. The Estimated Item Characteristic Curve for Item Number 15.

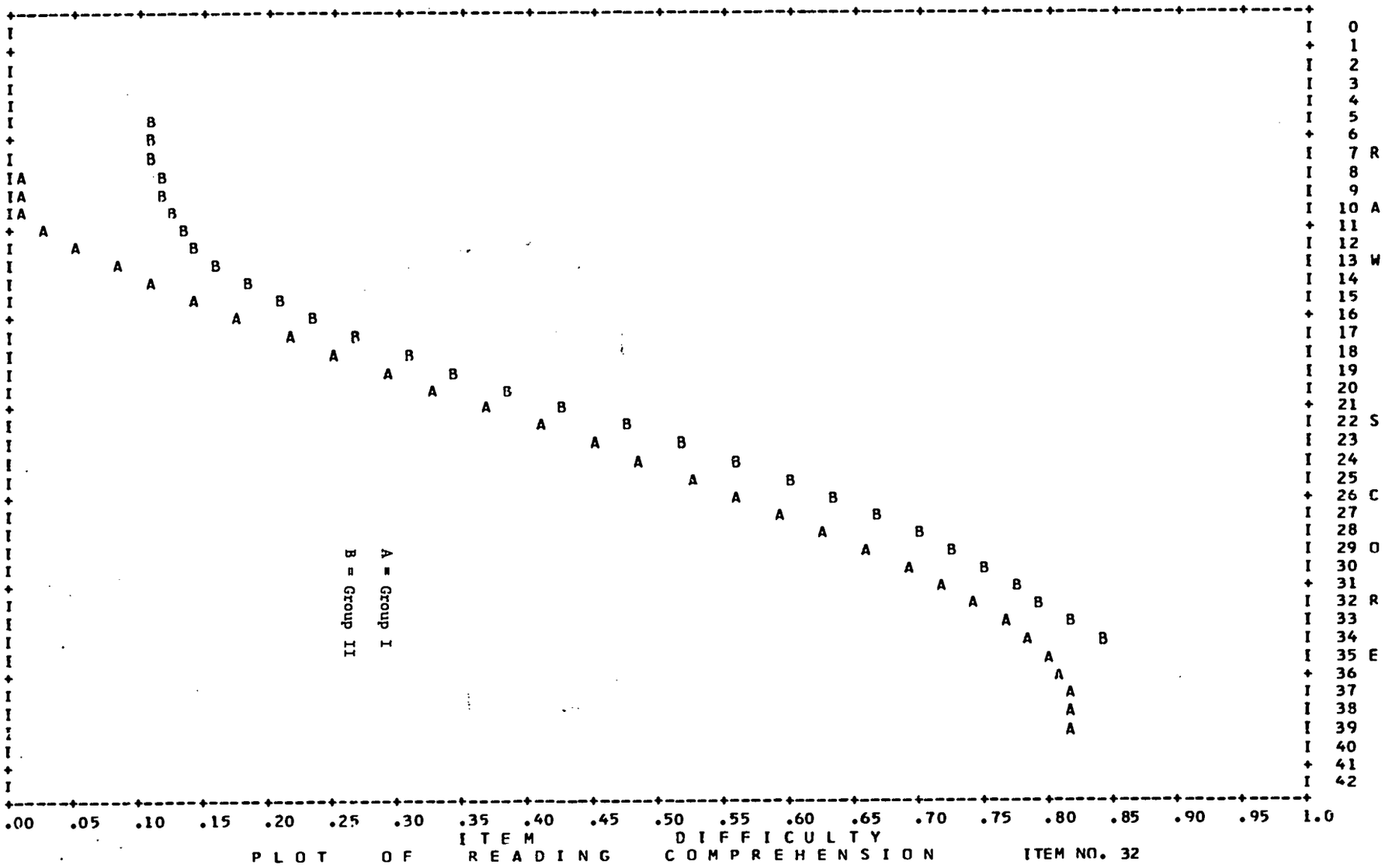


Figure 7. The Estimated Item Characteristic Curve for Item Number 32.

