DOCUMENT RESUME

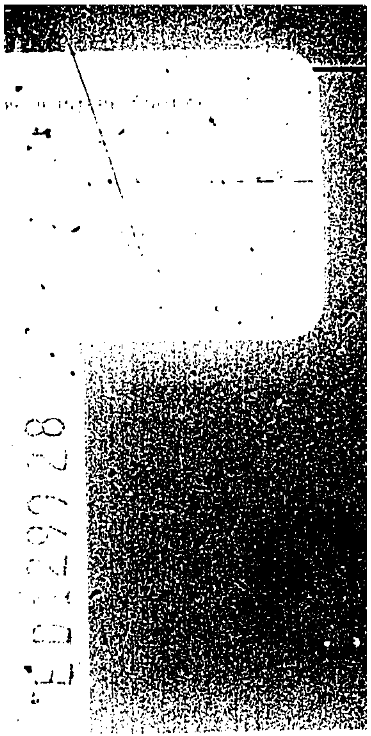
ED 129 928 TM 005 806

AUTHOR Green, Donald Ross
TITLE Racial and Ethnic Bias in Test Construction.
INSTITUTION CTB / McGraw-Hill, Monterey, Calif.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE [71]
CONTRACT OEC-9-70-0058(057)
NOTE 19p.; For a related document, see ED 056 090

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Achievement Tests; Caucasians; Data Analysis; Disadvantaged Youth; Elementary Secondary Education; Ethnic Groups; *Item Analysis; *Mexican Americans; Minority Group Children; *Negroes; *Racial Discrimination; Rural Youth; Standardized Tests; Suburban Youth; *Test Bias; *Test Construction; Test Reliability; Test Validity; Urban Youth
IDENTIFIERS *California Achievement Tests

ABSTRACT

To determine if tryout samples typically used for item selection contribute to test bias against minority groups, item analyses were made of the California Achievement Tests using seven sub-groups of the standardization sample: Northern White Suburban, Northern Black Urban, Southern White Suburban, Southern Black Rural, Southern White Rural, Southwestern Mexican Urban and Southwestern Anglo-American Suburban. The best half of the items in each test were selected for each group. Typically about 30 percent of the items in the upper half of the distribution of item-test correlations for a group on a test did not meet this criterion with another group. By this criterion minority groups were relatively similar as were the three suburban groups. The resulting unique item tests did not correlate well with each other. Scores of minority groups were relatively better on the selected items. Thus, standard item selection procedures produce tests best suited to groups like the majority of the tryout sample and are therefore biased against other groups to some degree. This degree varies. Ways to minimize this bias need to be developed. (Author/MS)



DESCRIPTION OF THE STUDY

Statement of the Problem

The standardized achievement tests used in schools are often said to be biased against, and thus inappropriate for, children belonging to disadvantaged racial and ethnic minorities. If this is so, then there are two possible sources of such bias. The first may originate in the preconceptions and thought patterns of the test item writers. The second may result from the customary item tryout and selection procedures used in test construction. This second possible source of bias is the general topic investigated in this study.

A number of problems occur when trying to consider bias in achievement tests because the criteria of bias are not completely clear. When most recent writers (Cardall & Coffman 1964; Potthoff 1966; Cleary & Hilton 1968; Messick & Anderson 1970; Green 1971) speak of bias, they say something about tests which measure different things when used with different groups. How might item tryout and selection procedures produce such a result?

The typical procedure in building standardized achievement and aptitude tests has remained essentially unchanged over many years (cf. Lord & Novick 1968, Chapter 15; Ruch 1929, Chapter 2). The first step is to develop a pool of items meeting various specifications as to form and content. Next these items are given to a sample of individuals — the step in question here. Various item statistics, such as point biserial correlations (item vs. total score), are calculated and the "best" items are then chosen; "best" is customarily characterized first and foremost by a high relationship of the item to the total score. Other characteristics such as difficulty and the effectiveness of distracters (in multiple-choice tests) are also considered. Most of these latter item characteristics are related to the item-test correlation to some degree. Therefore, the items which "discriminate" best (i.e., show the highest relationship to total score) are the ones usually chosen. This in turn means that the characteristics, or attributes, of the individuals in the tryout sample which are most responsible for differences in total score determine which items tend to be chosen and determine, in effect, what the test measures within the range of possibilities available in the item pool. That is, to deal with the items, the individuals tested call upon certain qualities, attitudes, knowledge, or skills found in widely varying degrees in their group. The items most sensitive to these attributes of the tryout sample then get selected.

