Psychological Bulletin
1976, Vol. 83, No. 6, 1053-1071

Critical Analysis of the Statistical and Ethical Implications of Various Definitions of *Test Bias*

John E. Hunter and Frank L. Schmidt
Michigan State University

This article defines three mutually incompatible ethical positions in regard to the fair and unbiased use of psychological tests for different groups such as blacks and whites. Five statistical definitions of *test bias* are also reviewed and are related to the three ethical positions. Each definition is critically examined for its weaknesses on either technical or social grounds. We ultimately argue that the attempt to define *fair use* without recourse to substantive and causal analysis is doomed to failure.

In the last several years there has been a series of articles devoted to the question of the fairness of employment and educational tests to minority groups (Cleary, 1968; Darlington, 1971; Thorndike, 1971). Although each of these articles came to an ethical conclusion, the basis for that ethical judgment was left unclear. If there were only one ethically defensible position, then this would pose no problem. But such is not the case. The articles that we review have a second common feature. Each writer attempts to establish a definition on purely statistical grounds, that is, on a basis that is independent of the content of test and criterion and that makes no explicit assumption about the causal explanation of the statistical relations found. We argue that this merely makes the substantive considerations implicit rather than explicit.

In this article we first describe three distinct ethical positions. We next examine five statistical definitions of test fairness in detail and show how each is based on one of these ethical positions. Finally, we examine the technical, social, and legal advantages and disadvantages of the various ethical positions and statistical definitions.

Frank L. Schmidt is now at the U.S. Civil Service Commission.

Requests for reprints should be sent to John E. Hunter, Department of Psychology, Olds Hall, Michigan State University, East Lansing, Michigan 48823.

Three Ethical Positions

Unqualified Individualism

The classic American definition of an objective advancement policy is giving the job to the person "best qualified to serve." Couched in the language of institutional selection procedures, this means that an organization should use whatever information it possesses to make a scientifically valid prediction of each individual's performance and always select those with the highest predicted performance. From this point of view, there are two ways in which an institution can act unethically. First, an institution may knowingly fail to use an available, more valid predictor; for example, it may select on the basis of appearance rather than scores on a valid ability test. Second, it may knowingly fail to use a more valid prediction equation based on its available information; for example, it may administer a more difficult literacy test to blacks than to whites and then use a cutoff score for both groups that assumes they both took the same test. In particular, if in fact race, sex, or ethnic group membership were a valid predictor of performance in a given situation over and above the effects of other measured variables, then the unqualified individualist would be ethically bound to use such a predictor.

Quotas

Most corporations and educational institutions are creatures of the state or city in

which they function. Thus, it has been argued that they are ethically bound to act in a way that is "politically appropriate" to their location. In particular, in a city whose population is 45% black and 55% white, any selection procedure that admits any other ratio of blacks and whites is "politically biased" against one group or the other. That is, any politically well defined group has the "right" to ask and receive its "fair share" of any desirable product or position that is under state control. These fair share quotas may be based on population percentages or on other factors irrelevant to the predicted future performance of the selectees (Darlington, 1971; Thorndike, 1971).

Qualified Individualism

There is one variant of individualism that deserves separate discussion. This position notes that America is constitutionally opposed to discrimination on the basis of race, religion, national origin, or sex. A qualified individualist interprets this as an ethical imperative to refuse to use race, sex, and so on, as a predictor even if it were in fact scientifically valid to do so. Suppose, for example, that race were a valid predictor of some criterion, that is, assume that the mean difference between the races on the criterion is greater than that that would be predicted on the basis of the best ability test available. This would mean that the use of race in conjunction with the ability test would increase the multiple correlation with the criterion. That is, prediction would be better if separate regression lines were used for blacks and whites. To the unqualified individualist, on the other hand, failure to use race as a predictor would be unethical and discriminatory, since it would result in a less accurate prediction of the future performance of applicants and would "penalize" or underpredict performance of individuals from one of the applicant groups. The qualified individualist recognizes this fact but is ethically bound to use one overall regression line for ability and to ignore race. Thus, the qualified individualist relies *solely* on measures of ability and motivation to perform the job (e.g., scores on valid aptitude and achievement tests, assessment of past work experiences, etc.).

Definition of Discrimination

There is one very important point to be made before leaving this issue: The word *discriminate* is *not* ambiguous. The qualified individualist interprets the word *discriminate* to mean *treat differentially*. Thus, he will not treat blacks and whites differently even if it is statistically warranted. However, the unqualified individualist also refuses to discriminate, but he uses a different definition of that word. The unqualified individualist interprets *discriminate* to mean *treat unfairly*. Thus, the unqualified individualist would say that if there is in fact a valid difference between the races that is not accounted for by available ability tests, then to refuse to recognize this difference is to penalize the higher performing of the two groups. Finally, the person who adheres to quotas will also refuse to discriminate, but he will use yet a third definition of that word. The person who endorses quotas interprets *discriminate* to mean *select a higher proportion of persons from one group than from the other group*. Thus, the adherents of all three ethical positions accept a constitutional ban against discrimination, but they differ in their views of how that ban is to be put into effect.

Three Attempts to Define *Test Fairness* Statistically

In this section we briefly review three attempts to arrive at a strictly statistical criterion for a fair or unbiased test. For ease of presentation, the discussion uses comparison of blacks and whites. However, the reader should bear in mind that other demographic classifications, such as social class or sex, could be substituted with no loss of generality.

The Cleary Definition

Cleary (1968) defined a test to be *unbiased* only if the regression lines for blacks and whites are identical. The reason for this is brought out in Figure 1, which shows a hypothetical case in which the regression line for blacks lies above the line for whites and is parallel to it. Consider a white and a black subject, each of whom have a score of A on the test. If the white regression line were used to predict both criterion scores, then

the black applicant would be underpredicted by an amount Δy, the difference between his expected score making use of the fact he is black and the expected score assigned by the white regression line. Actually, in this situation in order for a white subject to have the same expected performance as a black whose score is A, the white subject must have a score of B.

That is, if the white regression line underpredicts black performance, then a white and black are only truly equal in their expected performance if the white's test score is higher than the black's by an amount related to the amount of underprediction. Similarly, if the white regression line always overpredicts black performance, then a black subject has equal expected performance only if his test score is higher than the corresponding white subject's score by an amount related to the amount of overprediction. Thus, if the regression lines for blacks and whites are not equal, then each person will receive a statistically valid predicted criterion score only if separate regression lines are used for the two races. If the two regression lines have exactly the same slope, then this can be accomplished by predicting performance from two separate regression equations or from a multiple regression equation with test score and race as the predictors. If the slopes are not equal, then either separate equations must be used or the multiple regression equation must be expanded by the usual product term for moderator variables. Thus, we can view Cleary's definition of an unbiased test as an attempt to rule out disputes between qualified and unqualified individualism.

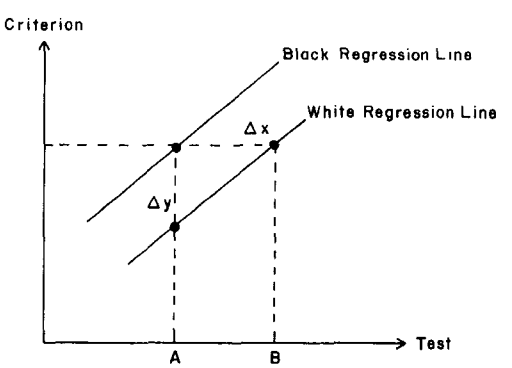

FIGURE 1. A case in which the white regression line underpredicts black performance.

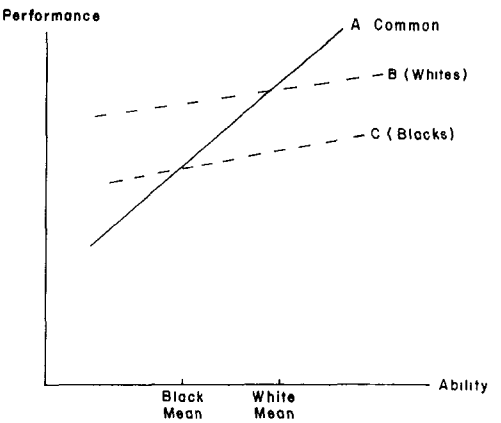

FIGURE 2. Regression artifacts produced by unreliability in a Cleary-defined unbiased test. (A is the common regression line for a perfectly reliable ability test. B and C are the regression lines for whites and blacks, respectively, for a test of reliability .50.)

