DOCUMENT RESUME

ED 084 285 TM 003 293

AUTHOR Green, Donald Ross
TITLE Racial and Ethnic Bias in Achievement Tests and What To Do About It.
NOTE 9p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Achievement Tests; *Criterion Referenced Tests; Cultural Differences; Ethnic Groups; *Norm Referenced Tests; Problem Solving; Program Evaluation; *Student Evaluation; *Test Bias; Test Construction; Test Results

ABSTRACT

A description of two proposals for alleviating the racial and ethnic bias in tests of achievement used in schools is presented. One of them entails adding steps to the construction procedures used in building norm referenced achievement tests; the second entails using criterion-referenced achievement tests rather than standardized tests for certain purposes. The principal uses of achievement tests are to: (1) evaluate the status of a student or a set of students in a class, school, or school system; (2) evaluate programs, curricula, and instructional materials; (3) diagnose problems; and (4) provide a basis for planning individual, class, or system programs. The bias built into tests arises in the minds of those who write and edit the tests and from the procedures used to improve the tests. It is suggested that members of each of the groups concerned with the test participate in constructing the examinations from the start and to use item writers and editors that represent all major ethnic and cultural groups in the population. Criterion-referenced tests should be designed to show exactly what the pupils have learned; these tests should be used for specific diagnosis of school and program problems. (CK)

RACIAL AND ETHNIC BIAS IN ACHIEVEMENT TESTS AND WHAT TO DO ABOUT IT

by
Donald Ross Green
CTB/McGraw-Hill

The purpose of this paper is to describe two proposals for alleviating the racial and ethnic bias in tests of achievement used in schools. One of them entails adding steps to the construction procedures used in building the usual standardized norm referenced achievement tests; the second entails using criterion-referenced achievement tests rather than standardized tests for certain purposes.

This discussion will be limited to educational achievement tests for two reasons. First, it seems likely that the problems associated with racial and ethnic bias in achievement tests can be substantially solved, partly because the issues concerning validity in achievement tests can be dealt with in a largely rational and logical manner. On the other hand, in the areas of aptitude tests, personality tests, and other sorts of tests, questions concerning bias require consideration of many more issues concerning values; hence, they cannot be dealt with as rationally. Therefore the problems of bias in these other areas are much less readily solved, and there does not seem to be any researched suggestions or solutions to offer although some of the procedures described here might apply. The second reason for limiting the discussion to achievement tests is that they constitute the majority of CTB's business, therefore, it is the topic about which we know most.

Standardized aptitude tests and achievement tests are often said to be one and the same thing, and the assertion is then made that the latter have all the bias problems of the former. Neither statement is true; they are not built to the same specifications and more important they are generally not used for the same purposes. In fact, there is substantial evidence recently available which demonstrate their difference.[1]

The principal uses of achievement tests are to: (1) evaluate the status of a student or a set of students in a class, in a school, or in a school system; (2) evaluate programs or projects, curricula, and instructional materials; (3) diagnose pupil, class, program, or system problems; and (4) provide a basis for planning individual, class, or system programs. Although achievement tests are usually published and distributed as separate entities, they may also be published and sold as parts of other instructional materials. Other achievement tests are produced by school systems or state personnel for their own use, although many of them end up being distributed widely. But published or unpublished, all these tests are almost certainly biased to some degree, large or small, against certain subgroups of the population they are intended to serve.

On this point the evidence is strong: there is bias in tests. The quantitative effects of this bias on test scores have not been adequately assessed. There is some evidence that these effects are not large for most minority groups taking the customary achievement test batteries (Green, 1972), but the same evidence demonstrates the bias does exist in the test. It is quite true also that there is bias in the use of tests and their misuse explains many of the objections to tests and testing now encountered; more will be said on that point later. However, it should be categorically stated that misuse is <u>not</u> the full explanation no matter how appealing that assertion may be to those who constitute the testing establishment, including, of course, test publishers. There is bias in the tests themselves, and it derives from the procedures used in the construction of these tests.

