DOCUMENT RESUME

ED 126 126 TM 005 382

AUTHOR Green, Donald Ross
TITLE Reducing Bias in Achievement Tests.
PUB DATE [Apr 76]
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Achievement Tests; *Item Analysis; Mathematical Models; Minority Groups; Racial Discrimination; Standardized Tests; Statistical Analysis; *Test Bias; *Test Construction; Testing Problems; *Test Validity

ABSTRACT

During the past few years the problem of bias in testing has become an increasingly important issue. In most research, bias refers to the fair use of tests and has thus been defined in terms of an outside criterion measure of the performance being predicted by the test. Recently however, there has been growing interest in assessing bias when such criteria are not available. In test construction in particular, where criterion measures are usually not collected until after the test is completed, assessment of bias, in the absence of criteria has become a vital issue. If unbiased tests are to be built, it is important to identify potentially biased items during the construction process when test content is still flexible and items may still be modified or eliminated. Presented here are the author's research efforts over the past six years on bias in the construction of achievement tests. A general overview of the problem and some of the difficulties involved in studying it are also presented. (Author/DEP)

Reducing Bias in Achievement Tests

by

Donald Ross Green

CTB/McGraw-Hill

Paper presented at

Annual meeting of American Educational Research Association

San Francisco, CA

April 20, 1976

In this paper I shall review the somewhat sporadic efforts I have made in the last six years to deal with bias in the construction of achievement tests. However, before proceeding I would like to make three points.

First, the topic of test bias usually arouses emotions. To most people it is obvious that bias is "bad" and it is widely assumed that biased tests are built by bad people trying to do bad things to others. These assumptions are sometimes accompanied by the view that all tests are bad especially if they are published. Those who feel this way should be warned that in this paper I will make exactly the opposite set of assumptions. I will assume that tests, especially published tests, are good and that they are built by good people trying to do good things. Now I won't go so far as to say that biased tests are good, but they are not necessarily bad.

I won't assume that biased tests are necessarily bad because test bias and unfairness are not the same thing. That is, unfairness is a function of how a test is used while bias is a characteristic of the test. We may say a test is biased when it systematically measures different things for one group than it measures for another. Note that if the biased test is used as though it is measuring the same thing, i.e., if it is used as though it was not biased, then it becomes unfair. A biased test is likely to be used in an unfair way but it doesn't have to happen. To prevent it one would have to know that the test was biased and either not use it with the wrong group, or if one could tell by how much, allow for or correct for the bias. If it is not known that a test is biased then it will presumably be used as though it is measuring the same thing for all groups and will probably lead to unfairness for one or more of them. Thus the problem is to detect bias in a test. In the absence of an outside criterion the task is not to judge fairness but to ascertain if the test is measuring the same things for the various groups under consideration.

No generally accepted set of procedures for doing this exists. One may of course have experts from various groups examine the test and try to judge the matter. I believe most test publishers do this. As I see it this is a useful procedure but hardly definitive. More empirical procedures are needed. Therefore, for lack of better, most of us here have turned to trying to detect bias in items rather than in test scores. It may seem at first blush that the two are the same thing, but unfortunately they are not. The hope is that we can build unbiased tests by eliminating biased items. However there are many problems and we have much to learn as yet.

Thus secondly I would note that bias in achievement test construction is an important topic sadly neglected in research. It appears that until just a few years ago most members of the measurement community did not consider it an important issue. To date aptitude tests and selection have received the lions share of attention, but even there construction has been a neglected aspect. Yet achievement testing is, I believe, the larger enterprise by far and, as the potential determiner of the fate of our various efforts to improve schooling, is more important to boot. The relatively small effort on the topic is illustrated by the fact - I believe it is accurate - that the work of the members of this panel encompass a substantial portion, if not the bulk, of the effort in this area. I am of course including the work of Coffman whose studies established one of the major lines of attack on the problem some years ago (Coffman, 1961; Cardall & Coffman, 1964) and that of Angoff who has continued and extended this work substantially in a series of studies. I should point out right now that Angoff wrote a fairly comprehensive review of the work on the topic of this symposium for an NIE conference last December (Angoff, 1975). It was, as you would expect, a good paper and makes unnecessary any literature summary now. In fact, it would seem doubly silly to do so since our roles were reversed at that conference.