If the predictors available to an institution are unbiased in Cleary's sense, then the question of whether to use race as a predictor does not arise. But if the predictors are *biased,* the recommended use of separate regression lines is clearly equivalent to using race as a predictor of performance. Thus, although Cleary may show a preference for tests that meet the requirements of both unqualified and qualified individualism, in the final analysis, her position is one of unqualified individualism.

A Cleary-defined unbiased test is ethically acceptable to those who advocate quotas only under very special circumstances. In addition to identical regression lines, blacks and whites must have equal means and equal standard deviations on the test, and this in turn implies equal means and standard deviations on the performance measure. Furthermore, the proportion of black and white applicants must be the same as their proportion in the relevant population. These are conditions that rarely occur.

Linn and Werts (1971) have pointed out an additional problem for Cleary's definition —the problem of defining the fairness when using less than perfectly reliable tests. Suppose that a perfectly reliable measure of intelligence were in fact an unbiased predictor

in Cleary's sense. But because perfect reliability is unattainable in practice, the test used in practice will contain a certain amount of error variance. Will the imperfect test be unbiased in terms of the regression equations for blacks and whites? If black applicants have lower mean IQs than white applicants, then the regression lines for the imperfect test will *not* be equal. This situation is illustrated in Figure 2. In this figure we see that if an unreliable test is used, then that test produces the double regression line of a biased test in which the white regression line overpredicts black performance. That is, by Cleary's definition, the unreliable test is biased against whites in favor of blacks.[1, 2]

Cleary's critics question whether the failure to attain perfect reliability (impossible under any circumstances) should be adequate grounds for labeling a test as biased. But suppose we first consider this question from a different viewpoint. Suppose there were only one ethnic group, whites, for example. Assume that Bill has a true ability level of 115 and Jack has an ability of 110. If ability is a valid predictor of performance in this situation, then Bill has the higher expected performance; and if a perfectly reliable test is used, Bill will invariably be admitted ahead of Jack. But suppose that the reliability of the ability test is only .50. Then the two obtained scores will each vary randomly from their true values, and there is some probability that Bill's will be randomly low while Jack's is randomly high—that is, some probability that Jack will be admitted ahead of Bill. If the standard deviation of the observed ability scores is 15, then the difference between their observed scores has a mean of 5 and a standard deviation of 21. The probability of a negative difference is then .41. Thus, the probability that Bill is admitted ahead of Jack drops from 1.00 to .59.

The unreliable test is in fact sharply biased against better qualified applicants. This bias, however, is not directly racial or cultural in nature. It takes on the appearance of a racial bias only because the proportion of better qualified applicants is higher in the white group. Thus, the bias created by random error works against more applicants in the white group, and thus, on balance, the test is biased against that group as a whole. But at the individual level, such a test is no more biased against a well-qualified white than a well-qualified black. The question then is whether Cleary's (1968) definition is defective in some sense in labeling this situation as biased. If so, it may perhaps be desirable to modify the definition to apply only to bias beyond that expected on the basis of test reliability alone.

While on the topic of reliability, we should

---

[1] This phenomenon would account for perhaps half of the overprediction of black grade-point average in the literature. In standard score units, the difference in intercepts due to unreliability is $\Delta Y = (1 - r_{XX}) \times (\mu_w - \mu_B)$, where r_{XX} is the test reliability and $\mu_W - \mu_B$ is the white–black mean difference on the criterion (about 1 *SD*). For $r_{XX} = .80$, this would be only .2 *SD*, whereas in the data reported in Linn (1973), the overprediction is about .37 *SD*.

[2] The reader may wonder why we show so much concern with the reliability of the test and no concern with the reliability of the criterion. Actually, despite its large effect on the validity coefficient, no amount of unreliability in the criterion has any effect on the regression line of criterion on predictor. Let the true score equations for X and Y be $X = T + e_1$ and $Y = U + e_2$, and let the regression true score equation be $U = \alpha T + \beta$. Then the observed regression line will not have the same coefficients. Let the observed regression line be $Y = aX + b$. The slope of the observed regression line will be

$$a = r_{XY}\frac{\sigma_Y}{\sigma_X} = (r_{TU}r_{TX}r_{UY})\frac{\sigma_Y}{\sigma_X} = r_{TU}\frac{\sigma_T}{\sigma_X}\frac{\sigma_U}{\sigma_Y}\frac{\sigma_Y}{\sigma_X}$$

$$= r_{TU}\frac{\sigma_U}{\sigma_T}\frac{\sigma_T}{\sigma_X}\frac{\sigma_T}{\sigma_X} = \left(r_{TU}\frac{\sigma_U}{\sigma_T}\right)\left(\frac{\sigma_T^2}{\sigma_X^2}\right)$$

$$= \alpha r_{XX} .$$

That is, the slope of the observed regression line is the slope of the true score regression line multiplied by the reliability of X. However, note that the slope of the observed regression line is completely independent of the reliability of Y. The intercept of the observed regression line is given by:

$$b = \mu_Y - a\mu_X = \mu_U - a\mu_T = \mu_U - r_{XX}\alpha\mu_T .$$

Thus, the intercept is also affected by the reliability of X, but is completely independent of the reliability of Y. Since the slopes of the true score regression equations are equal (assuming equal standard deviations, as we have in this article) any differences in the regression lines will be equal to the difference between the intercepts and hence independent of r_{YY}. In the case in which the true score regression lines are the same, the difference between the observed regression lines is $b_W - b_B = (1 - r_{XX})(\mu_{UW} - \mu_{UB})$.

note that as the reliability approaches .00, the test becomes a random selection device and is hence utterly reprehensible to an individualist of either stripe. On the other hand, a totally unreliable test would select blacks in proportion to the number of black applicants and hence might well select in proportion to population quotas. Ironically, the argument that tests are biased against blacks because they are unreliable is not only false, it is exactly opposite to the truth.

Let us consider in detail the comparison of whites and blacks on the unreliable test. We first remark that it is a fact that on the average, whites with a given score have a higher mean performance than do blacks who have that same score. Thus, the use of a single regression line will in fact mean that whites near the cutoff will be denied admission in favor of blacks who will, on the average, not perform as well. Cleary's (1968) definition would clearly label such a situation biased. Furthermore, in this situation the partial correlation between race and performance with observed ability held constant is not zero. Thus, race makes a contribution to the multiple regression because with an unreliable ability test, race is in fact a valid predictor of performance after ability is partialed out. That is, from the point of view of unqualified individualism, the failure to use race as a second predictor is unethical. If the test is used with only one regression line, then the predictors are in fact biased against whites. If two regression lines are used, then each person is being considered solely on the basis of expected performance.

Thus, in summary, we feel that Cleary's critics have raised a false issue. To use an unreliable predictor is to blur valid differences between applicants, and an unreliable test is thus, to the extent of the unreliability, biased against people or groups of people who have high true scores on the predictor. Thus, from the point of view of an unqualified individualist, an unreliable test is indeed biased. On the other hand, a qualified individualist would object to this conclusion. Use of separate regression lines is statistically optimal because the unreliable test does not account for all the real differences on the true scores. But the qualified individualist is ethically prohibited from using race as a predictor and therefore can employ only a single regression equation. He can, however, console himself with the fact that the bias in the test is not specifically racial in nature. And, of course, he can attempt to raise the reliability of the test.

Thorndike's Definition

Thorndike (1971) began his discussion with the simplifying assumption that the slope of the two regression lines is equal. There are then three cases. If the regression lines are identical, then the test satisfies Cleary's (1968) definition. If the regression line for blacks is higher than that for whites (as in Figure 1), then Thorndike labels the test "obviously unfair to the minor group." On the other hand, if the regression line for whites is higher than that for blacks, then he does ***not*** label the test as obviously unfair to whites. Instead, he has an extended argument that the use of two regression lines would be unfair to blacks. Clearly there must be an inconsistency in his argument and indeed we ultimately show this.