Consequently, the possibility exists that the items selected are biased and discriminate against groups unlike the modal group in the tryout sample. If some atypical group has traits not prominent in the tryout sample and if these traits interact more strongly with the items than do the attributes the group shares with the majority, then the tests will measure the distinctive characteristics of this group rather than the trait or traits measured in the more typical groups. Another possibility is that the atypical group is uniformly low on the measured traits, but not on other equally relevant, but unmeasured, attributes. In either of these cases, one could say the resulting test is biased. In the first instance, it is biased because it measures different things for different groups unbeknownst to the users; in the second instance, it measures only a portion of the relevant behaviors but is taken to measure them all.

Given circumstances such as those just described, then the use of "average" item tryout samples will result in the selection of item sets unsuited to one or more of the various racial, ethnic, cultural minority groups in our schools. From this, it may follow that the use of a single tryout group can never solve the problem — perhaps only the construction of separate tests would do so although this solution would have obvious drawbacks. Another alternative might be to use the same test but different item weights for different groups.

The need to consider such unattractive possibilities depends on how strongly the nature of a tryout sample determines the outcome of item selection. It is customarily assumed that the choice of people for item tryouts does not have much effect on item selection, although "atypical" groups (such as disadvantaged children) are usually avoided. This amounts to the assumption that the test items function much the same way with all kinds of people. Some evidence for evaluating this assumption is presented in this report.

Related Literature

Prior work on test bias does not seem to have dealt directly with these item tryout and selection procedures. In fact, as far as achievement tests are concerned, very little work of any sort on the matter of bias appears to be available. The work on bias in intelligence and aptitude tests is more extensive, but aspects of the bias issue other than the one considered here have dominated discussions.

That children's intelligence test scores are related to their social and economic status was reported by Binet and others more than 60 years ago and has been studied and argued about ever since. For a long time, these debates largely stayed within the bounds of the much older and highly emotional nature-nurture controversy, perhaps because many felt that the then new tests could settle the issue (Terman 1916, pp. 19-20). Since the intensity of the arguments shows no sign of diminishing after 50 years (consider, for example, the response to Jensen 1969), that hope may be considered unreasonable. In any case, the test score differences favoring the more privileged elements of society remain a fact (Coleman et al. 1966). It may be added that the accusations of the misuse and the misinterpretation of scores (Hunter & Rogers 1967; Mercer 1971) are also factual in some, if not most, instances.

However, the issue here is the nature of the tests themselves. This has not been as widely studied as it might be. Apparently, the first serious attempt to examine test items for bias was led by Allison Davis and his colleagues 20 years ago (Eells et al. 1951). They examined several existing group intelligence tests and the items in them in an attempt to determine the factors built into the tests which are related to differences in performance between cultural groups. They concluded: "Variations in opportunity for familiar cultural words, objects, or processes required for answering the test items seem . . . the most adequate general explanation. . ." (Eells 1951, p. 68). This sort of objection is also often made to achievement tests (Wasserman 1969) but is not a valid basis for asserting bias in an achievement test, unless the missing knowledge is irrelevant to what is being measured. Consider the finding reported by Chang and Raths (1971) that achievement test items which discriminate between middle- and lower-class groups reflect a different curricular emphasis on the part of the teachers. This is more nearly teacher bias than test bias. In an ability test, such objections have direct logical merit.

Interestingly, the subjects in the Eells study were all white and drawn from the schools of "a western industrial city of about 100,000 people." One result of the study was the publication of the Davis-Eells Games (1953) which was designed to eliminate this kind of cultural bias. Three things may be noted about this test, which is now out of print. First, the test proved to yield differences between middle and low socioeconomic status (SES) groups (Angelino & Shedd 1955) as substantial as those found using other group intelligence tests. Second, Davis and Eells eliminated the items that showed SES differences in difficulty only if they could rationalize the difference as a consequence of opportunity. Lastly, they apparently did not look at the differences between SES groups with respect to item discrimination. The common interpretation of the outcome of the Davis-Eells test and similar efforts by others has been that the task of building a "culture-free" or "culture-fair" test may be not only impossible but inappropriate because such a test would not be valid as a measure of general ability, as indeed was the case for the Davis-Eells Games (Lorge 1966).