Bias in the construction of tests deserves close attention because it is something that publishers can do something about. It is their principal responsibility. Misuse may or may not be a publisher's responsibility depending on the circumstances, but there is no question that the publisher of the test is responsible for the bias built into the test by the processes used in its construction.

As it happens, some bias is inevitable; there is no way to build a completely unbiased test that is of any use, any more than one can find a completely unbiased individual who has any values and opinions.

The bias built into tests has two principal sources. The first arises in the minds of those who write and edit the tests; the second stems from the procedures used to refine and improve the tests by trying them out and examining results. The first source of bias occurs simply because of cultural differences between users and producers of tests in styles of thinking, perceiving, and reasoning and in values and expectations. Another way to describe this phenomenon is to note that it is a result of a lack of congruence in perceptions of those producing the tests on the one hand and of some of those taking the tests on the other, as to what the task being presented is and what it means.

The most common recommendation for dealing with this source of bias is to have the materials reviewed by sophisticated members of the ethnic and cultural groups concerned. This procedure is often useful and should be followed whenever appropriate, but it is not adequate by itself. Such reviews certainly help eliminate the usually unconscious racism that sometimes has been visible in tests and other published materials, but the ability of anyone, no matter what his background, to really know what goes on in the minds of children when they face certain sets of materials is limited. None of us can simply look at materials and know precisely what thoughts will arise in a child's mind when he is in contact with these materials. Therefore, determination of bias must be an empirical procedure that includes direct examination of situations and data after materials have been prepared.

There is a possible earlier step that logically ought to be effective in reducing bias of this sort, i.e., the bias that occurs because of the differences of the styles of thinking among cultural groups. That procedure would require that members of each of the groups concerned participate in constructing the

examinations from the start. At least the initial drafts of the test materials would then have a heterogeneous set of biases built into them. The next step necessary to producing excellent tests is to try out the materials. Another part of the remedy for the first source of bias and the second reason that tests are biased relate to this tryout.

The second source of bias has its effect when data from the population, or a sample of it, are used to improve the effectiveness of the test by selecting, rearranging, and rewriting items. This procedure is essential to producing an effective achievement test but the improvement derived from it is not uniformly beneficial to all groups. Because the characteristics of the predominant group in the sample determine the results of this step (ordinarily called an item tryout), the test is usually sharply improved for that group (this is a desirable result), but relatively less improved for minority groups. The minority elements in the sample group do create noise in the data if they react to the materials in any way unlike the majority but this does not substantially affect the outcome. The characteristics of the majority group remain the determining factor in the process. The result is a better test for many children but a relatively more biased test for those minorities whose styles diverge from the majority of the tryout group. Note that majority and minority are defined here by the characteristics of the tryout group. If the tryout group were predominately black, blacks would be the majority group and the process would improve the test more for them than for others, i.e., it would tend to make the test biased against whites and other non-black groups.

The most promising solution to these dilemmas is to use item writers and editors that represent all major ethnic and cultural groups in the population, with each group producing a separate trial version of the test. The second step would be to try out <u>all</u> the materials on each subgroup <u>separately</u>. The third step would be to select items from all versions and edit them to best serve the interests of all groups.

At CTB we now believe, at least tentatively that one can build achievement tests that are less biased against minorities, but as adequate as ever for the majority by following these procedures. In other words, we believe that the divergence from the main stream or "middle America" view of the world of the major subcultural groups of the population that we are concerned about is not so great as to preclude the possibility of a common test that is reasonably fair to all concerned. Studies to confirm these assertions are in progress; available evidence supports the position

One report of a preliminary study of this matter is available (Green, 1972), and hopefully others will be forthcoming in 1974. Specific procedures for detecting biased items are given in a report by Green and Draper (1972). These reports refer to what to do with the data derived from the separate tryouts recommended above.

The purpose is to construct a test best for all groups; it is of course possible that "best" will require different tests for each group. If this occurs, logic and humanity both require the subsequent use of different and not comparable tests for each group. The information "lost" would be false and not worth collecting. It should be noted again that to date our evidence suggests that these untoward results are not likely on any large scale.