Thorndike noticed that whereas using two regression lines is the only ethical solution from the point of view of unqualified individualism, it need not be required by an ethics of quotas. In particular, if the black regression line is lower, then blacks will normally show a lower mean on both predictor and criterion. Suppose that blacks are one standard deviation lower on both and that validity is .50 for both groups. If we knew the actual criterion scores and set the cutoff at the white mean on the criterion, then 50% of the whites and 16% of the blacks would be selected. If the white regression line were used for both groups, then 50% of the whites and 16% of the blacks would be selected. However, if blacks are selected using the black regression line, then because the black regression line lies .5 standard deviations below the white regression line, blacks will have to have a larger score to be selected (i.e., a cutoff two sigmas above the black mean instead of one), and hence fewer blacks will be selected. Thus, if the predictor score is used with two regression lines, then 50% of the whites but only 2% of the blacks

will be admitted. Thorndike argued that this is unfair to blacks as a group. He then recommended that we throw out individualism as an ethical imperative and replace it with a specific kind of quota. The quota that he defined as the fair share for each group is the percentage of that group that would have been selected had the criterion itself been used or had the test had perfect validity. In the above situation, for example, Thorndike's definition would consider the selection procedure fair only if 16% of the black applicants were selected.

What Thorndike has rediscovered has long been known to biologists: Bayes's law is cruel. If one of two equally reproductive species has a probability of .49 for survival to reproduce and the other species has a probability of .50, then ultimately the first species will be extinct. Maximization in probabilistic situations is usually much more extreme than most individuals expect (Edwards & Phillips, 1964).

What then was Thorndike's contradiction? He labeled the case in which the black regression line was higher than the white line as "obviously unfair to the minor group." But his basis for this was presumably unqualified individualism. In effect, he said that if blacks perform higher than whites over and above the effects of measured ability, then this fact should be recognized and blacks should have a correspondingly higher probability of being selected. That is, if the black regression line is higher, then separate regression lines should be used. But if separate regression lines are used, then the number of whites selected would ordinarily be drastically reduced. In fact, the number of whites selected would be far below the Thorndike-defined fair share of slots. The mathematics of unequal regression lines (i.e., of Bayes's Law) is the same for a high black curve as for a high white curve: The use of a single regression line lowers the validity of the prediction but tends to yield selection quotas that are much closer to the quotas that would have resulted from clairvoyance (i.e., much closer to the selection quotas that would have resulted had a perfectly valid test been available). Thus, Thorndike's inconsistency lies in his failure to apply his definition to the case in which the white regression line underpredicts black performance. The fact that because of a technicality (i.e., racial equality on the performance measure), this effect would not manifest itself in Thorndike's (1971) Case 1 should not be allowed to obscure this general principle.

If only one regression line is to be used, then a test will meet the Thorndike quotas only if the mean difference between the groups in standard score units is the same for both predictor and criterion. For this to hold for a Cleary-defined unbiased test, the validity of the test must be 1.00—a heretofore unattainable outcome.

Once Thorndike's position is shown to be a form of quota setting, then the obvious question is, Why his quotas? After all, the statement that 16% of the blacks can perform at the required level would not apply to the blacks actually selected and is in that sense irrelevant. In any event, it seems highly unlikely that this method of setting quotas would find support among those adherents of quotas who focus on population proportions as the proper basis of quota determination. Thorndikean quotas will generally be smaller than population-based quotas. On the other hand, Thorndike-determined quotas may have considerable appeal to large numbers of Americans as a compromise between the requirements of individualism and the social need to upgrade the employment levels of minority group members.

There is another question that must be raised concerning Thorndike's position: Is it ethically compatible with the use of imperfect selection devices? We will show that Thorndike's selection rule is contradictory to present test usage, that is, that according to Thorndike we must fill N vacancies not by taking the top N applicants but by making a much more complicated selection. Consider the following example: Assume that one is using a test score of 50 ($\bar{X} = 50$, $SD = 10$, $r_{xy} = .50$) as a cutoff and that the data show that 50% of those with a test score of 50 will be successful. Applicants with a score of 49 would all be rejected, though for $r_{xy} = .50$ about 48% of them would have succeeded had they been selected (a score of 49 is $-.1$ sigmas on X and implies an average score

of −.05 on the criterion with a sigma of .87). Thus, applicants with a test score of 49 can then correctly state, If we were all admitted, then 48% of us would succeed. Therefore, according to Thorndike, 48% of us should be admitted. Yet we were all denied. Thus, you have been unfair to our group, those people with scores of 49 on the test. That is, strictly speaking Thorndike's ethical position precludes the use of any predictor cutoff in selection, no matter how reasonably determined. Instead, from each predictor category one must select that percentage that would fall above the criterion cutoff if the test were perfectly valid. For example, if one wanted to select 50% of the applicants and the validity were .60, then one would have to take 77% of those who lie 1 *SD* above the mean, 50% of those within 1 *SD* of the mean, and 23% of those who fall 1 *SD* below the mean. And Thorndike's definition could be interpreted, of course, as requiring the use of even smaller intervals of test scores.

There are several problems with this procedure. First, one must attempt to explain to applicants with objectively higher qualifications why they were not admitted—a rather difficult task and from the point of view of individualism, an unethical one. Second, the general level of performance will be considerably lower than it would have been had the usual cutoff been used. In the previous example, the mean performance of the top 50% on the predictor would be .48 standard score units, whereas the mean performance of those selected by the Thorndike ethic would be .29. That is, in this example, using Thorndike's quotas has the effect of cutting the usefulness of the predictor by about 60%. (These calculations are shown in the appendix.)

One possible reply to this criticism would be that Thorndike's definition need not be interpreted as requiring application to all definable groups. The definition is to be applied only to "legitimate minority groups," and this would exclude groups defined solely by obtained score on the predictor. If agreement could be reached that, for example, blacks, Chicanos, and Indians are the only recognized minority groups, the definition might be workable. But such an agreement is highly unlikely. On what grounds could we fairly exclude Polish, Italian, and Greek Americans, for example?

Perhaps an even more telling criticism can be made. In a college or university, performance below a certain level means a bitter tragedy for a student. In an employment situation, job failure can often be equally damaging to self-esteem. In the selection situation described above, the percentage of failure would be 25% if the top half were admitted, but one third if a Thorndikean admission rule were used. Furthermore, most of the increase in failures comes precisely from the poor-risk admissions. Their failure rate is two thirds. Thus, in the end, a Thorndikean rule may be even more unfair to those at the bottom than to those at the top.

Darlington's Definition

Darlington's (1971) first step was a restatement of the Cleary (1968) and Thorndike (1971) criteria for a "culturally" fair test in terms of correlation rather than regression. Let X be the predictor, Y the criterion, and C the indicator variable for *culture* (i.e., $C = 1$ for whites, $C = 0$ for blacks). He made the empirically plausible assumption that the groups have equal standard deviations on both predictor and criterion and that the validity of the predictor is the same for both groups (hence parallel regression lines). Darlington then correctly noted that Cleary's (1968) criterion for a fair test could be stated,

$$r_{CY \cdot X} = 0 .$$

That is, there is no criterion difference between the races beyond that produced by their difference on X (if any). If all people are selected using a single regression line, then Thorndikean quotas are guaranteed by Darlington's "definition two," that is,

$$r_{CX} = r_{CY} .$$

That is, the racial difference on the predictor must equal the racial difference on the criterion in standard score units. However, if people are selected using multiple regression or separate regression lines, then this equa-

tion would not be correct. Instead there are two alternate conditions:

$$R_{Y \cdot CX} = 1$$

or

$$r_{CY} = 0 .$$

That is, if *separate* regression lines are used, then the percentages selected match Thorndike's quotas only if the test has perfect validity or if there are no differences between the groups on the criterion.[3]

Darlington then attacked the Cleary definition on two very questionable bases: (a) the reliability problem raised by Linn and Werts (1971), which was discussed in depth above, and (b) the contention that race itself would be a Cleary-defined fair test. Actually, if race were taken as the test, then there would be no within-groups variance on that predictor and hence no regression lines to compare. Thus, Cleary's definition cannot be applied to the case in which race itself is used as the predictor test.[4] The nontrivial equivalent of this is a test whose sole contribution to predicting Y is the race difference on the mean of X, but for such a test the regression lines are perfectly horizontal and grossly discrepant. That is, in a real situation, Cleary's definition would rule that a purely racial test is biased.