My third point is one which I made at that time (Green, 1975). It is that we should not accept the absence of outside criteria. I have come to believe that achievement tests should be validated against outside criteria. Listen to the critics of the current crop of standardized achievement tests. They do not find the content and construct validity evidence offered for the tests very compelling (e.g., Levine, 1976), and many are becoming increasingly convinced that the tests are badly biased. It is true that in a number of cases these beliefs appear to be based more on distrust and suspicion than on evidence, but others have made some logical arguments for that viewpoint (see for example, Green, Nyquist & Griffore, 1975). I believe that the potential merit of the criticism is strong enough so that only evidence from external criteria will suffice to counter them.

This does not mean I think we are wasting our time here. In the first place, even if everyone were to agree immediately about this need it will take a long time to develop consensus about criterion measures. Furthermore as a practical matter during test construction one must usually function without an outside criterion and it is the test construction process we are discussing. Ultimately the validity of a test has to be established in use but it does have to be built and it would be most valuable if we knew what internal characteristics of tests and items tended to make a test biased. It was with this sort of question in mind that I started working on the topic about six years ago.

Most of this work on bias has been concerned with the item selection aspect of test development. To avoid misunderstanding let me describe where that fits in the test development process. After the rationale and specifications of the test are produced, items are written to fit these specifications. They are edited and assembled into tryout tests, which are then administered to a sample of the target population. From the results of this tryout an item analysis is produced which becomes the basis for

selecting the items for the final version of the test.

Item selection involves, first, eliminating defective items since no matter how experienced the item writers and how careful the editing, some of these always appear. If necessary, i.e., if there are not enough substitutes, these items are revised in hopes of improving them, but this is not desirable since then one is less certain about the difficulty level and discrimination power of the resulting test unless another tryout is done. Next comes selecting the most efficient and effective set of the remaining items. Efficiency and effectiveness relate to difficulty and discrimination. One would like to insure that the test contains a set of items with a suitable range of difficulties; that the items show growth over the period of time the topic is taught, and that the resulting test will exhibit adequate reliability. Thus items should discriminate and each should contribute to the reliability of its subtest. Other things being equal, items with good item test correlations are the ones to choose. In short, it it highly desirable to have a choice of items that one can use and still meet the rationale and content specifications for the test.

At CTB the practice is to tryout anywhere from 1 1/2 to 3 times as many items as are ultimately needed. A 2 to 1 ratio is typical. The higher ratio is used for those areas and item types about which relatively little is known or which have been found difficult to deal with in the past. Thus ordinarily there are several items to choose from for each content category.

When we first started thinking about test bias it occurred to us that this selection step might be accentuating the bias. The argument goes like this: there may be characteristics of the tryout sample which influence their ways of responding to the items in addition to those the test is intended to measure. Such things as general background knowledge, language styles and dialects, cultural values and motivations are likely candidates. Items responsive to these characteristics will tend to look

like good items and will tend to be selected. Since tryout groups usually more nearly represent the majority than any one minority in these things (i.e., minorities are in the minority!), majority group characteristics will determine selection. So in my first study the question asked was would one choose the same items if one had a minority tryout group instead of a "standard", i.e., representative group. Using the standardization data for CAT-70 I set up various tryout groups differing by race, SES, and region of the country. Using the point biserial as the basis for selection, the "best" half of the items in the several tests of the 1970 California Achievement Tests were chosen for each of the various ethnic and regional groups studied. The purpose was to see if the same items would be selected if minority groups were used for the item tryouts. The differences in choices varied from 20%-50% of the items chosen. Therefore it appears that tests built from minority group tryouts would differ from the tests created using the usual sort of tryout sample. Not only would the particular items chosen be different, but probably so would what the test is measuring be different even though all the items in the pool were written to the same specifications. This follows from the finding that the intercorrelations among the noncommon parts of the various item sets thus created (all data came from the standardization sample who took all the items) suggest that in many instances these items did measure different things for different groups. However, many of these "tests" were perforce very short and not very reliable, so the generality of this conclusion is open to question.

