DOCUMENT RESUME

ED 106 348 TM 004 464

AUTHOR Green, Donald Ross
TITLE What Does It Mean to Say a Test Is Biased?
PUB DATE [Apr 75]
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Achievement Tests; Groups; Individual Differences; Placement; Predictive Ability (Testing); *Test Bias; *Testing; *Testing Problems; Test Interpretation; Test Results; Test Selection; *Test Validity

ABSTRACT

Biased tests systematically favor some groups over others as a result of factors not part of what the test is said to measure. Bias is basically a problem of differential validity. Validity can be discussed in terms of either the procedures for establishing it or test use. Both ways clarify bias in any test. For content and construct validity, the question is "Does the test measure the same thing for each group? In criterion-related situations, an important question is "Fair to whom?" Fairness to the selector is not identical with fairness to the selectees. (Author)

AERA Session 7.01

WHAT DOES IT MEAN TO SAY A TEST IS BIASED?

by

Donald Ross Green

CTB/McGraw-Hill

A paper presented at the
American Educational Research Association in Washington, D.C.
March 31, 1975

INTRODUCTION

A biased test is generally understood to be a test that produces results that are systematically unfair to some group. For this to happen, the test must ordinarily measure variables for that group at least partly distinct from those it measures for other people in the population. There are two elements in this proposition: on the one hand, there is the matter of fairness and, on the other hand, there is the matter of measuring different things for different groups. Although the two are logically related, they are not identical in all instances. That is, a test must usually measure different things for different groups to be unfair, but it is possible to have a test that measures different things for different groups and yet does not produce unfair results because of the way it is used. This happens when it is known what the test is measuring for each group, and the test is used in a different but valid way for each group. This situation is apparently rare. Ordinarily a test that measures different things for different groups is likely to be used in a way that is at least somewhat unfair to many of those in at least one group. Such a test can be called biased (Merz, 1974). Thus, a biased test is one that is likely to be unfair in use, but the unfairness appears in the use of the test scores. Unfairness is not in the nature of the test, but bias is.

The condition of measuring different things is necessary for unfairness to occur, with the exception of one particular placement situation (see Bias in Placement below). In other words, bias is ordinarily a necessary condition for unfairness. This can be seen by considering the meaning of bias or unfairness when a test measures precisely the same thing for all. If that is the case, but bias is claimed, it must be unfair to measure that thing for one group but not another. There are

two cases in which that condition might seem reasonable on the surface but is not. One is if the scores are interpreted to mean something else, and it turns out that this additional inference is reasonable for some but not all groups. An example would be a test that measures word knowledge but is taken to measure intelligence. The use of vocabulary tests as intelligence tests have exactly this flaw as Williams has shown so well with his BITCH Test (Williams, 1972). Used only as an indication of word knowledge, the test is not biased (note that this does not mean that the test is valid). The other possibility for a test to measure apparently the same thing and be biased would occur if the test was less reliable for one group than another. Since this can be described as measuring more error for one group than another, it is merely a special case of measuring different things. Thus both cases turn out to relate to measurement of different things.

Consequently, the issues about bias concern, first, what a test measures for whom and, second, how the results are used. This means that bias issues concern validity. The issue is, in fact, test validity, and the only difference between this discussion of bias and an ordinary discussion of validity is that the question at hand is one of differential validity. The question asked in the title of this paper then can be restated as follows: Under what circumstances are there different degrees or amounts of validity for different groups for the same instrument? The groups must be defined on some basis other than scores on the test in question. Most discussions of unfairness refer to race, sex, nationality, or the like. When not otherwise specified, it will be assumed here that the case at hand concerns two groups, such as one minority and the rest of the larger population to which it belongs.

There are two traditional approaches to discussions of validity. One proceeds from the point of view of use, the other from the point of view of procedures for establishing validity. Since unfairness can only be evident in test use, that approach will be taken here. However, each approach has something to offer to the discussion and needs to be examined in the light of possible differential conclusions concerning validity for different groups. When such differential conclusions are obtained then one must determine how they impact on the question of fairness.