Supporting this view is work such as that of Lesser, Fifer, and Clark (1965). This study showed that patterns of ability are different for different ethnic groups. It also showed that within any one ethnic group, quantitative differences resulted from socioeconomic status, but the patterns for SES groups were very similar. That is, the lower-class and middle-class groups of any one ethnic group had similar patterns, but the latter had higher scores. Such data imply that any test measuring several abilities — as most ability and achievement tests do — is automatically stacking the cards against one ethnic group or another.

Furthermore, Williams (1970) reports that he has built a test biased in favor of blacks. His validation studies of the instrument as a measure of academic aptitude are not yet complete, but if Williams can produce a valid ability test favoring blacks, then it is probable that most ability tests are biased. In the meantime, many people are taking this to be established fact, and assertions that group intelligence tests necessarily discriminate against various minority and disadvantaged groups in our society have been increasing in number and vehemence. Some school systems (New York City, for example) have virtually abandoned the use of such tests (Gilbert 1966). Similarly, some college personnel now argue that the various placement and ability tests traditionally used are inappropriate (Brown & Russell 1964).

Many of the assertions made about bias in ability tests appear to be sound, but, as Anastasi (1968) has pointed out, bias in prediction involves a distinct set of issues. None of the preceding considerations necessarily apply if the test in question is meant to be used as a predictor of some criterion performance. For example, if one defines bias as systematic under-prediction, then the attacks on the aptitude tests used for college admissions appear largely unfounded. The claim that such tests fail to function among disadvantaged minority students in the way they do in other groups lacks supporting evidence. A series of studies at both the high school and college levels shows that academic aptitude tests frequently predict grades just as well for minority groups as they do for more privileged groups. Only the work of Green and Farquhar (1965) points to a different conclusion among a half dozen or so studies on this issue. In fact, some tests appear to over-predict the performance of lower-class and Negro students in contrast to middle-class and white students (Hewer 1965; Stanley & Porter 1967; Cleary 1968; Davis & Temp 1971).

Even in this relatively well explored area, much remains to be done, such as finding ways to deal with the possibility of bias in the criterion measure (Linn & Werts 1971). In addition, there is often more than one reasonable definition of bias in criterion-related validity situations (Thorndike 1971; Darlington 1971). As Potthoff (1966) has pointed out, the operational demonstration of bias is even more difficult and ambiguous when test validity cannot be defined as the relationship of scores to a directly measurable criterion. Any test yielding scores meant to be an indication of status — be it in achievement, in intelligence, or in what have you — creates such problems.

One approach consistent with the definition of bias offered at the start of this paper is to examine the items, rather than the whole test, for bias. Here, bias may be defined as an item by group interaction. Three studies (Cardall & Coffman 1964; Cleary & Hilton 1968; Angoff & Ford 1971) using this approach have been reported.

They each found statistically significant item by race interactions in the College Entrance Examination Board aptitude tests which they used (SAT and PSAT). Nevertheless, Cleary and Hilton concluded that "the PSAT is not biased for practical purposes," while Angoff and Ford suggested the "interaction was simply the difference in performance levels on the test shown by the two races." These studies were based largely upon a consideration of item difficulties.

Item interrelationships are also a relevant consideration. Data obtained by Kennedy et al. (1963) show that the grandfather of them all, the Stanford-Binet (Terman & Merrill 1960), produced equal or higher item-test correlations for an all black southern sample than was reported in either the 1937 or the 1960 standardization. Also, Merz (1970) has reported that the factor structure of the Goodenough-Harris-Drawing Test is substantially the same for samples of black, white, Mexican, and Anglo children in the southwest.

Incomplete as this research on bias in ability tests may be, it is way ahead of that on bias in achievement tests which is essentially nonexistent. The claims of bias in achievement tests (Wasserman 1969; Williams 1970; Houston 1971) need investigation. The approach of item by group interactions seems to be the logical place to begin. Certainly it seems reasonable to believe that a test based on items selected for a particular group (such as inner city black children) would be less biased against them and therefore more useful for them.