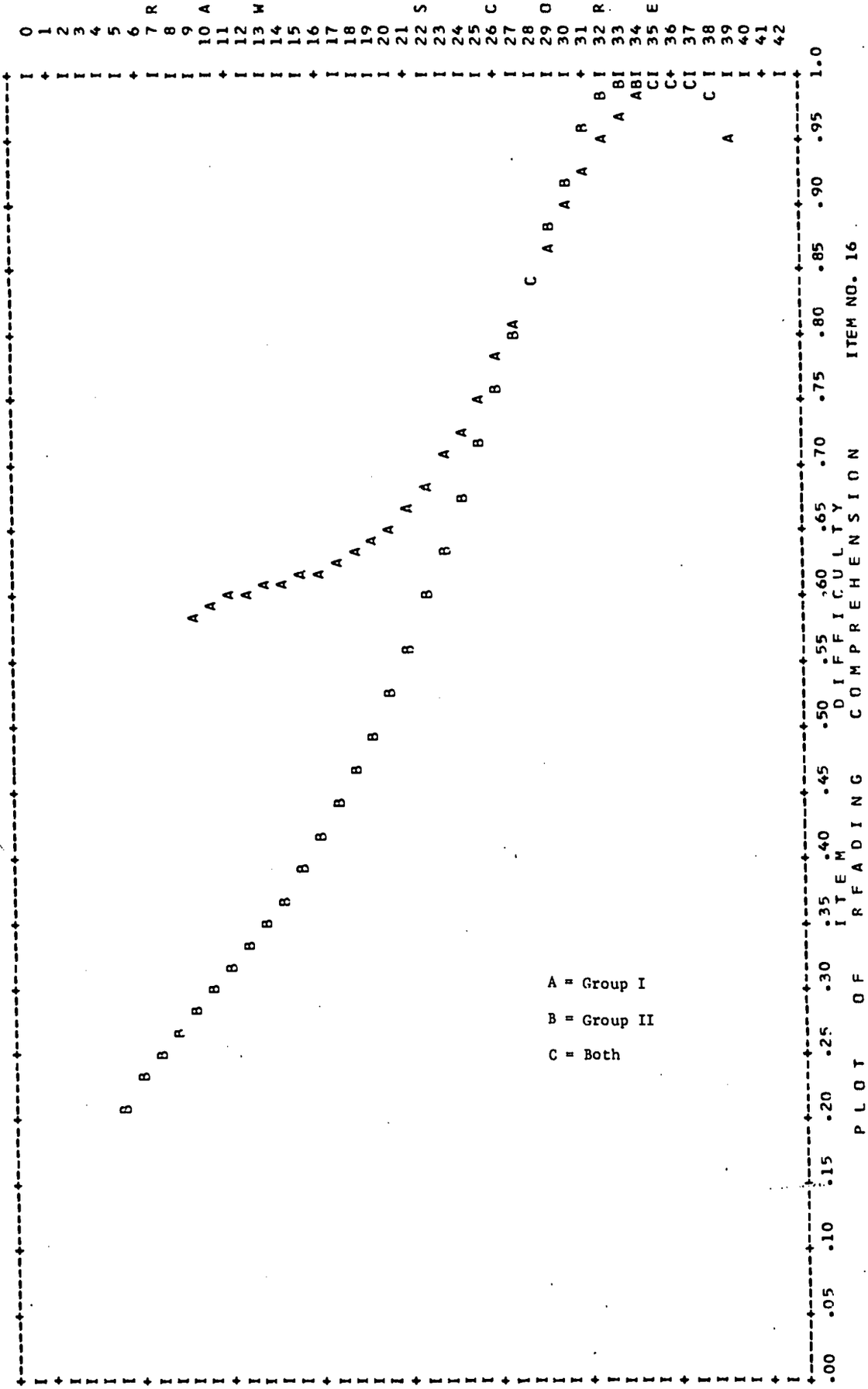


Figure 11. The Estimated Item Characteristic Curve for Item Number 16.

the two groups in responding to the test are not the same. Factor analytic techniques may be employed to estimate item-factor relational patterns for groups and these patterns may be compared.

A more complicated situation arises when the a priori domain of a test is not unitary or simple. In this case, the domain may be conceived of as a set of unitary or simple subdomains. Therefore, a set of item-subdomain patterns for each group being considered rather than a single pattern must be compared. As before, factor analytic methods may be employed to estimate such patterns.