CONTENT BIAS AND DELIBERATE MISUSE

First, it is necessary to dispose of two issues that often muddy the waters of the topic -- content bias and deliberate misuse.

Content Bias

Content bias is the expression in tests of racism, sexism, and other inappropriate attitudes. Such content may or may not affect scores. Occasionally, it may be seriously offensive and arouse negative emotions, thereby interfering with performance. Sometimes it may be merely irritating, and sometimes it may pass unnoticed and not affect scores. For example, there is no particular reason why a set of mathematic items that have girls in the kitchen and boys in the shop should lead women to score less well on that particular test. To be sure there may be cumulative negative effects on performance from prolonged contact with such attitudes, and in any case it is desirable to delete or alter material that is not only directly offensive, but also that unnecessarily denigrates and pigeonholes some group. There is some literature on this topic, and guidelines for dealing with it are available (see Dunfee, 1974; McGraw-Hill, 1973, 1974).

However only if the test measures different things for different groups can the test be called biased as defined here; therefore evidence of that fact is necessary. Content bias may in some instances cause a test to be biased but neither the absence of such content bias, nor its deletion if found, is evidence that test is or has become unbiased. In short, for our purpose here it is a side issue.

Deliberate Misuse

A more confusing issue is the matter of deliberate misuse of tests. In the abstract, one can easily dispose of this problem by noting that it is not really related to the instrument. However, it is a much harder issue to deal with than content bias because sorting out deliberate from inadvertent misuse is often impossible. The first represents malice, the second ignorance, but the two are not prelabeled. Furthermore, there are two kinds of ignorance. One is the ignorance of some people of facts known by others. The other is the ignorance common to all about either the nature of some instrument or about the causes of behavior, i.e., the kind of ignorance all of us here are working to dispel.

It has been claimed by some (e.g., Sommer, 1970) that misuse stemming from malice or unnecessary ignorance is the source of test bias and that consequently the only solution is to change people, i.e., educate them to use tests properly. The omission of the second kind of ignorance is a major flaw in this position since it takes test validity as given thus begging the question.

Furthermore the solution of better training of the people who use tests does not face the problem squarely. Tests are sometimes deliberately used to produce unfair results for some people; the people who use

them in this way are often biased. Almost any kind of technology can be misused by people if they set their mind to it. On that score, tests are not different from any other sort of instrument; they can be and sometimes are misused. While that does not make the instrument itself biased or defective, it also does not point to education as anything but an idealistic solution. Perhaps in the long run educating people about testing can improve this situation, but it really does not seem likely. Potentially more promising are the possibilities of (1) trying to build in safety devices, (2) making tests less susceptible to inadvertent misuse by reducing differential validity, or (3) controlling use. At the moment, I confess I do not see what can be done to accomplish the first of these. Some of us are working on the second (e.g., Green, 1971, 1973) but we have a long way to go as yet. For now the emphasis must be on trying to control use. To accomplish that, proper definitions of unfair use would be helpful. Note that for control purposes the distinction between deliberate misuse and misuse through ignorance is not important.

Some attempts to eliminate misuse do not depend directly on definitions. For example, one attempt to control the use of published psychological tests has appeared in the Standards for Psychological Tests, published by the American Psychological Association (1954, 1966, 1974). The earlier versions of the standards suggested three levels of qualifications to use tests and urged publishers to categorize tests according to the amount of knowledge and skill that is needed to make proper use of them. Many, if not most, publishers have done so and have refused to sell to those who could not show they met the standards. This approach may reduce inadvertent misuse, but it assumes people apply their knowledge

and that is known to be a sometimes affair. Also it obviously cannot prevent deliberate misuse effectively. Although this requirement in a less explicit form remains in the 1974 edition (the three categories are not used), it is noteworthy that a chapter on user responsibility has been added.