Darlington's Definition 3 and Cole's Argument

Darlington (1971) proposed a third definition of test fairness, his Definition 3. This definition did not attract a great deal of attention until Cole (1973) offered a persuasive argument in its favor. We first present Darlington's definition, his justification of it, and our critique of that justification. We then consider Cole's argument.

If X is the test and Y is the criterion and if C, the variable of culture, is scored 0 for blacks, 1 for whites, then Darlington's Definition 3 can be written as follows: The test is fair if

$$r_{XC \cdot Y} = 0 .$$

His argument for this definition went as follows: The ability to perform well on the criterion is a composite of many abilities, as is the ability to do well on the test. If the partial correlation between test and race with the criterion partialed out is not zero, then

---

[3] Because the groups have equal standard deviations on both predictor and criterion, assume for algebraic simplicity that the variables have been scaled so that all within-groups standard deviations are unity. This means that deviation scores are standard scores. Suppose that the selection ratio for whites has been determined. Then there is a corresponding standard score on Y, say Y^*, such that the standard score $Y^* - \bar{Y}_W$ would cut off that percentage of whites. To select that same percentage of whites, there is a predictor score on the test, X_W, such that

$$X^*_W - \bar{X}_W = Y^* - \bar{Y}_W .$$

If the multiple regression equation is

$$\hat{Y} = \alpha X + \beta C + \gamma ,$$

then the multiple regression cutoff score is

$$\begin{aligned} \hat{Y}^* &= \alpha X^*_W + \beta + \gamma \\ &= \alpha(X^*_W - \bar{X}_W + \bar{X}_W) + \beta + \gamma \\ &= \alpha(X^*_W - \bar{X}_W) + \alpha\bar{X}_W + \beta + \gamma . \end{aligned}$$

Because multiple regression always matches the group mean perfectly,

$$\bar{Y}_W = \alpha\bar{X}_W + \beta + \gamma ,$$

and hence

$$\hat{Y}^* = \alpha(X^*_W - \bar{X}_W) + \bar{Y}_W .$$

The predictor cutoff score for blacks is determined by

$$\begin{aligned} \hat{Y}^* &= \alpha X^*_B + \gamma \\ &= \alpha(X^*_B - \bar{X}_B + \bar{X}_B) + \gamma \\ &= \alpha(X^*_B - \bar{X}_B) + \alpha\bar{X}_B + \gamma . \end{aligned}$$

Because multiple regression matches means

$$\bar{Y}_B = \alpha\bar{X}_B + \gamma$$

and hence the black predictor cutoff satisfies

$$\hat{Y}^* = \alpha(X^*_B - \bar{X}_B) + \bar{Y}_B ,$$

Thorndike's quotas for blacks are obtained if the standard score for the predictor cutoff is the same as the standard score for the criterion cutoff, that is, if

$$X^*_B - \bar{X}_B = Y^* - \bar{Y}_B .$$

Now we have in general

$$\alpha(X^*_B - \bar{X}_B) = \hat{Y}^* - \bar{Y}_B .$$

Thus, Thorndike's quotas obtain only if

$$\begin{aligned} \alpha(Y^* - \bar{Y}_B) &= \hat{Y}^* - \bar{Y}_B \\ &= [\alpha(X^*_W - \bar{X}_W) + \bar{Y}_W] - \bar{Y}_B \\ &= \alpha(Y^* - \bar{Y}_W) + \bar{Y}_W - \bar{Y}_B . \end{aligned}$$

This is true only if

$$\alpha(\bar{Y}_W - \bar{Y}_B) = \bar{Y}_W - \bar{Y}_B .$$

it means that there is a larger difference between the races on the test than would be predicted by their difference on the criterion. Hence the test must be tapping abilities that are not relevant to the criterion but on which there are racial differences. Thus, the test is discriminatory.

Note that Darlington's argument makes use of assumptions about causal inference. If those assumptions about causality are in fact false, then his interpretation of the meaning of the partial correlation is no longer valid. Are his assumptions so plausible that they need not be backed up with evidence? Consider the time ordering of his argument. He is partialing the criterion from the predictor. In the case of college admissions, this means that he is calculating the correlation between race and entrance exam score, with GPA 4 years later being held constant. This is looking at the causal influence of the future on the past and is only valid in the context of very special theoretical assumptions. The definition would in fact be inappropriate even in the context of a concurrent validation study, since concurrent validities are typically derived only as convenient estimates of predictive validity. Thus, even when there is no time lag between predictor and criterion measurement, one is operating implicitly within the predictive validity model.

Let us explore this point more fully through the use of two concrete examples. First, consider a pro football coach attempting to evaluate the rookies who have joined the team as a result of the college draft. Since the players have all come from different schools, there are great differences in the kind and quality of training that they have had. Therefore, the coach cannot rely on how well they play their positions at that moment in time; they will undergo considerable change as they learn the ropes over the next few months. What the coach would like to know is exactly what their athletic abilities are without reference to the way they've learned to play to date. Suppose he decides to rely solely on the 40-yard dash as an indicator of football ability, that is, as a selection test. It is possible that he would then find that he was selecting a much larger percentage of blacks than he had using his judgment of current performance. Would this mean that the test is discriminatory against whites? That depends on the explanation for this outcome. Consider the defensive lineman on a passing play. His ability to reach the quarterback before he throws the ball depends not only on the speed necessary to go around the offensive lineman opposing him but also on sufficient arm strength to throw the offensive lineman to one side. Assume, for the sake of this example, that blacks are faster, on the average, than whites but that there are no racial differences in upper body strength. Since the 40-yard dash represents only speed and makes no measure of upper body strength, it cannot meet Darlington's substantive assumptions. That is, the 40-yard dash taps only the abilities on which there are racial differences and does not assess those that show no such differences.

How does the 40-yard dash behave statistically? If speed and upper body strength were the only factors in football ability and if the 40-yard dash were a perfect index of speed, then the correlations would satisfy $r_{YO \cdot X} = 0$. That is, by Cleary's definition, the 40-yard dash would be an unbiased test. Since $r_{YO \cdot X} = 0$, $r_{XO \cdot Y}$ cannot be zero and, hence, according to Darlington's definition the 40-yard dash is culturally unfair (i.e., biased against whites). (Since the number of whites selected would be less than the Thorndike quota, Thorndike too would call the test biased.) If the coach was aware that upper body strength was a key variable and was deliberately avoiding the use of a measure of upper body strength in a multiple regression equation, then the charge that the coach was deliberately selecting blacks would seem quite

---

Thus, Thorndike's quotas are obtained only if one of two things is true: either $\alpha = 1$ or both sides are zero; that is, either $\alpha = 1$ or $\bar{Y}_W - \bar{Y}_B = 0$.

Because the variables were all scaled to have equal within-groups standard deviations, the regression weight α is in fact the within-groups predictor–criterion correlation. Thus, $\alpha = 1$ means that the test has perfect validity.

The equation $\bar{Y}_W - \bar{Y}_B = 0$ is equivalent to $\bar{Y}_W = \bar{Y}_B$, that is, no group difference on the criterion and hence $r_{CY} = 0$.

[4] Darlington's error was subtle. He assumed that $r_{CY \cdot C} = 0$ when in fact $r_{CY \cdot C} = 0/0$, which is undefined.

reasonable. But suppose that the nature of the missing predictor (i.e., upper body strength) was completely unknown. Would it then be fair to charge the coach with using an unfair test?