As suggested earlier, many groups in the establishment (publishers are only one such group) prefer to consider misuse as the major source of bias in tests as used in schools. The problem is indeed real and solutions are needed. Amid the many recommendations for better teacher training, better supervision, better manuals and guides, and so forth, all of which appear to have been remarkably ineffectual in reducing misuse, there is a step that can be used in many situations to solve a variety of these problems directly. That step is to substitute criterion-referenced tests for typical standardized achievement tests in many of the situations in which the latter have been misused. There is a kind of bias or misuse of achievement test batteries that arises from a misunderstanding that has been around a long time.

Regular standardized achievement tests are built to measure broad skills such as reading, mathematics, and language which develop slowly in elementary school. They are designed to differentiate among pupils in these areas in a reliable and stable manner. These two criteria mean that the chances of reliably detecting any changes in score during, say a four-month period are small and are lowest for the students at the bottom end of the scale. Thus any assessment of progress over periods of less than a year is likely to show minimal gains, especially for those starting at a disadvantage. Because this is not widely understood, many pupils are discouraged, many teachers and programs are judged ineffective, and initially low scoring groups are almost certain to fail to show "significant" gains. Telling teachers and especially children that their efforts were futile when that is not true is plainly damaging. The pupils basically have learned things but the tests do not show it because they were not designed to do so.

Criterion-referenced achievement tests are, or should be, designed to show just that. Items in a criterion-referenced test should be written and selected to measure behaviors sufficiently specific to be taught directly in reasonable lengths of time and should reflect this change in behavior, i.e., learning. Sensitivity to instruction, not sensitivity to individual differences, is the standard for a good criterion-referenced achievement item (Roudabush, 1973). Logically such items should be less biased against minorities, but empirical evidence on this point is lacking and again it may be necessary to obtain separate tryout data for each ethnic group since new tryout procedures may introduce new sources of bias. Support for research on this topic is needed. In any case, criterion-referenced tests are not only directly useful in diagnosing instructional needs but are also the only reasonable way to evaluate progress and programs during an academic year.

For long term evaluation of the major academic goals of schools, the traditional achievement test (built to minimize bias) is by far the best source of information available. For example, such tests properly used have established

what a miserable job of education most schools are providing minority groups. However, for this purpose such tests need be given yearly at most and, in many cases, only to samples of pupils. For use in the classroom by teachers and for measuring progress toward short term goals, criterion-referenced tests are the best available answer. It seems probable that such a testing program would sharply reduce the often justified complaints of bias and lack of relevance in tests.

For several reasons, the reduction in bias resulting from this use of criterion-referenced tests should be direct and substantial in addition to elimination of that stemming from inappropriate assumptions about the meaning of standardized test data. First, the data are more direct because they refer to a set of relatively specific instructional objectives (e.g., "Can the student add two 2-digit numbers requiring regrouping?") rather than a more general trait (e.g., "arithmetic computation"). Inappropriate items are not only more obvious but they can also be ignored by either student or teacher since each objective is assessed separately. Scores are not derived from counting all different kinds of items in one domain. A sort of customized interpretation is immediately and directly available to all consumers of the data. Furthermore, inappropriate items can be spotted in advance and students can be told not to answer them with no adverse consequences on "scores." In fact there really are no scores, only a set of data about knowledge and skills that permit one to say "yes, he knows that" and "no, he still needs to learn this." Invidious comparisons are hard to come by (but of course possible) since norms are not routinely available. Of course class, school, district, or state norms or goals can be determined and evaluated but global comparisons and therefore negative labels are avoided because the large number of objectives, each of which is evaluated separately and independently, discourages generalization.

The principal strength of criterion-referenced tests is that they are built to reflect and respond to instruction so that if a teacher teaches something and a student learns it the test will show it immediately. In short, criterion-referenced

tests are suitable for classroom use and we believe that as they become used more widely, teacher and student disaffection with testing will be reduced because the distortions, misuse, and bias will be curtailed.

