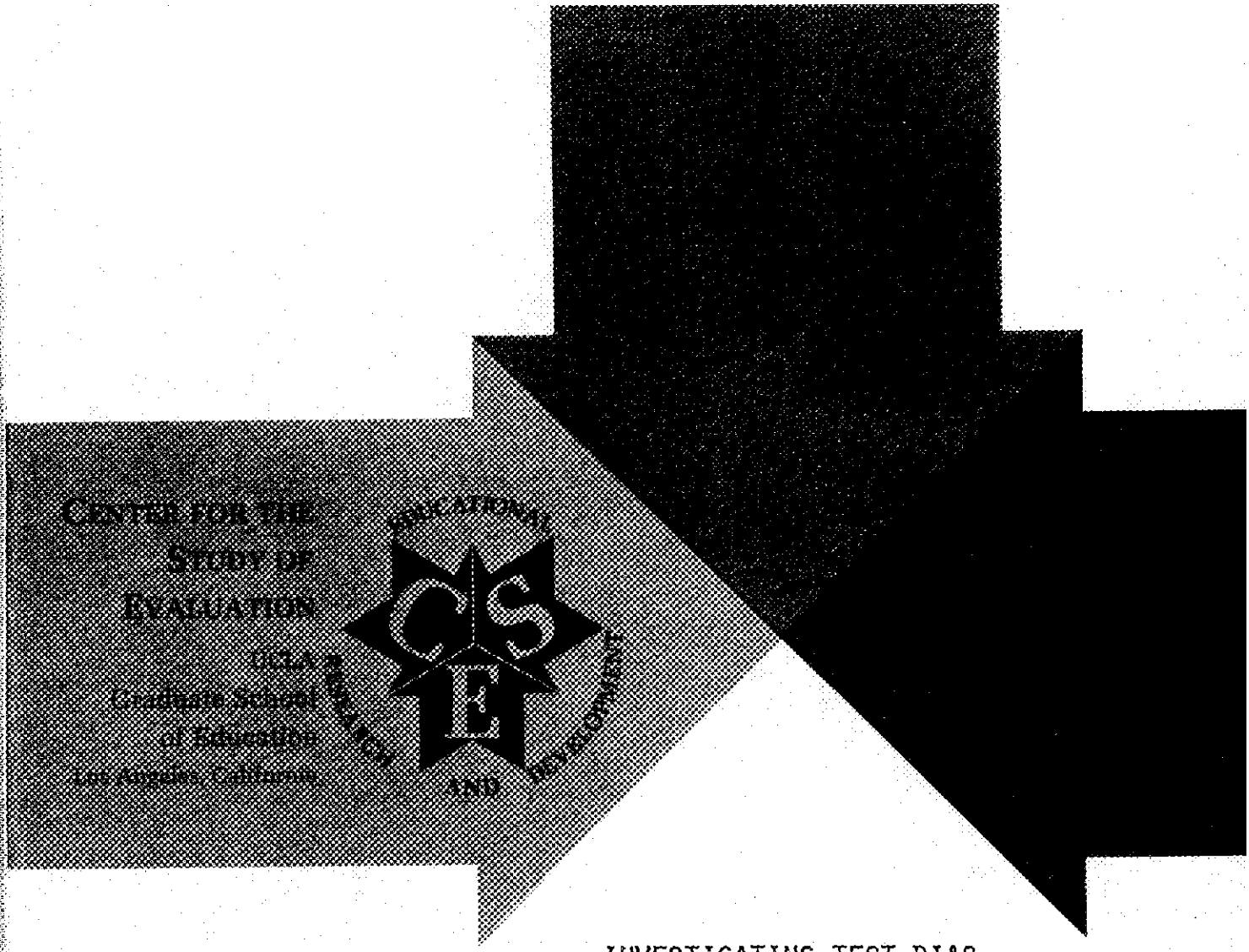

INVESTIGATING TEST BIAS

Ralph Hoepfner and Guy P. Strickland

CSE Report No. 74
February 1972

INVESTIGATING TEST BIAS

By

Ralph Hoepfner and Guy P. Strickland

CSE Report No. 74
February 1972

School Evaluation Program
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California

The Center for the Study of Evaluation is engaged in numerous studies involving the dissemination and application of evaluation methodology in the nation's classrooms. The Center has designed workshops and kits to educate school personnel in evaluation techniques and, at the same time, has tried to present those evaluation techniques to school personnel in a form both appropriate and useful to them. The most commonly used technique in educational evaluation, at this time, is the measurement of student performance through use of standardized test instruments. In the CSE Elementary School Test Evaluations (Hoepfner, R., Strickland, G., Stangel, G., Jansen, P., and Patalino, M., 1970), the Center has provided information on over 1,600 published tests and subtests for use in elementary schools. On the basis of this information, school principals are enabled to choose the tests most suitable for their needs in their schools.

The tests were evaluated and rated on 24 criteria; six of these criteria were concerned with the tests' appropriatenesses for the students being tested. This "appropriateness" was interpreted, for the sake of simplicity and generalizability, as appropriateness to an average class in each grade (tests for grades 1, 3, 5, and 6 were evaluated). At the same time, the Center acknowledged that some tests may be more or less appropriate for certain socioeconomic or ethnic groups.

Many school districts, particularly in lower socioeconomic and highly ethnic neighborhoods, are under pressure from teachers and/or parents to abandon testing completely because the tests are felt to be inappropriate for their communities. If one considers that testing has two functions, pupil placement and the measurement of program effectiveness, the abolition of testing can be seen as being potentially beneficial in ending the abuses

in pupil placement due to the use of inappropriate and biased tests; but such abolition would inhibit measurement of program effectiveness, thereby inhibiting program improvement and program change. This is particularly true when the test is appropriate (referenced) to the instructional program.

The Center felt it necessary to investigate the question of test bias, with the intention of developing an index of the appropriateness of a test to a particular socioeconomic or racial-ethnic group. This would be a valuable appendix to the CSE Elementary School Evaluations, expanding its usefulness to a broader range of school and pupil types. Our intent was to isolate either the tests or the test items that exhibited characteristics indicating bias, induce the aspects of the tests or items that are common to the "biased" measures, and from those aspects develop a quantitative index of what might be called "predicted bias" which could be generalized and applied to a wide range of test instruments.

Approaches to Measuring Bias

In attempting to develop its own bias indexes, the Center considered several approaches that have been described and employed in the literature.

External Explanations of Test Bias

Several of the procedures for explaining or establishing the existence of test bias depend upon criteria external to the test instrument itself. Characteristics of norms and validities are examples of this type.

Bias by Norm Sample. Certain tests may have different norm samples that cause differences among racial and socio-economic groups. If it is assumed that, relative to a particular objective of achievement or aptitude, there are not systematic and reliable differences in underlying standings

among the various subgroups in the population, then for a veridical test of that aptitude or achievement it shouldn't matter whether the norm sample systematically excludes members of any subgroup; the obtained norms would be more or less appropriate for all subgroups in the population.

But Millman and Lindlof (1964) have shown that different tests with different norm samples yield markedly different percentile norms by the equi-percentile method. The specific differences had been previously noted by several investigators and test users. More to the point, however, Eagle and Harris (1969) showed significant differences in white-non-white comparisons for two different tests. One test could be said to favor whites much more than the other. These findings may have been due to the norms used in assigning grade equivalents, to differences in test content or format, or to basic racial-socio-economic differences.