Objectives of the Study

To explore such a possibility, 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.

The study attempts to determine whether or not an item tryout using these different groups would lead to the selection of different items from the item pool and, if so:

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

(2) Are the resulting item sets "better" for the minority groups in the sense that they are more reliable and have better functioning items. (higher point biserial correlations)?

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

Limitations of the Study

The major limitation of this study is the restricted nature of the item pool: all items come from an already published test. They are therefore preselected and may be limited in their possibility of eliciting differential reactions from the sample groups. Also, it should be noted that grade and test level are not independent; the test levels were designed to be continuous and to articulate well, but they are different tests. Thus, the assumption made throughout the following material that grade differences are meaningful may not be justified. Finally, because of limitations of time and money not all relevant analyses of the data could be made.

METHOD

The basic data for this study 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. It was designed to measure educational attainment and to provide an analysis of learning difficulties. It is basically similar to the 1957 edition and generally measures:

(1) the ability to understand the meaning of the content material presented,

(2) the performance of the student in applying rules, facts, concepts, conventions, and principles to solve problems in the basic curricular material, and

(3) the level of performance of the student in using the tools of reading, mathematics, and language in progressively more complicated situations.

The tests in the battery which were investigated in this study are Reading Vocabulary, Reading Comprehension, Total Reading, Mathematics Computation, Mathematics Concepts and Problems, Total Mathematics, Language Mechanics, Language Usage and Structure, and Total Language. Total Reading, Total Mathematics, and Total Language were treated as tests separate from their parts. The standardization took place early in 1970 and involved over 200,000 students in about 400 schools. The sampling design called for obtaining a sample of school districts stratified by region (seven areas), school district size (three categories by average enrollment per grade), community type (urban, town, rural, other), and control (public or parochial). Within the districts, schools were chosen randomly for each test level, and all students in the selected schools who were in appropriate grades took the test.

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.

Sample

All schools participating in the CAT-70 standardization answered questionnaires which provided information on the basic character of the area served (e.g., residential suburb, inner part of a large city, etc.,), the percentage of white students, the percentage of children from homes where another language is spoken, and the percentage of children in families falling in each of four SES groups defined by parental occupation (professional-managerial, white collar, skilled, unskilled).

From the data on these questionnaires, seven groups of schools were drawn for this study. The characteristics and sizes of these groups are shown in Table 1. 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. Only at Grade 10 was it not possible always to find schools meeting these criteria in the standardization population; sufficiently segregated tenth grades were found only in the South.

The groups were paired for comparisons as follows:

(1) Northern, black, central city versus northern, white, suburban (II vs. I)

(2) Southern, black, rural versus southern, white, suburban (IV vs. III)

(3) Southern, black, rural versus southern, white, rural (IV vs. V)

(4) Southwestern, Mexican-American versus southwestern, Anglo-American, suburban (VI vs. VII)

Table 1

CHARACTERISTICS OF THE SAMPLE GROUPS

Group Number	Geographic[a] Region	Residential Type	Ethnic Group	Socioeconomic Status	Number of Cases by Grade				
					1	3	5	8	10
I	North	Residential Suburban	White (97%)[b]	High (81%)[b]	299	225	265	328	—
II	North	Central City	Black (99%)	Low (81%)	285	304	278	250	—
III	South	Residential Suburban	White (99%)	High (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-[c] American (87%)	Low (82%)	146	144	169	399	—
VII	Southwest	City and Suburban	Anglo-American (99%)	High (81%)	189	218	249	277	—

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

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

[c]81% speak mostly Spanish at home.

Enough schools meeting the appropriate criteria to provide between 150 and 300 students for each group at each of five grade levels were selected. Each of the grade levels (1, 3, 5, 8 and 10) corresponds to a different level of the CAT-70 battery.

Grade 10 comparisons were made in the South only. No analyses were made of the Total Language scores in Grades 1, 5, and 8 for the northern, white group and in Grades 1 and 8 for the northern, black group. Therefore, of the 315 possible analyses (7 groups x 9 tests x 5 grades), only 274 separate analyses were made.