Another popular approach is to ban tests. The California legislature, for example, has twice passed a bill outlawing the use of group intelligence tests in schools. The governor twice vetoed it, but it is about to pass again. Ironically, this leaves the field to the judgment of teachers and counselors or to the individual tests (which are not banned) in those cases where it is practical to use them. The unaided judgments of people have long since been proved to be less reliable and valid than those of people who use tests properly; furthermore, the two principal individual intelligence tests, the Stanford-Binet and the Wechsler, in sharp contrast to most of the group tests, were originally designed for and standardized on groups from which black, Spanish-speaking and other minority individuals were excluded. If any intelligence tests are biased, they are. It seems to me that this sort of law makes both deliberate misjudgments easier and inadvertent errors more likely, and in general that controls that avoid definitions will not do the job. Controls may be desirable, but to be effective they need to be based on precise definitions and on proper evidence.

It is of course entirely correct to say that when a test is biased against some group it is a misuse of that test to use it with that group in the same way as with any other group. Misuse is indeed the heart of the matter, but it is either naive or arrogant to assert that therefore the issue of test bias can be dealt with by assuming that one merely

needs to control the actions of the evil and enlighten the ignorant. That position assumes the existence of both established criteria of bias and definitive evidence on the matter for any use of any instrument. Unfortunately, either one or both of these is ordinarily lacking.

Talk of prejudice, unscrupulousness, or ignorance simply clouds the issue. What is needed is a means of determining that an instrument is or is not equally valid for different groups. It should be apparent that, in order to make sense, the notion of equal validity must refer to a particular use.

USES OF TESTS

It has long been recognized that the concept of the validity of a test applies to a use. Not only may a test have a different degree of validity for each use, but also the nature of that validity may vary. A test intended for one use may nevertheless serve many functions; its validity needs to be assessed for each.

The possible uses of tests are many, and the ways of describing them are more numerous still. Tests are used for selection, for placement, and for diagnosing individual strengths and weaknesses. They are used for prediction. They are used to describe or evaluate or assess the status of programs or materials or people or organizations. They are used for measuring change or growth. These and many other words can be used but they all tend to overlap in meaning.

For the purposes of this paper, three uses will be recognized: selection, placement, and description. These three uses are not fully parallel, mutually exclusive, or exhaustive; as a start on a taxonomy of uses they leave something to be desired. For example, selection can be

considered a special case of placement, i.e., the case in which those not selected are placed in discard or ignored. However, the three categories encompass most common test uses and they provide relatively clear contexts for considering the notion of fairness.

All these uses contain the idea that there are differences among people. Tests, all tests, and all uses of tests relate to this fact. Attempts to discuss fairness without acknowledging this fact soon founder. If one believes that treating some individuals differently from others is wrong, then one believes tests are unfair by definition -- in that case there is nothing to discuss. It is also true that if some treatments are believed to be more generally desirable than others, tests that are used somehow to assign only some people to the favored treatment will be judged unfair by at least some of those who are not assigned the desirable treatment. The three uses of tests differ with respect to how clearly this matter is obvious to those affected. It is most clear in the selection situation, least clear in description. This is one reason why bias and fairness need to be considered separately for these uses and why selection is the place to begin.

SELECTION

The use most commonly encountered in discussions of bias in tests is selection as the preceding comments suggest. It is the use for which bias issues have been discussed most adequately, although a definitive work on the topic has yet to appear. At first blush, selection appears to be the simplest of the three uses to discuss; however, a number of recent reports on bias in selection have made it apparent that the situation is more complex than had hitherto been realized (e.g., Cole, 1973; Darlington, 1971; Thorndike, 1971).

Selection using tests requires separating from the rest of some population those falling in a certain range of scores, such as all those either above or below a particular score. The purpose of this selection may be on the one hand to provide the selectees with some special treatment, service, or opportunity, or on the other hand to employ or reject them for the benefit of the selector. Admission to colleges or getting a job are the most common selection situations in which tests are used.

Validity

The validity of a test used for selection is ordinarily established by showing that the test can be used to predict some criterion measure such as course grades or an indication of job success.