At this point, we should note a related issue raised by Linn and Werts (1971). They too considered the case in which the criterion is affected by more than one ability, one of which is not assessed by the test. If the test assessed only verbal ability and the only racial differences were on verbal ability, then the situation would be like that described in the preceding paragraph: The test would be unbiased by the Cleary definition, but unfair according to Darlington's Definition 3. However, if there are also racial differences on the unmeasured ability, then the test will not be unbiased by Cleary's definition. For example, if blacks were also lower, on the average, in numerical ability and numerical ability was not assessed by the entrance test, then the black regression line and the test would be biased against whites by Cleary's definition. According to Darlington's Definition 3, on the other hand, the verbal ability test would be fair if, and only if, the racial difference on the numerical test were of exactly the same magnitude in standard score units as the difference on the verbal test. If the difference on the missing ability were less than the difference on the observed ability, then Darlington's definition would label the test unfair to blacks, whereas if the difference on the missing ability were larger than the difference on the observed ability, then the test would be unfair to whites. Furthermore, if the two abilities being considered were not the only causal factors in the determination of the criterion (i.e., if personality or financial difficulties, etc., were also correlated), then these statements would no longer hold. Rather, the fairness of the ability test under consideration would depend not only on the size of racial differences on the unknown ability, but on the size of racial differences on the other unknown causal factors as well. That is, according to Darlington's Definition 3, the fairness of a test cannot be related to the causal determination of the criterion until a perfect multiple regression equation on known predictors has been achieved. That is, Darlington's definition can be statistically but not substantively evaluated in real situations.

For the purpose of illustration, we now consider a simplified theory of academic achievement in college. Suppose that the college entrance test were in fact a perfect measure of academic ability for high school seniors. Why is the validity not perfect? Consider three men of average ability. Sam meets and marries "wonder woman." She scrubs the floor, earns $200 a week, and worships the ground Sam walks on. Sam carries a B average. Bill dates from time to time, gets hurt a little, turns off on girls who like him once or twice, and generally has the average number of ups and downs. Bill carries a C average. Joe meets and marries "Wanda the witch." She lies around the house, continually nags Joe about money, and continually reminds him that he is "sexually inadequate." As Joe spends more and more time at the local bar, his grades drop to a D average, and he is ultimately thrown out of school. In a nutshell, the theory of academic achievement that we wish to consider is this: Achievement equals ability plus luck, where luck is a composite of few or no money troubles, sexual problems, automobile accidents, deaths in the family, and so on. There are few known correlations between luck and ability and few known correlations between luck and personality, but for simplicity of exposition, let us ignore these and assume that luck is completely independent of ability and personality. Then luck in this theory is the component of variance in academic performance that represents the *situational* factors in performance that arise after the test is taken and during the college experience. Some of these factors are virtually unpredictable: just what girl he happens to sit beside in class, whether he arrives at the personnel desk before or after the opening for the ideal part-time job is filled, whether he gets a good advisor or a poor one, and so on. Some of these factors may be predicted: family financial support in case of financial crisis, probable income of spouse (if any), family pressure for continuing in college in case of personal crisis, and so on. However, even those factors that may be predicted are potentialities and will not

actually be relevant unless other, nonpredictable events occur. Thus, there will inevitably be a large random component to the situational factors in performance that is *not* measurement error, but that has the same effect as measurement error in that it sets an upper bound on the validity of any predictor test battery even if it includes biographical information on family income, stability, and the like.

According to this theory, a difference between the black and white regression lines (over and above the effect of test unreliability) indicates that blacks are more likely to have bad luck than are whites. Before going on to the statistical questions, we note that because we have assumed a perfect ability test, there can be no missing ability in the following discussion. And because we have assumed that nonability differences are solely determined by luck, the entity referred to as "motivation" is in this model simply the concrete expression of luck in terms of overt behavior. That is, in the present theory, *motivation* is assumed to be wholly determined by luck and hence already included in the regression equation.

Now let us consider the statistical interpretations of the fairness of our hypothetical, perfectly valid (with respect to ability) and perfectly reliable test. Because on the average blacks are assumed to be unlucky as well as lower in ability, the racial difference in college achievement in this model will be greater than that predicted by ability alone, and hence the regression lines of college performance compared with ability will not be equal. The black regression line will be lower. Thus, according to Cleary, the test is biased against whites. According to Thorndike, the test may be approximately fair (perhaps slightly biased against blacks). According to Darlington, the test could be either fair or unfair: If the racial difference on luck were about the same in magnitude as the race difference on the ability test, then the test would be fair; but if the race difference on luck were less than the difference on ability, then the test would be unfair to blacks. That is, the Darlington assessment of the fairness of the test would not depend on the validity of the test in assessing ability, but on the relative harshness of the personal–economic factors determining the amount of luck accorded the two groups. Darlington's statistical definition thus does not fit his substantive derivation in this context—unless one is willing to accept luck as an "ability" inherent to a greater extent in some applicants than in others.

The problem with Darlington's definition becomes even clearer if we alter slightly the theory of the above paragraph. Suppose that the world became more benign and that the tendency for blacks to have bad luck disappeared. Then, making the same assumptions as above (i.e., a perfect test and our theory of academic achievement), the regression curves would be equal and $r_{YG \cdot X} = 0$. Thus, according to Cleary's definition, the test would be unbiased against whites. Darlington's Definition 3 would then label the test unfair to blacks. This last statement is particularly interesting. In our theory we have assumed that exactly the same ability lay at the base of performance on both the test and later GPA. Yet it is not true in our theory that $r_{XG \cdot Y} = 0$. Thus, this example has shown that Darlington's substantive interpretation of $r_{XG \cdot Y}$ does not hold with our additional assumption (of a nonstatistical nature), and hence his argument as to the substantive justification of his definition is not logically valid.

We note in passing that this last example poses a problem for Cleary's definition as well as for Darlington's. If the difference between the regression lines were in fact produced by group differences in luck, then would it be proper to label the test biased? And if this model were correct, how many unqualified individualists would feel comfortable using separate regression lines to take into account the fact that blacks have a tougher life (on the average) and hence make poorer GPAs than their ability would predict? In the case of both definitions, this analysis points up the necessity of substantive models and considerations. Statistical analyses alone can obscure as much as they illuminate.

Darlington's Definition 3 received little attention until a novel and persuasive argument in its favor was advanced by Cole (1973). Her argument was this: Consider those applicants who would be "successful" if selected.

Should not such individuals have equal probability of being selected regardless of racial or ethnic group membership? Under the assumption of equal slopes and standard deviations for the two groups, the answer to her question is in the affirmative only if the two regression lines of test on criterion are the same (and hence $r_{XO \cdot Y} = 0$). That is, Cole's definition is the same as Cleary's with the roles of the predictor and criterion reversed. However, this similarity of statement does not imply compatibility—just the reverse. If there are differences between the races on either test or criterion, then the two definitions are compatible only if the test validity is perfect, so the two definitions almost invariably conflict.

Although Cole's argument sounds reasonable and has a great deal of intuitive appeal, it is flawed by a hidden assumption. Her definition assumes that differences between groups in probability of acceptance given later success if selected are due to discrimination based on group membership. Suppose that the two regression lines of criterion performance as a function of the test are equal (i.e., the test is Cleary-defined unbiased). If a black who would have been successful on the criterion is rejected and a white who fails the criterion is accepted, this need not imply discrimination. The black is not rejected because he is black but because he made a low score on the ability test. That is, the black is rejected because his ability at the time of the predictor test was indistinguishable from that of a group of other people (of both races) who, on the average, would have low scores on the criterion.

To make this point more strongly, we note that according to Cole's definition of a fair test, it is unethical to use a test of less than perfect validity. To illustrate this, consider the use of a valid ability test to predict academic achievement in any one group, say whites, applying for university admission. If the university decides to take only the people in the top half of the distribution of test scores, then acting under Cole's definition, applicants in the bottom half may well file suit charging discriminatory practice. According to Cole, an applicant who would be successful if selected should have the same probability of being selected regardless of group membership. That is, among the applicants who would have been successful had they been selected, there are two groups. One group of applicants has a probability of selection of 1.00 because their scores on the entrance exam are higher than the cutoff point. The other group of potentially successful applicants has a selection probability of .00 because their exam scores are lower than the cutoff point. According to Cole, we should ask: Why should a person who would be successful be denied a college berth merely because he had a low test score? After all it's success that counts, not test scores. But the fact is that for any statistical procedure that does not have perfect validity, there must always be people who will be incorrectly predicted to have low performance, that is, there will always be successful people whose predictor scores were down with the generally unsuccessful people instead of up with the generally successful people (and vice versa). In that sense, anything less than a perfect test will always be "unfair" to the potentially high achieving people who were rejected. It can be seen that lack of perfect validity functions in exactly the same way as test unreliability, discussed earlier.