Data Analyses

The basic procedure used for examining the data was an item selection routine. 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 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 per cent of items chosen for one group in the pair but not for the other was recorded. These items were labeled "biased." The number of these biased items in any one comparison indicates the degree to which the two groups interact in a distinct manner with the test items. All 21 possible pairs of groups were compared in this way; the remaining analyses were made only for the four pairs listed previously.

(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. From these correlations, estimates of the variance not common to the two biased item tests were made to judge how different the sets of items really are in what they measure. Thus, this analysis supplements the first.

(3) Another analysis consisted of examining and comparing full-test and half-test KR 20 reliability estimates since differential reliability would be a form of bias indicating that the test scores have a larger error component in one group than they do in another group.

(4) Finally, mean scores on the full-test, the half-test, and the biased item tests were examined for changes in relative status of the groups as a result of item selection.

RESULTS

Proportions of Biased Items

The item selection routine yielded a series of tests "best" for each group, half as long as the original test — when N was odd, the expression (N + 1)/2 was used to determine the length of the half-test. The next step was to identify those items selected for only one of the two members of a pair — the so-called biased items. Obviously, the number of biased items has to be the same for each group in a pair. This number as a proportion of the items in each half-test is an index of the degree to which the item selection procedure produces a different test for the two groups.

Table 2 exhibits these proportions for the four basic comparison groups. The proportions do not appear to vary systematically by grade or test. However, certain groups appear considerably more like each other than are others by the criterion of the relative size of these proportions. It can be readily seen from Table 2 that the differences between the Mexican-American and Anglo groups tend to be larger than those between the black and white pairs.

Table 2

PROPORTIONS OF BIASED ITEMS FOR COMPARISON GROUPS BY GRADE AND TEST

Test	Number of Items Selected	Comparison Groups II vs. I	IV vs. III	IV vs. V	VI vs. VII
Grade 1					
Vocabulary	46	.41	.33	.35	.59
Comprehension	12	.25	.58	.33	.42
Total Reading	58	.40	.36	.34	.69
Computation	20	.15	.25	.40	.25
Concepts & Problems	24	.42	.38	.42	.58
Total Mathematics	44	.16	.25	.23	.41
Mechanics	19	.42	.21	.21	.58
Usage & Structure	10	.30	.30	.40	.40
Total Language	39	—	.24	.27	.54
Grade 3					
Vocabulary	26	.30	.65	.35	.45
Comprehension	23	.22	.26	.22	.35
Total Reading	43	.28	.42	.28	.33
Computation	36	.17	.28	.22	.25
Concepts & Problems	23	.35	.48	.35	.43
Total Mathematics	59	.29	.32	.30	.32
Mechanics	33	.48	.42	.30	.45
Usage & Structure	13	.31	.46	.23	.46
Total Language	46	.41	.30	.28	.48
Grade 5					
Vocabulary	20	.50	.55	.35	.70
Comprehension	21	.48	.43	.29	.52
Total Reading	41	.46	.46	.37	.61
Computation	34	.41	.38	.21	.41
Concepts & Problems	20	.50	.40	.20	.55
Total Mathematics	54	.44	.46	.20	.46
Mechanics	40	.45	.35	.25	.53
Usage & Structure	21	.33	.48	.38	.33
Total Language	61	—	.30	.16	.26

Table 2 (Continued)

PROPORTIONS OF BIASED ITEMS FOR COMPARISON GROUPS BY GRADE AND TEST

Test	Number of Items Selected	Comparison Groups			
		II vs. I	IV vs. III	IV vs. V	VI vs. VII
Grade 8					
Vocabulary	20	.40	.15	.15	.45
Comprehension	23	.22	.39	.30	.39
Total Reading	43	.26	.23	.21	.44
Computation	24	.25	.46	.29	.29
Concepts & Problems	25	.36	.40	.36	.28
Total Mathematics	49	.29	.49	.35	.29
Mechanics	36	.42	.33	.42	.39
Usage & Structure	25	.36	.56	.32	.16
Total Language	61	—	.15	.15	.18
Grade 10					
Vocabulary	20	—	.55	.40	—
Comprehension	23	—	.22	.22	—
Total Reading	43	—	.42	.30	—
Computation	24	—	.33	.33	—
Concepts & Problems	25	—	.40	.32	—
Total Mathematics	49	—	.33	.24	—
Mechanics	40	—	.38	.35	—
Usage & Structure	27	—	.41	.30	—
Total Language	67	—	.21	.19	—
Median proportions for all tests and all grades	—	.36	.38	.30	.43