Factor analysis attempts to estimate parameters on the right of the regression expression

$$\underline{y} = A\,\underline{f} + \underline{e},$$

where only the vector of subject response evaluations $\underline{y}$ is "known", and where A is a matrix of sets of known variable to factor variable relationship patterns, $\underline{f}$ is a vector of factor variables, and $\underline{e}$ is a vector of residual and error terms. If we allow the factor variables $\underline{f}$ to represent locations in the subdomains and $\underline{y}$ to be the item response evaluations, then A will be the set of relational patterns to be compared. But suppose that for the individuals in some otherwise identifiable subgroup, the pattern of known variables to factor variables is not precisely the same as for all other subgroups. Let us then formulate a general model which expresses a relationship between known variables and two types of factor variables: 1) those that are common to all subgroups, and 2) those that are unique to a given subgroup. The following regression expression represents just such a relationship:

$$\underline{y} = A_c\,\underline{f}_c + A_u\,\underline{f}_u + e$$

where $\underline{y}$ and $\underline{e}$ are defined as before and where the subscripts c and u indicate for the patterns and factor variables those that are common to all

12 would seem to produce no significant change in total score.

Figure 8 concerns an item which provides an interesting example of how the different methods we have discussed can lead by themselves to different conclusions. The figure suggests that perhaps there is a group x score interaction but not a strong one. The χ^2 test for interactions is significant when the quintiles are established for the groups separately, but the test is not significant when the groups are pooled to determine the quintiles for both groups. The difference in difficulty is relatively large, but the difference in adjusted item difficulties is near zero. The point biserial approach indicated that the item was fair (see item 5, Table 7), yet the inter-group factor approach indicated that a large proportion of item variance within Group II is due to group specific sources. Thus there are some seeming contradictions. Some clarification results when it is realized that an item indicated as unrelated to a test's common inter-group factors may be a good item for another test. Perhaps an inspection of Figures 13 and 14 will clarify the situation further. Note how differently the quintiles would be formed if groups were pooled and then look again at Figure 8 and note how the different ways of forming the five subgroups can cause the groups to either appear or not appear to be interacting with score. Note also that if both groups had the same score distribution as Group II that there would be no group difference in the overall difficulty on this item. Further note that given the distributions in Figures 13 and 14 the group difference in item difficulty occurs because the distribution for Group I is concentrated under the area of the curves where the Group I curve is higher but that the concentration of the Group II curve occurs under the area where the curves cross.

Finally, note that there is a distinction between a conception of fairness of this item for a group and a conception of its fairness for an

subgroups with c and those that are specific to a particular subgroup with u. Note that this general model does not preclude either of the products $A_c \underline{f}_c$ or $A_u \underline{f}_u$ from forming a vector of all zeros; that is, either the common or unique part may be nonexistent. If on the other hand both unique and common parts do exist, the model can provide a measure of the overall fairness of a subtest by determining the proportion of variance accounted for by the common part and it can provide a means of identifying items which may be in a sense unfair. For just as we may determine the proportion of subtest variance accounted for by the common part of the model we may determine for each item the proportion of variance accounted for by the subgroup specific part of the model. If the amount of item variance accounted for by subgroup specific sources is large, then that item is probably unfair.

This conception of fairness rests on the assumption that if there is a large part of the variance accounted for in either a test score or an item score which is not due to sources specific to a particular subgroup, but is due to sources common to all subgroups of interest, then that test or item is probably fair.

The model which attributes test variance to common factors for all groups, to specific factors for groups, and to item specific and residual sources is based on the idea of inter-battery factor analysis offered by Tucker (1958). In the inter-battery model, variance is partitioned into factors common to test batteries, factors specific to batteries, and test specific and residual variance. The inter-battery model requires each subject to take all test batteries. The inter-group factor model presented in this paper requires that all groups take the same test.

The estimation of model parameters in the inter-battery model rests on the assumption that only the factors common to batteries are involved in