To the extent that the underlying ability or achievement is not equal among all the subgroups or the test is a poor measure of the underlying status, the norms are inappropriate for the excluded subgroup. Since the former case is not wholly consistent with notions of bias, we can look at some of the ways that tests can be poor measures of the underlying status.

In the construction of test instruments it is useful to distinguish between the norm sample utilized for a test and the pilot sample used in developing the test. If the pilot sample excludes subgroups (which it usually does, in an effort to increase the economies of test development), so that the statistical characteristics of the items reflect only certain subgroups, then the items selected for the test will be items appropriate (in terms of difficulty, external discriminability, and item content) for the subgroup utilized. As a result, the content validity of the test may be different for the different subgroups. But if we continue to assume

that the true performance of all groups over the objectives of the test is the same, then the process of item selection based on a pilot sample from one subgroup only is just as likely to yield higher or lower content validity for the excluded subgroups. The presence of bias is always possible when a test designed for one group is given to another; that is an experimental truism that transcends any issue of racial bias. But the bias is not a systematic one; it does not consistently or by design favor one group over another.

On the other hand, if it is assumed that some ethnic groups have lower ability on a particular objective than other groups, then poor norm sampling can create or accentuate the bias. Many studies have found that score means for racial minority groups are lower and standard deviations smaller than those for the majority white group. Were this phenomena constant for certain types of tests, then it would follow that any norming of the test must have appropriate subgroup representation if the bias is not to be enhanced or increased through use of the norming procedure. A disadvantaged minority student who takes a test normed on white children will be given a score that is too far below the mean because (1) the mean is artificially high, and (2) the distance in standard-score units is artificially inflated because the unit of measurement (standard-deviation) is too small. If the minority group is superior to the white norm group on the objective performance, the foregoing function still holds; only the bias will favor the minority group. Such phenomena would clearly not be cases of test bias, however; they would be cases of racial differences on the objectives under consideration, merely being reflected by the test scores.

Item selection based on data from a poorly normed pilot sample, in a situation where ethnic subgroups are assumed to be of unequal ability, does not by itself ensure that there will be bias. There is no reason to believe that the process will cause selection of items biased in favor of the pilot group in such a way that the pre-existing differences between ethnic groups are systematically increased or decreased. Again, the content validity may be different for different ethnic groups.

Bias by Predictive Validities. Certain tests may have different predictive validities for different racial groups by underpredicting or by overpredicting minority-group performance. Moderated prediction (Einhorn and Bass, 1971) may be called for for the different groups, yielding different regression functions. This differential validity proposition, however, is based upon the inequality (read differences) in the subgroups. In addition, certain tests may suffer in predictive validity due to inappropriateness of comprehension level or test content.

Temp (1971) found that, employing the Scholastic Aptitude Test (SAT) as a predictor of GPA, black and white regression equations were different and that black GPA, when estimated from the white regression equation, was overpredicted. This two-pronged approach addresses the issue of bias via prediction in two ways. First, Temp essentially considered his black sample and his white sample as independent populations (therefore, not necessarily equal, the same, or even similar), each giving rise to regression parameters that could be compared for differences. Second, the samples were considered to be subsamples of the same population; and,if race were the only selection variable (and also not influencing regression phenomena), the null hypothesis would be that utilizing one subgroup's regression equation on another's

scores and achievements would not result in systematic under- or over-prediction. Of course, it did result in overprediction, but race alone may not have been the cause, as observed race is confounded by many social and environmental variables.

Differences in predictive validities were noted for the Negro and white samples in the study by Goolsby and Frary (1970). Such large and systematic differences were not noted when the study sample was regrouped in terms of sex or school readiness. The validity differences, computed separately for black and white groups, were not, however, all in favor of higher predictability of the white sample, in fact, 46 of 70 of the validity coefficients were greater for the black sample. In a like manner, Mitchell (1967) found about an even split (26 to 19) in greater predictive validities for black and white samples, respectively.

Linn and Werts (1971) discuss some of the technical problems that may account for differences between subgroup regression equations. These problems are predictor unreliability and exclusion of certain predictors. Linn and Werts are the first to explicitly point out the necessary equality of criterion performance in order to compare regression data to uncover the effects of test bias. In a similar manner, Thorndike (1971) suggests several alternatives for handling test bias, depending upon how it is defined, but also suggests that criterion bias may play a larger role in the observed sub-group differences. Thorndike's conclusion closely parallels the one we shall reach:

> "If the criterion measure is itself biased on an unknown direction and degree, no rational procedure can be set up for "fair" use of the test. To determine that test scores in the two groups predict a given criterion rating is fruitless if the criterion rating does not really mean the same thing in the major and minor groups. And by the same token, setting up group quotas based on proportions in

> previous major and minor groups, that have achieved a specified criterion rating is fruitless if the criterion rating signifies different things in the two groups." (p. 70)

If, indeed, we reject the current definitions of test bias and are consequently forced to accept criterion bias, then Thorndike's conclusion is of far greater general consequence than it at first appears.

The investigator of test bias can determine the nature of the bias of a test, as a predictor, by examining those systematic over- or under-predictions of some criterion performance by the test. In order to uncover any bias in the predictor, however, we must assume that the distribution of criterion performance of the sample (the minority group that may be being detrimentally assessed or classified by the supposedly biased predictor) is not itself a biased sampling on the criterion performance on the population (the larger group upon which the validity of the test has been determined). What is meant here is essentially this: If we wish to determine whether or not the ACME Reading Test is biased for the educational assessment and placement of young black learners, or whether the ACE Typing Test is biased for selection or promotion of Mexican-American clerical job applicants, it is necessary to accept the unbiased nature of the "true" reading aptitude of the learners or the "true" typing skill of the job applicants. It is crucial to assume that their "true" performance (not necessarily faulty measures of it, for that compounds the problem) is in no way systematically different from the norm of the population, or else one has to "adjust" for those systematic differences (if they are known or can be confidently estimated). Whether or not the assumption of equal criterion performance is true is not known, but it is a reasonably safe null hypothesis from both a statistical and an egalitarian point of view.

We can understand these complex phenomena more clearly if they are illustrated through scatter plots that are commonly used to represent validation data. Figure 1 illustrates such a scatter plot for a hypothetical population, where each member's predictor score is cross-tabulated with his achievement-performance score. For the sake of simplicity, let's say that the correlation (validity coefficient) is +.60.