The medians of these proportions for all possible pairs are shown in Table 3. The overall median proportion is approximately .30. As expected, the white, middle-class groups are consistently more like each other (these pairs have lower medians) than they are like the minority groups. The latter also have more in common than they share with the three majority groups. The southern, rural, white group does not fully fit into this otherwise clear pattern; in general, they appear more like the three minority groups than they resemble the three suburban groups. Of course, economically they are undoubtedly more disadvantaged than the suburban groups, albeit much less so than the southern, black group.

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	—

Independence of Biased Item Tests

All groups differ 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 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 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. But even allowing for this, it appears that in many instances the majority and minority tests measure quite different things and as a rule do so for both groups involved.

Changes in Test Characteristics

A special 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 are .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 mostly higher for the majority group (97 of 162 comparisons). The data concerning half-test reliabilities also show a very small amount of bias.

The item-test correlations after item selection show only slight improvements and the uniformity of the increases prevents one from inferring the presence of substantial bias.

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 full-test. Each group's full- to half-test improvement on each of the nine tests in the battery was compared to the improvement shown by its comparison group. Table 4 reports the number of tests on which a group showed more full- to half-test improvement than was 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.

Table 4

NUMBER OF TESTS ON WHICH EACH GROUP SHOWED MORE FULL- TO HALF-TEST MEAN SCORE GAIN THAN ITS COMPARISON GROUP[a]

	Comparison Groups											
Grade	II & I		IV & III		IV & V		VI & VII		Totals		χ^2	p
									Min.	Maj.		
1	7	1[b]	1	8	0	9	7	2	15	20	0.7	NS
3	2	7	8	1	5	4	4	5	19	17	0.1	NS
5	7	1[b]	8	1	1	8	8	1	24	11	4.8	.05
8	8	0[b]	6	3	7	2	6	3	27	8	10.3	.01
10	—	—	6	3	7	2	—	—	13	5	3.6	.10
Totals	24	9	29	16	20	25	25	11	98	61	8.6	.01
χ^2	6.8		3.8		0.6		5.4		8.6			
p	.01		.05		NS		.02		.01			

[a] Let. $\overline{Y}$ = majority group mean, $\overline{X}$ = minority group mean, and let f and h represent full-test and half-test, respectively. Then $\overline{Y}_f - \overline{X}_f - 2(\overline{Y}_h - \overline{X}_h) > 0$ favors minority; $\overline{Y}_f - \overline{X}_f - 2(\overline{Y}_h - \overline{X}_h) < 0$ 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.

The majority biased item tests are almost uniformly more difficult for both groups than are the minority biased item tests. In addition, 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 5 shows the frequencies of this phenomenon. The biased tests are clearly biased in favor of the group used as the basis for selection and this result tends to hold for all groups at all grades. The disadvantaged group is less disadvantaged when tested with items selected as uniquely best for them. In other words, the data show that the relative advantage of majority groups is reduced when using items chosen as best for the minority group but is increased when using items chosen as best for themselves.

Table 5

NUMBER OF COMPARISONS IN WHICH MEAN DIFFERENCE ON BIASED ITEM TESTS FAVORS EACH GROUP[a]

	Comparison Groups											
Grade	II & I		IV & III		IV & V		VI & VII		Totals		χ^2	p
									Min.	Maj.		
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 $\overline{Y}_{n}$ = majority mean on majority test, $\overline{X}_m$ = minority mean on majority test, $\overline{Y}_n$ = majority mean on minority test, and $\overline{X}_n$ = minority mean on minority test. Then, $\overline{Y}_m - \overline{X}_m > (\overline{Y}_n - \overline{X}_n)$ favors minority; $\overline{Y}_m - \overline{X}_m < (\overline{Y}_n - \overline{X}_n)$ 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.

CONCLUSIONS

The four analyses of the data described previously permit the following conclusions:

(1) Different tryout samples lead to the selection of somewhat different sets of items. Considering the restriction on range and variety of points of view represented in the item pool, the 30% proportion of biased items, which was the average found in this study, seems large. That is, it seems likely that a majority of biased items would have been selected if the item pool had been more heterogeneous.

