DOCUMENT RESUME

ED 080 601 TM 003 120

AUTHOR Darlington, Richard B.

TITLE Is Culture-Fairness Objective or Subjective?

PUB DATE 73

NOTE 8p.; Paper presented at symposium of annual meeting of American Educational Research Association (New Orleans, Lousiiana, February 25-March 1, 1973)

EDRS PRICE MF-$0.65 HC-$3.29

DESCRIPTORS *College Entrance Examinations; *Culture Free Tests; *Definitions; *Evaluation Criteria; Speeches; Tests; *Test Validity

ABSTRACT

The search for a satisfactory objective definition of a culture-fair test is doomed to failure, except in the special case in which different cultural groups have the same mean scores on the criterion variable to be predicted by the test. In the general case, it can be shown that no test (except one with the rare quality of perfect validity) can meet all the criteria reasonably expected of a "culture fair" test. The search for an objective definition of culture fairness must herefore be replaced by a subjective judgment of the degree of validity a tester is willing to sacrifice in order to select more or fewer members of certain cultural groups. (Author)

5.

Is culture-fairness objective or subjective?

Richard B. Darlington

Cornell University

Final draft of a talk delivered on February 26, 1973, in a symposium, Models of bias for using tests in selection, chaired by Dr. Gary Hanson, at the 1973 annual meeting of the American Educational Research Association, New Orleans, Lousiana.

My discussion today consists of three parts.

The first part is background, consisting of a brief summary of a paper I published in the Journal of Educational Measurement in 1971, arguing that there can be no generally-applicable objective statistical definition of a culture-fair test, and stating how I think the problem of cultural bias in tests should be handled.

The second part is a further defense of this position, with a friendly critique made particularly of the attempts of Dr. Cole of this panel and her collaborators to promote one particular o' jective definition.

The third and last part describes the type of further research and thought which I consider most likely to advance our attempts to make our tests be the servants rather than the masters of the highest goals of a multi-faceted society.

I shall talk particularly about culture-fairness in college admissions, though my general points are relevant to many other selection problems.

In this discussion, I shall assume that the tests we are considering have been constructed to predict a particular criterion variable, such as college grade-point average, and the problem is to consider whether they are culture-fair when used as predictors of that criterion. If different cultural groups score equally on that criterion, then there is no problem. The problem arises only when they do not, and that is the case I shall consider today.

Three different statistical definitions of culture-fairness explicitly take into consideration the possibility that cultural groups may differ on the criterion variable. In terms of our concrete example, they all consider the fact that in many colleges, for whatever reasons, an average white student is more likely to get high grades than an average black student.

The first definition defines a test as culture-fair if a black student and a white student with equal scores on the test are equally likely to succeed in college. To my knowledge, this has been the standard definition for at least twenty years.

The second definition defines a test a culture-fair if the test results in the selection of the same number of black and white applicants as would be selected by a test with perfect validity. This definition was suggested by 1971 by Robert Thorndike.

The third definition defines a test as culture-fair if a black student who is genuinely able to succeed in college has the same chance of being admitted as a white student who is equally able to succeed. This definition was considered at some length in my 1971 paper, and then rejected. However, in a paper entitled Bias in selection, which will appear later this year in the Journal of Educational Measurement, Mrs. Cole has defended it vigorously. Mrs. Cole calls this definition the "conditional probability" method.

An examination of these three definitions, which I shall call Definitions 1, 2, and 3, reveals a remarkable paradox: if the cultural groups differ on the criterion, then it can be shown mathematically that no test of less-than-perfect validity can meet more than one of the definitions. Any test which is culture-fair by Definition 2--and in the cases I have examined most tests are approximately fair by Definition 2--any such test is defined as biased against blacks by Definition 3, and biased against whites by Definition 1.

This paradox can be explained as follows. In the same way in which factor analysts consider a test to be the sum of a common-factor component and a specific component, we can consider any test to be composed of three components: a valid component which correlates perfectly with the criterion variable, a second component which is affected only by culture, and a third, error component which is purely random error, or which at least correlates zero with both the criterion variable and culture. The first two components, the criterion component and the cultural component, may correlate with each other, while the error component is defined as uncorrelated with the other two. If the cultural component does correlate with the criterion component, then it is a valid predictor of the criterion, and contributes to the validity of the test.