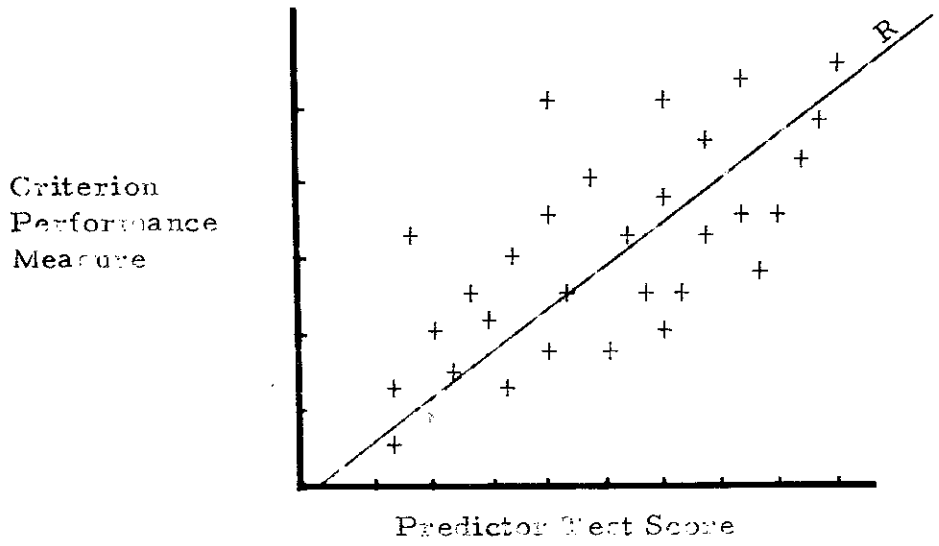

Figure 1

Scatter Plot of Predictive Validity Data for a Population

The regression line, R, has been drawn in Figure 1 to indicate the best estimate (in a least-squares sense) of the criterion score if only the predictor score is known (one merely locates the predictor score on the horizontal axis, projects that point straight up until it intersects R, and then projects the intersection left to the criterion score axis). If one were to follow this procedure with any random or stratified-random sample

from the population, the predicted scores would, of course, fall on the regression line and the actual criterion scores would be randomly distributed above and below the line. While it is, moreover, safe to say mathematically that no matter what kind of sample we select the predicted scores will fall on the regression line, it is not so safe to say that the actual criterion scores would also be randomly distributed above and below that line.

But the assumption of the null hypothesis --- that criterion performance is essentially equal among all groups --- leads to the conclusion that non-random deviations of the predicted scores from the regression line will not generally occur, except through faulty or unlikely sampling. Figure 2 illustrates the impossibility of this phenomenon by posing a case where there appears to be over-prediction of success by the regression equation developed from Figure 1.

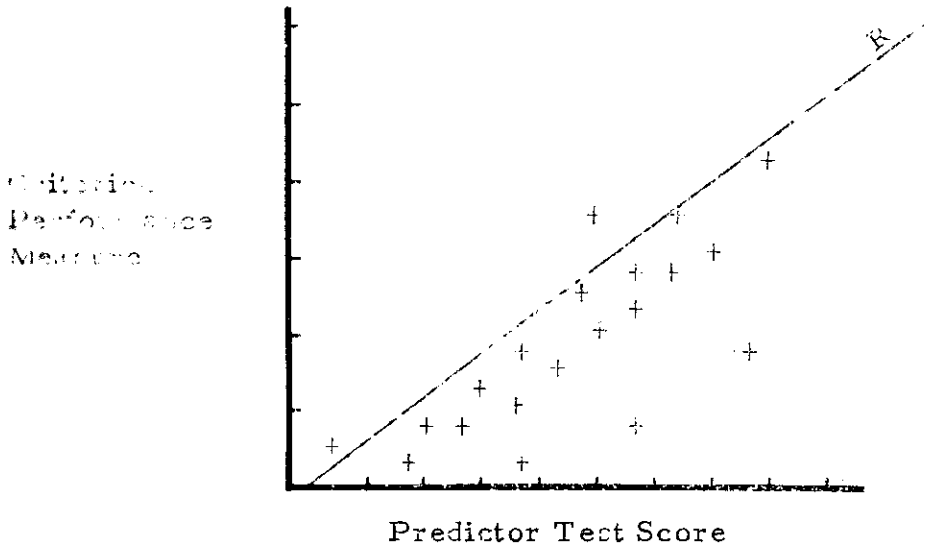

Figure 2

Theoretical Scatter Plot of Predictive Validity Data for a Minority Group

The reason that Figure 2 (or a figure with a similar underpredictor bias) cannot be expected to occur is that either demands, contrary to the null hypothesis, restriction on the range of criterion scores; after all, in the first case (overprediction) the sample exhibits few high criterion scores. (Cases of restriction of range of the predictor scores are different and are treated in a later section.) It becomes obvious that for cases of either systematic over- or under-prediction, we have cases of biased criterion performance, not of biased test! It can be seen that Cleary's (1968) definition of test bias:

> A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias, there may be a connotation of "unfair", particularly if the use of the test produces a prediction that is too low. If the test is used in selection, members of a subgroup may be rejected when they were capable of adequate performance.

is really a definition of performance superiority or deficit or of criterion bias, not of test bias.

Two alternative phenomena could occur, however, that might indicate a biased test; first, differential prediction of the criterion, and second, truncation of the predictor score distribution. Since our null hypothesis demands scores throughout the range of criterion performances, there are only three types of misprediction that could occur with a predictor having a full range of scores. These types are called "dulled" prediction, "misprediction", and "reversed" prediction, and are graphically illustrated in Figures 3, 4, and 5, respectively.