Given an acceptable criterion measure, there is a standard process for establishing this validity. It runs roughly as follows: (1) test sample X; (2) select (admit, hire) all of them so as to get a criterion measure for each; (3) determine the regression equation relating the two scores; (4) test a new sample, sample Y; (5) again select (admit, hire) everyone; (6) determine the relationship or validity coefficient for the test, i.e., find the correlation coefficient between the obtained criterion scores for sample Y and their predicted criterion scores based on their test scores and the regression equation established with group X. Plainly this can be done separately for various groups. If the groups all have the same regression equation and validity coefficient, most people would say there is no bias.

Unfortunately, this last statement makes an assumption about predictive validity that is often not true. That is, it assumes that establishing predictive validity fully satisfies the need for validity evidence for selection, when in fact it may not. Take for example a

college using a set of predictive measures for admission and using the freshman grade point average as the criterion measure. If the concern is only to prevent failure and the plan is to admit all who have some specified probability of passing, perhaps the relationship to the criterion is sufficient validity evidence. However, there are other possible concerns such as the degree to which the student will profit from attending, and the payoff for society. These concerns are probably not identical with success as defined by a high GPA.

For these reasons, predictive validity in a selection situation may not be sufficient. In any case, it is also desirable to establish as much construct and content validity as possible. One might ask: "Does the test indicate individual traits that mean that he or she can learn more of the things intended than others?" For example, is it a measure of learning rate under academic conditions? Or again: "Does it measure too much of what the college teaches?" (Since if the student knows it already, who profits from his selection?)

Bias in Selection

The preceding discussion suggests the need to look at the possibility of bias when prediction of the criterion measure is sufficient separately from when it is not. When it is sufficient, the first question is, "Does the test measure different things for different groups as indicated by different regression equations and validity coefficients for these groups?" There is reason to believe this will happen more often than not (Linn, 1973). Given a difference, one can examine the effect of this on fairness. This is the situation that has been considered at length by Thorndike, by Darlington, and by Cole. Since Linn (1973) has published a useful review and analysis of this work, the

details can be omitted here. Suffice it to say that there are many ways in which the equations can differ but each instance can be analyzed and judgments of fairness can be made. Linn's final point is that this work makes "...it quite clear that there is more than one reasonable definition of test fairness and that these definitions in conflict." The selecting agency, those selected, those rejected, and the membership of the different groups may each have distinct views on what is a reasonable choice of definition in a given situation.

A further problem is that this procedure assumes an unbiased criterion measure. There is no reason to make this assumption. Thus one must look at the criterion measure itself for both validity and bias. To do this, one should proceed as one would for a test to be used for description (see below). If the criterion instrument does appear to measure different things for different groups, one might as well assume that the test does also.* Under these circumstances, there does not appear to be any way to proceed fairly using the predictive relationship. One could instead try to assess the construct validity of the test for each group as discussed in the section on description but no clear guidelines for using the test for selection are likely to follow. Without a fair criterion measure the choices are either to ignore the problem, i.e., use some test, that is probably biased, or to select without reference to any relevant criterion, e.g., random selection.

A similar conclusion follows for the other selection situation, i.e., where prediction of the criterion measure is insufficient. Again one may either ignore the problem and proceed as though the criterion

---

*If the test was not biased, it would still appear to be so -- a case of pseudo-bias.

measure were both unbiased and sufficient, or select the first applicants, select at random, or perhaps use "professional judgment". As before, one can resort to procedures for assessing construct validity to determine if the test is biased but the result will probably not provide a good basis for selection much less one that is fair.

This seems to be the situation in selecting students for college. Grade point averages and other customary criteria give probably biased and almost certainly insufficient information about what the student accomplished. Therefore, the claim of fairness to blacks for the SAT is suspect. The fact that the test apparently tends to overpredict their grade point averages (Temp, 1971) does not by itself constitute solid evidence of fairness. Rather, it suggests a biased criterion measure and leaves the question of fairness ambiguous.