As noted earlier in the case of Thorndike's definition, this problem can be partly overcome in practice if social consensus restrictions could be put on the defining of *bona fide minority groups*. But given the almost unlimited number of potentially definable social groups, it is unlikely that social or legal consensus could be reached limiting the application of this definition to blacks, Chicanos, American Indians, and a few other groups.

Basically Cole has noted the same fact that Thorndike noted: In order for a test with less than perfect validity to be fair to individuals, the test must be unfair to groups. In particular, in our example, the group of applicants who score below average on the test will have none of their members selected despite the fact that some of them would have shown successful performance if selected. It is thus unfair to this group. However, it is fair to each individual, since each is selected or rejected based on the best possible estimate of his future performance. It is perhaps im-

portant to note that this is not a problem produced by the use of psychological tests; it is a problem inherent in reality. Society and its institutions must make selection decisions. They are unavoidable. Elimination of valid psychological tests will usually mean their replacement with devices or methods with *less* validity (e.g., the interview), thus further increasing the unfairness to individuals and/or groups.

Darlington's Definition 4

The fourth concept of test bias discussed by Darlington (1971) defines a test as fair only if $r_{ox} = 0$. That is, by this definition, a test would be unfair if it showed any mean difference between the races at all, regardless of the size of difference that might exist on some criterion that is to be predicted. If the same cutoff score is to be used for blacks and whites, then this statistical criterion corresponds to the use of population-based quotas. If separate regression lines and hence separate cutoff scores are to be used, then mean differences on the test are irrelevant to the issue of quotas.

A Fifth Definition of Test Fairness

After defining and discussing four different statistical models of test fairness, Darlington (1971) turned to the commonly occuring prediction situation in which there is a difference favoring whites on both the test and the criterion and the black regression equation falls below that for whites. This situation is shown in Figure 3a. Noting that the use of separate regression equations (or the equivalent, use of multiple regression equations with race as a predictor), as required by Cleary's (1968) definition, would admit or select only an extremely small percentage of blacks, Darlington introduced his concept of the "culturally optimum" test. Darlington suggested that admissions officers at a university be asked to consider two potential graduating seniors, one white and the other black, and to indicate how much higher the white's GPA would have to be before the two candidates would be "equally desirable for selection" (p. 79). This number is symbolized K and given a verbal label such as "racial adjustment coefficient." Then in determining the

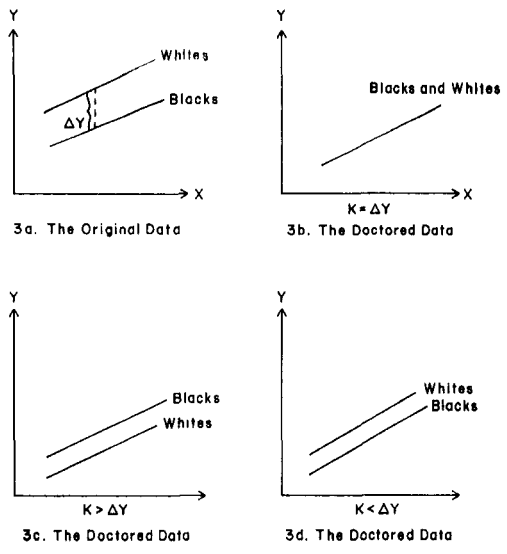

FIGURE 3. Darlington's (1971) method of doctoring the data to define a culturally optimum test.

fairness of the test, K is first subtracted from the actual criterion scores (GPAs) of each of the white subjects. If these altered data satisfy Cleary's (1968) definition of a fair test, the test is considered culturally optimum.

Figure 3 illustrates the geometrical meaning of Darlington's doctored criterion. If the admissions officer chooses a value of K that is equal to ΔY in Figure 3a, then the altered data will appear as in Figure 3b, that is, there will be a single common regression line and the test as it stands will be culturally optimum. If, however, an overzealous admissions officer chooses a value of K greater than ΔY, then the doctored data will appear as in Figure 3c, that is, the test will appear to be biased against blacks according to Cleary's definition and will thus not be culturally optimum. Similarly, should an uncooperative admissions officer select a value of $K < \Delta Y$, then the altered data would look like Figure 3d and would thus appear to be biased against whites by Cleary's criterion and hence show the test *not* to be culturally optimum.

Thus, a cynic might well assume that in practice Darlington's definition of *culturally optimum* would simply lead to the selection of a value of K that would make whatever test was being used appear to be culturally optimum. But suppose that admissions officers

were willing to choose K without looking at the effect that their choices would have on the data. How then are the nonoptimum tests to be used? One might suppose that since Darlington is eager to doctor the criterion data, he would also be willing to doctor the predictor data as well. For example, the situation in Figure 3d could be remedied by simply subtracting a suitable constant from each black's test score. However, Darlington permits only the doctoring of criterion scores; he is opposed to doctoring the predictor scores.

If the predictor scores are not manipulated by a direct mathematical rescoring, then the same effect must be obtained by constructing a new test. Consider then the situation shown in Figure 3d. The new test must have the ultimate effect of giving blacks lower scores than they would have gotten on the original test, while leaving white scores untouched. Thus, the test constructor is in the awkward position of adding items that are biased against blacks in order to make the test fair!

Darlington was not unaware of this problem, though he dealt with it in a different place. Darlington would not label the test fair unless the two regression lines using the doctored criterion were equal. But what if we do not yet have a culturally fair test? What did Darlington recommend as an interim procedure for using an unfair test? He stated that the unfair test can be *used* in a fair way if it is combined with race in a multiple regression equation based on the doctored criterion (i.e., if separate doctored regression lines are used). Thus, he used the unequal doctored regression lines in much the same way as Cleary recommended use of the unequal regression lines for undoctored regression lines. What does this procedure come to in the case in which the administrator has chosen a value of K that is too low to label the existing test culturally optimum? If the doctored regression line for blacks is still below the doctored regression line for whites, then the beta weight for blacks will still be negative and the multiple regression equation will implicitly subtract that weight from each black's score.

What would an administrator do if this were pointed out to him? We believe that he would react by increasing the value of K to make the doctored regression lines equal. That is, we think that the actual consequence of Darlington's recommendation for the fair use of an unfair test would be to further increase the likelihood of using a value of K that makes the doctored regression lines equal. That is, we believe that Darlington's definition of *fair use,* like his definition of *fair test,* is most likely to result in a harried administrator choosing K to eliminate the difference between the doctored regression lines. If we are right in this, then it means that Darlington's recommendations for fair use will lead in practice to simply labeling existing tests culturally optimum. We feel this bolsters our earlier argument that this is also the likely result of his basic procedure for defining a test as fair.

What is the upshot of Darlington's suggestion? From a mathematical point of view, adding or subtracting a constant to the criterion is exactly equivalent to adding or subtracting constants to the predictor, and this in turn is equivalent to using different cutoff scores for the two groups. Thus, in the last analysis, Darlington's method is simply an esoteric way of setting quotas, and hence the expense of constructing culturally optimum tests to do this is a waste of time and money.

Ethical Positions, Statistical Definitions, and Problems

In this section, we briefly relate each ethical position to its appropriate statistical operation and point out some of the advantages and disadvantages of each approach.

Unqualified Individualism

The ethical imperative of individualism is to apply to each person that prediction procedure which is most valid for that person. Thus, white performance should be predicted by that test which has maximum validity for whites. Black performance should be predicted using that test which has maximum validity for blacks. The person with the highest predicted criterion score is then selected.

There is no reason why the test used to select blacks need be the same as that used to select whites. Indeed, if there is a more

valid test for blacks, then it is ethically wrong *not* to use it. Furthermore, in situations in which the mean black criterion performance is lower than that for whites, the number of blacks admitted is maximized by using that test which has maximum validity for blacks.