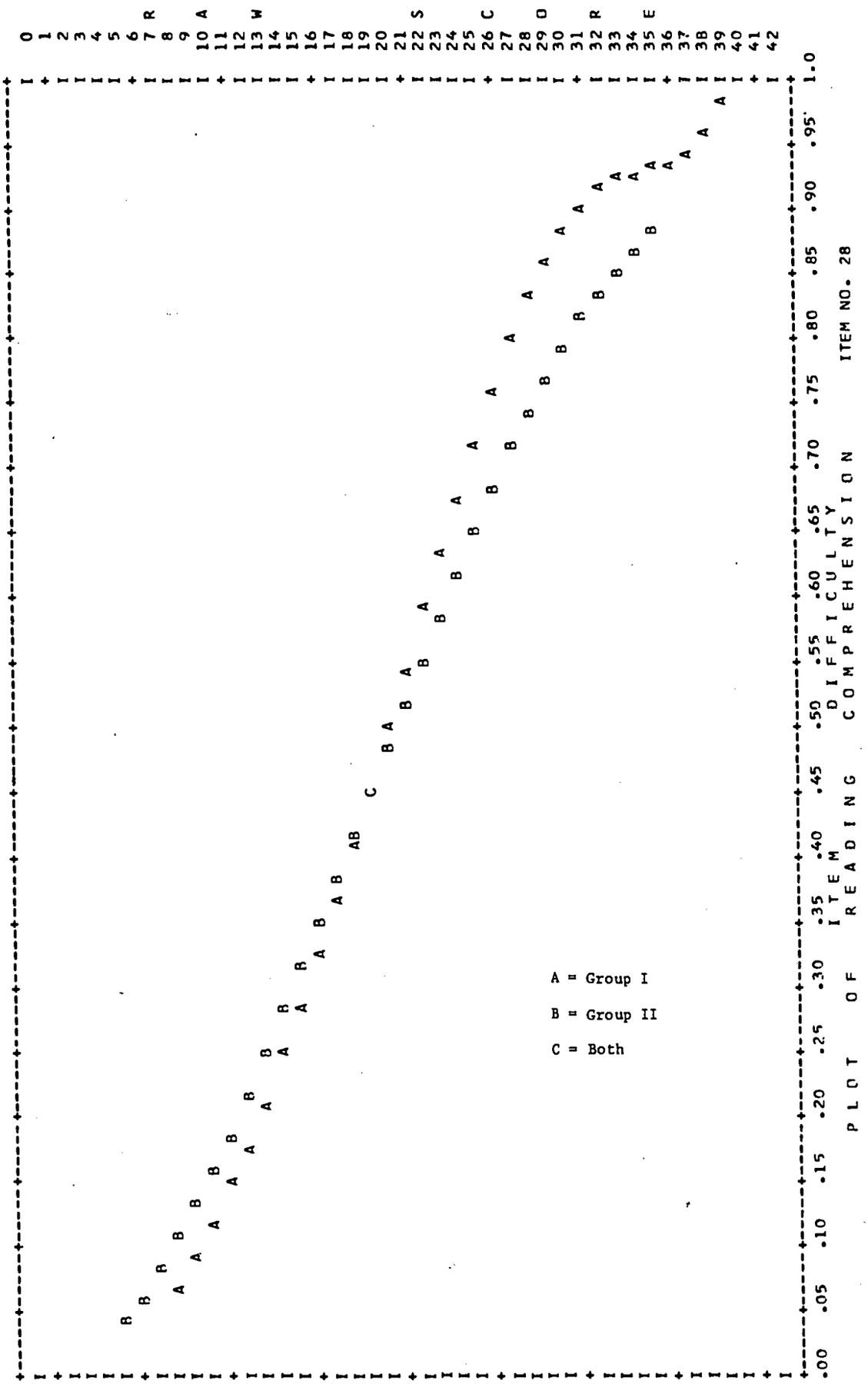


Figure 12. The Estimated Item Characteristic Curve for Item Number 28.

between battery cross products or convariances. Analagously, the intergroup model's parameter estimation is based on the assumption that only the factors common to groups are involved in between group cross products or convariances. In the handout, there is an outline of the procedure for estimating the parameters of an intergroup model which is for only two groups, followed by a demonstration of the extension of the procedure for the estimation of parameters for a model for m groups.

Method of determination of model parameter estimates. The model for a subgroup, as opposed to the model for an individual in a subgroup, may be written as

$$Y_i = A_c F'_{ci} + A_i F'_i + E_i$$

where Y_i is a matrix of item response evaluation with p rows for each of the p items and n_i columns for each of the n_i subjects in subgroup i, A_c is the common pattern matrix which is p by k the number of common factor variables, F'_{ci} is the matrix of common factor scores which is k by n_i, A_i is the matrix of subgroup specific patterns and is p by q_i the number of specific factor variables, F'_i is the matrix of specific factor scores, q_i by n_i and E is the matrix of error and residuals. The conception of A_i as subgroup specific allows the definition of A_i as orthogonal to A_j, $i \neq j$, so that $A'_i A_j = 0$ the zero matrix. Further note the definition $E'_i E_j = 0$. Thus for the pair of subgroup evaluation matrices Y_i and Y_j we may write the expression for the product $Y'_i Y_j$ as the identity

$$Y'_i Y_j = F_{ci} A'_c A_c F'_{cj} + F_{ci} A'_c A_j F'_j + F_i A_i A_c F'_{cj} + F_i A'_i A_j F'_j.$$