Two final comments, one positive and one negative, can be made on the selection problem. The negative one is that few selection agencies are willing or able to do a proper validation study. Schools will not usually admit a random sample of applicants, and most employers will not hire people randomly either. Without the evidence these procedures can provide, any discussion of bias is academic.

The positive comment is that sufficient and unbiased criterion measures are usually possible; they may be complex, involve multiple scores, and expensive to develop and use but they *are* usually possible. If the selector cares enough or is responsible enough to do a proper validity study, then taking the effort to develop an adequate criterion measure is a reasonable expectation. In some cases, it is a still more responsible step to consider the task one of placement rather than selection, since the unfairness is often to those not selected.

PLACEMENT

When the purpose of testing is to determine *which* treatment, *which* course of studies, *which* job assignment, or the like, one talks of placement rather than selection. Logically, it can be treated as an extension of the selection situation with those falling in each score range being selected into a different treatment. In a way, this conception illustrates one of the defects of the notion that establishing predictive validity is adequate for examining validity in selection; it implicity assumes that not being selected is either irrelevant or is the optimum choice for the rejectees. In other words, it considers prediction of only one criterion when there may be more than one kind of outcome. In this sense, selection is unfair to somebody almost by definition; if the end is the good of all, placement is the proper approach.

Validity

Establishing the validity of an instrument to be used for placement is obviously difficult even if there are only two possible treatments. Ordinarily a placement test can be validated only against some criterion measure common to the two treatments (Cronbach, 1971). Validity is established by looking at the regression lines and using the score where they cross (if they do) as the cutting point for decisions. To get these data an experiment is needed in which assignment to treatment is random.

However, few instruments used for placement have been validated in this way. More commonly, either simple selection, which is called placement to conceal the inequity, or a multiple selection process is

used. An example of the first would be when a test is used to select students for algebra, and those not chosen are assigned to a course in typing simply on the grounds they cannot succeed in algebra. However, a better alternative is to use a test predicting success in typing also and to make placements based on the higher probability of success. Note that the two treatments require two different criterion measures, thus one has a double selection problem. Of course, all the difficulties in selection just discussed arise in double in addition to the problem of putting the two together.

Bias in Placement

Bias in placement occurs in some respects just as it does in selection. If the test measures different things for different groups, the regression lines will differ, probably for each treatment, and all the same sorts of judgments about fairness must be made. Also there are the same problems of sufficiency, of bias in the criterion measures, and of conflicting interests among the people concerned.

There are, however, some differences. On the positive side is the fact that the conflict in interests should be less since nobody is rejected. On the other hand, operational problems are multiplied. For example, what a test measures can be altered by a treatment and therefore the criterion measure may be biased for those having one treatment but not for those having the other. Similarly, since treatments may affect different groups differently, placement may have to proceed differently for one group than another to be fair. If the test did not measure the characteristic leading to this differential response to the treatment, the result could be unfair.

Cronbach (1975) recently noted that this sort of multiplication of interactions usually throws our generalizations into question, and this is an example. The test could measure the same things for different groups yet be unfair because it does not measure a trait that interacts with the treatment and that is not equally represented in the two groups. Thus introducing additional interactions into a test-criterion relationship can, but does not necessarily, create an exception to the general proposition that unfairness stems from tests that measure different things for different groups. Please note that the rule does hold in most placement situations and completely for both other uses. In particular, it holds by definition for description.

DESCRIPTION

The use of tests for description includes all those uses that do not directly lead to some differential treatment. There are many of these. For example, achievement tests are often given simply to track the progress of individuals or groups or to judge programs. The programs may be judged poor and be changed, but this subsequent course of action is usually not treated as the direct use of the test that needs validation. The inference that needs validation is the judgment about program quality. Another example would be a test that leads to a description of the interest patterns of a person, such as high or low in each of several areas without any bad-good connotations. Again a counselor might advise action on the basis of these scores but probably without any expectation of validity evidence for that advice as part of the test validation. The variety of descriptions that tests may yield is large since people differ in many complex ways.