Consider the alternative. Suppose there is a group for which the test used has low validity. For simplicity assume no validity at all. Then, the predicted criterion score for everyone in that group is the same—the mean criterion score for that group. Thus, either everyone in that group is accepted or everyone in that group is rejected. If that group is in fact highly homogeneous on the criterion, then this is perfectly reasonable. But if the zero validity group has the same degree of spread on the criterion as other groups, then this lack of discrimination poses ethical problems: either a great many poor prospects are being admitted, or a great many excellent prospects are being overlooked. Because selection ratios are typically low (say 50% or less), this means that the use of a low validity test for some groups is likely to mean that that group is virtually eliminated. Thus, indeed, it is important to seek the maximum validity test for any group. But there is little evidence to suggest that different demographic groups will in fact require different tests. In an age of mass culture, this seems a very implausible hypothesis for most such groups. For example, the research evidence strongly indicates that differential validity by race is no more than a chance phenomenon (Schmidt, Berner, & Hunter, 1973). The same may later be shown with respect to other population subgroups, thus greatly reducing the scope of this problem. The problem would not thus be eliminated, however; although the same tests may be valid across population subgroups and regression slopes may be equal, there is much research evidence (Reilly, 1973; Schmidt & Hunter, 1974; Ruch, Note 1) that intercepts often differ significantly. That is, the same test may be a maximum validity test for many groups, but it need not therefore be unbiased by Cleary's definition. Thus, some adjustment for differences in group intercepts would still have to be made.

Qualified Individualism

For the most part, the qualified individualist is also concerned with maximum validity. However, should there be a subgroup for whom there was low validity, it would pose greater problems for the qualified individualist because he cannot give different tests to different groups. Thus, should such a case ever be found, the qualified individualist would presumably respond by searching for a less valid test (for the population as a whole) that had less variability in subgroup validity.

There is another, more subtle but perhaps more real, problem that advocates of qualified individualism must face. The ethical imperative here requires that the prediction equation that has maximum validity for the entire population—without regard to group membership—be identified and employed. But there is a problem with this solution. Suppose, for example, that for a certain city college the black regression line falls below the white regression line, that is, race is a valid predictor for that college. Use of race as a predictor is, of course, forbidden to the qualified individualist, but there may be alternative ways of increasing the overall validity of the prediction equation that are equally objectionable. For example, if race is a valid predictor, then a properly coded version of the student's address may also be a valid predictor and increase overall validity. This "indirect indicator of race" would probably be detected and rejected, but a more subtle cue might not be properly identified. In particular, the most subtle problem is the one facing the test constructor: If the black regression line falls below the white regression line, then the introduction of items whose content is biased against blacks would increase the overall validity of the test. If the separate regression lines of the unqualified individualist are used, then racially biased test material would have no effect on the selection of applicants. But if that is forbidden, then material biased against blacks would lower the black scores on the predictor and hence make their scores using the white regression line more accurate. That is, the introduction of material biased against blacks

would reduce the overprediction of black performance and hence raise the validity of a one-regression-line use of the test.

The problem, in its general form, is that any measured variable which correlates with race, sex, religion, and so on (i.e., shows group differences), can be considered to be an indirect (and imperfect) indicator of group membership. Because he is forbidden to use group membership itself as a predictor even if valid, the qualified individualist may be tempted to substitute indirect indicators of group membership that may be unfair. How can he decide whether a given race-correlated predictor is fair or unfair? We discuss two such criteria: (a) Is the relation between predictor and criterion an "intrinsic" one? (b) Is the within-groups validity high enough?

The first criterion is the apparent intrinsicness of the relationship between the predictor and performance. If the predictor is a job sample test (e.g., a typing test) assessing the skills actually required on the job, there is little doubt that the relation is intrinsic. Scores on a written achievement test could also easily pass this test, as would a face-valid aptitude test. Scores on a weighted biographical information inventory, on the other hand, would be allowed only if they were able to meet the second, less subjective standard: high validity coefficients for both groups separately. Thus, the qualified individualist's answer to the question posed in the example above is that if the material to be added to the test appears to have an intrinsic relation to performance, it is ethically admissible. It is not biased against blacks as blacks but merely against applicants (of whatever race) who are less capable of performing well on the criterion. The fact that there happen to be more blacks with low ability (percentage-wise) than whites is an ethically irrelevant fact.

While the preceding distinctions are regarded as crucial among qualified individualists, they receive short shrift from those committed to other ethical positions. Those who are *un*qualified individualists will argue that parental income is an indicator of motivation to do well in school or on the job and that such motivation is surely intrinsic to high performance. That is, the unqualified individualist says that the question of whether a variable is intrinsically related to performance is subject to empirical test: If it is correlated with performance, then it is intrinsically related, whereas if it is not correlated, then it is not. Those who promote quotas will also reject validity of the intrinsic–extrinsic distinction. They argue that *ability* just means whether or not you went to a good school and is thus highly contaminated with extrinsic elements. The qualified individualist may offer scientific theories in arguing his case, but his opponents will simply argue that the theories are wrong. And the data we have in 1975 will *not* decide the argument.

Is there a less subjective way to test for an intrinsic relation between the predictor and the criterion? Certainly one test is within-groups validity. If the relation is intrinsic, then there should be a correlation for each group separately. But how high should that correlation be? Certainly statistical significance is no answer. If a college were gathering data on a new test on an entering class of 4,000, then it would only take a within-groups validity of .01 to be significant. On the other hand, if we set some standard, such as .10, we run into another problem. If the within-groups validity of some piece of biographical information were .10, while the correlation with race were .70, it would be clear that most of the validity of the test would lie in its serving as an indirect indicator of race. Indeed, one might well consider requiring that the correlation with race be less than the within-groups validity.

But rather than continue searching for arbitrary standards, let us consider a purely statistical definition of an indirect indicator of race: X is an indirect indicator of race to the extent that it correlates more highly with race than is required by its relation with the criterion. By this definition X would be safe from a charge of being an indirect indicator if $r_{XC \cdot Y} = 0$—that is, if it satisfied the Darlington–Cole definition of a fair test! And indeed the Darlington–Cole definition relates the within-groups validity to the test race difference. If the within-groups validity were .10, then by this definition, the biographical item would be an indirect indicator of race unless the standard score difference on the

test were less than one tenth the standard score difference on the criterion. But for all practical purposes, this means that there cannot be any difference on the test at all!

But that brings us full circle on the statistical question. The concept of an intrinsic relation is inherently a matter of causal arguments, and as we noted above, it cannot thus be assessed by any statistical procedure. Thus, as we noted in our discussion of the theoretical objections to the intrinsic–extrinsic distinction, most such arguments between adherents of different ethical positions will come down to scientific issues that cannot be resolved on the basis of the data at hand today.

Quotas

The main technical question for an adherent of quotas is, Whose? Once the quotas have been set, the only remaining ethical question is how to select from within each group. Although some would use random selection within groups, most would evoke individualism at this point. With this assumption, the optimal strategy for filling quotas can be stated. For each group, obtain predicted criterion performance using that test which has maximum validity for the given group. If the test with maximum validity for blacks is not the test with maximum validity for whites, then it is unethical to use the same test for both.

The major problem for a quota-based system is that the criterion performance of selectees as a whole can be expected to be considerably lower than under unqualified or even qualified individualism. In college selection, for example, the poor-risk blacks who are admitted by a quota are more likely to fail than are the higher scoring whites who are rejected because of the quota. Thus, in situations in which low-criterion performance carries a considerable penalty, being selected on the basis of quotas is a mixed blessing. Second, there is the effect on the institution. The greater the divergence between the quotas and the selection percentages based on actual expected performance, the greater the difference in mean performance in those selected. If lowered performance is met by increased rates of expulsion or firing, then the institution is relatively unaffected, but (a) the quotas are undone and (b) there is considerable anguish for those selected who don't make it.[5] On the other hand, if the institution tries to adjust to the candidates selected by quotas, there may be great cost and inefficiency. Finally, there is the one other problem that academic institutions must face. Quotas will inevitably lower the average performance of graduating seniors, and hence lower the prestige rating of the school. Similar considerations apply in the case of the employment setting. In both cases, the effect of these changes on the broader society must also be considered. These effects are difficult to assess, but they may be quite significant.