(2) The more economically dissimilar the groups contrasted, the less likely it is that they will produce data leading to the selection of the same set of items.

(3) If a biased test is a test that contains a substantial proportion of items that would not have been selected had they been tried on some other particular group, then probably most tests are biased against most groups.

(4) By this criterion of bias, the tests used here are more biased against minority groups than against middle-class, white children. This is probably true for most published batteries of standardized tests.

(5) The proportion of biased items is a fairly good, but uneven, criterion of bias since in most cases the biased item tests do measure different things. What is measured depends on which group is used for selection and which group is being tested. This conclusion is not uniformly true and varies widely according to test, grade, and tryout group.

(6) The psychometric quality of the half-tests was only very slightly better than that of the originals. That is, the effect of the item selection procedure was small, presumably because all the items were already a product of an item selection procedure and because the battery is rather homogeneous in style and point of view.

(7) The half-tests were barely more reliable for the minority groups than for the majority groups, but this improvement is small in both kinds of groups and suggests minimal bias of this sort in the battery.

(8) The use of items particularly suited to a tryout group will improve the chances of good scores among individuals from similar groups. This outcome may be more likely in the upper grades.

(9) The amount of relative improvement in score that a minority group could expect to gain by using tests built with tryout groups like itself does not appear to be very large. This relative improvement is most unlikely to overcome any large discrepancy between typical scores in that group and those in more favored groups.

(10) It should be possible to build tests somewhat biased in favor of any group by using a fair sample of that group for item selection data.

RECOMMENDATIONS AND QUESTIONS

The conclusions strongly suggest that there should be some changes and additions to the test construction procedures commonly used whenever there is a possibility that the resulting instrument will be used with people belonging to a group ethnically or culturally different from the test builder's principal reference group. Clearly, the first additional step is to obtain data on all relevant groups separately. It is important to note that if a set of items is likely to measure different attributes in different groups, the majority group in a tryout sample will determine which attributes are most strongly measured and the odds are that the inclusion of one or more minorities will merely obscure the issue. Just as the degree of minority representation in standardization samples can have only a small influence on norms, minority group presence in tryout samples dominated by some solid majority will not accomplish much.

What is needed is a way either to (1) select unbiased items, (2) compensate for known bias by establishing alternate weighting and scoring schemes, (3) interpret scores according to the group membership of the examinee, or at least (4) acknowledge and document the existence of the bias and its effect on scores. Until more experience is available in using various kinds of separate tryout groups, it is not reasonable to state a preference among these options; a number of questions need to be answered first, such as:

(1) What proportion of items tried can one expect to find "unbiased" by each of various criteria?

(2) Can one expect simple scoring and weighting schemes to reduce bias?

(3) Are the same criterion measures appropriate for all groups?

(4) What sort of indices of bias could one offer that would be readily interpretable?

If the only favorable procedure turns out to be the last option, a test constructor could choose to build alternate versions, each biased toward a different group; the problems created by adopting this procedure are large and many but not necessarily insoluble.

In addition to exploration of the effects of variations in tryout groups, studies are needed on the role of points of view, cognitive style, and/or ethnic background among those contributing to the item pool. Would blacks tend to create items more useful for black children? Many blacks believe so. It seems obvious that Spanish-speaking item writers can produce better items for Spanish-speaking children than could someone who could not write in that language. Yet, we still often use English language tests with children whose native language is not English and claim to be measuring something other than facility with English. It is, of course, less obvious if the children are fully bilingual. Are black children bilingual?

The answers to these and many other questions one might raise are not obvious. What is obvious is that it is no longer adequate for those who build tests to argue that bias is largely a matter of misuse or to say that they cannot see why a particular test would be biased and thus ignore the matter. All tests are not necessarily biased, but any test may be. Until there are good answers to these questions, research on the matter should be a standard part of producing a test.

REFERENCES

Anastasi, A. *Psychological Testing* (3rd ed.) New York: Macmillan, 1968.

Angelino, H., & Shedd, C. L. An initial report of a validation study of the Davis-Eells Test of General Intelligence or Problem-Solving Ability. *Journal of Psychology*, 1955, 40, 35-38.