Validity

Establishing validity for descriptive use is correspondingly complex. The traditional validity question is appropriate for description, i.e., does the test measure what it purports to measure? The effort is to determine what characteristics the test does measure. The principal procedure is that set of operations that provide evidence of construct validity, although for achievement tests, assessment of content validity is traditional. On occasion, criterion-related validity is considered appropriate and sufficient evidence of construct validity, but in general a more thorough demonstration is necessary for any test intended to be used descriptively including achievement tests (Messick, 1974). Construct validity procedures require trying to confirm theoretical inferences about variations in the traits that the test is intended to measure. It is necessarily a multistep process and in ordinary circumstances is never complete, i.e., there are always further inferences that could be checked. Content validity is determined by examining the adequacy of the coverage of the intended domain of tasks.

Bias in Description

For this use, unfairness is identical with bias, i.e., with not measuring the same thing. The unfairness occurs because the resulting description is more erroneous for one group than another. The scores do not mean the same thing for the different groups. This may not always be serious but it certainly can be. For example, a reading comprehension test may be biased because some group knows relatively little about the content of the passages; their test scores are more dependent on their knowledge than is the case for other groups who all are thoroughly familiar with the material. The result is an unfair description of the reading skills of the first group.

Although reading tests are usually validated by content procedures, a content validity analysis ordinarily cannot provide evidence of differential validity (it may show what I earlier called content bias). There is the possibility, perhaps remote, that different groups given the same test would describe its content in a systematically different way. This would be evidence of bias, but it would be more telling to show that empirically, since it is ultimately the responses to the test that count. In the reading comprehension example, one could demonstrate that the test questions were more passage-dependent for one group than another (e.g., Pyrczak, 1974; Tuinman, 1974). Evidence of bias can be seen better through construct validity studies of this sort than through the use of a content approach. In short, bias in description is best demonstrated by showing the test scores related to other variables in a differential fashion.

A concurrent relationship between the test and some other measures known to be valid can often be considered evidence that the test is measuring some or all of the traits intended. As before, different regression equations for different groups would indicate measurement of different things; if the regression lines have the same slopes, it is probably measuring mostly the same thing. Since somewhat different slopes may not mean very large differences, fairness has to be determined by finding out which components are the same and which are different.

So even when another valid measure is available, one usually has to undertake a long series of studies, each one of them comparing the relationships of the test scores to other variables for the different groups. Another major problem is how to determine whether the differences in relationships found in any such studies are large. In a study

just completed, three different procedures were used and yielded three different estimates of the amount of bias in a single test (Green & Roudabush, 1975). Only a small handful of tests have more than one study of bias, and it is, I believe, fair to say that establishing differential construct validity has proved too large a task for most tests.

The practical difficulties in obtaining adequate data for assessing differential construct validity have made examination of internal test structure a popular procedure. Many of the internal structure studies are directed toward detecting biased items rather than assessing test bias overall (e.g., Angoff & Ford, 1973; Cardall & Coffman, 1964; Cleary & Hilton, 1968; Green & Draper, 1972). Since these procedures begin with the assumption of overall validity for each group, they are primarily useful in test construction. There is some limited evidence indicating that they can help build less biased tests (CTB/McGraw-Hill, 1974). Use has been made of factor analyses of the items for the different groups in an attempt to identify score variance specific to the different groups (Green & Draper 1972). This line of attack is continuing and, although still in a developmental stage, it seems promising. Others are also working along these lines (Merz, 1973). Any approach dealing only with item data has limitations since even if a different internal structure can be proved it does not necessarily follow that the scores have different meaning. After all, Cole and his colleagues have shown that it can require different tests to measure the same thing for groups of children whose basic education occurred in very different cultural traditions (Cole, Gay, Glick, & Sharp, 1971).

Consequently, its popularity notwithstanding, examinations of the internal structure of tests have limited value in assessing bias. It may be argued by some that this is all that can be done, especially for achievement tests. That is not entirely true; achievement tests usually have some underlying theoretical constructs that can be investigated empirically in addition to those that apply to its internal structure (Cronbach, 1971; Messick, 1974).