But $A'_c A_j$, $A'_i A_c$, and $A'_i A_j$ are all equal to zero matrices. Thus,

$$Y'_i Y_j = F_{ci} A'_c A_c F'_{cj}.$$

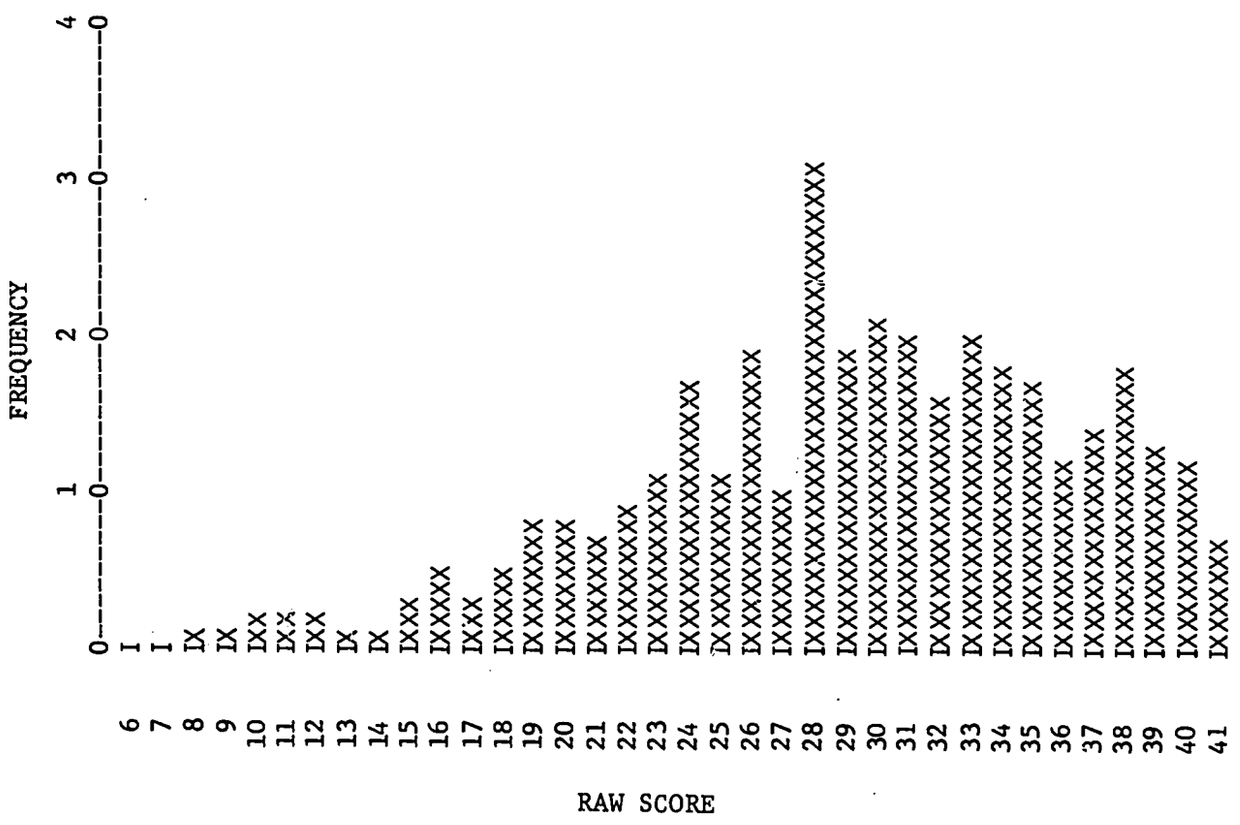


Figure 13. Frequency Distribution of Raw Scores for Group I on the Grade 5 Reading Comprehension Test.