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.

Brown, W. M., & Russell, R. D. Limitations of admissions testing for the disadvantaged (letter). *The Personnel and Guidance Journal*, 1964, 43, 301-304.

Cardall, C., & Coffman, W. E. A method for comparing the performance of different groups on the items in a test. *College Entrance Examination Board Research and Development Reports*, 1964, RB 9.

Chang, S. S., & Raths, J. The schools' contribution to the cumulating deficit. *The Journal of Educational Research*, 1971, 64, 272-276.

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

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.

Darlington, R. B. Another look at "culture fairness." *Journal of Educational Measurement*, 1971, 8, 71-82.

Davis, J. A., & Temp, G. Is the SAT biased against black students? *College Board Review*, 1971, 2-9.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., Tyler, R. W. *Intelligence and cultural differences*. Chicago: University of Chicago Press, 1951.

Gilbert, H. B. On the IQ ban. *Teachers College Record*, 1966, 67, 282-285.

Green, D. R. Biased tests. 1971 (unpublished manuscript).

Green, R. L., & Farquhar, W. W. Negro academic motivation and scholastic achievement. *Journal of Educational Psychology*, 1965, 56, 241-243.

Hewer, V. H. Are tests fair to college students from homes with low socio-economic status? *Personnel and Guidance Journal*, 1965, 43, 764-769.

Houston, S. Cultural disadvantages: creativity, cooperation. *Behavior Today*, 1971, 2, (24), 3.

Hunter, L. B., & Rogers, F. A. Testing: politics and pretense. *The Urban Review*, 1967, 2 (3), 5-6, 8, 25-26.

Jensen, A. R. How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, 1969, 39, 1-123.

Kennedy, W. A., Van De Reit, V., & White, J. C. A normative sample of intelligence and achievement of Negro elementary school children in the southeastern United States. *Monographs of the Society for Research in Child Development*, 1963, 28, No. 6.

Lesser, G. S., Fifer, G., & Clark, D. H. Mental abilities of children from different social-class and cultural groups. *Monographs of the Society for Research in Child Development*, 1965, 30 (4, Whole No. 102).

Linn, R. L., & Werts, C. E. Considerations for studies of test bias. *Journal of Educational Measurement*, 1971, 8, 1-4.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

Lorge, I. Difference or bias in tests of intelligence. In A. Anastasi (Ed.) *Testing problems in perspective*. Washington, D.C.: American Council on Education, 1966, 456-471.

Mercer, J. R. Pluralistic diagnosis in the evaluation of black and chicano children. Paper presented at the American Psychological Association, Washington, D.C., September 1971.

Merz, W. R. A factor analysis of the Goodenough-Harris Drawing Test across four ethnic groups. *Dissertation Abstracts International*, 1970, 31, 1627 A.

Messick, S., & Anderson, S. Educational testing, individual development, and social responsibility. *The Counseling Psychologist*, 1970, 2, 80-88.

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.

Ruch, G. M. *The objective or new-type examination.* Chicago: Scott, Foresman and Co., 1929.

Stanley, J. C., & Porter, A. C. Correlation of Scholastic Aptitude Test scores with college grades for Negroes versus whites. *Journal of Educational Measurement*, 1967, 4, 199-218.

Terman, L. M. *The measurement of intelligence.* Boston: Houghton Mifflin, 1916.

Terman, L. M., & Merrill, M. A. *Stanford-Binet Intelligence Scale: Manual for the Third Revision, Form L-M.* Boston: Houghton Mifflin, 1960.

Thorndike, R. L. Concepts of culture-fairness. *Journal of Educational Measurement*, 1971, 8, 63-70.

Wasserman, M. Planting pansies on the roof. *The Urban Review*, 1969, 3 (3), 30-35.

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

ACKNOWLEDGMENTS

Special thanks are due to my colleague, Dr. George Burket, who first suggested this use of item selection procedures and who offered valuable advice throughout. I would also like to thank Dr. Joseph L. Dionne who gave encouragement and who ensured that CTB/McGraw-Hill provided the facilities and personnel needed at a cost substantially in excess of that provided by the grant. Finally, I would like to thank Lois Maclin who kept track of the entire project, made a number of statistical analyses, constructed all the tables, and edited the entire report.