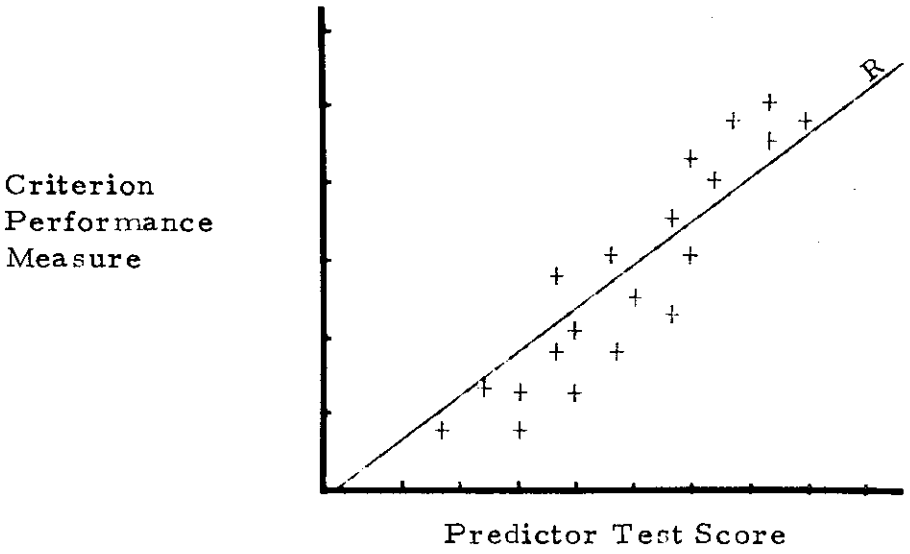

Figure 3

Scatter Plot of "Dulled" Prediction Validity Data for a Minority Group

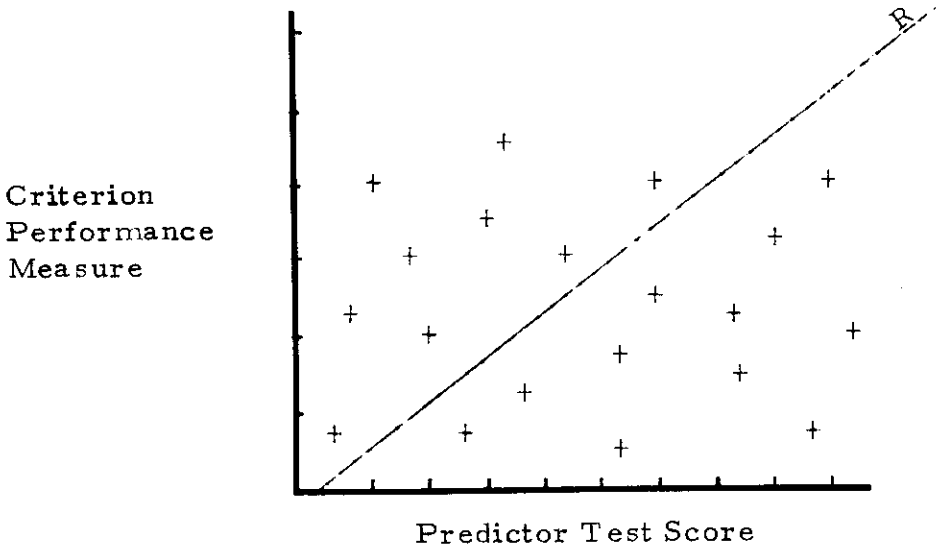

Figure 4

Scatter Plot of "Misprediction" Validity Data for a Minority Group

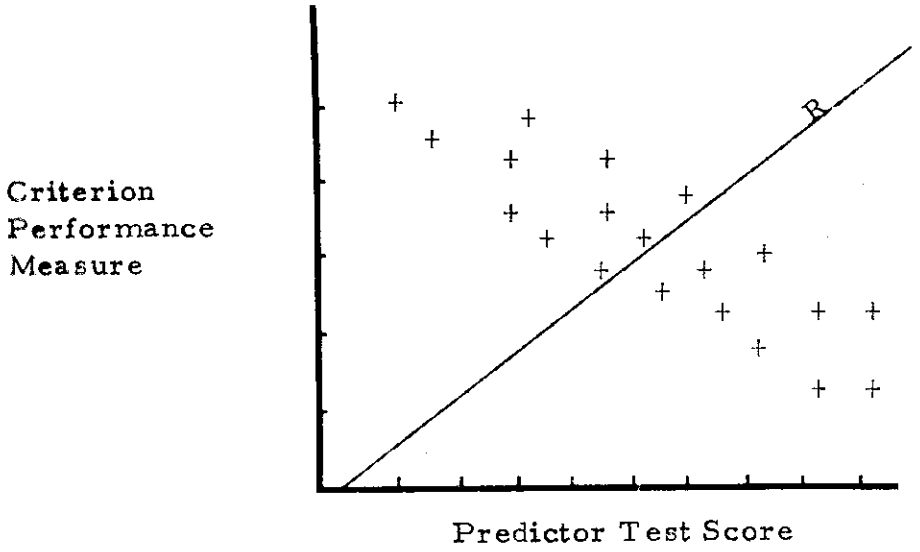

Figure 5

Scatter Plot of "Reversed" Prediction Validity Data for a Minority Group

It could be said that the case of dulled prediction (Figure 3), where low scores are systematically overestimated and high scores are systematically underestimated, is not really what we think of as bias at all; merely a case where prediction based upon population data is not as good as if based upon sample data. The same is true for misprediction as illustrated in Figure 4, where no bivariate distribution (or one with a correlation smaller than +.60) appears to support the population regression. Reversed prediction is only a little closer to our notion of bias. Here low scorers are underpredicted and high scorers are overpredicted. Selection or prediction in this case will maximize personal injustice and personnel inefficiency, of course.

The phenomena that are more obviously examples of bias occur when there is truncation of the distribution of predictor scores, either at the higher extreme (more commonly expected) or at the lower extreme. The

effects of such truncations on prediction are potentially widely varied. But one doesn't have to look at validity data to determine such bias; a simple test of the differences of the predictor scores for the minority subgroup will reveal this bias more sensitively and more accurately.

Going back to Cleary's (1968) definition of bias, above, we can look at regression lines for subgroups separately and compare them, both as to how they predict the unbiased criterion performance and as to how each subgroup's predicted performance systematically under- or over-estimates its and other subgroups' actual performances.

In the first case, where the subgroup regression is compared to the population regression, the phenomena described above obtain, the only difference being that regressions are computed and compared instead of noting scatterplot differences. Likewise, in the second case, where subgroups' regressions are compared and systematically cross-validated on other subgroups, the same phenomena occur.

What emerges from this exhaustive review of the possibilities of bias in prediction is this: if criterion performance is unbiased and equal across subgroups, predictor studies can reveal only trivial examples of test bias. The key to the meaning of this conclusion is, of course, the assumed unbiased and equal criterion performance. To the extent that the "true" criterion performance is not equal among subgroups, or measurements of the performance are not equal (biased), then all the classical approaches to predictor bias take on meaning --- but usually the wrong meaning; the criterion is biased, not the predictor.

In terms of describing what happens when there is bias in the criterion performance or its measures, there is no value in making distinctions between

the population and one (majority) subgroup (the distinction is certainly crucial, but not to the development of the prediction events to be described). For this reason, "population" will be used to represent either the real population or the majority subgroup, and "subgroup" will apply only to one of the minority subgroups. Given these definitions, two criterion performances are possible (equality being discussed above); either the population performance is better then that for the subgroup, or it is worse.

If the population performance (or its measure) is better than that for the subgroup (caused by hereditary or environmental selection or effects), scores on a valid predictor for the population can appear to be biased in two ways. When the predictor is also valid for the subgroup, Figure 6 shows what will occur. When the prediction equation developed from the population is used for members of the subgroup, the procedure will overpredict the success

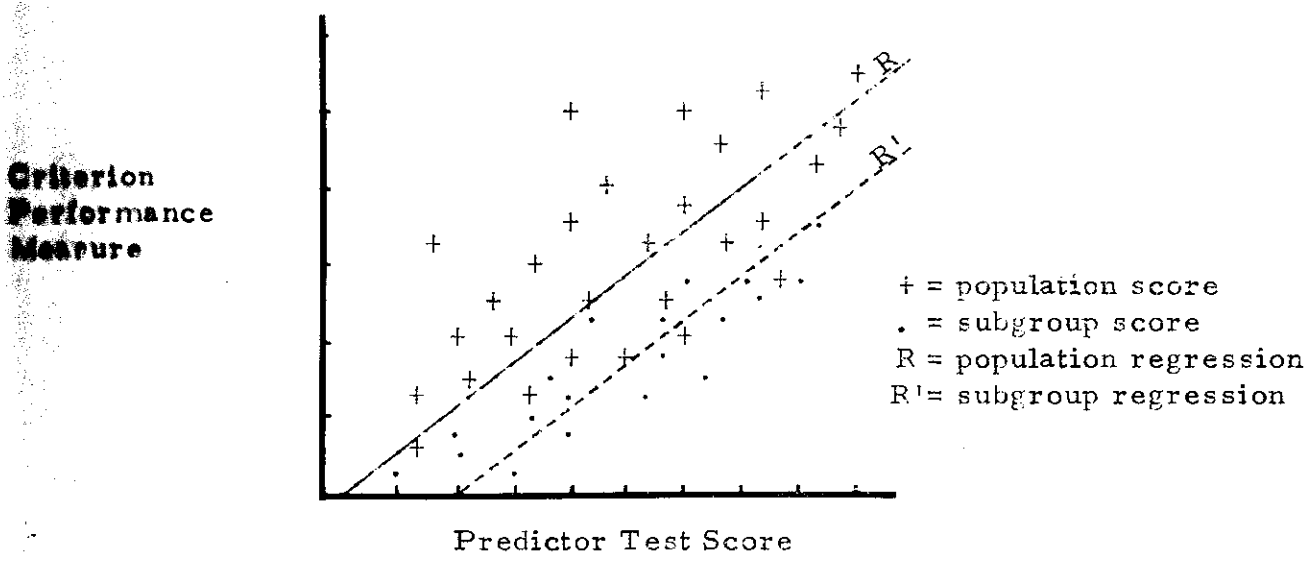

Figure 6

Scatter Plot of Predictive Validity Data for a Test Valid for the Criterion Performance

of the subgroup, causing the selection of subgroup people who would not be selected from the population and who have a high probability of not meeting performance criteria standards. When the prediction equation developed from the subgroup is applied to members of the population, the success of population members will be underpredicted, causing the rejection of the population people who would be selected from the population and who have a high probability of meeting performance criterion standards.