In these terms, the user of Definition 3 essentially says: even if the cultural component of the test is a valid component and contributes to the validity of the test, I refuse to use it. I will attempt to design a test in which the direct influence of this component is minimized. The user of Definition 2 says: I will use the cultural component to increase validity, but only up to a certain point. I will not allow the test to correlate higher with culture than the criterion itself does. The user of Definition 1 says: I will use the cultural component to whatever extent produces maximum validity.

All three definitions agree that if a black student is predicted to do just as well as a white student in college, the white should not be given preference over the black. A test which is fair by Definition 1 will consider the two students equally desirable candidates for admission; a test fair by Definition 2 will consider the black student preferable to the white, and a test which is fair by Definition 3 will favor the black student even more.

As mentioned above, I advocate a method which is different from any of these three, and which I shall call the corrected-criterion method. This method is suggested by going back to fundamentals, and asking: why does the term culture-fairness even exist? Why don't we just talk about fairness, regardless of culture? The very fact that the term culture-fairness is used, in place of simple fairness, implies that we are more concerned with some types of errors of prediction than with others. That is, if two students, one black and one white, are equally able to succeed in college, most college admissions officers would say that mistakenly rejecting the black student is a more serious error than rejecting the white student. In other words, grade point average, the thing that has generally been accepted as the criterion in the above discussion, is not the only criterion. Once it is admitted explicitly that culture-group membership, particularly race, is one of the criteria for college admissions, it is clear that the question of how much weight should be given to race must be decided subjectively, at a policy level, rather than by a testing technician. This is the essence of the corrected-criterion method. Its name derives from its explicit recognition that the variable which is usually accepted as the criterion, such as college grade-point average, is not the only criterion. Culture is also part of the criterion, and thus the traditional criterion must be modified or corrected for culture.

When the race of each applicant is known, the corrected-criterion method is very simple. It works as follows. The testing technician designs admissions batteries that will predict the traditional criterion accurately for each cultural group. The batteries may be the same for black students as for white, or they may be different. Once the batteries have been constructed and validated, the technician informs the admissions officers how likely each applicant is to succeed in college. It is then up to the admissions officers, together with other groups concerned with admissions but <u>not</u> the testing technicians, to decide how much weight to give to race in the final admissions process.

This kind of procedure is in fact followed routinely in many colleges with respect to a variety of variables. If a student is a swimming champion, or if he would make an unusual contribution to the extracurricular college orchestra, or if he has an unusual background in some way that would add to the life of the college, it is routine practice now in private colleges for admissions officers to give these factors some weight in the admissions process. It has never occurred to testing technicians to try to tell the admissions officers exactly how much weight they should give these factors. The corrected-criterion method does nothing more than handle cultural factors in the same way.

* * * * * * * * * * * * * * * * * *

This concludes a brief description of the corrected-criterion method, which makes the treatment of race in admissions a policy-level decision, and of Definitions 1, 2, and 3, which attempt to keep the treatment of race in the hands of the testing technicians. This brings us to the second part of this paper, which is a further defense of the corrected-criterion method against the alternative positions with special emphasis on Definition 3, which Dr. Cole calls the conditional probability model.

Before discussing in detail the relative merits of the different positions, we should see whether the practical differences among the methods are large enough to warrant careful discussion. In her <u>Bias in selection</u>, which I shall quote several times today, Dr. Cole has very sensibly pointed out that if the practical differences turn out to be small, then we shouldn't lose too much sleep in choosing among them.

The most interesting comparison is between Definitions 3 and 1. The corrected-criterion method can be equivalent to either 3 or 1, depending on the importance assigned to cultural factors in the admissions process. Definition 1 is the one which essentially recommends attempting to maximize test validity and thus select the best people, regardless of what race they turn out to be. I compared Definitions 1 and 3 using the data on undergraduate admissions published by Cleary in the <u>Journal of Education Measurement</u> for 1968. I found, as Cole had suggested might occur, relatively little difference between the two methods. However, I also compared the two methods using data on the prediction of first-year grade-point averages in five law schools, using data collected by Dr. Schrader and others at ETS. In these data I found very large differences between the two methods. The analysis was done separately for each law school. In all five law schools, a test defined as fair by Definition 3 considered a black applicant and a white applicant to be equally desirable candidates for admission when the predicted grade-point average of the black student was between 1 1/4 and 1 1/2 standard deviations below that of the white student. A white student predicted to score at the mean in law school grade-point average would be considered an equally desirable candidate for admission with a black student whose chance of scoring at or above the mean is between 1% and 12% in the different law schools, averaging less than 5%. We thus conclude that the differences among the methods are at least sometimes large enough to consider very seriously.