Having concluded that using a separate minority tryout group would change the test, it seemed reasonable to try having both a black and a "standard" (i.e., representative) tryout group for Form S of the CTBS which we were then constructing. We did this and we have learned many things not the least of which is that the use of multiple tryout groups is not only expensive, but procedurally most awkward. I should note that the editorial

policy we set up was that the black tryout data would be used as a screening device to detect items markedly bad for blacks. We emphasized low point biserials as the screening criterion, but unlike the procedure in the first study, all the item data were available for use. Actually the procedure varied widely from level to level and test to test both because content validity considerations were supposed to prevail when there was a conflict and because each of the many different editors seemed to be able to find ways of making unique interpretations of the data they were given. I can say with certainty only that all of them looked at the item analysis data for both groups and tried to produce the best test they could.

In any case, whatever the practice, the resulting sets of point biserials were affected. The kind of effect is illustrated by Table 1 which shows frequency distributions of point biserials for the reading comprehension test of Level 3 of CTBS/S/ (CTB/McGraw-Hill, 1974). As the table illustrates, it appears from the tryout data that in most subtests at most levels (something over 60 separate tests are involved) the items finally chosen had higher point biserials for blacks relative to those for the standard groups than would have been expected if the black data had been ignored. From the standardization data is appears that there are fewer items with really low point biserials for blacks (i.e., under 0.20) than otherwise might have been expected. Whether the resulting tests are in fact better for blacks than might have been expected remains to be seen. The only other thing I can say at this point is that based on standardization data, the number of very low point biserials is generally higher for the Spanish speaking group than for blacks. Since there were no data for Spanish background groups obtained in the tryout, perhaps that result means something. I hope in due course to be able to make stronger statements.

Please note that while offering these data as weak indications that our test construction activities did reduce bias against blacks a little, I do not claim that such data mean anything very positive about the validity of

the tests. However remember we are talking about choosing items from among those that fit the content requirements and thus the elimination of items not related to total score among blacks should mean more adequate measurement for that group.

Still it is all based on the assumption that overall the test is valid, which is the assumption of goodness I mentioned at the beginning. In his review Angoff describes any such procedure as a bootstrap operation because of this assumption. All of the procedures being considered in this symposium are directed at item bias rather than test bias and most of them find this assumption necessary. Perhaps, given that content validity is given first consideration, the proposition is reasonable. But if it is not we are in trouble. However we are in trouble anyway because the various procedures have problems which tend to lead to conflict with this assumption.

A basic one is that minority groups usually score lower than the majority. On the one hand, given content validity, this may merely mean the low scoring group has achieved less on the average. On the other hand, these lower scores are the starting point of the suspicion of bias, and consequently one cannot accept the assumption until it is demonstrated valid. Without some way to talk about the relationship between item bias and test bias it seems that we are going around a very nearly circular path.

It became apparent to us some time ago that we needed a way to talk about the amount of bias in a test and to relate procedures for identifying biased items to that. In a 1972 APA paper with John Draper I had proposed thinking of test bias as consisting of those factors in scores unique or specific to particular groups. John Draper had been exploring ways of using factor analysis to identify these "group specific" factors. While not entirely successful he suggested that the amount of "group specific" variance in contrast to the variance common to the groups could be considered an index of amount of bias. He further set up a model of this and proposed we do a

simulation of various item selection procedures to see if they did affect the proportions of group specific variance or bias. At this point John departed for the greener grass of SRI. Fortunately for me my colleague Wendy Yen stepped in, finished developing the model and worked out all the procedures. Although she has been my guide through the printout piles, she is not responsible for my conclusions* However you should assume that all the good ideas are hers.

It took a while to get started but we are now into the endless games that simulations lead to. So far we have only looked at point biserials and even that is not all done. Still I would like to discuss briefly what I think we have found so far.

Our model consists of 10 items and two groups with three factors common to the two groups and 10 factors having variance in only one or the other group, i.e., five factors specific to each group. Groups 1 and 2 were assumed to have 170 and 670 members respectively. Obviously Group 1 was meant to represent the minority group.