When the predictor is not valid for the subgroups and the population performance is better than that for the subgroup (illustrated in Figure 7), we overpredict population performance by utilizing the subgroup regression; the error of prediction being systematically larger for high scorers than for low scorers.

In the event that the subgroup performance (or its measure) is better than that for the population (an infrequently observed phenomenon) the errors

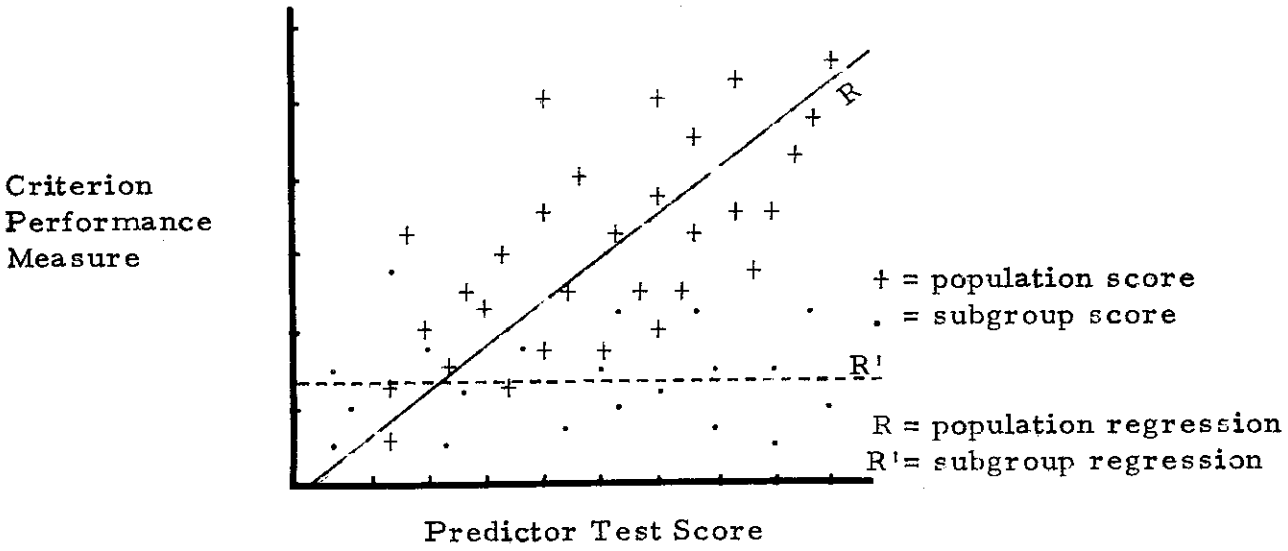

Figure 7

Scatter Plot of Predictive Validity Data for a Test Valid for the Population and Invalid for the Subgroup, with the Subgroup being Lower on the Criterion Performance.

of prediction will be in the opposite direction of those discussed above.

What emerges from this review of the possibilities of bias in prediction is this: If the criterion performance (or its measures) is biased for one group, prediction studies will invariable reveal the apparent and opposite bias in the predictors for that group.

The final set of cases in which tests, as predictors, can be investigated for bias occurs when both the criterion performance and the predictor scores are different between groups. Four such sets of inequalities can occur, in degrees, but consideration of them will be limited to extreme examples for purposes of clarity.

When criterion performance (or its measure) is higher for the population than for the subgroup and predictor scores are also higher for the population, the situation is illustrated in Figure 8. In part this situation will not

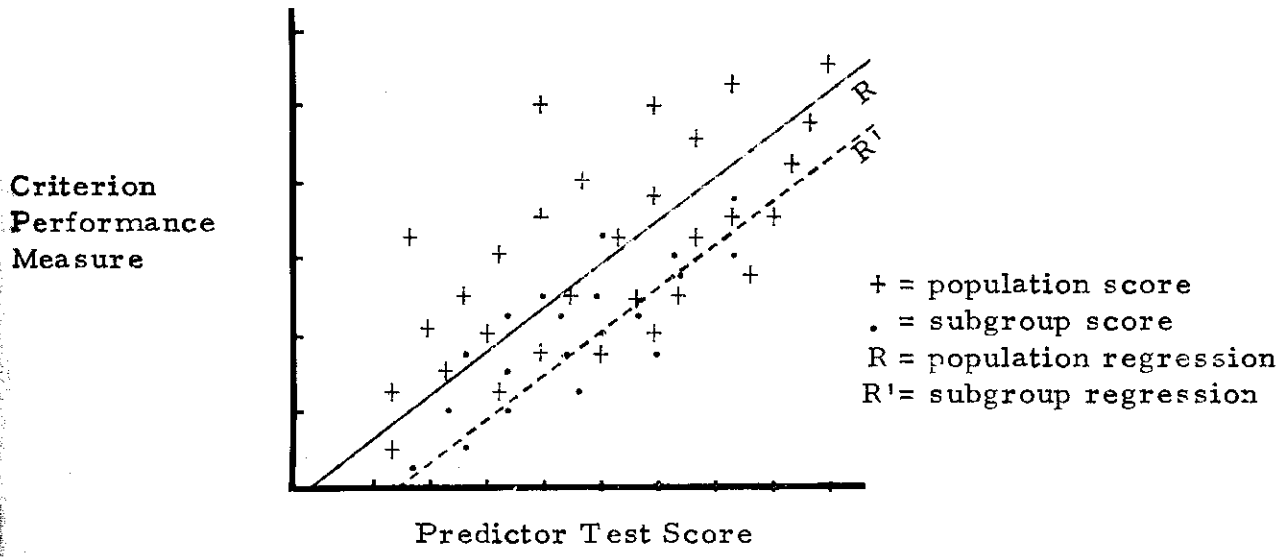

Figure 8

Subgroup, with the Population Superior to the Subgroup on both the Predictor and the Criterion Measures

yield effects different from those associated with Figure 5, with the exception that still fewer subgroup individuals will be selected (although it is also true that fewer of the selected subgroup individuals will fail). The opposite, of course, is true in the unlikely situation that subgroup standings on both criterion performance and predictor measure are higher than those for the population.

In the case where the subgroup's predictor score is higher than that for the population, but its criterion measure is lower (as illustrated in Figure 9), there would be a great overprediction of the success of the subgroup members and a large number of failures in performance; a case of charitable disservice.

In the opposite case, where the subgroup's predictor scores are lower but its criterion measures are higher than the population's, we have the ultimate in uncharitable disservice --- too many subgroup individuals who would be successful will not be selected.