These data can also be used to consider another problem raised by Dr. Cole. She wrote that, when methods other than Method 1 are used, "the importance of the criterion may be overlooked." In the case just mentioned, the use of Definition 3 could very well lead to the worst possible combination of circumstances in this regard; people will feel that they don't need to worry about the criterion since they are told the formula takes care of it, and yet use of the formula in fact leads to a very substantial deviation from the admissions policy that would be followed if the criterion is considered paramount.

In the corrected-criterion method, on the other hand, it is hard to see how the importance of the criterion can be overlooked, since the test user must make an explicit judgment about it.

From the above discussion it is clear that either the corrected-criterion method or Definition 3 is open to the charge of favoritism, since either one will generally admit some black students whose predicted performance in college is below that of some white students who are rejected. The difference between the methods is in how they respond to this charge.

The user of the corrected-criterion method admits that the charge is true, and defends the favoritism on social and political grounds.

The user of Definition 3, on the other hand, uses a psychometric argument to deny that any favoritism is occurring. In other words, the user of the corrected-criterion method uses primarily a sociopolitical argument, while the user of Definition 3 uses primarily a psychometric.argument.

But how can you defend a procedure which does not attempt to select the best applicants on the grounds that it is "fairer" in some psychometric sense? After all, what is fairness except selecting the best people? And what is validity but the same thing? Thus if only psychometric arguments are used, how can a test which is less valid be called fairer? When confronted with competing psychometric arguments, we have to ask what the basic point of the whole testing enterprise.is, and the answer has to be that the basic point is validity. For example, if a psychometrician charges that our test is not as reliable or as factor-pure as his test, but admits that our test is more valid, then the proper response is that among competing psychometric considerations, validity is the name of the game. I think the same response must be given in the present instance.

I have said above that the user of Definition 3 defends his position primarily on psychometric grounds. Actually, the user of Definition 3 will also get involved in socio-political arguments. Dr. Cole advocates the use of Definition 3 primarily in situations in which test validity is not, in her words, the "primary concern." Thus, if I understand this phrase correctly, Dr. Cole would not advocate the use of Definition 3 in, say, a test for the certification of airline pilots, where validity is clearly the factor of primary concern, but she would advocate its use in undergraduate college admissions.

Thus the defenses of the corrected-criterion method and Definition 3 are not as different as I have been suggesting; each will involve a sociopolitical discussion about the relative importance, in any given instance, of test validity and cultural factors. Under Definition 3, however, there are three reasons why such a discussion is likely to be more confused and less satisfactory than under the corrected-criterion method.

First, the discussion is not purely sociopolitical, as it is in the corrected-criterion method; consideration of sociopolitical factors will inevitably by intermixed with a psychometric argument about the properties of Definition 3 and its appropriateness or inappropriateness for any situation. I can well imagine such a discussion; some people are talking psychometrics, some are talking politics, and nobody is understanding anybody else. I consider one of the great strengths of the corrected-criterion method to be that a discussion of sociopolitical factors is clearly separated from discussions of psychometric factors.

Second, this discussion must clearly involve the test users, and perhaps representatives of various political groups, since the testing technicians have no right to decide on the relative importance of test validity and cultural factors. But the nontechnical participants in the discussion will never understand the psychometric argument which swirls around the conflicts among Definitions 1, 2, and 3. Thus the people who must participate in the discussion are not equipped to do so.

Third, the discussion will be made more heated and acrimonious by the fact that the choice which must be made is a dichotomy; either cultural factors are given no weight at all in a given situation, and simple validity is maximized, or cultural factors are actually considered "primary", that is, outranking tast validity in importance. This is the kind of discussion that is going to leave the losers--either the advocates of test validity or the advocates of cultural factors--not merely unhappy but furious, since there is no possibility of compromise.