Plainly these investigations can be conducted to permit inferences about differential validity. For example, one characteristic of a criterion-referenced test that is indicative of validity is the "sensitivity to instruction" shown by its scores. Since the test is designed to discriminate as sharply as possible between those who have mastered some concept or skill and those who have not, scores should change sharply given competent relevant instruction. If this change occurs more clearly in one group than the other, it may be that the test is biased. Of course, it may instead be the fault of the instruction, and some controls for that possibility should be included in the design. Such an approach is certainly rather clumsy and messy, but it can provide relevant data.

CONCLUSIONS

As a rule, when a test measures different things for different groups, it creates the liklihood of an unfair result and the test may be considered biased; a test measuring the same things for different groups is not biased. A biased test can be used fairly but to do that it must be used in different ways with each group; separate validity studies for

each group would be required. Also contrary to the rule is the placement case in which an unbiased test produces an unfair result because the treatments are biased. Both these departures from the rule, that a biased test is unfair and an unbiased test is fair, appear to be rare. However, they do point to the fact that unfairness appears in use.

To demonstrate that a test is not biased for any given use, it is sufficient to show that it is equally valid for different groups. An examination of criterion-related validity can indicate bias but it is not ordinarily sufficient to indicate lack of bias. An assessment of content validity will not indicate bias or lack of it. An adequate demonstration that a test is not biased almost always requires exploration of its construct validity regardless of use.

References

American Psychological Association. Technical recommendations for psychological tests and diagnostic techniques. Supplement to the Psychological Bulletin, 1954, 51 (2, Pt. 2).

American Psychological Association. Standards for educational and psychological tests and manuals. Washington, DC: Author, 1966.

American Psychological Association. Standards for educational and psychological tests. Washington, DC: Author, 1974.

Angolff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-106.

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

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

Cole, M., Gay, J., Glick, J. A., & Sharp, D. W. The cultural context of learning and thinking. New York: Basic Books, 1971.

Cole, N. S. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational Measurement. Washington, DC: American Council on Education, 1971.

Cronbach, L. J. Five decades of public controversy over mental testing. American Psychologist, 1975, 30, 1-14.

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

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

Dunfee, M. (Ed.) Eliminating ethnic bias in instructional materials: Comment and bibliography. Washington, DC: Association for Supervision and Curriculum Development, 1974.

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

Green, D. R. Racial and ethnic bias in achievement tests and what to do about it. Monterey, CA: CTB/McGraw-Hill, 1973.

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.

Green, D. R., & Roudabush, G. E. An investigation of bias in a criterion-referenced reading test. Paper presented at the meeting of the American Educational Research Association, Washington, DC, 1975.

Linn, R. L. Fair test use in selection. Review of Educational Research, 1973, 43, 139-161.

McGraw-Hill. Guidelines for equal treatment of the sexes in McGraw-Hill book company publications. New York: Author, 1974.

McGraw-Hill. Multiethnic publishing guidelines. New York: Author, 1973.

Merz, W. R. Factor analysis as a technique in analyzing test item bias. Paper presented at the meeting of The California Educational Research Association, Los Angeles, 1973.

Merz, W. R. A biased test may be fair, but then what does that really mean? Paper presented at the meeting of The California Educational Research Association, San Francisco, 1974.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. Research Bulletin 74-44. Princeton, NJ: Educational Testing Service, 1974.

Pyrczak, F. Passage-dependence of multiple-choice items designed to measure the ability to identify the main idea of a paragraph: Implications for validity. Educational and Psychological Measurement, 1974, 34, 343-348.

Sommer, J. Response to Robert Williams. The Counseling Psychologist, 1970, 2, 92.

Temp, G. Validity of the SAT for blacks and whites in thirteen integreted institutions. Journal of Educational Measurement, 1971, 8, 245-251.

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

Tuinman, J. J. Determining the passage dependency of comprehension questions in five major tests. Reading Research Quarterly, 1973-1974, 9, 206-223.

Williams, R. L. The BITCH Test. St. Louis: Author, 1972.