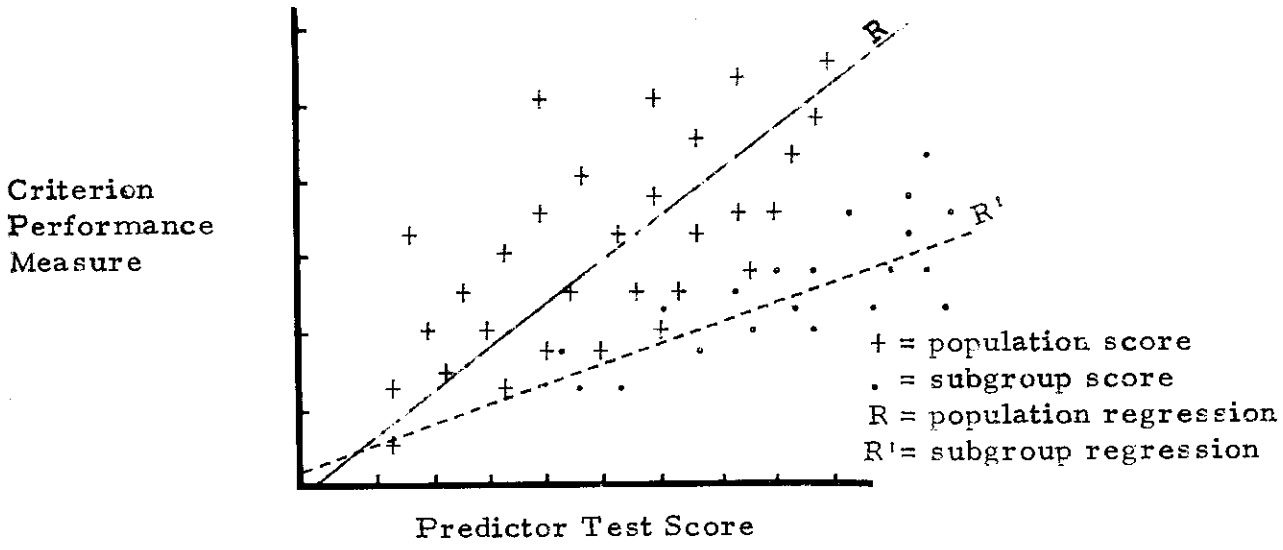

Figure 9

Scatter Plot of Predictive Validity Data for the Population and the Subgroup, with the Population Superior to the Subgroup on the Criterion Measure, but Inferior on the Predictor Measure

From these reviews it can be seen that test bias can only be demonstrated through prediction when the criterion itself is unbiased; bias of the predictor alone will not be uncovered by predictive-validity studies for different subgroups of the population.

Internal Explanations of Test Bias

A second approach to the identification of bias is through the analysis of test-internal characteristics. The tests themselves are analyzed against racial-ethnic groups to determine inherent test characteristics that may cause or explain the bias effect.

Differential Factor Structures. Certain tests may have different factor structures for different ethnic groups which could account for differences in observed scores. Goolsby and Frary (1970) found just such factor structure differences, but their factors were largely determined by achievement measures and their predictors, predictive validities known to be different for the two groups of white and black school children.

On the other hand, numerous studies indicate that there is no difference in the factor structure of intellectual abilities among the racial groups. These studies find relatively invariant factors among groups, usually with one or two culturally-bound factors of a verbal sort (indicating language fluency or common experiences) also found. Vandenberg (1959), in his systematic study of racial differences and similarities, found that Chinese students in the U.S.A. exhibited the same factor structors on Thurstone's PMA tests as American students. In a later study (1967), Vandenberg found high agreement of factor structure between South American and Chinese students using the same test battery.

Guthrie (1963) found that Philippine women college students exhibited factors very similar to those found in western culture. Johnson (1969)

found educational abilities and aptitudes factorially invariant with scores from subjects in Rhodesia and Zambia, and El Abd (1970) found similar factors of intelligence in samples of East African students and American samples. El Abd concluded that there are no basic differences in the structure of intelligence across the races.

Addressing the issue of socio-economic level, which is a potential cause for any observed differences among the races, McGaw and Jöreskog (1970) utilized a large sample from the Project TALENT study, and divided the high school students into four groups on the basis of high and low intelligence and high and low socio-economic status. Factor analysis of 21 aptitude test scores revealed similar factor scores across all four groups, but factor interrelatedness was found to be higher in the high intelligence groups.

Differential Test Scores. Test bias may be defined as the interaction of race and test in an analysis of variance or by other tests for the significance of differences in means or standard deviations among groups. These statistical techniques use the scores of individuals from a variety of races on a variety of tests, determine the amount of variation in scores that is accounted for by differences in the tests alone, the amount accounted for by differences in the races alone, and the amount attributable to some combination or interaction of both race and test differences. The approaches do not address questions of equality of test or equality of races; they look instead to significant differences among test scores and races and therefore will determine only which tests are relatively more or less biased for which groups. The ultimate question of whether or not there is a bias goes unanswered. The logistics involved in this approach are fairly staggering; each test under consideration should be given to each subject. The number

of subjects would have to be quite large in order that other variables (IQ, socioeconomic status, parent education level) be controlled. Furthermore, as noted by Eagle and Harris (1969), who used this approach in a small-scale study:

> "Though this study strongly suggests the operation of significant Test x Culture interaction effects, the specific test characteristics (content, cognitive function, technical features such as speed, etc.) and specific socio-cultural characteristics (ethnicity, economic class, attitudes, mental ability level) which may be entering into these interactions, have not been examined."

Such control problems as these do, however, inhere in all studies where racial-socio-cultural characteristics are examined. The effects of lack of control are simply more apparent when small samples of examinees are employed.

A simplified version of differential score approach is, of course, to merely compare test scores between groups. The control problems will disallow incisive investigation into any nature-nurture issues, but will concentrate instead upon the observable phenomena of score differences. Such total score differences have been observed for a long time and in a fairly consistent manner, but they do not answer questions of bias.

Differential Item Scores. In the item approach to bias (exemplified by Cleary and Hilton, 1968), subjects from various races would take one common test, and the variations in their item scores would be analyzed; some of the variations would be due to difference in items (some items are more difficult than others); some variations would be due to race (one group might have higher apparent intellectual or achievement status than another); and some variation would be due to interaction of item and race (some items might be relatively more difficult for one race than for another), and would therefore show bias.

It should be noted here that this approach adopts a unique definition of what bias is; the definition almost solely dictated by the intent of Cleary and Hilton's study. That study was directed, like the one to be reported later, at finding the parameters of biasedness, an approach which, at best, can be relative. The relative nature is most evident in the fact that items can only show bias in relation to other items. For example, if item A exhibited highly significant differences in difficulty over two racial groups, it would not be considered biased (in the relative sense) unless that difference is different from the other item differences.

Defining bias in this manner necessarily excludes some portion (possibly most) of what it is that bias is (i.e. the overall and consistent unfair differences in scores over the racial groups). The method will uncover distinctive or unique bias effects, but will leave unnoticed the overriding bias effect, which is, incidentally, the subject of major social concern. The hope is, of course, that by isolating many of the unique bias effects over many studies, there will be convergence upon the overall bias effect.

Once biased items are identified (items contributing to the significant item x race interaction) it would be possible to make hypotheses about why they are biased. These hypotheses could be verified by field testing items with the same kind of characteristics, to see if they yield high item-by-race interactions in analysis of variance.

Isolating the biased or differentially differentiating items should lead to the study of their characteristics that might underly the observed bias. Bornstein and Chamberlain (1969) have investigated one such item characteristic-difficulty of the language of items. Their results indicated that mere simplification of item language does not significantly reduce the observed racial differences.