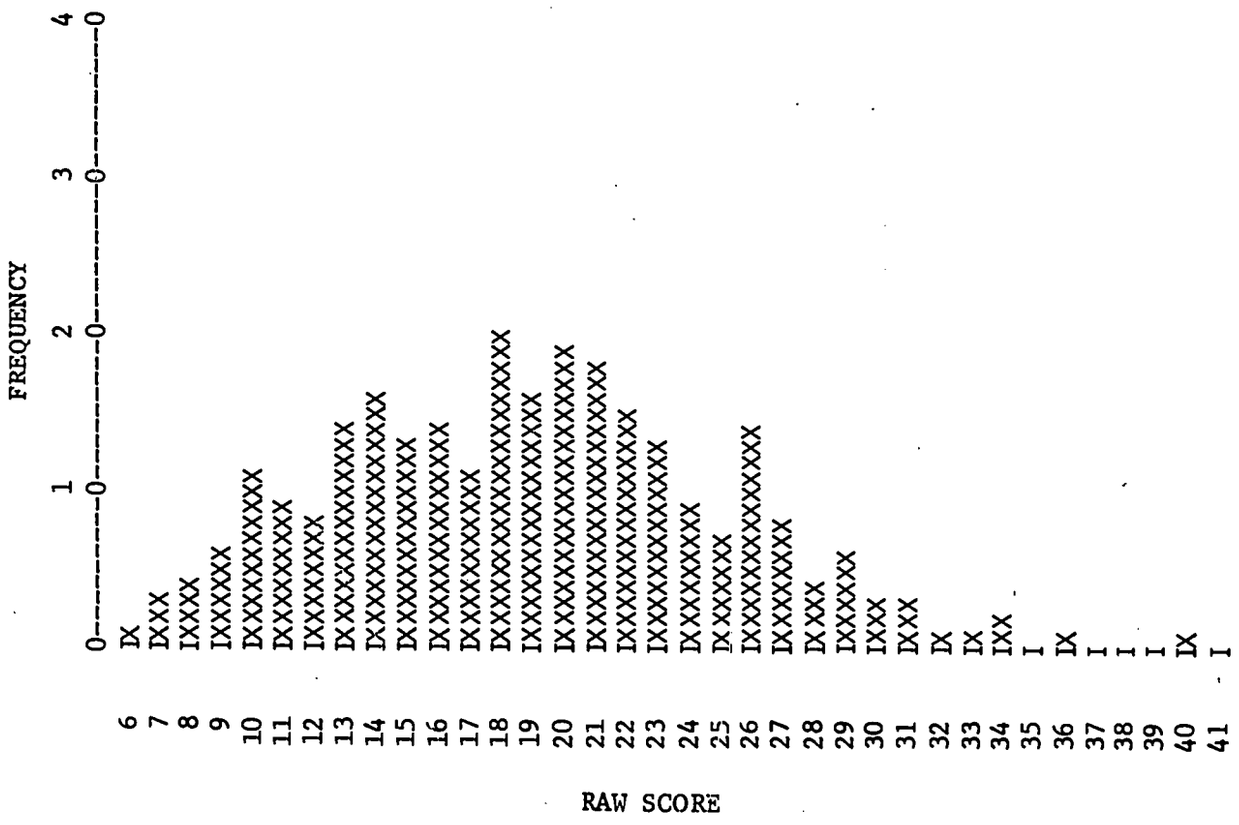


Figure 14. Frequency Distribution of Raw Scores for Group II on the Grade 5 Reading Comprehension Test.

individual in a group. For while the unconditioned probability of correct response to the item is greater for Group I, the probability conditioned on a score of 15 is not higher for either group. Thus if you were a fifth grader of either group who was likely to score a 15 on the test, the item may be more fair for you than for others.

ITEM DIFFICULTY AND BIAS

Do the preceding statements mean that item difficulty is the heart of test bias? Not by our definition although plainly the notion of an unusual difference in difficulty between groups is a useful indication that an item may be biased against one of the groups. Also it is clear that items that are too easy or too hard for only one of a pair of groups cannot measure the same thing in equal amounts in both groups. Thus some items and some tests are biased only because they are inappropriately difficult or easy for one or the other group. The effect is a set of scores too high or too low and can be classified as bias of Type A. Nevertheless the item will not prove to be biased for members of the group for whom it is not too hard or easy because it is measuring the appropriate trait. Thus if extreme difficulty is the only factor involved the bias is merely inappropriate use. Also note that it is entirely possible for bias of Type A to occur among items of just the right difficulty for a group.