Two sets of difficulties for the 10 items were arbitrarily chosen as were the loadings of these items on the common factors. An equal amount of error variance for each variable in each group was assumed; errors were assumed independent of each other and of other factors. These are displayed in Table 2. Also postulated were the several sets of covariances shown in Table 3. From these data the common variance, which was used throughout, was determined. Next, starting with zero loadings, the group specific loading was randomly incremented in Group 1. Then, not altering Group 1 further, thirty iterations (increments in group specific loadings) were made for Group 2. For each of these thirty iterations, item point biserials were calculated.

*I must confess to near total ignorance of factor analysis. I did take a course once from Karl Holzinger but it had no effect. I would like to blame Karl but Henry Kaiser took the same course at the same time and as far as I know it is the only such course Henry has taken also.

At the end of these thirty iterations, all the group specific loadings in Group 2 were set back at zero. A second randomly chosen group specific loading was incremented in Group 1. Again, not altering Group 1 further, thirty iterations were performed on Group 2. In this way a total of 30 x 30 combinations of different amounts of group specific variance in Groups 1 and 2 were examined.

Using the point biserials calculated for each iteration, the "best" five item tests were chosen. As a first stab we proceeded in two ways. One way was to select those items whose point biserials for the two groups differed least, the other was to rank the point biserials within each group and choose those with highest average rank. As might be expected these two selection procedures produced very different sets of "best" items.

To determine the effect on test bias the ratio of group specific variance to the total variance for the 10 items in each pair of iterations was determined and compared to the same ratio for the five items selected. If the latter figure is smaller one can say that the selection reduced the amount of test bias. The outcome of that comparison can be seen in Table 4. The proportion of the group specific variance in the selected item set does seem to be reduced when the amount of that variance is small to begin with. As it gets larger the selection procedure appears to become less effective. When the bulk of the variance in Group 1 is group specific this selection procedure tends to increase it still further.

The second selection procedure appears to be ineffective in reducing bias when the amount overall is low but quite effective in increasing it when group specific variance is the majority.

Ordinarily Groups 1 and 2 would not be separated and the item statistics would be calculated for the total group. Therefore after each iteration

References

Angoff, W. H. The investigation of test bias in the absence of an outside criterion. Paper presented at the NIE Conference on Test Bias, Washington, D.C., December, 1975.

Cardall, C., & Coffman, W. E. A method for comparing the performance of different groups on the items in a test (ETS RB 64-61). Princeton, NJ: Educational Testing Service, 1964.

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

CTB/McGraw-Hill. Comprehensive tests of basic skills, form S: Technical bulletin no. 1. Monterey, CA: Author, 1974.

Green, D. R. Racial and ethnic bias in test construction. Monterey, CA: CTB/McGraw-Hill, 1971.

Green, D. R. Procedures for assessing bias in achievement tests. Paper presented at the NIE Conference on Test Bias, Washington, D.C., December, 1975.

Green, R. L., Nyquist, J. G., & Griffore, R. J. Standardized achievement testing: Some implications for the lives of children. Paper presented at the NIE Conference on Test Bias, Washington, D.C., December, 1975.

Levine, M. The academic achievement test. American Psychologist, 1976, 31, 228-238.

Table 1. Frequency Distributions of Point Biserials for the Tryout and Standardization by Ethnic Group for Reading Comprehension, CTBS/S, Level 3.

	TRYOUT				STANDARDIZATION			
	Items Rejected		Items Accepted					
Pt. Bis.	Standard	Black	Standard	Black	Standard-	Black	Spanish	Oth
.800-.899								
.700-.799								
.600-.699			2	1	2			2
.500-.599	4	1	7	3	12	5	7	11
.400-.499	11	9	14	18	17	18	15	18
.300-.399	14	6	16	12	12	15	16	11
.200-.299	9	11	4	9	2	3	5	3
.100-.199	2	8	1	3		4	2	
.000-.099	2	7						
-.001-.099								
-.100-.199								
-.200-.299								
Median	.357	.253	.417	.391	.461	.401	.396	.464

Table 2. Simulation Model

MODEL 3 30 ITERATIONS

INITIAL PARAMETERS

NUMBER OF OBSERVED VARIABLES =	16
NUMBER OF COMMON FACTORS =	3
NUMBER OF SPECIFIC FACTORS FOR GROUP ONE =	5
NUMBER OF SPECIFIC FACTORS FOR GROUP TWO =	5
NUMBER OF ITERATIONS FOR RUN =	30
VALUE OF INCREMENT IN GROUP SPECIFIC LOADING =	0.240