The Present Study

CSE chose to define bias as an item by race interaction in an analysis-of-variance design. Subjects from various races would take one common test, and the variations in their item scores would be analyzed; some of the variations would be due to differences in items (some items are more difficult than others); some variation would be due to race (one group might have higher apparent intellectual or achievement status than another); and some variation would be due to interaction of item and race (some items might be relatively more difficult for one race than for another), and would therefore show bias.

In order to obtain the initial information about item by race interaction, the Stanford Achievement Test, Paragraph Meaning subtest (Form W, Primary II Battery) was administered to 172 third graders at two integrated elementary schools in a large California school district. The sample included 26 white students, 20 blacks, 64 Mexican-Americans, and 37 Orientals. Fifteen others were deleted from the analysis. This grouping of children is anthropologically, sociologically, and economically impure, but does account fairly well for the constellation of characteristics frequently cited as involved in bias. The data would be used to validate the predictions of item by race interaction and to uncover the item characteristics that appear to be involved in effecting the apparent bias.

F-ratios were obtained to determine which items showed significant differences between ethnic groups. Of the 60 items on the subtest, 21 items showed differences between ethnic groups significant at the .001 level; thirteen other items were significant at the .01 level. Since there were four groups (with six possible comparisons between pairs of groups), Duncan's Multiple Range Test (Duncan, 1955; 1957) was applied to see which of the pairs of relationships

were significant. The results, presented in Table 1, show that most of the significant differences are due to differences between the scores of Orientals vs. Mexican-Americans (37 items) and/or Orientals vs. Blacks (28 items). On only 3 items are there significant differences between black and white.

The results for this particular sample indicate that there is considerable item by race interaction. This interaction may be due either to item bias or to race characteristics (confounded by the social and economic peculiarities of this in-vivo sample), or to both. An assumption that significant differences are due to bias inherent in the items would lead us to the implausible conclusion that the test was drawn up to be biased in favor of Orientals, for few significant differences are found involving whites, and many are found involving Orientals.

It is more likely that the results of the study are influenced unduly by the characteristics (racial, social, or economic) of the sample used. There was no control for socioeconomic status, intelligence, or any variable other than observable race. Because these variables were not controlled, we are dealing with an item by (race + IQ + SES) interaction, with no way to separate the effects of the variables.

Linear regressions (used to predict scores of one group, given the equations from the other) were computed for each of the six pairs of racial groups. Results are reported in Table 2. The raw data were item difficulties for the racial group on the 60 items in the test. Each of the regressions involving the Oriental group had a relatively large intercept; this may have been due to the ceiling effect, limiting Oriental scores at the upper end. The mean score for the Oriental group was a perfect score on 12 percent of the items. Both the slope and the intercept are affected by this ceiling effect, limiting the interpretability of the regression equation.

TABLE 1

Item Means for Racial Groups, F-Ratios among the Means, and Post-Hoc Pair Comparisons between Racial Groups

Item	Item Means				F-Ratio	Significant Differences					
	W[a]	B	M	O	DF=153.3	W/B	W/M	W/O	B/M	B/O	M/O
1	.85	.73	.73	.95	2.70						
2	.85	.60	.72	.97	5.63*					x	x
3	.92	.73	.69	.89	3.27						
4	.96	.70	.77	1.00	6.18**	x				x	x
5	.88	.83	.83	.89	.35						
6	.69	.60	.69	.89	2.74						
7	.85	.77	.67	.84	1.67						
8	.85	.83	.75	.87	0.84						
9	.96	.83	.77	.95	3.10						
10	.81	.64	.67	.92	3.55					x	x
11	.73	.67	.61	.89	3.23						x
12	.92	.67	.72	.87	2.82						
13	.81	.80	.70	.95	2.94						x
14	.96	.83	.88	1.00	2.62						
15	.85	.70	.78	1.00	4.29*					x	x
16	.96	.73	.75	1.00	5.93**					x	x
17	.81	.43	.61	.89	7.14**	x				x	x
18	.77	.60	.66	1.00	6.91**					x	x
19	.85	.60	.64	.97	6.82**					x	x
20	.85	.70	.63	.89	3.68						x
21	.69	.60	.47	.84	5.10*						x
22	.92	.70	.69	.97	5.82**					x	x
23	.81	.67	.53	.92	6.81**		x				x
24	.77	.70	.70	1.00	4.96*					x	x
25	.77	.47	.53	.92	8.21**					x	x
26	.65	.47	.59	.95	7.36**					x	x
27	.69	.50	.67	.89	4.33*					x	
28	.69	.47	.61	.89	5.29*					x	x
29	.73	.47	.52	.76	3.38						
30	.52	.47	.47	.65	1.16						
31	.58	.40	.55	.70	2.11						
32	.46	.40	.36	.62	2.32						
33	.62	.40	.50	.78	4.18*					x	x
34	.77	.53	.69	.73	1.46						
35	.73	.60	.59	.95	5.65*					x	x

[a]W = white, N = 26; B = black, N = 30; M = Mexican-American, N = 37; O = Oriental, N = 37

Table 1 (continued)

Item	Item Means				F-Ratio DF=153.3	Significant Differences					
	W[a]	B	M	O		W/B	W/M	W/O	B/M	B/O	M/O
36	.58	.53	.39	.84	7.06**						x
37	.62	.37	.42	.87	9.01**					x	x
38	.77	.63	.73	1.00	5.45*					x	x
39	.77	.47	.66	.95	7.38**					x	x
40	.81	.53	.61	.95	6.93**					x	x
41	.77	.53	.47	.84	6.21**		x				x
42	.69	.53	.42	.73	3.92*						x
43	.69	.40	.55	.89	7.40**					x	x
44	.73	.47	.56	.89	6.13**					x	x
45	.46	.23	.33	.60	3.89					x	x
46	.65	.57	.39	.73	4.43*						x
47	.54	.43	.33	.76	6.55**					x	x
48	.35	.33	.39	.68	3.95*			x		x	x
49	.54	.40	.34	.51	1.72						
50	.54	.30	.41	.62	2.87						
51	.46	.23	.36	.40	1.18						
52	.58	.33	.34	.65	4.28*						x
53	.92	.43	.55	.68	5.98**	x	x				
54	.65	.33	.38	.76	7.25**					x	x
55	.58	.37	.44	.92	10.99**			x		x	x
56	.38	.30	.14	.38	3.31						
57	.23	.30	.33	.32	0.30						
58	.27	.17	.19	.32	1.14						
59	.54	.47	.39	.62	1.81						
60	.23	.17	.27	.65	8.72			x		x	x