Criterion-referenced tests conceivably could produce new sources of problems with bias. The items could turn out to be just as biased and misleading as those from the more traditional achievement tests and that possibility needs study. However not only do the item specifications and selection criteria seem less likely to permit bias to operate strongly, but also since large numbers of items are not summed, the bias, if any, does not accumulate. Therefore it seems reasonable to predict that the bias found in criterion-referenced tests will be minimal and will have a relatively negligible effect on children.

SUMMARY

Typical achievement tests are biased to some degree and are often used inappropriately and in biased ways. Two kinds of remedies are proposed. One entails procedures for building less biased tests; the other entails differentiating among the uses of achievement tests by using criterion-referenced tests and regular achievement batteries for different purposes.

To build less biased tests, members of all relevant population groups should participate in their construction from the start. Items should be tried and evaluated in separate samples of these groups to enable one to build a test appropriate for all. These procedures should be followed for both criterion-referenced tests and the traditional norm referenced achievement batteries. The latter instruments should be used for evaluation of programs and general long term (e.g., year-to-year) progress and status of schools and districts. For specific diagnosis of school and program problems and especially for individual instructional guidance, criterion-referenced tests are needed. They should prove to be relatively unbiased.

Footnote

[1] One source of evidence comes from a recent study done at CTB by Burket (in press). He has shown that, given adequate quantities of data, one can usually distinguish between aptitude tests and achievement tests scaled to have the same means and variances simply by looking at these test scores without knowing ahead of time which set of scores are which. One can examine the pattern of test scores over a period of time or across groups of students at different grade levels, and by looking at these patterns say this has to be the achievement test and this has to be the aptitude test. Typically they do not behave the same way, they are not alike. Another example comes from a recent study reported by Carroll (in press). Carroll was able to show that students at the beginning of a course of study in a foreign language knew absolutely nothing about that foreign language and had zero scores on a test of knowledge of the language. Nevertheless, their performance during the course was successfully predicted by a language aptitude test. At the end of the course, predictions as to who would do well and who would not do so well were verified. Furthermore, the aptitude test was then given again and the scores on it had not changed. Thus, the scores on the achievement tests had changed from a uniform zero at the beginning to a predictable set of different scores at the end. The aptitude test predicted final outcome on the achievement test but the reverse prediction was not a possible event since all pretest achievement scores were zero. Clearly the two tests were different.

In short, one cannot argue rationally that aptitude tests and achievement tests are the same; they are different in their intent and their purpose, they are built in different ways, and they differ in the degree of abstraction of the meaning of their scores and in the number of assumptions that one has to make to interpret those scores. For example, a major assumption usually made about an aptitude test, which is not made for an achievement test, is equality or at least equivalence of opportunity and experience among those performing at any given score level. Achievement tests are ordinarily used differently than aptitude tests, in particular, they are not selection and prediction instruments, but that is not the only difference. They are also different in their construction, and although both kinds of tests may be and usually are biased, the achievement tests' bias problem can probably be solved to a substantial degree, whereas the problem in aptitude tests appears much more difficult. When tests built to be achievement tests are used for selection and prediction as though they were aptitude tests, that use introduces all the bias problems that go with aptitude tests and perhaps others as well.

REFERENCES

Burket, G. Empirical criteria for distinguishing and validating aptitude and achievement measures. In D. R. Green (Ed.), The aptitude-achievement distinction. Monterey, Calif.: CTB/McGraw-Hill 1973, in press.

Carroll, J. B. The aptitude-achievement distinction: The case of foreign language aptitude and proficiency. In D. R. Green (Ed.), The aptitude-achievement distinction. Monterey, Calif.: CTB/McGraw-Hill, 1973, in press.

Green, D. R. Racial and ethnic bias in test construction. Monterey, Calif.: CTB/McGraw-Hill, 1972.

Green, D. R. & Draper, J. F. Exploratory studies of bias in achievement tests. Paper presented at the meeting of the American Psychological Association, Honolulu, 1972; ERIC ED 070 794.

Roudabush, G. E. Item selection for criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, 1973.