ITEM DIFFICULTIES FOR THE GROUPS -

1	0.480	0.200	0.290	0.390	0.450	0.200	0.450	0.360	0.520	0.500
2	0.780	0.700	0.520	0.610	0.710	0.250	0.850	0.590	0.720	0.710

Z SCORES FOR THE GROUPS -

1	-.050	-.843	-.555	-.275	-.125	-.843	-.125	-.355	0.050	0.0
2	0.773	0.525	0.050	0.275	0.553	-.675	1.035	0.225	0.580	0.553

LOADINGS FOR COMMON FACTORS

1	0.496	0.001	0.135
2	0.069	0.046	0.008
3	-0.041	0.041	0.115
4	0.051	0.404	-0.166
5	0.020	-0.045	0.109
6	0.213	0.101	0.119
7	0.235	0.067	0.072
8	0.411	-0.027	-0.008
9	0.010	0.067	0.528
10	0.228	0.157	-0.074

COVARIANCE OF ERROR FOR ALL GROUPS

1	0.120	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.120	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.120	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.120	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.120	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.120	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.120	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.120	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.120	0.0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.120

Table 3. Model Covariances

COVARIANCE OF COMMON FACTORS

1	1.000	0.474	0.437
2	0.474	1.000	0.505
3	0.437	0.505	1.000

COVARIANCES OF COMMON + GROUP SPECIFIC FACTORS FOR GROUP 1

1	0.140	0.145	0.096	0.103	0.122
2	0.135	0.146	0.114	0.092	0.138
3	0.145	0.179	0.108	0.090	0.148

COVARIANCES OF COMMON + GROUP SPECIFIC FACTORS FOR GROUP 2

1	0.139	0.143	0.112	0.158	0.137
2	0.144	0.129	0.136	0.137	0.149
3	0.159	0.153	0.149	0.148	0.134

COVARIANCE OF GROUP SPECIFIC FACTORS FOR GROUP 1

1	1.000	0.548	0.305	0.358	0.543
2	0.548	1.000	0.373	0.395	0.595
3	0.305	0.373	1.000	0.225	0.324
4	0.358	0.395	0.225	1.000	0.384
5	0.543	0.595	0.324	0.384	1.000

COVARIANCE OF GROUP SPECIFIC FACTORS FOR GROUP 2

1	1.000	0.515	0.466	0.503	0.499
2	0.515	1.000	0.387	0.493	0.524
3	0.466	0.387	1.000	0.417	0.417
4	0.503	0.493	0.417	1.000	0.522
5	0.499	0.524	0.417	0.522	1.000

Table 4. Effect of two item selection procedures on the relative amount of group specific variance in the test.

Selection procedure: Smallest differences between point biserials

Group 1 Iteration	% Group Specific Variance in Group 1	Number of times in 30 the proportion of group specific variance was: increased	unchanged	decreased
1	1	0	5	25
5	8	1	6	23
10	20	3	7	20
15	33	2	9	19
20	43	5	15	10
25	51	6	12	12
30	58	13	6	11

Selection procedure: Highest average rank of point biserials

1	1	7	20	3
5	8	6	12	12
10	20	4	22	4
15	33	6	24	0
20	43	1	28	1
25	51	24	6	0
30	58	28	2	0

Table 5. Effect of item selection procedures using two groups in contrast to a selection procedure using pooled data.

Selection procedure: Smallest differences between point biserials

Group 1 Iteration	% Group Specific Variance in Group 1	Number of times in 30 the proportion of group specific variance was increased	unchanged	decreased
1	1	0	5	25
5	8	2	5	23
10	20	3	10	17
15	33	2	14	14
20	43	5	13	12
25	51	2	12	16
30	58	0	20	10

Selection procedure: Highest average rank of point biserials

1	1	0	22	8
5	8	0	15	15
10	20	6	21	3
15		14	16	0
20	43	5	23	2
25	51	15	15	0
30	58	21	9	0