Difficulty enters in some fashion into all approaches to assessing bias. Point biserials are necessarily low for really extreme difficulty values but one would ordinarily reject such items anyway. More to the point is that items that are really too difficult do not measure what they are meant to because too many people are responding to aspects of them not meant to be determining elements. Accordingly such items have low point biserials.

This sort of thing happens often to at least a few people on almost any item. The distributions by fifths demonstrate this well. Figure 15 shows some of the reading comprehension items and it can be seen that ceiling and floor effects for the top or bottom of one or the other groups is common and perhaps explains the high frequency of significant interactions. The two highest chi square values for reading comprehension are for items 20 and 26 which demonstrate just such effects. They are, nevertheless, good items by other criteria. The interaction approach appears to be unduly influenced by difficulty and will lead to faulty conclusions about bias when the groups compared differ appreciably in mean score. In such a case items are frequently too easy or too hard for some substantial portion of one or both groups. To have it otherwise is not easily accomplished and it is not always desirable to restrict the range that a test can cover.

Still, it is our experience that at least for some topics at some levels it is possible to reduce the differences in difficulty between groups similar to I and II without any apparent decrease in content validity. With the exception of the science tests, we think we have accomplished this substantially for the tests listed in Table 7 although proof of success awaits standardization.*

Finally we would like to draw attention to the contrast between the item characteristic curves of Figures 5-12 and the mean difficulties of the fifths displayed in Figure 15. This contrast emphasizes how the distribution of the scores of a particular group may conceal the characteristics of the item for that group over the full range of scores. In the item characteristic curves the role of relative difficulties is displayed at each score

* No attempt was made with the CAT tests because the tryout data were not available.

point. Under these circumstances the relationship between differential conditional difficulties becomes a criterion of bias. However, if one must rely on a statistic, the point biserials are more likely to suggest what the characteristic curves would be like than any other statistic, because it represents a linear approximation of such curves.

CONCLUSIONS

It may be apparent that we have some preferences among the approaches examined for determining item bias and test bias. We believe use of item characteristic curves of the sort described here and intergroup factor analysis will permit test builders to build both fairer and more generally effective tests. However, we find merit and value in all approaches since they each provide some relevant information and none of them are completely redundant. We will, however, continue to look for better ways to proceed since the efficiency of such an eclectic approach is rather obviously low.

In the meantime, we would like to make six points we believe our data support.

1) Bias against various groups in achievement tests occurs but it may well be small and unimportant in amount for most groups; we simply cannot say from our data and procedures, nor do we know of any other data that can answer that question adequately.

2) Item bias and test bias are not quite one and the same thing. Thus a demonstration that a test has some biased items does not necessarily prove that the test score overall is biased since some items may balance others.

3) Nevertheless finding biased items and fixing or eliminating them appears to be important at least until one can find ways to demonstrate that

the amount of bias is unimportant. Furthermore biased items are often bad items generally.

4) Nor is group bias identical with bias against all members of a group, for the first can exist in the absence of the second.

5) Most of the ambiguities in determining bias that we have noted stem from the lack of appropriate external criteria. Work of the sort being reported here by Caylor and by Williams should be emulated by all. External criteria ought to be found for all tests even if their relevance is indirect.

6) Thus we are asking for a reconsideration of the construction and validation procedures used with achievement tests. The internal characteristics of a test along with armchair decisions about content validity (however expert the judges may be) do not provide an adequate basis for judging validity. Content validity procedures probably obviate bias only for the item writers.

The issue of test bias tends to arouse the emotions of a number of people. Perhaps as a consequence of this concern and attention people will be willing to undertake and support efforts to carry out the full program of research any published test should have.

REFERENCES

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude. *College Entrance Examination Board Research and Development Reports*, 1971, RB 71-59.

California Achievement Tests, 1970 Edition, Monterey: CTB/McGraw-Hill, 1970.

Cleary, T. A., & Hilton, T. L. An investigation of item bias. *Educational and Psychological Measurement*, 1968, 28, 61-75.

Coleman, J. S. et al. *Equality of Educational Opportunity*. U.S. Dept. of Health, Education, & Welfare, 1966.

Green, D. R. *Racial and Ethnic Bias in Test Construction*. Monterey: CTB/McGraw-Hill, 1972.

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

Tucker, L. R. An inter-battery method of factor analysis. *Psychometrika*, 1958, 23, 111-136.

Williams, R. L. Black pride, academic relevance and individual achievement. *The Counseling Psychologist*, 1970, 2, 18-22.