* significant at .01 $F \geq 3.91$

** significant at .001 $F \geq 5.70$

Table 2

Item Regression Data for Six Racial Pairings

Whites (W) and Blacks (B)	
Regression:	B = .789 W = .030
Standard error of regression coefficient:	.067
$Correlation_{WB}$:	.841
Items more than 2.58 standard error units from regression line:	8, 13, 17, 53
Whites (W) and Mexican-Americans (M)	
Regression	M = .797 W = .004
Standard error of regression coefficient:	.059
$Correlation_{WM}$:	.870
Items more than 2.58 standard error units from regression line:	53, 56
Whites (W) and Orientals (O)	
Regression:	O = .764 W = .279
Standard error of regression coefficient:	.074
$Correlation_{WO}$:	.805
Items more than 2.58 standard error units from regression line:	51, 53, 55, 56, 60
Blacks (B) and Mexican-Americans (M)	
Regression:	M = .821 B = .118
Standard error of regression coefficient:	.067
$Correlation_{BM}$:	.851
Items more than 2.58 standard error units from regression line:	46, 56
Blacks (B) and Orientals (O)	
Regression:	O = .727 B = .429
Standard error of regression coefficient:	.090
$Correlation_{BO}$:	.727
Items more than 2.58 standard error units from regression line:	56, 57
Mexican-Americans (M) and Orientals (O)	
Regression:	O = .846 M = .346
Standard error of regression coefficient:	.079
$Correlation_{MO}$:	.816
Items more than 2.58 standard error units from regression line:	51, 57

In linear regression, a graph of the scores for each item should show most scores lying near the regression line. The scores for some items, however, will be far from the regression line because one group's score on an item may be much greater or less than expected. When this occurs, we must assume that there are forces at work affecting the group's score on that item, which are not affecting (as greatly) scores by other groups on that item. The criterion for deciding that an item score is "far" from the regression line is that it be distant from the regression line greater than 2.58 times the standard error of the regression coefficient. This criterion eliminates 99% of the items in a normally distributed sample.

The Stanford Achievement Test - Paragraph Meaning consists of items that involve the examinee's making of logical implications from a reading selection, sometimes by referring back to specific wordings, and then choosing words (from four relatively equally attractive alternatives) to complete sentences continuing from the given selection. Since all the items rather consistently conform to this description, it seemed useless to analyze bias effects in terms of item intellectual processes. Instead, the item contents were inspected to see if the subject matter of the item content, either through knowledge, relevance, or interest, might meaningfully correspond to score differences. Utilizing the regression approach that underlies Table 2, item contents for each racial group were inspected.

White/Black Item Differences. Black students score higher than expected on items 8 and 13. These items are concerned with television and cowboys on television. The white students score higher than expected on items 17 and 53, concerned with sledding weather and supermarkets, respectively. The differences uncovered in this comparison could be related to experience and relevance to the separate groups.

White/Mexican-American Item Differences. White students do better than expected on items 53 and 56 of the test. The first item is concerned with supermarkets and the second with comparisons of physical sizes of boys. The physical size item is, however, slightly tricky --- the wording leads one to an incorrect alternative (it is possible that more socially threatened students will respond in the way that appears correct - and consequently be misled). The group experience and relevance hypothesis could also be operating in this comparison.

White/Oriental Item Differences. Oriental students score higher than expected on items 55 and 60. The first item is highly implicational and is concerned with the misnomer of Greenland. Item 60 is concerned with copyrights, but also involves the making of abstract implications. White students score higher than expected on items 51, 53, and 56; items concerned with farming, supermarkets, and comparing of physical sizes of boys. While the experience-relevance hypothesis is reasonable for explaining the whites' item-relative superiority, it is not very compelling as an explanation of the Orientals' performance.

Black/Mexican-American Item Differences. Black students do better than expected on items 46 and 56; items concerned with Mt. Vernon and comparing physical sizes of boys, respectively. With the exception that blacks may be more size-conscious than Mexican-Americans (doubtful, because of the pressures of "machismo" in the latter culture) the experience-relevance hypothesis does not hold up in this comparison.

Black/Oriental Item Differences. Blacks score higher than expected on items 56 and 57, both concerned with comparing physical sizes of boys. It is difficult to hypothesize a unique relationship between racial size differences and item response consistencies.

Mexican-American/Oriental Item Differences. Mexican-Americans score higher than expected on items 51 and 57; concerned with farming and comparing physical sizes of boys, respectively. If the experience-relevance hypothesis holds here, we would have to note the consistent tendency for other groups to score higher than predicted on items involving size comparisons, when they are compared to the Orientals.

The evidence from this study is not overwhelmingly in favor of the hypothesis that the differential familiarity, relevance, and interest arousing aspects of items underlie the observed group differences.

Summary

This report had as its original intent, the development of procedures to codify the amount and nature of bias that inheres in standardized tests, so that Center evaluations of the tests could be modified for different racial-ethnic groups. Various methods for establishing the existence and nature of test bias are discussed, with the conclusion that test bias cannot be conclusively demonstrated in a wholly satisfactory manner. One method was nonetheless selected and applied to test items administered to two field-test schools for the purpose of investigating bias. The results of that small-scale study are discussed, but do not offer compelling reasons for the observed racial-ethnic differences.

References

Bornstein, H. & Chamberlain, K. An investigation of some of the effects of "Verbal Load" in achievement tests. Research Bulletin RB - 69 - 94. Princeton, N. J.: Educational Testing Service, 1969.

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

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

Duncan, D. B. Multiple range and multiple F tests. Biometrika, 1955, 11, 1-42.

Duncan, D. B. Multiple range tests for correlated and heteroscedastic means. Biometrika, 1957, 13, 164-176.

Eagle, N. & Harris, A. S. Interaction of race and test on reading performance scores. Journal of Educational Measurement, 1969, 6, 131-135.

Einhorn, H. J. & Bass, A. R. Methodological considerations relevant to discrimination in employment testing. Psychological Bulletin, 1971, 75, 261-269.

El Abd, H. A. The intellect of East African students. Multivariate Behavioral Research, 1970, 5, 423-434.

Goolsby, T. M. & Frary, R. B. Validity of the Metropolitan Readiness Test for white and Negro students in a southern city. Educational and Psychological Measurement, 1970, 30, 443-450.

Guthrie, G. M. Structure of abilities in a non-western culture. Journal of Educational Psychology, 1963, 54, 94-103.

Hills, J. R. & Stanley, J. C. Easier test improves prediction of black students' college grades. Journal of Negro Education, 1970, 39, 320-324.

Hoepfner, R., Strickland, G., Stangel, G., Jansen, P. & Patalino, M. CSE elementary school test evaluations, 1971. Los Angeles: Center for the Study of Evaluation, UCLA.

Johnson, M. Factorial univariance of African educational abilities and aptitudes. Research Bulletin 69-3. Princeton, N. J.: Educational Testing Service, 1969.

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

McGaw, B. & Jöreskog, K. G. Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status. Research Bulletin, RB - 70 - 63. Educational Testing Service, Princeton, N. J., 1970.

Millman, J. & Lindlof, J. The comparability of fifth-grade norms of the California, Iowa, and Metropolitan Achievement Tests. Journal of Educational Measurement, 1964, 1, 135-137.

Mitchell, B. C. Predictive validity of the Metropolitan Readiness Tests and the Murphy-Durrell Reading Readiness Analysis for white and for Negro pupils. Educational and Psychological Measurement, 1967, 27, 1047-1054.

Temp, G. Test bias: Validity of the SAT for Blacks and Whites in thirteen intergrated institutions. College Entrance Examination Board Research and Development Reports - 70 - 71, No. 6. Berkeley, California: Educational Testing Service, 1971.

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

Vandenberg, S. G. The primary mental abilities of Chinese students: A comparative study of the stability of factor structure. The Annals of the New York Academy of Sciences, 1959, 79, 257-304.

Vandenberg, S. G. The primary mental abilities of South American students: A second comparative study of the generality of a cognitive factor structure. Multivariate Behavioral Research, 1967, 2, 175-197.