Concluding Remarks

We have presented three ethical positions with respect to the use of tests and other psychological devices in selecting people for entry into various kinds of institutions, and we have shown these ethical positions to be irreconcilable. We have also reviewed a number of attempts to define the fair or unbiased use of tests or other devices and have shown them to be related to different ethical positions. Moreover, we have shown that the scientific principles used to justify the statistical procedures vary considerably in their plausibility from one concrete selection situation to another. Indeed we feel that we have shown that any purely statistical approach to the problem of test bias is doomed to rather immediate failure.

The dispute reviewed in this article is typical of ethical arguments—the resolution depends in part on irreconcilable values. Furthermore, even among those who agree on values there will be disagreements about the validity of certain relevant scientific theories that are not yet adequately tested. Thus, we

[5] Furthermore, the public image of the institution may suffer as much from the higher rate of expulsion as from the charge of discrimination in hiring. For example, if we read the trial records correctly, there is a company that deliberately reduced its entrance standards to hire more blacks. However, these people could not then pass the internal promotion tests and hence accumulated in the lowest level jobs in the organization. The government then took them to court for discriminatory promotion policies!

feel that there is no way that this dispute can be objectively resolved. Each person must choose as he sees fit (and in fact we are divided). We do hope that we have clarified the issues to make the choice more explicitly related to the person's own values and beliefs.

REFERENCE NOTE

1. Ruch, W. W. A re-analysis of published differential validity studies. In R. E. Biddle (Chair), *Differential validation under Equal Employment Opportunity Commission and Office of Federal Contract Compliance testing and selection regulations.* Symposium presented at the meeting of the American Psychological Association, Honolulu, September 1972.

REFERENCES

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

Cole, N. S. Bias in selection. *ACT Research Report* No. 51, May, 1972. Iowa City, Iowa: The American College Testing Program, 1972. (Also in *Journal of Educational Measurement,* 1973, *10,* 237–255.)

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

Edwards, W., & Phillips, L. D. Man as transducer for probabilities in Bayesian command and control systems. In M. W. Shelley II & G. L. Bryan (Eds.), *Human judgements and optimality.* New York: Wiley, 1964.

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

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

Reilly, R. R. A note on minority group test bias studies. *Psychological Bulletin,* 1973, *80,* 130–132.

Schmidt, F. L., Berner, J. G., & Hunter, J. E. Racial differences in validity of employment tests: Reality or illusion? *Journal of Applied Psychology,* 1973, *58,* 5–9.

Schmidt, F. L., & Hunter, J. E. Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. *American Psychologist,* 1974, *29,* 1–8.

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

Appendix

This appendix contains the mathematical calculation of the expected achievement level of the group that would be selected by the full application of Thorndike's criterion, that is, a group selected so that for each test score x, the number of people selected is proportional to the probability that persons at that test level would in fact be successful. The definition of *successful* used below is performance above average on the criterion. That is, the calculations below assume a 50% selection ratio.

For simplicity, both test and performance have been assumed to be measured in standard scores. The symbol $\varphi(x)$ is the standard normal density function and the symbol $\Phi(x)$ is the standard normal cumulative distribution function. The symbol A is used for *accepted* (or selected for admission).

If the criterion of success is the top 50%, then in terms of standard scores, the success criterion is $Y > 0$. Thus, the conditional probability of being accepted is:

$$P(\mathrm{A}|X) = P[Y > 0|X] = \mathrm{P}\left(\frac{y - rx}{\sqrt{1 - r^2}} > -\frac{rx}{\sqrt{1 - r^2}}\right).$$

Let us simplify the following expressions by defining the parameter α by

$$\alpha = \frac{r}{\sqrt{1 - r^2}}.$$

In particular, if $r = .6$, then

$$\alpha = \frac{.6}{\sqrt{1 - .36}} = \frac{.6}{.8} = .75.$$

We can then write

$$\mathrm{P}(\mathrm{A}|X) = P(z > -\alpha x),$$

where z has a standard normal distribution. Thus,

$$\mathrm{P}(\mathrm{A}|X) = 1 - \Phi(-\alpha x) = \Phi(\alpha x).$$

If the number selected at each test score is $\mathrm{P}(\mathrm{A}|X)$, then the overall selection ratio will be

$$\mathrm{P}(\mathrm{A}) = \int \Phi(\alpha x)\varphi(x)dx = \tfrac{1}{2}.$$

The distribution of the test score among those selected is

$$f_A(x) = \frac{\mathrm{P}(\mathrm{A}|X)\mathrm{P}(X)}{\mathrm{P}(\mathrm{A})} = 2\Phi(\alpha x)\varphi(x).$$

Since X and Y are in standard score form, the regression of Y on X is given by $\mathrm{E}(Y|X) = rx$.

Thus, the mean criterion score among those selected will be

$$\begin{aligned} E(Y) &= E[E(Y|X)] \\ &= \int rx f_A(x)dx \\ &= \int rx\, 2\,\Phi(\alpha x)\varphi(x)dx\,. \end{aligned}$$

This is not an easy integral to calculate, and the calculation below will thus be broken into five steps. The formula for the mean criterion performance among those selected can then be written

$$E(Y) = 2r\int \Phi(\alpha x)\varphi(x)dx\,.$$

Step 1. First we apply the method of integration by parts:

$$\begin{aligned} &\int x\Phi(\alpha x)\varphi(x)dx \\ &\quad = \int \Phi(\alpha x)[x\varphi(x)]dx \\ &\quad = \Phi(\alpha x)u(x) - \int u(x)\{\Phi'(\alpha x)\alpha\}dx\,, \end{aligned}$$

where

$$u(x) = \int x\varphi(x)dx = -\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,.$$

Step 2. Thus, we have the definite integral:

$$\begin{aligned} &\int_{-\infty}^{+\infty} x\Phi(\alpha x)\varphi(x)dx \\ &\quad = \Phi(\alpha x)u(x)\Big|_{-\infty}^{+\infty} - \int u(x)\Phi'(\alpha x)\alpha dx \\ &\quad = 0 - \int u(x)\frac{e^{-\alpha^2x^2/2}}{\sqrt{2\pi}}\alpha dx \\ &\quad = \frac{\alpha}{2\pi}\int e^{-x^2/2}e^{-\alpha^2x^2/2}dx \\ &\quad = \frac{\alpha}{2\pi}\int e^{-(1+\alpha^2)x^2/2}dx \\ &\quad = \frac{\alpha}{2\pi}\int e^{-\beta^2x^2/2}dx\,, \end{aligned}$$

where

$$\beta = \sqrt{1+\alpha^2}\,.$$

Step 3. Using the substitution $x = (1/\beta)y$, we can calculate the following integral:

$$\begin{aligned} \int e^{-\beta^2x^2/2}dx &= \int e^{-y^2/2}\frac{1}{\beta}dy \\ &= \frac{1}{\beta}\sqrt{2\pi}\,. \end{aligned}$$

Step 4. Thus, we can finally calculate the main integral:

$$\begin{aligned} \int_{-\infty}^{+\infty} x\Phi(\alpha x)\varphi(x)dx &= \left(\frac{\alpha}{2\pi}\right)\left(\frac{1}{\beta}\right)(\sqrt{2\pi}) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)\left(\frac{\alpha}{\sqrt{1+\alpha^2}}\right) \end{aligned}$$

Because α was defined to be

$$\alpha = \frac{r}{\sqrt{1-r^2}}\,,$$

we have

$$\frac{\alpha}{\sqrt{1+\alpha^2}} = \frac{r\sqrt{1-r^2}}{\sqrt{1+\frac{r^2}{1-r^2}}} = r\,.$$

Step 5. Finally, we can use the main integral to calculate the expected achievement level:

$$\begin{aligned} E(Y) = 2r\ \text{Integral} &= (2r)\left(\frac{1}{\sqrt{2\pi}}\right)\left(\frac{\alpha}{\sqrt{1+\alpha^2}}\right) \\ &= \frac{2r^2}{\sqrt{2\pi}}\,. \end{aligned}$$

For $r = .6$, this formula yields $E(Y) = .288$.

(Received June 2, 1975)