Putting the three problems together then, we have a discussion in which the issues are never clearly defined, in which many of the participants can't understand half the arguments, and which somebody has to lose since there is no room for compromise. It's hard to imagine a better recipe for destroying relations between the testing profession and the public.

The importance of this argument is multiplied by the fact that for each testing situation there is not just one "cultural" variable, race, to be considered. To name just a few, there are sex, height, socioeconomic status, religion, and homosexuality. Thus the testing technician who has just explained in psychometiic terms why it is "fair" to use Definition 3 with respect to race must either agree to treat each of these other "cultural" variables in the same way, with the potential that each additional cultural variable considered will lower validity still further, or else he must explain why a formula which is "fair" for blacks isn't fair for women, short people, poor people, Catholics, or homosexuals.

The user of the corrected-criterion method must also consider each of these variables; the advantage is that consideration of each variable isn't hampered by the unclarity of terms, lack of understanding, and impossibility of compromise described above.

Another problem arises when we consider the possibility that some minority cultural groups may score higher on some criterion variable than the WASP majority. Some psychologists have argued that there is evidence to show that Jews and Orientals are such groups. Whether they are or not, there is certainly the possibility that some minority group might someday be found which is above the majority. If Definition 3 were applied to this minority, we would have cases in which a minority applicant scores higher than a majority applicant on all the standard admissions criteria, and yet the majority applicant is accepted and the minority applicant rejected.

In summary, the use of Definition 3 may sometimes lead to a serious loss of test validity: the defense of Definition 3 has been phrased in such a way as to lead users to ignore that loss in the mistaken impression that it has been considered by a formula; discussions of its appropriateness for particular testing situations are almost certain to end in disaster· and additional awkward problems arise when it is applied to cultural variables other than the black-white dichotomy.

* * * * * * * * * * * * * * * * * *

Having spent the second part of my talk disputing one of the lines of work of the American College Testing Program, I shall spend the third and last part encouraging them in another line of work which they and others have begun, and which I feel is far more promising. This is a further investigation of the criterion problem.

One of my students, Robert Billings, has become interested in this problem, and what I shall say in this section is almost entirely a summary of his very persuasive arguments.

Billings accepts my emphasis on validity,and argues that any deviation from maximum correlation with an observed criterion must be justified on the grounds that the observed criterion is not the ultimate criterion. In other words, he argues that the criterion problem is not one aspect of the problem, it is the whole problem.

Billings accepts the corrected-criterion method for some cases. He argues, however, that there are many cases--and college admissions is probably one of them--in which the ultimate criterion is not a simple weighted average of an observed criterion and race, but is a hypothetical construct which cannot be measured directly. Thus our tests must be validated not by standard correlational validation studies, but by the more complex techniques involved in construct validation.

Perhaps I am biased toward this approach, because it turns the culture fairness problem into a special case of a problem I have been thinking about for almost ten years--the problem of how to construct a test with maximum construct validity. I described several techniques for doing this in a paper in the Journal of Educational Measurement for 1970. Since I must finish in one minute, I will not describe these techniques. I w_ll merely illustrate them by mentioning one possible set of assumptions and the conclusions which, according to the techniques, follow from those assumptions.

First, suppose we assume that whatever we call the ultimate criterion, there is no difference between blacks and whites on it. This may be because we feel that the ultimate criterion is some sort of ability or intelligence, and we are willing to assume, contrary to Professor Jensen, that the races do not differ on this trait. Or it may be because we feel the ultimate criterion is the increase in a student's potential contribution which would result from attending college, which we might assume is as great for blacks as for whites. Or the same conclusion may be reached from other conceptions of the ultimate criterion.

Second, suppose we assume that whatever we call the ultimate criterion, measures can be constructed which correlate just as highly with it within the black population as other measures correlate with it within the white population.

From these two assumptions it follows directly, using the techniques in the paper just mentioned, that maximization of test validity, as measured against this ultimate criterion, is achieved by selecting the same proportion of blacks as whites. This would lead to the selection of many more black students than would be selected even under the supposedly liberal Definition 3 if it started by using grade-point average as the criterion.

Other conclusions follow from other assumptions. Thus our future work should concentrate not on the purely psychometric aspects of the problem, but on the substantive problem of what the criterion really is.